T.C.
İNÖNÜ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

METİN ÇİZGELERİNDE BAĞIMSIZ KÜMELERE DAYALI ÇIKARIMSAL METİN ÖZETLEME

TANER UÇKAN

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

OCAK 2020

T.C.

İNÖNÜ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

EXTRACTIVE TEXT SUMMARIZATION BASED ON INDEPENDENT SETS IN TEXT GRAPHS

TANER UÇKAN

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

OCAK 2020

Tezin Başlığı : Metin Çizgelerinde Bağımsız Kümelere Dayalı Çıkarımsal Metin Özetleme

Tezi Hazırlayan : Taner UÇKAN

Sınav Tarihi : 28.01.2020

Yukarıda adı geçen tez jürimizce değerlendirilerek Bilgisayar Mühendisliği Ana Bilim Dalında Doktora Tezi olarak kabul edilmiştir.

Sınav Jüri Üyeleri

Tez Danışmanı: Prof. Dr. Ali KARCI .....

İnönü Üniversitesi

Prof. Dr. Davut HANBAY .....

İnönü Üniversitesi

Doç.Dr. M. Fatih TALU .....

İnönü Üniversitesi

Prof. Dr. Ahmet Bedri ÖZER .....

Turgut ÖZAL Üniversitesi

Dr. Öğretim Üyesi İhsan TUĞAL .....

Muş Alparslan Üniversitesi

Prof. Dr. Kazım TÜRK

Enstitü Müdürü

ONUR SÖZÜ

Doktora Tezi olarak sunduğum "Metin Çizgelerinde Bağımsız Kümelere Dayalı Çıkarımsal Metin Özetleme" başlıklı bu çalışmanın bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurmaksızın tarafımdan yazıldığını ve yararlandığım bütün kaynakların hem metin içinde hem de kaynakçada yöntemine uygun biçimde gösterilenlerden oluştuğunu belirtir, bunu onurumla doğrularım.

Taner UÇKAN

ÖZET

Doktora Tezi

METİN ÇİZGELERİNDE BAĞIMSIZ KÜMELERE DAYALI ÇIKARIMSAL METİN ÖZETLEME

Taner UÇKAN

İnönü Üniversitesi

Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

78 + xi sayfa

2020

Danışman: Prof. Dr. Ali KARCI

Bu tez çalışması kapsamında genel, denetimsiz ve çıkarıcı metin özetleme problemine iki yeni çizge tabanlı yaklaşım sunularak katkıda bulunulmuştur. Her iki yaklaşımın veri işleme aşamasında da kullanılan KUSH (Karcı,Uçkan,Seyyarer,Hark) aracı önerilmiş ve denenmiştir. Önerilen yöntemlerden ilki, üç ana adımdan oluşan CatSumm (Cengiz, Ali, Taner Summarization) modelidir. İlk adımda KUSH aracı ile normalizasyon gerçekleştirildi. Modelin ikinci adımında spektral çizge bölmeleme ile çizgeler kümelenerek özetlerin alt çizgelerde bulunan cümle sayısı oranları ölçüsünde üretilmesi sağlanmıştır. Son aşamada düğüm ağırlıklandırma yöntemleri kullanılarak, merkezilik değerleri yüksek olan cümlelere özette yer verilmektedir. Bağımsız kümelerde yer alan düğümlere karşılık gelen cümlelerin, özette yer almaması gerektiği öngörüsünden yola çıkılan ikinci çalışmada ise düğümlerin genel çizge üzerindeki etkisi sayısal olarak belirlenmeden önce, özetlenecek belgeler üzerinde bir sınırlamaya gidilmiştir. Her iki yaklaşım da DUC (Document Understanding Conference, DUC-2002 ve DUC-2004) veri seti üzerinde ve ROUGE (Recall-Oriented Understudy for Gisting Evaluation) değerlendirme metrikleri kullanılarak test edilmiştir. 100, 200 ve 400 kelimelik özetler için deneysel süreçler tekrarlanmıştır. Önerilen modeller ile rapor edilen değerler, yenilikçi yöntemlerin katkılarını ortaya koymaktadır.

ANAHTAR KELİMELER:Otomatik metin özetleme, Çıkarıcı metin mzetleme, Çizge teorisi, Metin ön işleme, Maksimum bağımsız kümeler, Spektral çizge kümeleme

ABSTRACT

Ph.D. Thesis

EXTRACTIVE TEXT SUMMARIZATION BASED ON INDEPENDENT SETS IN TEXT GRAPHS

Taner UÇKAN

İnönü University

Graduate School of Natural and Applied Sciences

Department of Computer Engineering

78 + xi pages

2020

Supervisor: Prof. Dr. Ali KARCI

Within the scope of this thesis, two new graph-based approaches have been contributed to the general, unsupervised and extractive text summarization problem. The KUSH (Karcı, Uçkan, Seyyarer, Hark) tool used in the data processing stage of both approaches has been proposed and tried. The first proposed method is the CatSumm (Cengiz, Ali, Taner Summarization) model, which consists of three main steps. In the first step, normalization was performed with the KUSH tool. In the second step of the model, the graphs were clustered with spectral graph partitioning, so that the summaries were produced in accordance with the number of sentence ratios in the subgraphs. In the last stage, using the node weighting methods, sentences with high centrality values are included in the summary. In the second study, based on the prediction that the sentences corresponding to the nodes in the independent clusters should not be included in the summary, a limitation was made on the documents to be summarized before the effect of the nodes on the general graph was determined numerically. Both approaches were tested on the DUC (Document Understanding Conference, DUC-2002 and DUC-2004) data set and using ROUGE (Recall-Oriented Understudy for Gisting Evaluation) evaluation metrics. Experimental processes were repeated for summaries of 100, 200 and 400 words. The values reported with the proposed models reveal the contributions of innovative methods.

KEYWORDS: Automatic text summarization, Extractive text summarization, Graph theory, Text preprocessing, Maximum independent sets, Spectral graph clustering

TEŞEKKÜR

Tez çalışmamın her aşamasında bilgi ve tecrübelerinin yanı sıra ilgi ve teşvikleriyle de bana destek olan danışman hocam Sayın Prof. Dr. Ali KARCI'ya; çalışmalarım boyunca, bana destek olan Anabilim Dalındaki tüm değerli hocalarıma; kendisiyle ortak çalışmalar yaptığımız ve fikir alışverişinde bulunduğum arkadaşım Cengiz HARK'a teşekkürlerimi sunarım.

Ayrıca doktora çalışmalarım süresince bana hep sabırla davranan ve yanımda olan sevgili eşime, aileme ve iş arkadaşlarıma teşekkür ederim.

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

G	Çizge (Graph)
V	Düğüm (Vertice)
E	Ayrıt (Edge)
i	i'inci düğüm
j	j'inci düğüm
$e_{i,j}$	i ve j'inci düğüm arasındaki ayrıt
$w_{i,j}$	i ve j'inci düğüm arasındaki ağırlık değeri
a_{ij}	Komşuluk matrisi
n	Düğüm sayısı
DC_i	i düğümünün derece merkeziliği
w_{ij}	i ve j'inci düğüm arasındaki ağırlık değeri
BC_i	i düğümünün arasındalık merkeziliği
KUSH	Karci ,Uçkan ,Seyyarer ,Hark
MBK	Maksimum Bağımsız Kümeler
MIS	Maximum Independent Set
SD	Standart Deviation
sp_{jk}	j ve k düğümleri arasındaki kısa yolların sayısı
$sp_{jk}(i)$	i düğümünden geçen j ve k düğümleri arasındaki kısa yolların sayısı
SSC	Spectral Sentence Clustering
CC_i	i düğümünün yakınlık merkeziliği
CatSum	Cengiz,Ali,Taner Summarization
d_{ij}	i ve j derece matrisi
A	nxn boyutunda benzerlik matrisi
x_i	i düğümünün özvektör merkeziliği
λ	A'nın en büyük özdeğeri
C_C_i	i düğümünün kümelenme katsayısı
N_i	i düğümünün komşularının birbiri ile bağlantıda olanlarının sayısı
k_i	i'inci düğümünün düğüm derecesi
$\in$	Elemanıdır
$\notin$	Elemanı değildir
$\emptyset$	Boş küme
∞	Sonsuz

ŞEKİLLER LİSTESİ

ÇİZELGELER LİSTESİ

1. GİRİŞ

İnternet teknolojisinin hızlı gelişimi ile internet üzerindeki elektronik dokümanların olağanüstü bir hızla arttığı görülmektedir. Bu durum şüphesiz insan yaşamını zenginleştirmektedir; ancak aynı zamanda hızlı ve doğru bilgiye ulaşmayı zorlaştırmaktadır [1]. Sayısal ortamlar yoluyla artan verilerin beraberinde getirdiği birçok avantajı olmakla beraber bir o kadar da dezavantajı bulunmaktadır. Herhangi bir konu hakkında bilgi zenginliğinin olması ,birden fazla kaynaktan daha detaylı bilgilere ulaşılması gibi farklı avantajları sayılabilir. Bunun yanında bilgi fazlalığı beraberinde bilgi yanlışlığını, yararsız bilgi fazlalığını getirmektedir. İnternet üzerinden yapılan bir arama sonucunda konuyla alakalı veya alakasız binlerce ya da milyonlarca sonuca ve veriye ulaşmak mümkündür. Günümüzde teknolojinin ilerlemesi ile bu tarz problemlerle sürekli karşılaşılmaktadır. Teknolojinin gelişimi ve veri oranın üstel bir seviyede artıyor olması doğru bilgiye erişim sorunlarınıda beraberinde getirmektedir. Sürekli artan ve sayısal sistemlere kaydedilen verilerin büyük bir çoğunluğunu yapılandırılmamış veriler oluşturulmaktadır. Düzenli bir yapıya sahip olmayan ve üzerinde herhangi bir işlem yapılmamış olan verilere ham veri denilmektedir.

Ham veri, bir olgu hakkında birbirleri ile olan ilişkileri henüz tam olarak ortaya konmamış bilgi toplulukları olmakla birlikte sayısal formatlarla ifade edilebilen, taşınabilen dizeler olarak da tanımlanabilmektedir. Bu ham verilerin bilinen bir amaç doğrultusunda anlamlı birlikteliklerinin elde edilmiş hali ise bilgidir. Her geçen gün büyüyen veri miktarı yeni yaklaşımlar ile çözülmeyi bekleyen problemlere kapı açmaktadır. İşlenmeyi bekleyen dijital veriler çok büyük oranda belirli bir formatı olmayan ham metinlerdir. Sözü edilen verilerin anlamlı ve faydalanılabilir veri kaynaklarına dönüşmesi için üzerinde çalışılarak analiz edilmeleri gerekmektedir. Bilgiye erişim zamanının kısaltılarak kabul edilebilir erişim sürelerinin elde edilmesi için yeni yöntemlerin geliştirilmesi gerektiği kadar verinin analiz edilmesi de gerekmektedir [2]–[4].

Kaydedilen veriler farklı türlerde bilgi içerebilmektedirler. Bu veri türleri arasında ses, görüntü ve metin gibi sayısallaştırabilir veriler bulunmaktadır. Metin verilerine erişim yapılmak istendiğinde doğru bilgiye erişmek isterken aranılan bilgiden çok farklı anlam taşıyan, ilişkili olmayan veya çok fazla boyuttaki verilerle karşılaşmak mümkün olmaktadır. Bilgi aranırken genellikle çok fazla zaman kaybı yaşanmadan doğrudan en uygun sonuca erişilmek istenmektedir. Kısa zamanda en uygun ve en doğru bilgiye erişim sağlanması zaman kaybını büyük oranda ortadan kaldırmaktadır. Böylelikle insanlar zamanı daha

verimli kullanabilme imkanına sahip olmaktadır. Zaman kaybının önlenmesi ve incelenen dökümanın içerdiği genel bilgiye ulaşmak için metin özetleme yöntemleri kullanılmaktadır. Son yıllarda araştırmacılar metin dokümanlarının özetlenmesi amacı ile yeni algoritma ve yöntemler geliştirmek için çalışmaktadırlar. Zira uzun metin belgelerinin insanlar tarafından özetlenmesi hem çok zor hem de gerçekten fazla zaman gerektirmektedir. İş dünyası, akademi, sağlık gibi birçok alanda özetleme oldukça gerekli görülmektedir.

Otomatik metin özetleme teknolojileri giderek daha çok araştırmacı tarafından çalışılmakta ve yeni öneriler ile daha verimli hale getirilmeye çalışılmaktadır [5]. Doküman özetleme alanında gerçekleştirilen tüm çalışmalara rağmen yeni çalışmalara olan ihtiyaç artarak devam etmektedir [6]. Otomatik metin özetleme, uzun metin dökümanlarını sıkıştırılmış ve anlaşılır bir biçimde temsil etmeyi amaçlayan önemli bir doğal dil işleme (natural language processing - NLP) dalıdır [7]. Doküman özetleme teknikleri çıkarıcı ve yorumlayıcı doküman özetleme olarak iki kategoriye ayrılabilmektedir. Çıkarıcı özetler metinlerin temsili, cümle skorlama ve cümle seçimi gibi üç aşamadan oluşurlar. Yorumlayıcı özetler ise doğal dil üretim yaklaşımlarını kullanarak dökümanların ana içeriklerini yorumlarlar ve yeniden ifade ederek özetleri oluştururlar [8], [9]. Ayrıca özetlenecek dökümanların sayısına bağlı olarak stek dökümanlı ve çoklu doküman olarak da kategorize edilebilirler. Özetleme ile ilgili ilk çalışmalar Tek dökümanlı özetleme kapsamında yapılmıştır. Devam eden çalışmalarda çoklu döküman özetleme çalışmaları yapılmaya başlanmıştır ve birden fazla metne uygulanan yöntemler geliştirilmiştir [7], [10], [11]. Ayrıca özetler ya genel ya da sorgu odaklı özetleme olarak kategorize edilirler [12]. Araştırmacılar tarafından gerçekleştirilen çalışmaların büyük çoğunluğu genel özetleme bağlamında olmuştur. Özeti oluşturma amacı hakkında birkaç sınırlı varsayım ile gerçekleşen bu özetleme türünde genel konu içeriği korunarak mümkün olduğunca fazla bilgi kapsanmaya çalışılırken sorgu odaklı özetlemelerde ise kullanıcı tarafından belirlenen sorgulara bağlı kalınarak bir özetleme amaçlanır ve belirtilen konu ile ilgili metindeki bilgiler elde edilir [13]–[15].

Kelime öbeklerinin veya cümlelerin puanlandırılarak özetlerin elde edilmesi otomatik çıkarıcı özetleme sistemlerinde kullanılan en yaygın yöntemdir. Günümüzde uygulanan yöntemlerin büyük çoğunluğunda ise cümle puanlama yöntemi benimsenmiştir. Metin analiz yöntemlerinde çizge tabanlı temsiller içeren metotlar sıklıkla kullanılmışlardır ve oldukça etkili çözümler üretmişlerdir. [10]'da yazarlar, özetleme için çizge tabanlı temsil içeren TextRank'ı önermişlerdir. Sunulan TextRank yaklaşımında metin içeriklerinin kesişimleri

kullanılarak çizge tabanlı bir metot sunulmuştur. [14]'te, benzer şekilde cümlelerin temsil edildiği LexRank tanıtılmıştır. LexRank algoritması düğüm merkeziliği yöntemlerinden biri olan özvektör (eigenvector) düğüm merkeziliğini temel alan bir algoritma önerilmiştir. Hem LexRank hem de TexRank algoritması PageRank [16] algoritmasından ilham almışlardır. Terim ve cümle kümeleri arasındaki karşılıklı bilgileri kullanarak bir dokümandaki merkezi cümleleri elde edecek bir doküman özetleme modeli sunulmuştur[17]. Daha farklı bir temsil ile [18]'de katmanlarını dokümanlar, cümleler ve kelimelerin oluşturduğu çok katmanlı bir temsil kullanılmıştır. [19]'da yazarların, otomatik doküman özetleme için bağlantı oluşturmadan (link generation) faydalandıkları çalışmalarında dokümanları çizgelerle ifade etmiştir. Dokümanlardaki metin ilişkilerini ortaya koyarak bu yapıyı tanımlamıştır. Elde ettikleri özetleri insanlar tarafından oluşturulan özetlerle karşılaştırarak değerlendirmiştir. Anlamsal devamlılığın sağlanması için çizge tabanlı bir yaklaşımın sunulduğu [20]'da düğümler dokümanlara ait terimleri tutmakta, ayrıtlar ise bu düğümler arasındaki anlamsal (semantic) ilişkileri yansıtmaktadır. Temel olarak çizgelerde yer alan tüm düğümlere ait çizge çapı (graph diameter) hesabı gerçekleştirmekte ve en kısa ve en uzun yolları en zayıf ve en güçlü bağlar olarak nitelendirmektedir. [21]' de ki çalışmada çizge yapıları ile dökümanlar tanımlanırken düğümler ve ayrıtlar yerel benzerliklerden yola çıkılarak oluşturulmaktadır. Ana dökümanlara ait özetleri elde etmek içinde Random Walk yöntemi uygulamıştır. [22]' de ise biyomedikal alan için bir özetleme sistemi önerilmektedir. Landauer et al. [23], [24] çalışmalarında herhangi bir bilgi ön kabulü kullanmadan sadece genel matematiksel yöntemleri kullanarak yeni bir yaklaşım sunmuştur. Unified Medical Language adı verilen bir sistem kullanılarak yarı sözlük tabanlı bir uygulama ile kavram ve ilişkilerden yola çıkılarak bir graph elde edilmekte ve akabinde PageRank algoritması kullanılmaktadır.Benzer şekilde [25]'te yazarlar Reinforced Random Walk tabanlı yeni bir çizge önerisinde bulunmaktadırlar. Literatürde sıklıkla kullanılan temel bir yöntem olan SumBasic [26], özette yer alması gereken top N cümleyi etkili bir olasılık fonksiyonu kullanarak seçme yoluna gitmektedir. [27]'de yazarlar çoklu doküman için greedy yöntemi ile belirlenen bir olasılık modeli ile özete cümleleri eklemektedir.

Yukarıda bahsedilenlerle birlikte literatürde önerilen çizge tabanlı özetleme sistemlerinin özetleme türleri, doküman sayıları ve kullanılan metotlara ait bilgiler Çizelge 1.1 'de gösterilmektedir.

Çizelge 1.1. Çizge tabanlı metin özetleme çalışmaları

Yazar,Yayınlanma Yılı	Özetleme Tipi	Döküman Sayısı	Kullanılan Model
[28] Luhn, 1958	Çıkarıcı Özt.	Tekli	İstatistiksel Metot
[19] Salton et al., 1997	Çıkarıcı Özt.	Çoklu	Bağlantı Tahmini Algoritmaları
[14] Erkan et al., 2004	Çıkarıcı Özt.	Çoklu	Çizge tabanlı Metot & Özvektör merkeziliği
[29] Mihalcea et al., 2005	Çıkarıcı & Genel Özt.	Tekli	Çizge tabanlı Metot & PageRank
[20] Medelyan, 2007	Anahtar sözcük Çıkarımı	Tekli	Çizge tabanlı Metot & Bağlantı Tahmini
[30] Fattah et al., 2009	Çıkarıcı Özt.	Çoklu	Ga, Mr, Ffnn, Pnn, Gmm
[31] Shardan et al., 2010	Çıkarıcı Özt.	Tekli	Semantik Nets, Bulanık Mantık & Gelişim Algoritmaları
[32] Chen et al., 2011	Anahtar sözcük Çıkarımı	Tekli	Çizge tabanlı Metot & Random Walk
[22] Plaza et al., 2012	Çıkarıcı Özt.	Çoklu	Çizge tabanlı Metot & Umls
[33] Abuobieda et al., 2012	Çıkarıcı Özt.	Tekli	Özellik Seçimi & GA
[34] Moawad et al., 2012	Yorumlayıcı Özt.	Tekli	Semantik Çizge indirgeme
[35] Gupta, 2013	Çıkarıcı Özt.	Çoklu	Makine Öğrenmesi
[18] Canhasi et al., 2014	Sorgu Odaklı Özt.	Çoklu	Çizge tabanlı Metot & Matrik Ayrışımı
[36] Linqing Liu et al., 2014	Yorumlayıcı Özt.	Çoklu	Reinforcement Öğrenme
[37] Student et al., 2015	Çıkarıcı Özt.	Tekli	Cüme Çıkarımı, GA & NN
[17] Parveen et al., 2015	Çıkarıcı Özt.	Tekli	Çizge tabanlı Metot & ILP
[38] Nallapati, et al., 2016	Yorumlayıcı Özt.	Çoklu	Nöral Ağlar
[13] Xiong & Ji, 2016	Sorgu Odaklı Özt.	Çoklu	Hyper Çizge-bazlı ağırlıklandırma
[39] Nasr Azadani et al., 2018	Çıkarıcı Özt.	Tekli	Çizge Kümeleme & Frekans kümeleri

Bu tez kapsamında çizge tabanlı metin özetlemeye yönelik iki özgün çalışma ile metin özetleme ve sınıflandırma çalışmalarında en önemli aşamalardan biri olan metin ön işleme (Text Preprocessing) aşamasında kullanılmak üzer yeni bir araç geliştirilmiştir.

- **KUSH Metin işleme aracı*:***

  Yazım şekilleri bakımından birbirlerinden farklı olan ancak anlamsal olarak ortak bir kelime kökünden türeyen yakın anlamlı kelimelerin çizgeler oluşturulurken farklı kelimeler gibi işleme alınması, cümleler arasındaki bağlantı ve ilişkilerin tespitini zorlaştırmaktadır. Sezgisel olarak sözü edilen problemin sınıflandırma sonrasında elde edilen başarımı ciddi bir şekilde etkileyeceği öngörülmüştür ve bu tez kapsamında metin ön işleme adımında kullanılan bir metin işleme aracı geliştirilmiştir

- **CatSumm: Extractive Text Summarization with Spectral Sentence Clustering and Node Centrality(Spektral Metin Kümeleme Yöntemi ve Düğüm Merkezilik Değerleri Kullanılarak Çıkarıcı Metin Özetleme)*;*** Bu çalışmada CatSumm (Cengiz, Ali, Taner Summarization), adını verdiğimiz yeni ve denetimsiz bir otomatik çoklu döküman özetleme modeli önerilmiştir. Önerilen yöntem tarafımızdan belirlenen üç ana adıma göre özet oluşturmaktadır: Giriş metinlerinin temsili, CatSumm modelinin ana aşamaları ve cümle puanlandırma şeklindedir. Giriş metinlerinin temsili aşamasında cümleler arasındaki anlamsal bağlılığın korunması için KUSH (Karcı, Uçkan, Seyyarer, Hark) adını verdiğimiz bir metin işleme aracı kullanılmaktadır. CatSumm modelinin ana aşamalarından biri olan Spectral Sentence Clustering (SSC) ile spektral çizge bölmeleme sonrasında elde edilen alt çizgelerde bulunan cümle sayısı oranları ölçüsünde özetin elde edilmesi işlemidir. Standart sapma (SD) yolu ile hesaplanan bir eşik değerinin altında değer alan ayrıtın silinmesi ile zayıf bağlantılı cümlelerin özette yer alamayacağı varsayımı ile süper ayrıtların elde edilmeside yöntemin ana aşamaları arasında yer almaktadır. CatSumm yaklaşımının farklı düğüm merkeziliği yöntemlerini kullanarak önemli düğümleri ve dolayısı ile önemli cümlelerin belirlenmesi önerilen özetleme sisteminin cümle puanlandırma aşamasını oluşturmaktadır

- **Extractive Multi-Document Text Summarization based on Graph Independent Sets(Maksimum Bağımsız Kümeye Dayalı Çoklu Doküman Özetleme)*;*** Bu çalışma kapsamında daha önce hiçbir özetleme çalışmasında kullanılmamış olan

çizgelere ait maksimum bağımsız kümeler (Maximum Independent Set) yaklaşımı kullanılmıştır. Ayrıca giriş metinlerinin temsili aşamasında cümleler arasındaki anlamsal bağlılığın korunması için KUSH metin işleme aracından faydalanılmaktadır. Bağımsız kümelerde de yer alan düğümlere karşılık gelen cümlelerin, özette yer almaması gerektiği öngörüsünden hareket edilmiştir. Sözü edilen öngörü ile çizgeler üzerinde bağımsız kümeyi oluşturan düğümler belirlenerek çizgeden çıkarılmıştır. Böylece düğümlerin genel çizge üzerindeki etkisi sayısal olarak ölçülmeden önce, özetlenecek dokümanlar üzerinde bir sınırlamaya gidilmiştir. Bu sınırlama ile oluşacak özette yer alacak kelime gruplarının tekrarı engellenerek çok daha kapsayıcı özetler elde edilebilmiştir.

1.1. Amaç ve Kapsam

Tanımsal olarak özet, bir dokümanın içeriğini doğru bir şekilde ifade edecek, tercihen yazarları tarafından oluşturulan kısaltılmış metindir. Oluşturulan bu kısa metnin ana işlevi dokümanın içeriği, yapısı, konusu, fikri hakkında bilgi vermektir. Ancak bir özet ne kadar iyi olursa olsun, dokümanda kısaltma işlemi yapıldığı için bilgi kaybı kaçınılmazdır. Burada dikkat edilmesi gereken, bilgi kaybının dokümanın önemli kısımlarından değil önemsiz kısımlarından oluşmasıdır. Bu şekilde kullanıcıya dokümanın temel olarak ne ile ilgili olduğu ve okunmaya değer olup olmadığı hakkında yol gösterir [40]. Kısaca metin özetleme, bir belgeyi girdi olarak alan ve çıktı olarak daha kısa, aslının yerine geçen ve onun en önemli içeriğini barındıran bir süreç olarak tanımlanabilir [41]. Otomatik metin özetleme temel olarak çıkarıcı ve yorumlayıcı olmak üzere iki ana kısma ayrılmaktadır. Çıkarıcı metin özetlemede girdi olarak verilen cümleler çeşitli puanlandırma teknikleri kullanılarak en önemli cümle bulunur. En önemli cümle ise, anlamsal olarak ana metinin taşıdığı ana fikre en yakın cümle olarak tanımlanabilir. Yorumlayıcı özetleme yöntemlerinde ise, metinler analiz edilerek belirlenen en önemli kelimelere en yakın alternatif kelimer ile farklı cümleler oluşturulmaktadır.

Bu tezin amacı çıkarıcı metin özetleme alanında yapılan çalışmaları daha ileri bir noktaya taşıyarak otomatik metin özetleme alanında daha tutarlı özetler oluşturmak ve bu alandaki çalışmalara katkıda bulunmaktır. Bu tez'de metinlerin ağırlıklı çizgeler olarak temsil edilmesi ,çizgelerin ayrıt ağırlıklarının belirmesi ve elde edilen çizgelerden en uygun özetin bulunmasına yönelik çalışmalar amaçlanmaktadır. Bu amaç doğrultusunda metinlerde bulunan cümlelerin kapsayıcılık ve fazlalık olma durumları da dikkate alınarak gerekli

analizlerin nasıl yapılacağı üzerine öneriler sunulmaktadır. Kapsayıcılık ve fazlalık analizleri yapılırken bir çizgeye ait maksimum bağımsız kümeler kullanılarak cümlelerin önem dereceleri ve özette bulunup bulunamayacağı üzerine yapılan çalışmaların sonuçlarının analizleri yapılacaktır. Ayrıca kullanılan metinlerde spektral kumeleme analizi ile cümleleri kümeleme sonrasında özete konulup konulmayacağı üzerine çalışmalar yapılacaktır. Ek olarak istatistiksel yöntemler ile ayrıt seyrekleştirme işleminin özete etkisi incelenecektir.

Ham veriler ile yapılması düşünlen bütün işlemler bir ön işlem aşamasına ihtiyaç duymaktadır. Ön işlem aşamasında üzerinde çalışılacak metinlerde bulunan hatalı yazımlar, gereksiz kelimeler, kelime yalınlaştırma gibi işlemlerin yapılması gerekmektedir. Bu çalışmada yukarıda belirtilen çalışmalara ek olarak ham bir metin üzerinde ön işlem aşamaları analizin daha verimli yapılması için geliştirilen bir ön işleme aracı ve algoritması hakkında bilgiler verilecektir.

1.2. Tezde Yapılan Çalışmalar ve Literatüre Katkısı

Bu tez kapsamında otomatik metin özetlemeye yönelik literatüre katkıda bulunacağı düşünülen bir birinden farklı Çizge tabanlı iki özetleme yöntemi ile bir metin ön işleme aracı önerilmektedir.Yöntemlerin literatüre bulunduğı katkılar maddeler halinde aşağıda belirtilmektedir.

a) KUSH Metin İşleme Aracı

i. Metinlerde bulunan cümleler arasındaki anlamsal bağımlılığı doğru belirlemek , gereksiz kelime fazlalığını önlemek ve cümlelerde bulunana gizli bağlılıkları tespit edebilmek adına KUSH ismi verilen yeni bir metin önişleme aracı geliştirilmiştir.

ii. Bu araç ile önerilen algoritmalar sayesinde metinlerin çizgeye dönüşüm aşamaları daha az zaman almakta ve daha tutarlı olmaktadır.

iii. Metinde bulunan kelimelerin kök haline dönüşüm aşamasında herhangi bir sözlük veya dil kütüphanesi kullanılmamaktadır.

iv. Herhangi bir dil kısıtlaması olmadan verilen ham metini kendi içerisinde bulunan kelimeler ile yalınlaştırma işlemi yapmaktadır.

b) Özetleme Çalışmalarınında Bulunana Ortak aşamalar

i. Girdi olarak verilen metinlerden metinsel çizgeler (Textual Graph) elde edilmektedir.

ii. Cümleler arasındaki ortak kelime sayıları hesaplanıp oluşturulan ağırlıklı çizgeye ait ayrıt ağırlığı olarak belirlenmektedir.

c) CatSumm: Extractive Text Summarization with Spectral Sentence Clustering and Node Centrality

i. Metinler spectral kümeleme yöntemi ile alt çizgelere ayırılmakta ve böylelikle yakın anlamı taşıyan cümleler kümelenerek birbiri ile güçlü bağlılığı olan cümleler aynı kümede toplanmaktadır.

ii. Özdeğer Ayrışımı ve Fielders vektör kullanılarak çizgeler bölmelenmektedir.

iii. Elde edilen kümelerden yeni alt çizgeler oluşturulup çeşitli istatiksel işlemler yapılarak en önemli cümlelerin tespitleri sağlanmaktadır.

iv. Önemli cümleler belirlenirken Derece merkeziliği, Arasındalık merkeziliği, Yakınlık merkeziliği ve Özvektör merkeziliği ölçümlerinden faydalanılmıştır

d) Bağımsız Küme Özetleme Yöntemi: Extractive Multi-Document Text Summarization based on Graph Independent Sets

i. Daha önce hiç bir özetleme çalışmasında kullanılmamış olan maksimum bağımsız kümeler kullanılarak özetleme işlemi yapılmıştır.

ii. Elde edilen metinsel çizgelere ait maksimum bağımsız kümeler tespit edilerek ana metinden uzaklaştırılmaktadır.

iii. Böylelikla anlam bakımından zayıf olan cümleler yerine daha güçlü olan cümleler üzerinde çalışmaktadır

iv. Anlam olarak en güçlü cümleler tespit edildikten sonra düğüm merkezilik ölçüm değerlerinden olan Özvektör merkezilik değerleri ile cümleleri temsil eden düğümler belirlenerek özet elde edilmektedir.

1.3. Tezin Yapısı

Doktora tezi olarak hazırlanan bu çalışma altı bölümden oluşmuştur. Bölümlerin özeti aşağıdaki gibidir.

Birinci bölümde, teze giriş yapılmış, problem tanımlanmış, geliştirilecek çözümün amacı, etkileri, kapsamı, yapılan çalışmalardan ve literatüre katkısından bahsedilmiştir.

İkinci Bölümde otomatik metin özetleme açıklanmış ve çeşitleri hakkında ayrıntılı bilgiler verilmiştir.

Üçüncü bölümde, Doküman özetleme konusu açıklanmıştır. Özetlemenin tarihsel gelişiminden bahsedilmiş; bu çalışmada kullanılan özetleme yöntemleri ele alınarak literatürde yapılan çalışmalara değinilmiştir.

Ardından Çizge bölmeleme, Spektral Çizge Bölmeleme ve Fielders vektörü ile Özdeğer Ayrışımı açıklanmıştır. Ek olarak önerilen yöntemlerde düğüm önemini belirlemek için kullanılan yöntemlerden düğüm merkezilik ölçümleri açıklanmıştır.

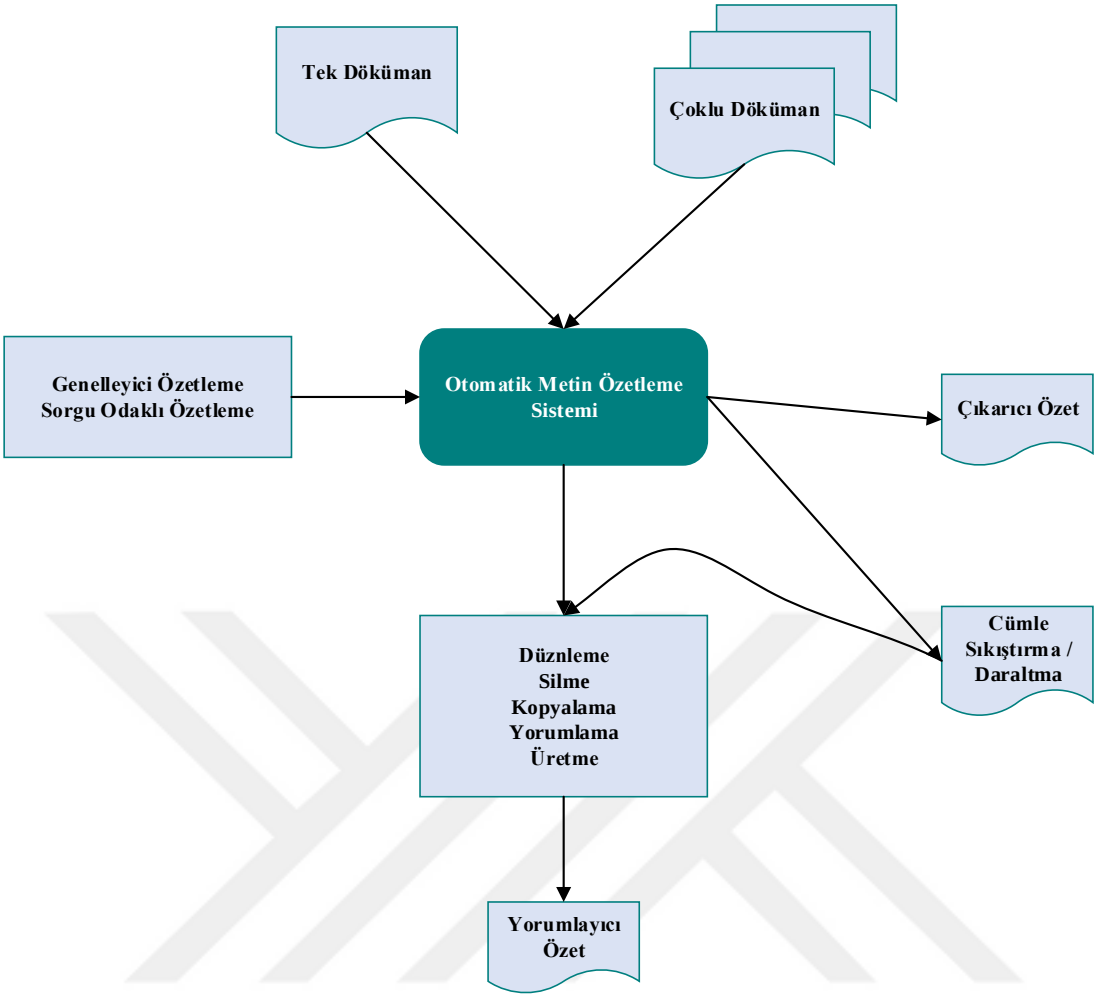
Sonrasında dördüncü bölümde Çizgelere ait Maksimal ve Maksimum bağımsız kümeler hakkında bilgiler verilmiştir.

Beşinci bölümde önerilen her iki yöntemden ayrıntılı olarak bahsedilmiş, algoritma akış diyagramı ve hesaplama adımları açıklanmıştır. iki farklı veri seti kullanılarak önerilen yöntemin uygulanabilirliği, geleneksel yöntemlerle kıyaslaması ve doğruluğu gösterilmiş, yorum ve açıklamalar yapılmıştır. Sonuçlar tablolar ve görsel şekillerle gösterilmiştir. Her bir veri seti için literatürde kullanılan önemli yöntemler olan Random, Luhn, LSA, TextRank, LexRank, SumBasic ve KL-Sum değerleri ile üç farklı boyuta sahip özetler bulunarak ROUGE metrik ölçümleri kullanılarak karşılaştırmalar yapılmıştır. Son olarak altıncı bölümde ise tez çalışması ile elde edilen sonuçlar ve öneriler yer almaktadır.

2. METİN ÖZETLEME

Metin özetlemenin amacı orijinal dokümanın genel anlamını kapsayan ve önemli noktaları içeren kısa bir halinin üretilmesi işlemidir. Özetleme kullanıcılara büyük boyutlu verilerle uğraşmak yerine küçük boyutlu verilerden ana fikri anlama kolaylığı sunar [42]. Doküman özetleme alanına yapılan çalışmalar ilk olarak 1958 yılında Luhn tarafından terim frekansları dikkate alınarak yapılan çalışma ile başlamıştır [43]. Bu çalışmada, Dokümanlarda bulunan terimlerin tekrarlanma sayıları ve pozisyonlarının metinin temsili konusunda önem arzettikleri düşünülmüştür. Ayrca Luhn tarafından yapılan bu çalışma çıkarıma dayalı metin özetleminin de temelini oluşturmaktadır. Çıkarıma dayalı metin özetleme yönteminden sonra metin özetleme çok farklı alanlarda ve farklı amaçlarla kullanılmaya başlanmış ve bu alanlarda daha tutarlı özetler oluşturmak için çeşitli çalışmalar yapılmıştır.

Öyleki otomatik metin özetleme ilk ortaya çıktığı dönemdeki amacından farklı olarak şu anda bilgisayar bilimi, yapay zeka, istatistik, bilişsel bilimler, doğal dil işleme, makine öğrenmesi, dilbilim analizi gibi bir çok farklı bilim alanında kullanıldığı için disiplinler arası bir araştırma alanı olmuştur [44]. Herhangi bir yapıya sahip olmayan tamamen ham verilerden oluşan metin dökümanlarının özetlenmesi oldukça zor bir görevdir. İnsanlar bir metini okuyarak metinde anlatılmak istenen ana fikri tam olarak anlayabilmek için dökümanı sonuna kadar okuma ihtiyacı duymakta ve gerek dilbilimsel gerekse kelimelerin taşıdığı anlamlarıda analiz ederek bir özete gidebilmektedir. Oysaki bilgisayarlar insanların sahip oldukları gerek dilbilgisi gerekse anlam bilgisi gibi özelliklere sahip değillerdir. Bu da ham bir dökümandan doğru bir özete ulaşmanın oldukça zor bir araştırma alanı olmasına neden olmaktadır [45]. Otomatik metin özetlemeye ait kavramsal diyagramı Şekil 2.1 gösterilmektedir.



Şekil 2.1. Otomatik metin özetlemeye ait kavramsal diyagram

2.1. Metin Özetleme Türleri

Otomatik Metin Özetleme işlemi, farklı durumlara göre çeşitli kategorilerde sınıflandırılabilir. Bu durumlar, kullanılan doküman sayısını, istenilen çıktıyı, amacı veya içeriği, eğitim verilerinin kullanılabilirliğini, dili vb. içerir [46]. Aşağıdaki bölümler ve kategoriler, metin özetlemesinin farklı faktörlerini açıklar.

2.2. Çıkarıcı ve Yorumlayıcı Metin Özetleme

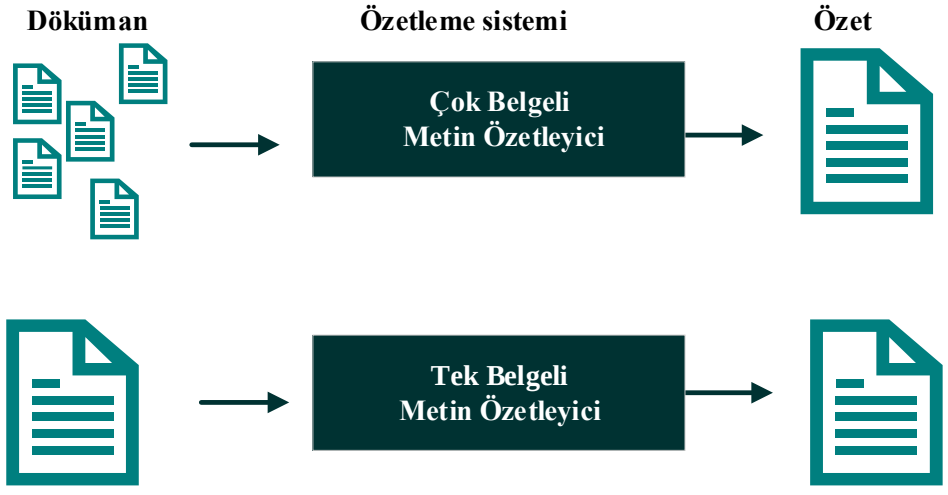
Çıkarıma dayalı özetleme sistemleri genellikle cümleleri terim frekansı ve ters doküman frekansı (TF-IDF), cümlelerin metin içerisindeki pozisyonu, anahtar kelime sayısı gibi önceden tanımlanan özelliklerin bir kümesine bağlı kalarak cümleleri bir puanlamaya tabi tutmaktadır. Puanlanan cümle veya cümlecikler sıralanır ve en yüksek puana sahip metin

birimleri ile istenen boyuta sahip özet oluşturulur. Sözü edilen sistem yorumlayıcı özetleme sistemine göre daha yalın adımlardan oluşmaktadır[47],[1], [12], [48].

Yorumlayıcı metin özetleme sistemlerinde ise çıkarıma dayalı özetleme sitemlerinden farklı bir durum söz konusudur. Bu tür sistemlerde oluşturulan özette bulunan cümleler orijinal metinden tamamen farklıdır. Çıkarıcı metin özetlemede temel amaç öncelikle orijinal dokümanın konseptini veya taşıdığı ana fikri anlamaktır. Bu işlemlerin yapılması için çeşitli dilbilimsel analizler yapılmaktadır [34], [38], [49]. Yorumlayıcı metin özetleme çalışmalarından elde edilen özetlerin tutarlılıklarının çıkarımsal metin özetlemeye oranla daha iyi olmasının yanında çıkarıcı metin özetlemede özeti oluşturma aşamalarında oldukça karmaşık doğal dil işlme süreçleri uygulanmaktadır. Çıkarımsal özetleme yöntemlerinde ise, böyle bir durum söz konusu olmamaktadır.

2.3. Tek belgeli ve Çok Belgeli Özetleme

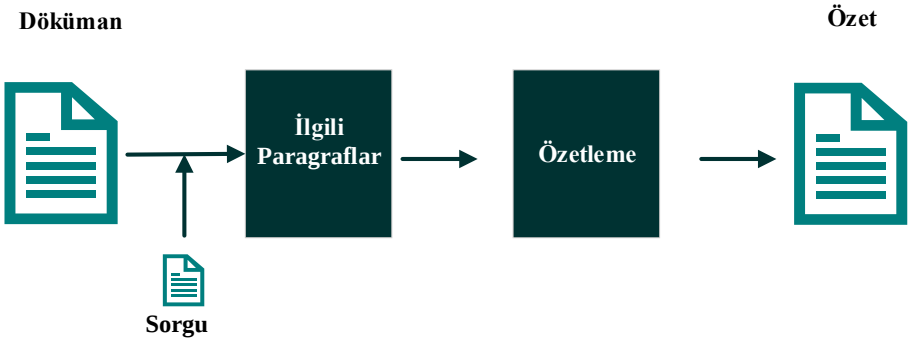
Şekil 2.2 de görüldüğü üzere, metin özetleme sistemleri girdi olarak aldığı doküman sayısına görede sınıflandırılabilmektedir. Tek belgeli özetlemelerde özetleme sisteminde girdi olarak tek bir belge verilmekte ve özet elde edilmekte iken çok belgeli özetleme sistemlerinde birden fazla doküman özetlenmek üzere sisteme girdi olarak verilebilmektedir. Çoklu metin özetlemede sistemin tutarlı sonuçlar üretebilmesi için girdi olarak verilen dokümanların aynı konu başlığı altında toplanmış veriler olması gerekmektedir.



Şekil 2.2. Belge sayısına göre metin özetleme

2.4. Genel ve Sorgu Odaklı Özetleme

Metin özetleme sistemlerinde en önemli zorluklardan biri de doküman içerisinden istenilen herhangi bir sorguya uyan bir özet elde etmektir. Bu probleme yönelik olarak sorgu odaklı metin özetleme sistemi önerilmektedir. Sorgu odaklı çoklu belge özetleme, otomatik çoklu belge özetleme sisteminin özel bir çeşididir. Şekil 2.3'te görüldüğü gibi bu tür özetlemelerde kullanıcı tarafından verilen bir sorguya yönelik orijinal metinde bulunan yanıtlar aranmaktadır [50]. Bu tür özetleme sistemleride çıkarıcı metin özetleme yapısını kullanmaktadır. Sorguya en uygun kelime, kelime öbekleri veya cümleler orijinal dokümandan alınarak özete konulmaktadır. Sorgu odaklı özetleme sistemlerinin aksine genel özetleme sistemleri verilen orijinal dosyanın ana fikrini içeren genel bir özet oluşturmakadır [51]. Genel özetleme sitemlerinde bütün dosyaların ana fikirlerini bulup özete koymak için zorlu süreçler gerekmektedir. Bu tür yöntemlerde genel olarak iki temel kısıtın karşılanması gerekmektedir. Bu kısıtlardan ilki minimum fazlalık değeridir. Minimum fazlalık değeri özette bulunan cümlelerin en az düzeyde birbirine benzemesini sağlayarak özet içinde aynı anlamı taşıyan cümlelerin tekrarının önlenmesini amaçlamaktadır. Diğer bir kısıt ise, maksimum kapsama değeridir. Maksimum kapsama değerinde özette bulunacak cümlelerin orijinal cümlenin taşıdığı ana fikre mümkün mertebede en yakın cümleleri gerektirmektedir. Bu iki kısıtın en iyi şekilde sağlanabilmesi için optimizasyon tabanlı çalışmalar yapılmakta ve kayda değer iyileştirilmeler elde edilmektedir [52].



Şekil 2.3. Sorgu odaklı metin özetleme

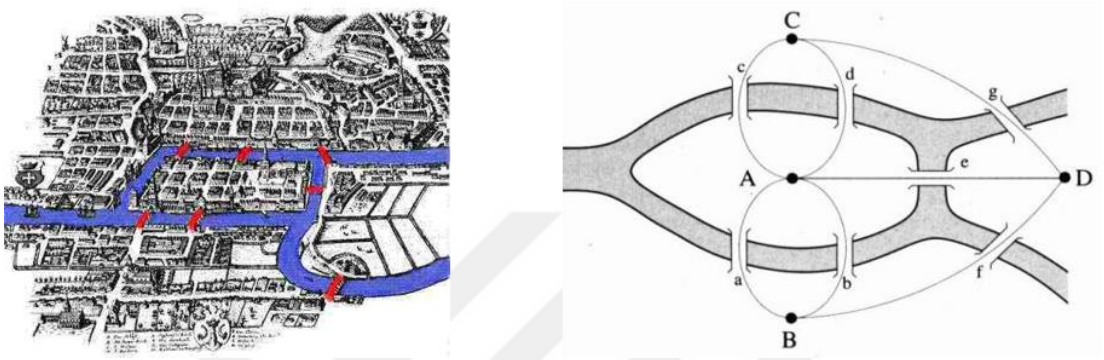
2.5. Denetimli ve Denetimsiz Özetleme

Özetleme sistemleri ayrıca denetimli veya denetimsiz özetleme sistemleri olarak iki kısıma ayırılabilmektedirler [53]. Denetimli metin özetleme sistemlerinde, belgelerden önemli olan içeriği seçmek için eğitim verisine ihtiyaç duyulmaktadır. Özetleme sistemini

eğitmek için büyük miktarda eğitim verisine ihtiyaç duyulmaktadır. Küçük boyutlu eğitim verilerinde sistem cümleler ve kelimelerin özetler ile ilişkilerini yeterli şekilde öğrenemeyeceği için doğru sonuçlar üretmekte zorluk çekecektir. Bu yüzden öğrenme tekniklerinin uygulabilmesi için büyük miktarda etiketlenmiş veya açıklamalı veri gereklidir. Bu sistemler cümle düzeyinde, özete ait cümlelerin pozitif örnek olarak ve özette bulunmayan cümlelerin negatif örnek olarak adlandırıldığı iki sınıflı sınıflandırma problemi olarak ele alınmaktadır [54]. Denetimsiz özetleme sistemlerinde ise, herhangi bir ön bilgiye ya da eğitim verisine de ihtiyaç duyulmamaktadır. Bu sistemlerde sadece orijinal doküman dikkate alınarak özetleme yapılmaktadır. Bu tür özetleme sistemleri genellikle önceden kullanılmamış ve yeni veri setleri için uygun görülmektedir.

3. ÇİZGELER

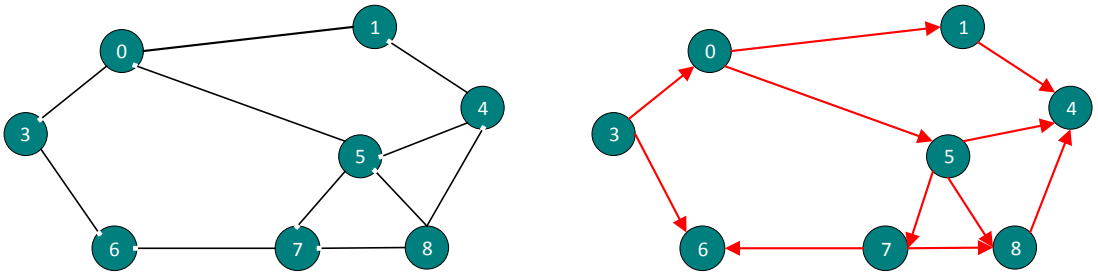
Çizge teorisi, ünlü matematikçi Leonhard Euler tarafından 1736 yılında Şekil 3.1'de gösterilen Kneiphof adasında bulunan 7 adet köprüden sadece birer defa geçmek koşulu ile bütün adanın gezilip gezilemeyeceği sorusu üzerine ortaya çıkmıştr [55]. Euler, problemi çözerken somut bir olay modelleyip soyut bir şekle dönüştürerek teorinin temellerini atmış ve ardından teori Hamilton, Heawood, Whitney, Tutte, Ore, Erdös, Katona, Polya, Seshu ve Frank Harary gibi matematikçilerin eşsiz katkılarıyla bugünkü durumuna gelmiştir [56].



Şekil 3.1. Königsberg köprüleri

Çizge teorisi ile ilgili uygulamlar günlük hayatın karmaşık ve birbirinden farklı alanların problemlerinin çözümünde sıklıkla kullanılmaktadır. Çizgelerin kullanıldığı alanlar arasında ekonomi, sosyal bilimler, bilgi iletimi, sosyal ağ analizi, ağ (network) tasarımları gibi çeşitli disiplinler bulumaktadır. Çizge teorisi yapısal olarak matematiksel bir alt yapıya sahip olduğu için problemlerin tanmlanması ve nesneler arasındaki ilişkilerin belirlenmesinde kullanılmaktadır.

Çizgeler ayrıt ve düğüm çiftlerinden oluşan matematiksel yapılardır. Düğümler nesnelerin temsili için kullanılmakta ve literatürde "node" veya "vertex" olarak isimlendirilmektedirler. Ayrıtlar ise, $V \; x \; V$ nin (düğümler kümesinin kartezyen çarpımının alt kümesi veya düğümler kümesinden düğümler kümesine olan bağıntılardan biri şeklinde tanımlanabilir) bir alt kümesi olarak ifade edilmektedirler yani iki düğüm arasında oluşan bağlantı ayrıt olarak tanımlanmakta ve literatürde "Edge" olarak isimlendirilmektedirler [57]. Oluşturulan çizgelerdeki düğümler arasında bulunan ayrıtların yapılarına göre farklılık gösterebilmektedirler. Şekil 3.2'de gösterildiği üzere çizgelerde bulunan ayrıtlar sadece bir düğümden diğer bir düğüme gidildiğini belirtiyorsa, bu tur çizgelere yönlü (directed) çizgeler denilmektedir. Aksi durumda herhangi bit yön söz konusu değil ise yönsüz (undirected) çizgeler olarak isimlendirilmektedir.



Şekil 3.2. Yönlü ve yönsüz çizgeler

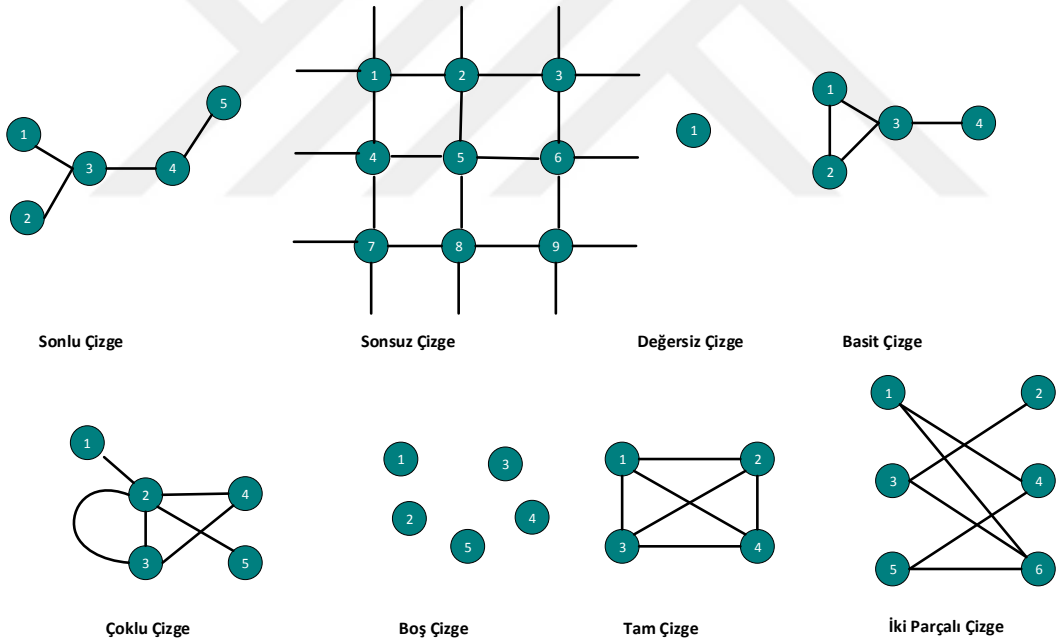
3.1. Çizge Çeşitleri

Bir çizgeyi tanımlayabilmek için öncelikle düğümlerin ve ayrıtların kümesini belirlemek gerekmektedir. Sonrasında ayrıt (ayrıt) kümesinde bulunan her bir ayrıtın sorumlu olduğu düğümler belirlenmeli ve düğümler bu ayrıtlar yolu ile bir birine bağlanmalıdır. Her bir ayrıtın iki ucu da bir düğümle sonlandırılmak zorundadır. Bu yüzden ayrıtlar için öncelikle sorumlu oldukları düğüm kümeleri belirlenir.

Basit bir çizge $G = (V, E)$ şeklinde tanımlanmaktadır [58], [59]. Burada V düğümler kümesini E ise ayrıtlar kümesini belirtmektedir. Düğüm ve ayrıtlar kullanılarak ayrıtların yönlerine, sayısına ve taşıdıkları değerlere göre çizgeler bir birinden farklılaşmaktadırlar.

$G = (V, E)$ şeklinde tanımlanan bir çizge sınırlı sayıda düğüm ve ayrıttan oluşuyorsa, bu tür çizgelere **sonlu çizgeler** denilmektedir. Sonlu çizgeler, düğüm kümesinin ve ayrıt kümesinin sonlu kümeler olduğu bir çizgedir. Aksi takdirde, buna **sonsuz çizge** denir. Yaygın olarak çizge teorisinde kullanılan çizgelerin sonlu olduğu kabul edilir. Çizgeler sonsuzsa, bu özellikle belirtilir [60]. Sonlu bir çizgede yalnızca bir düğüm bulunuyorsa ve ayrıt yoksa bu çizgenin **önemsiz** çizge olduğu kabul edilir [61]. Çizge teorisinde ilk olarak ortaya atılan çizge modelleri basit çizgeler olarak isimlendirilmektedir. Var olan bir çizgede bulunan düğümler arasında birden fazla ayrıt yok ise, bu tur çizgelere **basit** çizgeler denilmektedir. Günlük hayatta karşılaşılan bir çok problemin çözümünde ve matematiksel modellemesinde basit çizgelerin yetersiz kaldığı görülmektedir. Bu yüzden çözümlerin modellenmesinde daha başarılı sonuçlar ürettiği görülen **çoklu** çizgeler kullanılmaktadır. Çoklu çizgelerde iki düğüm arasında birden fazla ayrıt bulunabilmektedir fakat kendi kendine bir ayrıt oluşturan döngü ayrıtı bulunmamaktadır [59]. Bu çizgelerin tersi olarak paralel ayrıtları bulunmayan ve kendi kendine bir döngü şeklinde arıtları bulunan çizgelere ise **sahte** çizgeler

denilmektedir. Çizgede bulunan bütün düğümlerin bir ayrıt ile bir birlerine bağlantısı bulunuyor ise, bu tur çizgelere **tam çizge** denilmektedir [62]. Benzer şekilde bir çizgedeki bütün düğümler eşit dereceye sahip ise, bu çizgeye **düzenli çizge** denilmektedir. Yukarıda bahsedildiği gibi bütün düğümlerin bir birlerine bağlı olma zorunluluğu yoktur. Düğüm kümesindeki düğümler iki ayrı kümeye ayrılabiliyor ve her kümedeki herhangi bir düğüm sadece diğer kümedeki düğümlerle ortak bir ayrıta sahipse, bu tür çizgelere iki **parçalı** çizgeler denilmektedir [63]. Çizgeler çoğunluka ayrıt ve düğümlerin bir birleri ile olan ilişkileri dikkate alınarak sınıflandırılmaktadır. Bu ilişkilerin dışında ayrıt özellikleri dikkate alınarak yapılan sınıflandırmalar vardır. Öyleki bir ayrıtın yönünde ve taşıdığı değere göre çizgeler farklılık göstermektedir. Düğümler arasında bulunan bir ayrıtınn herhangi bir düğümden diğer düğüme bir yön bilgisi içeriyorsa, bu tür çizgelere **yönlü** çizgeler denilmektedir. Benzer şekide ayrıtların herhangi bir ağırlık değeri yani o ayrıtın gücünü belirten bir özellik mevcut ise, bu tür çizgelere ise **ağırlıklı** çizgeler denilmektedir. Çizge çeşitlerine ait grafiksel gösterim Şekil 3.3 te ayrıntılı olarak sunulmaktadır.



Şekil 3.3. Çizge çeşitleri

3.2. Metinsel Çizge

Özellikler ve bu özellikler arasındaki ilişkilerin şekilsel olarak temsil edilmesi ile problemlerin çizgeler ile modellenmesi gerçekleştirilebilmektedir. Çizgeler, kavramsal olarak düğümlerden ve düğümler arası ilişkileri temsil eden ayrıtlardan oluşmaktadır. Düğümler ve ayrıtlar iki sonlu kümedirler. Düğümler, çizgenin temsil ettiği topluluğu

oluşturan esas bireylerdir. Ayrıtlar ise, esas bireyler arasındaki söz konusu ilişkilerdir. Genel olarak çizge $G = (V, E)$ ile gösterilmektedir. Düğümler kümesi V={v_1, v_2,...,v_n}, ayrıtlar kümesi ise E={e_1,e_2,…e_n} (E⊆V**X**V) dir. e_i={v_j,v_{j+1}} ayrıtı için uç düğümler v_j ve v_{j+1} düğümleridir. v_j ve v_{j+1} komşu düğümleri için e_1=(v_j,v_{j+1})∈ E ve (v_{j+1},v_j)∈ E ise, bu ayrıtlar yönsüz ayrıtlardır. Bu tarz ayrıtların oluşturduğu çizgeler yönsüz çizgelerdir. Eğer v_j ve v_{j+1} komşu düğümleri e1=(v_j,v_{j+1})∈ E ve (v_{j+1},v_j)∉ E ise, bu tür ayrıtlar yönlü ayrıtlardır. Yönlü ayrıtların oluşturduğu çizgeye yönlü çizge adı verilmektedir [5], [64]. Bir G=(V,E) çizgesi üzerinde V={v_1, v_2,...,v_n} düğümleri arasındaki ilişkileri temsilen iki düğüm arasındaki ayrıtlar E={e_1,e_2,…e_n} eksi olmayan ağırlıklar taşıyor ise, bu tarz çizgeye ağırlıklı çizge adı verilmektedir [65]. Bu çalışmada metinleri temsilen oluşturulan çizgeler yönlü ve ağırlıklı çizgelerdir.

Metinsel çizgelerin temsil edilmesi amacıyla literatürde farklı yaklaşımlar söz konusudur. Tam bir cümle veya bir cümlecik düğümler ile temsil edilebilirken, düğümler arasındaki kesişim, birlikte oluşumlar gibi farklı ilişki şekilleri ise, ayrıtlar ile temsil edilebilmektedirler.

Genel olarak bir çizgede yer alan düğümlerin metinleri temsili noktasında;

- ✓Kavramlar
- ✓Terimler
- ✓Cümleler
- ✓Paragraflar
- ✓Belgeler

düzeyinde karşılıkları olabilmektedir.

Düğümler arasındaki ilişkileri temsil etmesi için ise;

- ✓Anlamsal ilişkiler
- ✓Kelime birliktelikleri
- ✓Çerçeve modeli (belli boyutlara sahip dizelerdeki yer alma sıklıkları)

gibi farklı yaklaşımlara sahip temsil tiplerine rastlanmaktadır.

Bu çalışmada metinleri temsil eden çizgeler oluşturulmuştur. Temsili çizgelerdeki sonlu kümelerden ilkini yani düğümleri cümleler temsil etmekte iken, diğer sonlu küme olan ayrıtları ise, ortak kelime sayıları temsil etmektedir. Böylece metinlerin yönlü ve ağırlıklı çizgeler ile temsili sağlanabilmektedir. İfade şekilleri bakımından ise;

- ✓ Dinamik olarak ifade

- ✓ Matrislerle ifade

gibi farklı temsil tipleri mevcuttur.

Bu farklı ifade şekillerinin kendilerine göre avantaj ve dezavantajları söz konusudur. [64] de Karcı'nın da ifade ettiği gibi çizgeleri dinamik olarak ifade etmek hafıza kullanımı noktasında avantaj sağlar ve çalışma anında çizgelere ait düğüm ve ayrıt sayılarına müdahale edebilme imkânı sağlar. Ancak bu avantajlarına karşılık kullanım açısından bazı zorlukları beraberinde getirmektedir. Her düğüm ve ayrıt için hafızada yer ayrılmaktadır. İlgili yöntemde bir düğümler kümesi ve bu düğümlerin komşuluklarının mevcut olduğu düğümler listesi veya ayrıtlar listesi şeklinde tanımlanmaktadır. Herhangi bir düğümün komşuluk kümesi için bağlı liste kullanılır. Tüm bu durumlar dinamik yaklaşımı hafıza kullanımı noktasında avantajlı kılmaktadır ancak dinamik yaklaşımlar kullanım açısından çok komplekstirler.

G=(V,E) gibi bir çizgenin düğüm etiketlerinin matris indisleri olarak kullanıldığı komşuluk matrisi $A(G)=(a_{i,j})$ için *v x v'*lik simetrik bir matristir ve Denklem 3.1'de ki gibi tanımlanır. Burada $w_{i,j}$, iki düğümü birbirine bağlayan ayrıt ağırlığıdır.

$$A(G) = \begin{cases} w_{i,j} & , \quad (i,j) \in E \\ 0 & , \quad (i,j) \notin E \end{cases} \tag{3.1}$$

G=(V,E) gibi bir çizgenin derece matrisi $D(G)=(d_{i,j})$, *vxv'*lik köşegen bir matristir ve Denklem 3.2 'de belirtildiği gibi tanımlanır.

$$D(G) = \begin{cases} d_{i,j} & , \quad i = j \\ 0 & , \quad i \neq j \end{cases} \tag{3.2}$$

Metinler içerisindeki her cümle için temsili çizgeye bir düğüm eklenmiştir. Düğümler arasındaki ayrıtlar için ise, metinde bulunan cümlelere dair kesişen kelime sayıları dikkate alınarak ayrıt ağırlığı eklenmiştir. Şayet cümleler arasında kesişen kelime bulunmamakta ise, ilgili düğümler arasına ayrıt eklenmemektedir. Dilbilimsel bir süreç izleyen KUSH algoritması cümleler arasındaki ilişkinin anlamsal doğruluğunu düzenlemekte, böylelikle düğümler anlamlı ilişkiler ile birbirlerine bağlanmaktadırlar. Böylelikle bütün cümleler, tüm kombinasyonları kapsayacak şekilde birbirleri ile olan ortak içerikleri bakımından ilişkilendirilmiş ve cümleler arasındaki ilişkiler çizgeye doğru bir şekilde aktarılabilmiştir.

(0) Erich Honecker, the former GDR head of state, died at his house in Santiago, Chile on Sunday morning [29 May], according to his lawyer.

(1) Lawyer Nicolas Becker, who had represented 81-year-old Honecker before the Berlin court in 1992 and early 1993, told DPA on Sunday afternoon that Honecker had rejected an operation.

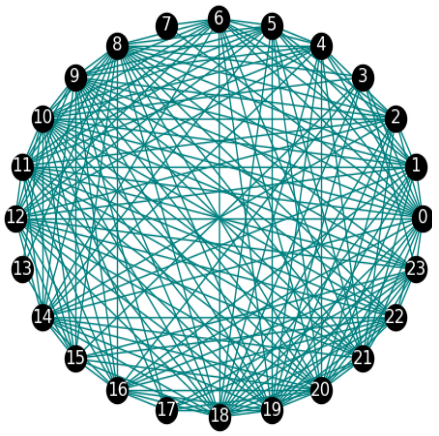
(2) Becker could not comment on the exact cause of Honecker's death.

.....

.....

(22) "You traitor!" 82-year-old former security chief Erich Mielke shouted at Honecker, according to Bild newspaper.

(23) Mielke also screamed at Honecker "I'll break your neck!" the newspaper quoted witnesses as describing the scene when the two met in an East German jail.

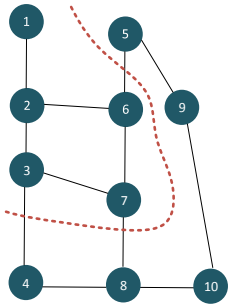


Şekil 3.4. Metinsel çizge gösterimi

Şekil 3.4'te veri önişleme ve hazırlık aşamalarının sonrasında oluşturulan örnek bir paragraf ve bu paragrafa karşılık gelen çizge görülmektedir. Örnek paragraf DUC 2002 Veri seti'nde yer alan örnek metinden oluşmaktadır.

3.3. Çizge Bölmeleme

Bir çizge kümesi içindeki benzerliklerin mümkün olduğunca büyük ve kümeler arasındaki benzerliklerin ise, mümkün olduğunca küçük olduğu optimal bir bölüm bulmayı amaçlayan yönteme çizge bölmeleme adı verilmektedir. Çizge bölmeleme devre tasarımı, yük dengeleme, optimizasyon uygulamaları, paralel hesaplamalar gibi konularda uygulama alanı bulan önemli bir problemdir. Çizgelerle ifade edilen veriler üzerinde yapılan bilimsel hesaplamalarda cebir, olasılık kuramı, ergodik teori, geometri gibi dallarla ilişkili olan problemler mevcuttur. Bu tarz problemlerin çözümünde verimliliğin artması, sürecin yönetilebilirliğinin sağlanması için problem uzayının ayrık alt kümelere bölünmesi gerekmektedir. Bu amaç doğrultusunda verinin temsil ettiği çizgenin Şekil 3.5'de görüldüğü gibi en maliyetsiz şekilde bölmelenmesi gerekmektedir. Problemi oluşturan verilerden yola çıkılarak çizge oluşturulmakta ve elde edilen çizgenin bölmeleme süreçleri farklı yöntemlere göre gerçekleştirilmektedir. Çizge bölmeleme problemi üstel (exponential) matematiksel karmaşıklığa (complexity) sahiptir yani NP-zor (NP-hard) tipi bir problemdir. Farklı çizge bölmeleme algoritmaları mevcut olmakla birlikte bu algoritmaların hiçbiri ideal çözümü garanti edememektedir. Buna rağmen sunulan çizge bölmeleme yöntemlerinin büyük kısmı iyi sonuçlar verebilmektedirler [64], [66]–[68].



Şekil 3.5. Çizge bölmelenmesine dair temsili gösterim

Spektral kümeleme fikri spektral çizge bölmeleme teorisinin yöntemlerini kullanır. Donath ve Hoffman benzerlik matrislerinin öz değer vektörleri ve çizge bölmeleme problemleri arasında bir ilişki olduğunu ileri sürmüştür. Fiedler ise çizge bölmelemenin, Laplacian matrisinin ikinci en küçük öz değer vektörü ile yakın bir şekilde ilişkili olduğunu ispatlamıştır [68]. Benzer şekilde [69]'da yer verilen Rayleigh-Ritz teorisine göre de, Laplacian matrisinin öz değer vektörleri çizge bölmelemenin optimal çözümüne yaklaşmak için kullanılabilir olduğuğu belirtilmiştir.

Bir veri seti veya topluluğu temsilen yönlü ve ağırlıklı bir *G=(V,E)* çizgesi oluşturulur. *V={v₁, v₂,...,vₙ}*, düğümler kümesi *E={e₁,e₂,...eₙ}* ise düğümler arasındaki ağırlıklı ayrıtlar kümesidir. G çizgesine ait çizge bölmeleme problemi Denklem 3.3 ve Denklem 3.4'te ki gibi tanımlanmaktadır.

$$kesim(A,B) = \sum_{i\in A, j\in B} w_{ij}\,, \bar{A} = \{V_p \mid V_p \in V \quad ve \quad v_p \notin A_i\} \tag{3.3}$$

$$kesim(A_1 \dots A_k) = \sum_{i=1}^{k} cut(A_i, \bar{A}_i) \tag{3.4}$$

İyi bir çizge bölmeleme işlemi amaç fonksiyonunu minimum yapar. Ancak bu amaç fonksiyonu kümelerdeki yoğunluk dağılımını hesaba katmaz ve sadece kümelerin harici bağlantılarını dikkate alır. Bu durum veri setindeki aykırı ve dengesiz verilerden ötürü, dengesiz veri kümelerinin oluşmasına neden olabilmektedir. Ancak [70]'de Hagen ve Kahng daha dengeli veri kümeleri elde edebilmek amacı ile Denklem 3.5'te gösterilen oransal kesim yontemini önermişlerdir. Oransal kesim amaç fonksiyonu elde edilecek kümeler arasındaki benzerlikleri en aza indirgemek için kesim fonksiyonuna bir dengeleme değişkeni olarak küme boyutunu ($\bar{A}_i$) eklemeyi ileri sürmüştür. Böylelikle dengesiz bir şekilde bölünmenin önüne geçmeye çalışılmaktadır.

$$Oransal\ kesim(A_1 \dots A_k) = \sum_{i=1}^{k} \frac{kesim(A_i, \bar{A}_i)}{\bar{A}_\iota} \tag{3.5}$$

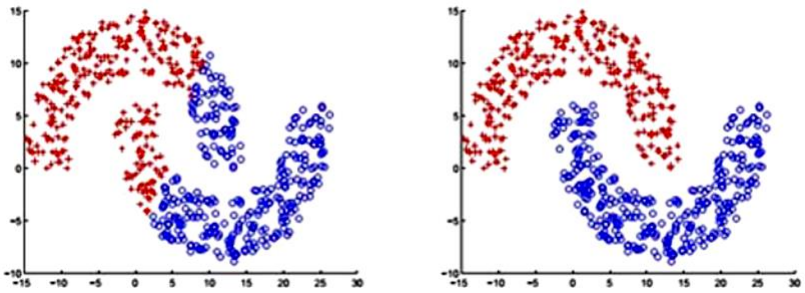
Ancak oransal kesim kümelerin benzerliklerini daha yüksek başarım ile elde etmek için yeterli olmamaktadır. Bu nedenle Shi ve Malik, Denklem 3.6'daki gibi tanımladıkları normalleştirilmiş kesimi önermektedirler. Amaç fonksiyonu minimuma indirgemek, alt çizgeler arasındaki toplam ağırlıkları minimuma indirmek ve her bir alt çizgenin iç ağırlıklarını aynı anda en üst düzeye çıkarmak anlamına gelir; bu nedenle yaklaşım daha iyi sınıflandırma sonuçları üretebilmektedir [68], [70], [71].

$$vol(A_i) = \sum_{p \in A, q \in V} w_{pq}\,, Nkesim(A_1 \dots A_k) = \sum_{i=1}^{k} \frac{kesim(A_i, \bar{A}_\iota)}{vol(A_i)} \tag{3.6}$$

Bu tez kapsamında yapılan ilk çalışma çizge bölmeleme yöntemlerinden olan spektral çizge bölmeleme yöntemi kullanılmakta, çizgenin öz değerleri ve bu öz değerlere karşılık gelen öz değer vektörleri elde edilerek çizge bölmeleme işlemi gerçekleştirilmektedir.

3.4. Spektral Çizge Bölmeleme (Spectral Graph Partitioning)

Son zamanlarda çizgelerin genel yapısal özelliklerini karakterize etmek için spektral çizge teorisinin kullanımına olan ilgi artmıştır. Spektral çizge teorisi, komşuluk matrislerinin veya Laplacian matrislerinin öz değer vektörlerini kullanarak çizgelerin genel ve yapısal özelliklerini özetlemeyi amaçlamaktadırlar. Birçok nümerik algoritma için algoritmanın önemli bir parçasını spektral yöntemler oluşturmaktadır. Spektral yöntemler matris bölmelemede, çizge bölmelemede ve devre simülasyonlarında sıklıkla kullanılmaktadır [64], [67]. Spektral kümeleme algoritmaları, Şekil 3.6'da görüldüğü üzere k-ortalamalar ve diğer kümeleme algoritmalarının aksine bağlantısallık noktasında düzensiz biçimlere sahip olan nesneleri gruplayabilmektedir.



Şekil 3.6. K-ortalamalar ve Spektral Kümeleme yönteminin karşılaştırılması [72]

Spektral kümeleme, spektral çizge teorisindeki kavramları kullanmaktadır. Kümeleme problemi, uygun bir amaç fonksiyonunu optimize ederek bir çizge kesim problemi olarak yapılandırmaktadır. Temel fikir hazırlanan veri setinden bir çizge inşa etmektir. Veri setindeki her düğüm bir ilişkiyi temsil eder ve her ağırlıklı ayrıt basit bir şekilde iki ilişki arasındaki benzerliği hesaba katar. Bu yapıda kümeleme problemi, bir çizge kesim problemi olarak düşünülebilir ve bu da aslında spektral çizge teorisi olarak ele alınmaktadır. Bu teorinin temeli veriden elde edilen ağırlıklı çizgenin Laplacian matrisinin öz değer vektörlerinin ayrışımının gerçekleştirilmesidir [73]. Literatürde Basit Laplacian, normalleştirilmiş Laplacian, genelleştirilmiş Laplacian, gibi çok sayıda Laplacian hesaplama yöntemi mevcuttur. Bu Laplacian türlerinin tümü düğümlerdeki dereceleri ölçen köşegen derece matrisini kullanmaktadırlar. İlişki matrisini kümelere göre dengelemek için bir normalleştirme etkeni olarak derece matrisi kullanılmaktadır [74], [75]. Herhangi bir çizgeye ait Laplace matrisi reel ve simetrik bir matristir. İlk olarak [76]'de Fiedler'in önerdiği yönteme göre matrisin ikinci en küçük öz değerine karşılık gelen öz değer vektörleri çizgenin cebirsel bağlılığı olarak düşünülmekte ve çizgenin iki parçaya bölünmesinde kullanılabilmektedir. Fiedler'in bu önerisinden ötürü ikinci en küçük öz değeri matrisin Fiedler değeri olarak adlandırılmaktadır. Çizgenin minimum ayıracını tutan öz değer vektörüne de Fiedler vektörü adı verilmektedir.

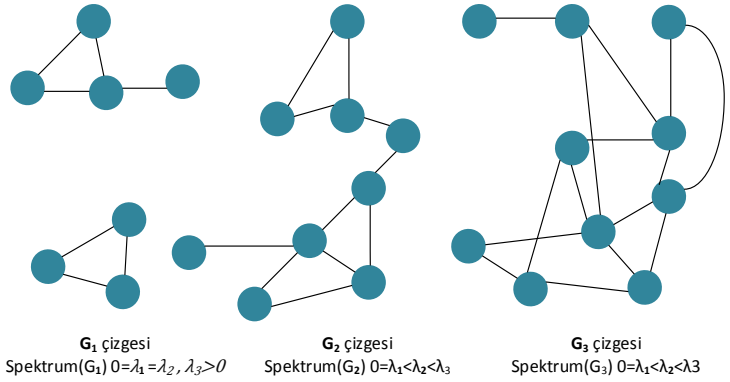
3.5. Laplace Matrisi Ve Fiedler Vektörü

G=(V,E) gibi bir çizgenin *V* düğümler kümesi, *E* ayrıtlar kümesidir. *G* çizgesinin */V/=n* için *nxn* boyutunda komşuluk matrisi Denklem 3.1'de ifade edilmektedir. G çizgesinin nxn boyutuna sahip olan derece matrisi Denklem 3.2'de olduğu gibi hesaplanmaktadır. Derece matrisi köşegen bir matristir. Basit Laplace hesabı çizgenin komşuluk matrisi ile köşegen derece matrisi kullanılarak Denklem 3.7 'de ki gibi hesaplanırken Laplace matrisinin ayrı ayrı bileşenleri ise Denklem 3.8'deki gibi hesaplanmaktadır.

$$L = D - A \quad , \quad d_i = \sum_{\{j|(i,j)\in E\}} w_{ij} \tag{3.7}$$

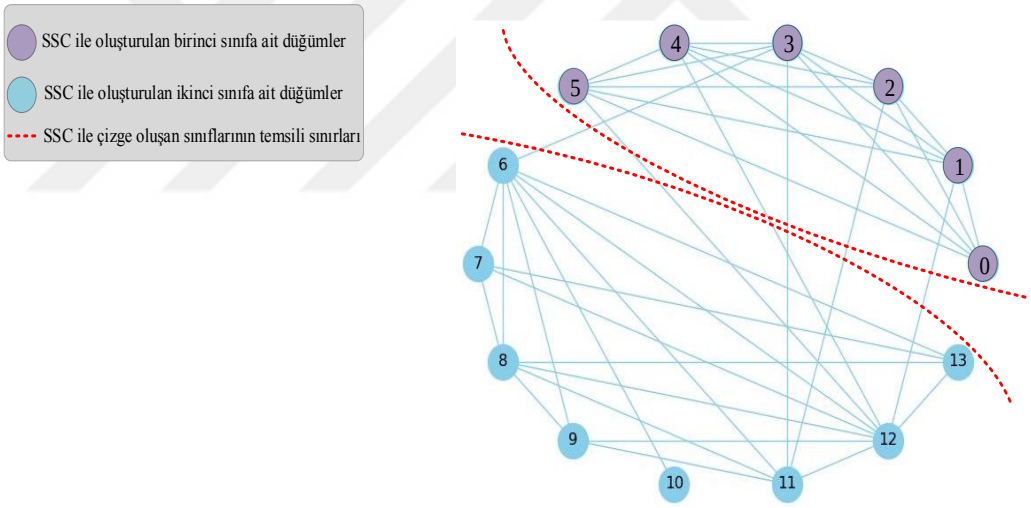
$$LaplaceG(i,j) = \begin{cases} \sum_{(i,j)\in E} A(i,j) & , \quad if \ i = j \\ -A(i,j) & , \quad if \ i \neq j \\ 0 & , \quad diğer \end{cases} \tag{3.8}$$

Laplace matrisi birçok önemli özelliğe sahip olan bir matristir. Laplace matrisi sıfırdan büyük ve yarı tanımlı aynı zamanda simetrik bir matristir. Laplace matrisinden öz değerler ve karşılık gelen öz vektörleri elde edilir. Öz değerler azalan şekilde sıralanarak *G* çizgesine ait olan çizge spektrumu elde edilmektedir. Aynı zamanda çizgeye ait spektrum çizgenin bağlantısallığı hakkında önemli bilgiler taşır. *G* çizgesinin Laplace matrisi olan *L(G)*'ye ait öz değerler (*0= $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$)* dir. *L(G)* nin öz değerler kümesi genellikle *L(G)*'nin spektrumu olarak veya G çizgesinin spektrumu olarak adlandırılmaktadır [64], [67], [77]. İlaveten *L(G)*'nin spektrumu, G çizgesinin bağlantı yolları ile ilgili bilgiler taşımaktadır. Öz değerler ayrıt yapılarında açıkça görülemeyen, genel çizge özelliklerini ortaya çıkarmak için kullanılabilmektedirler. Örneğin k farklı öz değer için *L(G)*'nin öz değerlerinin sıfır olarak hesaplanması durumunda *0= $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_k$ G* çizgesinin *k* adet bağlı bileşenden oluştuğu anlaşılmaktadır. Çizge bağlılık durumlarına dair sözü edilen örnek durum Şekil 3.7'de şematize edilmektedir. *L(G_1)*'in spektrumu *(0= $\lambda_1 = \lambda_2$, $\lambda_3 > 0$)* incelendiğinde, λ_1, λ_2 öz değerlerinin sıfıra eşit olduğu ve teoriye [76] uygun olarak çizgesinde iki parçadan oluştuğu açıkça görülmektedir. *L(G_2)* ve *L(G_3)* spektrumlarının yalnızca λ_1 öz değerlerinin sıfıra eşit olduğu ve yine teoriye uygun olarak *G_2, G_3* çizgelerinin tüm düğümlerinin birbirlerine bağlı oldukları, tek parçalı bir çizge oldukları görülmektedir. G_3 çizgesi G_2 çizgesinden daha yüksek bağlantısallık taşımaktadır. Bu durumu λ_2 değerlerinin karşılaştırarak Fiedler teorisine uygun olarak öngörmek mümkün olabilmektedir *L(G3)(λ_2)>L(G2)(λ_2)*.



Şekil 3.7. Çizge spektrumu ve çizge bağlılıkları arasındaki ilişki

SSC(Spectral Sentence Clustering) yaklaşımının çalışma şekli örnek bir çizge üzerinde şematize edilmiştir. Şekil 3.8'de elde edilen temsili kümelere ait düğümler gösterilmektedir. SSC modeli ile aynı kategoride yer alan cümleler anlam bakımından farklı kategorilerde yer alan cümlelere göre birbirleri ile daha yakın ilişkilidirler.



Şekil 3.8. Örnek metin üzerinde Spektral çizge bölmeleme

3.6. Çizgelerde Merkezilik

Merkeziyet kavramı, Bavelas tarafından bir bireyin sosyal ağ içindeki yapısal konumunun bu kişinin grup çapında süreçlerdeki etkisini nasıl belirlediğini anlamak için ortaya atılmıştır [78]. Sonrasında yapılan çalışmalarda çok sayıda merkezilik ölçümü önerilmiştir. Bu önlemlerin her biri bir ağda "merkezi" olmanın ne demek olduğuna dair farklı fikirleri savunmaktadır. Merkezilik ölçümleri bir çok farklı alanda merkez noktaları veya merkez düğümleri tespit etmek amacı ile kullanılmıştır[30], [79], [80]. Bu tez

kapsamında metinlerden elde edilen çizgelerde cümleleri temsil eden düğümler içerisinden en değerli olanları belirlemek amacı ile düğüm arasındalık merkeziliği, yakınlık merkeziliği, derece merkeziliği ve özvektör merkeziliği ölçümlerinden faydalanılmıştır.

3.6.1. Derece merkeziliği (Degree centrality)

Tarihsel olarak ilk ve kavramsal olarak en basit olanı, bir düğüm üzerinde meydana gelen bağlantıların sayısı (yani bir düğümün sahip olduğu bağların sayısı) olarak tanımlanan derece merkeziyettir. Freeman tarafından bir merkez düğümünün derecesinin bir ağdaki bitişiklerin sayısı, yani merkez düğümünün bağlı olduğu düğümlerin sayısı olduğunu iddia etmiştir. Düğüm derecesi temel bir göstergedir ve genellikle ağları incelerken ilk adım olarak kullanılır [81]–[83]. Verilen bir $G = (V, E)$ çizgesi N adet düğümden oluşmuş olsun ve ν düğümünün derece merkeziliği değeri C_D (ν) Denklem 3.9'daki gibi hesaplanmaktadır.

$$C_D(v) = \frac{derece(v_i)}{N-1} \tag{3.9}$$

3.6.2. Arasındalık merkeziliği (Betweenness centrality)

Bir çizgedeki merkeziliği ölçmek adına kullanılan yöntemlerden biri arasındalık merkeziliğidir. Bu ölçütte temel fikir x ve y düğümlerinin çoğunda V düğümü yer alıyorsa V düğümünün önemli bir düğüm olduğudur. Bu yöntem önemli düğümlerin diğer düğümlere bağlandığı varsayımına dayanır [5]. N çizgede bulunan düğümler kümesidir. V düğümünün arasındalık merkeziliği değeri Denklem 3.10'daki gibi hesaplanmaktadır:

$$C_b(v) = \sum_{x,y \in N} \frac{G_{x,y}(v)}{G_{x,y}} \tag{3.10}$$

3.6.3. Yakınlık merkeziliği (Closeness centrality)

Yakınlık merkeziyeti, uzaklığın tersi, yani bir düğüm ile tüm diğer düğümler arasındaki en kısa mesafelerin toplamı olarak tanımlanmaktadır [84]. Yakınlık merkeziliği hesaplanan düğümün, çizgeyi oluşturan tüm diğer düğümler ile arasındaki en kısa yol mesafelerinin toplamı ile yakınlık ölçüsü değeri elde edilebilir [5]. Çizge teorisinde yakınlık, sofistike bir merkeziyet ölçüsüdür. Bir düğüm ν olmak üzere ve kendisinden erişilebilen diğer düğüm noktaları arasındaki ortalama jeodezik mesafe (en kısa yol) olarak tanımlanır [85]. V_i, V_j

düğümleri arasındaki en kısa mesafe $mesafe(V_i, V_j)$ olmak üzere yakınlık merkeziliği değeri Denkem 3.11'deki gibi hesaplanmaktadır [84] :

$$C_c(v_i) = \frac{N-1}{\sum_{v_j \in v} mesafe(v_i, v_j)} \tag{3.11}$$

3.6.4. Özvektör merkeziliği (Eigenvector centrality)

Özvektör merkeziliği, bir ağdaki bir düğümün önemini belirleyen diğer bir ölçüm yöntemidir. Bu yöntemde yüksek önem değerine sahip düğümlere olan bağlantıların düşük önem değerine sahip düğümlere olan bağlantılarla aynı puanlamayı önem değerini vermekten ziyade yüksek değere sahip düğümlerin bağlantılı olduğu düğümlere daha fazla önem verdiği prensibine dayanmaktadır [85]. Google tarafından önerilen PageRank, Özvektör merkeziliği merkeziyet ölçümünü kullanan bir diğer merkeziyetçi yöntem olarak sunulmuştur[86]. V_j Ve V_i düğümler arasındaki ayrıtlarının ağırlığı ve λ bir sabit olmak üzere, V_i düğümünün özvektör merkezliliği Denklem 3.12'deki gibi hesaplanmaktadır [84]:

$$C_e(V_i) = \frac{1}{\lambda} \sum_{V_j \in N(V_i)} V_{ji}\ x\ C_e(V_j) \tag{3.12}$$

4. MAKSİMUM BAĞIMSIZ KÜMELER

NP-Zor karmaşıklık sınıfına ait çizge problemlerinden biri olan Maksimum Bağımsız Küme (MIS) problemi günlük hayatta görüntü işleme, harita işaretleme, moleküler biyoloji, zaman çizelgesi planlama(scheduling) gibi bir çok alanlarda kullanılmaktadır[87].

Çizge teorisinde karşılaşılan birçok problem, belli kısıtlamaları karşılayan alt çizgelerin bulunmasına dayanmaktadır. Bu problemlerden en fazla üzerinde durulanlardan biri de bir çizgeye ait maksimum bağımsız kümenin bulunmasıdır. Bir G çizgesinde, bulunmak istenilen bağımsız düğüm kümeleri $V(G)$ ye ait düğüm değerlerini içeren S ile gösterilmektedir. Öyle ki S kümesinde bulunan düğüm değerleri arasında G çizgesi üzerinde doğrudan herhangi bir bağlantı bulunmamaktadır. Benzer şekilde bir maksimum bağımsız küme G çizgesi üzerinde başka bir bağımsız küme uygun bir alt kümesi değildir. Başka bir deyişle bir bağımsız küme olan S te bulunan her bir düğüm S'te olmayan en az bir uç noktaya sahiptir ve S'te olmayan herbir düğüm S'te bulunan en az bir düğüme komşudur [88].

İlk olarak Erdös ve Moser genel olarak maksimum bağımsız küme değerinin maksimum sayısını, n düğümleri ve bu düğümlere ait çizgelerele belirleme problemini gündeme getirmiştir [89]. Bu problemin gündeme gelmesinden sonra çeşitli yaklaşımlar sunulmaktadır. Maksimum elemana sahip bağımsız kümeyi bulma problemi sinyal işleme analizi, sınıflandırma teorisi, optimizasyon gibi bir çok farklı alanda yer alamaktadır [87], [90]–[92]. Benzer şekilde, Chan ve Har-Peled çalışmalarında maxsimum ağırlıklı maksimum bağımsız kümeyi bulmaya yönelik bir yakınsama algoritması sunmuşlardır [93].

Tanım: Düğüm (V) ve ayrıtlardan (E) oluşan bir $G = (V\ ;\ E)$ çizgesi olmak üzere, bu G çizgesinin tamlayan i ve j düğüm olmak üzere $\bar{E} = \{(i,j) \notin E\ ; i,j\ \in\ V, i \neq j\}$ koşullarını sağlayan $\bar{G} = (V, \bar{E})$ olarak kabul edilmektedir. Bir bağımsız küme çizgede bulunan düğümler kümesidir. Öyle ki bu kümede bulunan düğümler arasında doğrudan bir komşuluk söz konusu değildir. Bir $S \subset V$ altkümesi bağımsız küme ise $\forall\ \mathrm{i}, j \in S$ ve bu düğümlere ait ayrıtlar $(i,j) \notin E$ koşullarını sağlamalıdır. Bir maksimum bağımsız kümeyi bağımsız bir maksimum kapsam kümesidir. Bir maksimum bağımsız kümeyi bulmak , bir maksimum grup veya minimum düğümü bulmakla eşdeğerdir ve bu üç problemde NP-hard problemi olarak sınıflandırılmaktadır [94].

Matematiksel Model:

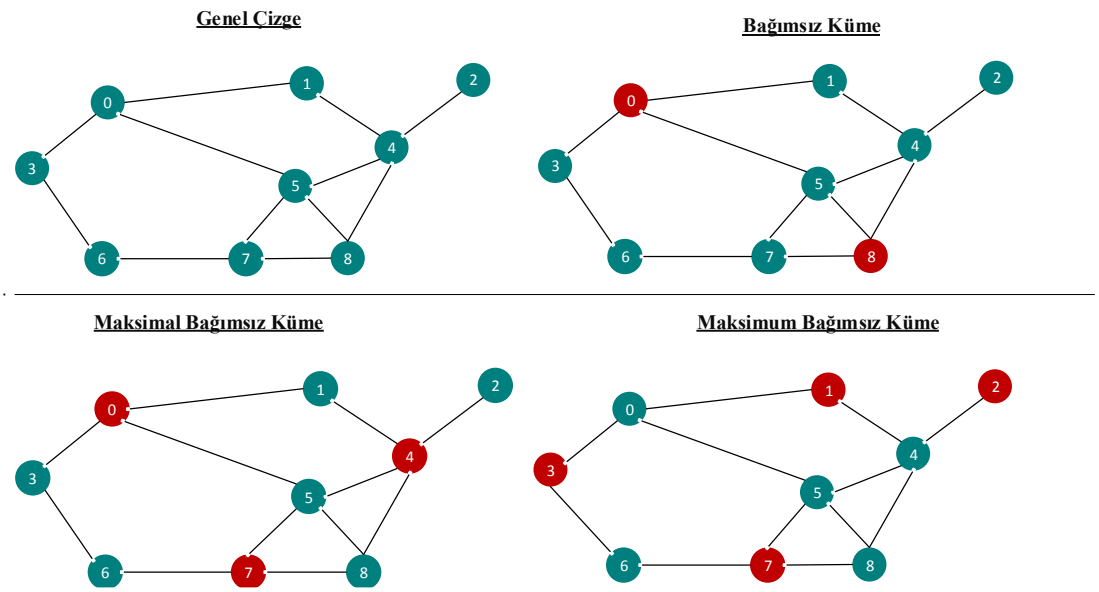
Maksimum bağımsız küme problemi, bir tamsayı programlama problemi ve sürekli dışbükey olmayan bir optimizasyon problemi olarak birçok eşdeğer formülasyona sahiptir. Bir tam sayı problemi olarak formülasyonu Denklem 4.1'deki gibidir [90]:

$$\begin{aligned} \max f(x) &= \sum_{i=1}^{n} w_i x_i \\ x_i + x_j &\leq 1, \forall (i,j) \in E, \\ x_i &\in \{0,1\}, i = 0, \ldots, n \end{aligned} \tag{4.1}$$

Shor, maksimum ağırlıktan bağımsız küme problemine sürekli yeni bir formülasyon önermiştir [95]. Denklem 4.2'de gösterilen bu fonksiyon sürekli fonksiyonlar olarak isimlendirilmektedirler.

$$\begin{aligned} \max f(x) &= \textstyle\sum_{i=1}^{n} w_i x_i, \\ x_i\, x_j &= 0, \forall (i,j) \in E, \\ x_i^2 - x_i &= 0\,, i = 0, \ldots, n \end{aligned} \tag{4.2}$$

Maksimum bağımsız küme (MIS) çizge teorisinde birçok alanda kullanılmasına rağmen, NP-Hard bir sorun olmaya devam etmektedir. Bağımsız kümeler bir çizgede bulunan düğümler arasından bir birleri ile hiçbir durumda bir ortak ayrıtı yani ortak ayrıtı bulunmayan düğümlerin bulunduğu kümedir. Bir çizgede birden fazla maksimal bağımsız kümeler bulunabilir. Maksimal bağımsız kümeler farklı sayılarda düğümlerden oluşan bağımsız kümelere verilen isimdir. Maksimal bağımsız kümeler içerisinden en fazla düğüme sahip olan kümeye o çizgenin maksimum bağımsız kümesi denilmektedir. Maksimal bağımsız kümelerde bulunan düğümlere farklı bir ayrıt eklenmesi sonucu o ayrıtın sonlandırlacağı yeni bir düğüm çizgeden seçilebilir. Fakat bir maksimum bağımsız kümede bulunan düğüme yeni bir ayrıt eklenmesi ile o düğümün başka bir düğümle sonlandırılması ve sonlandırılan düğümünde maksimum bağımsız küme içinde bulunan herhangi diğer bir düğüm ile komşuluğunu doğurmaktadır. Böyle bir durumda kümenin bağımsızlığı bozulmaktadır. Bu tanımdan yola çıkarak bir birleri ile doğrudan komşulukları bulunmayan ve yeni bir düğümün eklenmesi durumunda bağımsız küme özelliği bozulabilecek en yüksek boyutlu maksimal bağımsız kümelere maksimum bağımsız küme denilmektedir.



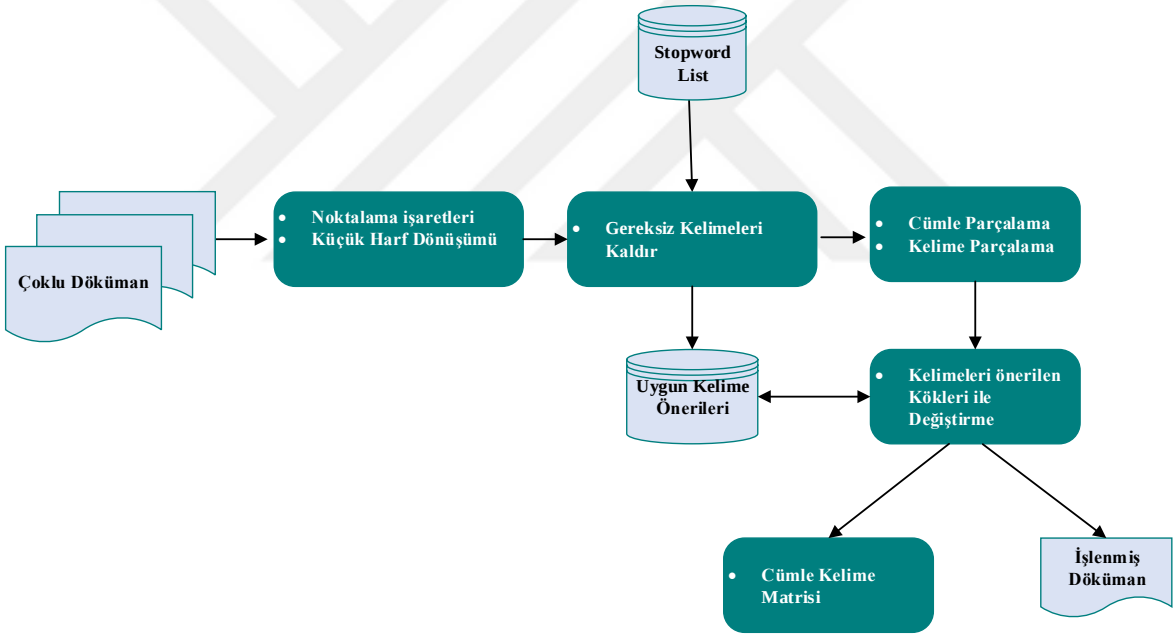
Şekil 4.1. Maksimal ve maksimum bağımsız kümeler

Şekilde 4.1'de görüldüğü üzere, ilk olarak genel küme gösterilmektedir. Ardından birbirleri ile herhangi bir ayrıt bağımlılığı olmayan iki rastgele düğüm seçilerek bağımsız bir küme elde edilmiştir. Sonrasında yine birbirleri ile doğrudan komşuluğu bulunmayan üç adet düğüm seçilerek yeni bir bağımsız küme oluşturulmuştur. Bu bağımsız kümelere yeni bir ayrıt eklenerek başka bağımsız kümeler elde etmek mümkün olduğundan dolayı bu tür kümelere maksimal bağımsız kümeler denilmektedir. Sonraki adımda yine bir birinden bağımsız fakat yeni bir düğümün eklenmesi söz konusu olmayan yani daha fazla düğümün bulunacağı bir küme olmadığından dolayı bu tür bağımsız kümelerede maksimum bağımsız kümeler denilmektedir.

5. YÖNTEM VE UYGULAMALAR

5.1. Metin Önişleme Aracı (KUSH)

Veri madenciliği, büyük miktarda veriden yararlı bilgileri bulmak için kullanılır. Veri madenciliği teknikleri farklı araştırma problemlerini uygulamak ve çözmek için kullanılır. Veri madenciliği ile ilgili araştırma alanları metin madenciliği, web madenciliği, görüntü madenciliği, sıralı model madenciliği, mekansal madencilik, tıbbi madencilik, multimedya madenciliği, yapı madenciliği ve çizge madenciliğidir [96]. Metin madenciliğinda ön işleme aşaması hayati bir öneme sahip olmakla beraber sıklıkla kullanılmaktadır. Basit olarak belirtmek gerekirse, metin ön işleme işlemi verilen bir metnin cümleler ve kelimelerin ayrıştırılması yolu ile sayısallaştırılması olarak tanımlanabilir [97], [98]. Ön işleme aşaması temel olarak 4 adımdan oluşmaktadır. Şekil 5.1'de KUSH aracına ait ön işleme adımları gösterilmektedir.



Şekil 5.1. KUSH metin ön işleme aracına ait iş akış modeli

Bu bağlamda metin ön işleme aracı için tez kapsamında kullanılmak üzere bir araç geliştirilmiştir. Özetlenecek metinlerin temsili aşamalarından biri olan KUSH metin işleme adımı veri önişleme ve hazırlık aşamalarını oluşturmaktadır. Bu aşamada, yazım şekilleri bakımından birbirlerinden farklı olan ancak anlamsal olarak ortak bir kelime kökünden türeyen yakın anlamlı kelimelerin çizgeler oluşturulurken farklı kelimeler gibi işleme alınması, cümleler arasındaki bağlantı ve ilişkilerin tespitini zorlaştırmaktadır. Sezgisel olarak sözü edilen problemin sınıflandırma sonrasında elde edilen başarımı ciddi bir şekilde

etkileyeceği öngörülmüştür ve bu tez kapsamında metin ön işleme adımında kullanılan bir metin işleme aracı geliştirilmiştir. KUSH adını verdiğimiz bu yazılım aracı .NET platformunda C# kullanılarak geliştirilmiştir.

KUSH Metin Ön işleme Algotirması

İnput: {text} – Ham Metin

Output: {kush_text} – İşlenmiş Metin

```
1   "text read";
2   sentences=[], words=[], alternative =[];
3   best_alternative =0, kush_text={}
4   sentences []=text.Split(.);
5   for c to len(sentences) do
6       sentences [c]=clear(c);
7   end for
8   words[]=sentences.Split(.);
9   for k to boyut(words) do
10      alternative [k]= alternative_search(words [k], words);
11      for u to alternative do
12          best_alternative = best_alternative_search (k, alternative);
13      end for
14      words [k]= best_alternative;
15  end for
16  for k to words do
17      kush_text+=k+" ";
18  end for
19  return kush_text;
```

Çizelge 5.1 DUC 2002 Veri Seti'nde bulunan d070f isimli örnek metin dokümanının KUSH yazılım aracı ile gerçekleştirilen cümle dönüşümlerini içermektedir. İlgili tablo durak kelimeler çıkarıldıktan sonra elde edilen metinler üzerinde KUSH yazılım aracının işlevinin detaylandırılması amacı ile gerçekleştirilen dönüşüme dair kelime ve cümle bazında bazı örnekleri içermektedir. KUSH yazılım aracı, algoritması gereği dinamik bir formda çalışmaktadır. Değişikliğe gidilecek kelimeler her özetleme işleminde yeniden belirlenmektedir. Sadece özetlenecek metinlerin ortak olarak içerdikleri kelimeler üzerinde

sadeleştirmeler yapılmaktadır. Metinler üzerinde herhangi bir kesişime sahip olmayan kelimeler için bir sadeleştirmeye gidilmemektedir. Bu dinamik çalışma prensibi ile KUSH yazılım aracı her defasında sınıflandırılacak metinlere bağlı olarak sadeleştirmeye gideceği metin parçalarını yeniden belirlemektedir. Bu sayede metinler değiştikçe farklı kelimeleri sadeleştirmektedir. Çizelge 5.1'de değişikliğe uğrayan veya silinen kelimelerin koyu renkle gösterilmektedir. Cümleler bazında değişiklikleri içeren iki numaralı örnekte program tarafından "*the*" ve "*in*" gibi kelimeler durak kelimeler oldukları için tamamen silinmiş iken yine aynı cümlede "*Lawyer*" kelimesinin yalnızca taşıdığı yapım eki silinmiştir. Bazı kelimeler üzerinde hiçbir işlemin yapılmadığı görülmektedir. KUSH yazılım aracının herhangi bir değişikliğe gitmediği kelimeler metinde yalnızca bir adet bulunan kelimelerdir. Dolayısı ile anlamsal bakımından benzeştiği veya yakın anlamlı olduğu bir ifade ile değişimi mümkün olamamaktadır. Buna karşılık cümle bazında değişiklikleri içeren bir, iki, beş numaralı cümlelerde yer alan "*Lawyer*" ve "*lawmaker*", gibi kelimeleri, tüm metinde bulunan anlamlı en yalın halleri ile yani "*Law*" kelimesi ile değiştirilmiştir. Ekli ifadelerin yerini alan "*Law*" kelimesi yine metnin tamamında aranmış ve belirlenmiştir. Böylelikle sözü edilen ekli kelimelerin farklı anlamlar taşıyormuş gibi çizgeye aktarılmalarının önüne geçilmektedir. Özetle değişikliğe uğrayan kelimenin yerini alacak olan en yakın ve anlamlı kelime KUSH yazılım aracı tarafından yine metin içinde aranmakta ve değişim yapılmaktadır.

Çizelge 5.1 d070f isimli örnek metin dokümanının dönüşümü

		KUSH öncesi	**KUSH sonrası**
Kelime Bazlı Gösterim	*1*	*Lawyer*	*Law*
	2	*powerful*	*power*
	3	*husband's*	*husband*
	4	*Russian*	*Russia*
	5	*Chilean*	*Chile*
	6	*newspaper*	*news*
		Örnek metin	
Cümle Bazlı	*1*	*Erich Honecker, the former GDR head of state, died at his house in Santiago, Chile on Sunday morning [29 May], according to his* ***lawyer****..*	*erich honecker former gdr head died house santiago chile sunday morning [29 may] according* ***law.***

	2	***Lawyer*** *Nicolas Becker, who had represented 81-year-old Honecker before the Berlin court in 1992 and early 1993, told DPA on Sunday afternoon that Honecker had rejected an operation.*	***Law*** *nicolas becker represented 81 year old honecker berlin court 1992 1993 told dpa sunday afternoon honecker rejected operation.*
	3	*However, everything that was sensible had been done with regard to his cancer.*	*sensible regard cancer.*
	4	*Becker could not comment on the exact cause of Honecker's death. His colleague, Wolfgang Ziegler, had said earlier that the former most* ***powerful*** *man in the GDR had obviously died from his liver cancer.*	*becker comment exact cause honecker death. colleague wolfgang ziegler earlier former* ***power*** *gdr obviously died liver cancer*
	5	*Federal* ***lawmaker*** *Johannes Gerster said refusal to turn over thedeposed East German leader for trial could "burden" Bonn and* ***Moscow's*** *"outstanding relations," the Neue Osnabruecker Zeitung* ***newspaper*** *reported.*	*federal* ***law*** *johannes gerster refusal deposed east german leader trial burden bonn* ***moscow*** *outstanding relations neue osnabrueck zeitung* ***news*** *report.*
	6	*The* ***Russian*** *Foreign Ministry said it had not yet formulated a response.; Outside the* ***Chilean*** *Embassy, several dozen demonstrators* ***continued*** *their vigil, demanding humane treatment for Honecker.*	***russia*** *foreign ministry formulated response. outside* ***chile*** *embassy dozen demonstrators* ***continue*** *vigil demand humane treatment honecker.*

Bu tezde açıklandığı gibi ham ve günlük konuşma dili özelliklerini taşıyan metinler, metin önişleme ve hazırlık süreci ile evvela gereksiz ve istenmeyen verilerden arındırılmıştır. Daha sonra geliştirilen KUSH aracı kullanılarak benzer veya çok benzer anlamlar taşıyan ifadelerin çizgeler oluşturulurken farklı anlamlar taşıyor gibi algılanmasının önüne geçilmiştir. Cümleler arasındaki bağlantıların ve ilişkilerin tespitinin kolaylaştırılması hedeflenmiştir. SSC modelinin cümleler arasında anlamsal olarak ayırt

edilebilir ve daha sağlıklı ölçülebilir ilişikler sağlaması bakımından KUSH aracı ile yapılan testlerin bu aracın ilişkileri yakalamada etkili olduğunu göstermektedir. Sunulan modelin elde edilen özetleme başarım değerleri üzerinde, KUSH yazılım aracı ile azımsanamayacak oranlarda artışlar sağlamaktadır.

5.2. Kullanılan Veri Setleri

Bu tez kapsamında önerilen yöntemlerin doğruluklarını test etmek amacı ile literatürde en çok kullanılan Document Understanding Conference (DUC)[99] veri setinden faydalanılmıştır. DUC veri setinde 2001 ile 2007 arasında farklı datasetler bulunmaktadır[100]. DUC 2002 veriseti yorumlayıcı ve çıkarıcı metin özetlemede kullanılması için tasarlanmış veri kümelerine sahiptir. Extractive adı altında toplanmış dosyalar çıkarıcı metin özetlemede kullanılmakta benzer şekilde abstractive dosyası altında bulunan metin dosyaları ve özet dosyaları ise yorumlayıcı metin özetleme çalışmalarında kullanılmaktadır. NIST tarafından üretilen bu verisetinde 60 adet referans veri kümesi bulunmaktadır. Farklı alanlarda haber yazılarından oluşan bu metin dosyalarında hem tek dökümanlı hemde çoklu dokümanlı model özetleri bulunmaktadır.

Yapılan çalışmalarda kullanılan diğer bir veriseti ise, yine NIST tarafından oluşturulan DUC 2004 veri setidir. Bu veriseti boyutu, kullanılacak özetleme türü ve içerdiği metin türleri dikkate alınarak 5 farklı görev olarak tanımlanmıştır.

Görev 1 500 adet gazete ve köşe yazısından oluşan 75 karakterlik çok kısa özetleri içermektedir. Bu özetler kullanıcının kullanım şekline bağlı olarak özetlenecek metinlerin başlıkları olarak tanımlanabilmektedir. Görev 2'de ise her biri 10 dosyadan oluşan 50 adet kümeden oluşmaktadır. Bu görev dahilindeki özetlemelerde 665 karakter boyutunda özetlerin oluşturulması ön görülmektedir. Kullanıcı tarafından yapılacak özetlemelerde elde edilecek aday özetin 665 karakteri geçmemesi gerekmektedir. Görev 3 ise, görev 1'e benzemekle beraber görev 1'den farklı olarak her biri 10 dosyadan oluşun 25 farklı özetleme kümesi sunulmaktadır. Görev 4 ile görev 3 aynı işlemi yapmakla beraber görev 4 çoklu metin özetleme için kümelenmiş model özetlerden oluşmaktadır. Benzer şekilde görev 5'te görev 2'de olduğu gibi özetleme için herbiri 10 dosyadan oluşan 50 adet veri kümesi sunmaktadır. Bu tez kapsamında DUC 2002 veri seti üzerinde 200 ve 400 kelimelik aday özetler oluşturulmakta aynı zamanda DUC 2004 veri seti kullanılarakta 665 karakterlik aday özetler oluşturulmaktadır. Oluşturulan aday özetler yine DUC veri setlerinde bulunan model özetler

ile karşılaştırılmaktadır. Kullanılan veri setlerinin karakteristik bilgileri Çizelge 5.2'de verilmektedir.

Çizelge 5.2. Kullanılan veri setlerine ait karakteristik bilgiler

Tanım	DUC-2002	DUC-2004
Küme sayısı	59	50
Her kümedeki döküman sayısı	~10	10
Toplam döküman sayısı	567	500
Özet uzunluğu	200 ve 400 kelimelik	665 bytes

5.3. Ölçüm Metrikleri

Özetleme sistemlerinde kullanılan en popüler değerlendirme ölçütleri ROUGE (Recall-Oriented Understudy for Gisting Evaluation) performans ölçütleridir. ROUGE, özetlerin otomatik olarak değerlendirildiği bir performans metriğidir. Bu ölçütler, n-gram, kelime dizileri bazında değerlendirme yapmaktadır. Özetleme sistemleri tarafından oluşturulan özetler ile insanlar tarafından oluşturulan ideal özetler arasında meydana gelen kesişimlerin baz alındığı bir ölçüttür [101],[102]. Tez kapsamında önerilen her iki modelinde performansının değerlendirilmesi amacı ile n-gram istatistiklerine dayanan ve insan değerlendirmeleri ile yüksek oranda ilişkili olan ROUGE değerlendirme araç seti kullanılmaktadır. ROUGE ölçütleri tarafından üretilen puanlar 0 ile 1 arasındadır. Üretilen puanlar ne kadar yüksek olursa, model özeti ile o kadar fazla paylaşılan içerik mevcut demektir ve bu durumda otomatik özet daha iyi ve daha bilgilendirici olarak kabul edilmektedir. [101]'de Lin, ROUGE puanları ile insanlar tarafından verilen puanlar arasında yüksek bir korelasyon olduğunu ortaya koymaktadır. Bu tezde, önerilen her yaklaşımının performansını değerlendirmek için ROUGE-N (N-1, N-2), ROUGE L, ROUGE-W-1.2 ve ROUGE-SU ölçülerini kullanılmıştır. ROUGE-N, önerilen özet ve model özet arasında paylaşılan n-gram sayısını değerlendirir.

$$ROUGE - N = \frac{\sum_{C \in \{Referans Özetler\}} \sum_{gram_n \in S} sayı_{eşleşme}(gram_n)}{\sum_{C \in \{Referans Özetler\}} \sum_{gram_n \in S} sayı(gram_n)} \quad (5.1)$$

Formülde yer alan n, $gram_n$ n-gram ın uzunluğuna eşittir. $Count_{match}(gram_n)$, aday özet ve referans özette kesişen maksimum n-gram'ların sayısıdır. Denklem 5.1'de açık bir şekilde görüldüğü üzere Rouge-N recall ile ilişkili bir ölçümdür. Çünkü eşitliğin paydası referans özette oluşan n-gramların sayılarının toplamıdır [101]. Sözgelimi (N-1) iki özet arasında paylaşılan uni-grams sayısını ölçer. Benzer şekilde (N-2) önerilen özet ve model özet arasında paylaşılan bi-grams sayısına odaklanır. Benzer şekilde ROUGE -L değerinde verilen iki alt kelime dizgesinde en uzun ortak kelime dizisini hesaplamaktadır. X ve Y verilen iki kelime dizisi olmak üzere Denklem 5.2, Denklem 5.3 ve Denklem 5.4'te bu dizilere ait ROUGE-L değerine ait hesaplamalar yapılmaktadır.

$$R_{lcs} = \frac{LCS(X,Y)}{m} \tag{5.2}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \tag{5.3}$$

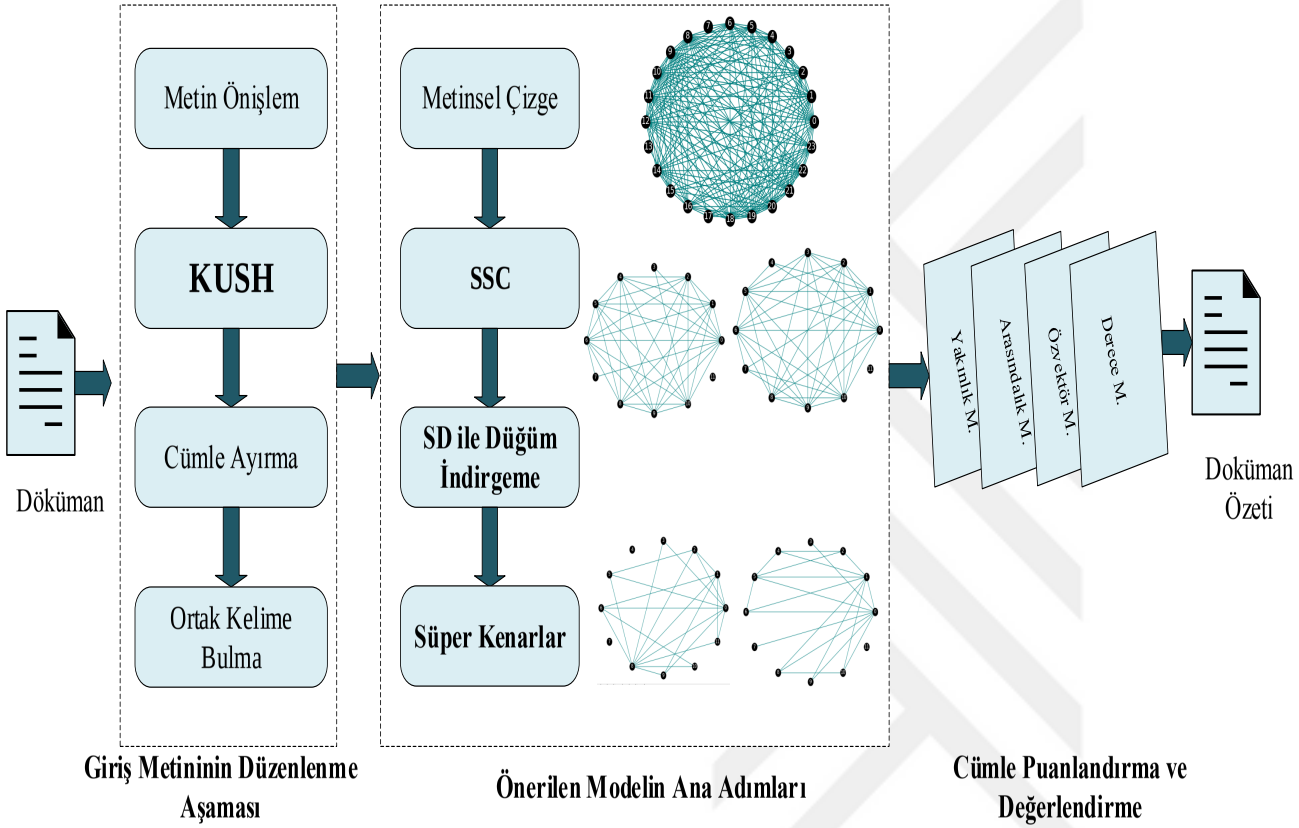
$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \tag{5.4}$$

ROUGE-W-1.2 önerilen özet ve model özet arasında art arda meydana gelen en uzun eşleşmeleri hesaplar. ROUGE-SU4 ise iki özet arasında ortak olan skip-bigrams sayısını ölçmektedir.

5.4. CatSumm: Spektral Metin Kümeleme Yöntemi ve Düğüm Merkezilik Değerleri Kullanılarak Çıkarıcı Metin Özetleme

Metin özetleme için önerilen CatSumm modelinin süreçlerine dair blok diyagramı Şekil 5.2'de gösterilmektedir. Bu çalışmada, verilen metinlerden uygun özetler çıkarılması amacıyla çizge tabanlı genel, çıkarıcı ve ve çok belgeli bir özetleme yöntemi sunulmuştur. Önerilen CatSumm otomatik özetleme yöntemi temelde 3 ana aşamadan oluşmaktadır. İlk aşamada tarafımızdan geliştirilen KUSH metin ön işleme aracı kullanılarak birtakım önişlemler yapılmaktadır. Yazım şekilleri bakımından birbirlerinden farklı olan ancak anlamsal olarak ortak bir kelime kökünden türeyen yakın anlamlı kelimelerin çizgeler oluşturulurken farklı kelimeler gibi işleme alınmasının önüne geçilmektedir. İkinci aşamada cümleler arasındaki ilişkiler şekilsel ve matematiksel olarak temsil edilmekte, önerdiğimiz SSC yöntemi ile oluşturulan çizgenin düğümleri gruplara ayırmaktadır. İlgili aşamanın son adımında standart sapma ile ayrıt indirgeme yöntemi uygulanmakta böylelikle güçlü

ilişkilerin daha belirgin olduğu süper ayrıtları içeren çizgeler elde edilmektedir. Üçüncü ve son aşamada süper ayrıtları içeren çizge ye ait farklı düğüm merkezilik yaklaşımları uygulanarak önemli düğümlere ve dolayısı ile önemli cümlelere ulaşılmaktadır.



Şekil 5.2. CatSumm blok diyagramı

5.4.1. Metinlerin Çizge Bölmeleme ile Kümelenmesi

Daha önce de belirtildiği gibi Cümle-Kelime Çizgesi'nin düğümleri, metinlerin cümlelerini temsil etmektedir. SSC metodu çizgenin düğümlerini gruplara ayırmaktadır. Bu yaklaşım ana çizgeyi oluşturan her bir cümleyi diğer cümlelerle ilişkilendirir. Elde edilen bu ilişki kullanılarak cümleler kümelendirilirler. KUSH yazılım aracı ile bu kümelerin daha doğru ve belirgin bir yapıya ulaşması hedeflenmektedir. Farklı alanlarda kullanılmakla birlikte, çizge bazında, çizgenin genel yapısal özelliklerini analiz etmeyi ifade eden spektral çizge teorisi tekniklerini kullanmaktayız. Bu teknik düğümler arasındaki mesafe veya ağırlıkları dikkate alarak (an az ya da en fazla olacak şekilde) çizge düğümlerini en az kesim maliyeti ile kesişmeyen, mümkün olduğunca denk gruplara bölmeyi amaçlamaktadırlar.

Bu çalışmada, spektral çizge bölümleme yöntemini uygularken Laplacian matrisi kullanılmıştır. Bir çizgenin Laplacian matrisi, tıpkı komşuluk matrisi gibi, graf hakkında çok fazla bilgi taşır, ancak birçok farklı kullanım ve farklı özelliğe sahiptir. Çizge bölmeleme işlemi yapılırken literatürde bulunan Laplacian matris tiplerinden biri olan basit laplacian matrisi kullanılmıştır.

Önerilen algoritmada temel fikir, cümlelere karşılık gelen metin belgelerinin düğümlerini karşılıklı ilişkiler temelinde gruplandırmaktır. Dolayısıyla, aynı kategorideki cümlelerin benzer olabileceği ve farklı kategorilerdeki cümlelerin daha az benzer olabileceği veya hiç benzemeyeceği öngörüsüne dayanmaktadır.

Spektral Metin Kümeleme Sözde Kodu	
Adım 1	***(Veri Girişi)*** KUSH aracından elde edilen işlenmiş metinler.
Adım 2	***(Ön Hazırlık)*** Başlangıç değişkenleri ve matrislerinin tanımlanması.
Adım 3	***(Cümleleri Vektörize Etmek)*** Giriş verileri parçalanarak cümle vektörleri oluşturulur.
Adım 4	***(Kelimeleri Vektörize Etmek)*** Cümle vektörleri parçalanarak kelime vektörleri oluşturulur.
Adım 5	***(Kesişim (cümle(i) ∩ cümle(j))*** Cümleler arası ortak kelimeler analiz edilerek kesişim sayılarından komşuluk matrisi ve derece matrisi oluşturulur.
Adım 6	***(Laplacian Dönüşümü)*** Basit Laplacian dönüşümü ile Laplacian matrisi elde edilir.

Adım 7 ***(Özvektör Ayrışımı)*** Özdeğer ayrışımı ile Laplacian matrisine ait Özvektör ve Özdeğerler elde edilir.

Adım 8 ***(Fiedler's Teorisi)*** Özdeğerler büyükten küçüğü doğru sıralanır vee n küçük ikinci Özvektör değerine ait Özdeğer matrisi seçilir [66].

Adım 9 ***(Kümeleme)*** Özdeğer vektöründeki değerler sayısal işaretlerine göre sınıflandırılır ve ilgili indisteki cümleler aynı kümeye dahil edilir.

Bir kare matris olan $A \in R^{nxn}$ ve x vektörü $A \in R^{n}$'de tanımlı olmak üzere Ax çarpımı R^{n}' de yeni bir vektör oluşturmaktadır. A matrisi ile x vektörünün çarpımı sonucunda oluşan yeni vektör, x vektörü ile aynı yönde fakat farklı büyüklükte olmaktadır. Yani x vektörü bir λ değeri ile yeniden ölçeklendirilmiş olmaktadır. Bu durumdu Ax çarpımındaki x değerlerine, A matrisinin öz vektör değerleri ve aynı zamanda λ değerlerine ise her bir vektör değerinin öz değeri denilmektedir. Matematiksel olarak $Ax = \lambda$x şeklinde gösterilir. nxn Boyutlu bir A matrisinin öz değerlerini hesaplamak için I birim matris olmak üzere $det(A - \lambda I_n)x = 0$ eşitliğinin çözülmesi gerekmektedir [66].

Spektral Metin Kümeleme Algoritması

Giriş: Metin ön işlem sonrasında elde edilen işlenmiş metin.

Çıktı: { C_1, C_2 } | $C_1 \cap C_2 = ø$ - Metinlere ait ayrık kümeler

```
1   degree_matrix=[][], adjacency_matrix =[][] , laplace_matrix=[][];
2   relationship_value=0, degree_value=0, eigen_value =0, eigen_value _vec=0
3   C1={}, C2={}, ssc_cluster_list={}
4   for i to len(text) do
5       for j to len(text) do
6           if (sentence(i) ∩ sentence(j)) > 0 and  i!=j  then
7               relationship_value= len(sentence(i),sentence(j))
8               degree_value+=1
9           else
10              relationship_value=0
11          end if
12       end for
13      adjacency matrix[i,j]=relationship_value, degree_matrix[i,j]=degree_value
14  end for
15  laplace_matrix = (degree_matrix -  adjacency matrix)
16  eigen_value, eigen_value_vec = eigen_decomposition (laplace_matrix)
```

```
17  eigen_value = sort(eigen_value)
18  for i to len(eigen_value _vec) do
19          if eigen_value _vec[i,2] > 0 then
20              C1.add(i), ssc_cluster_list.add(0)
21          else
22              C2.add(i), ssc_cluster_list.add(1)
23          end if
24  end for
25  return C1,C2, ssc_cluster_list
```

5.4.2. Standart Sapma ile Çizge Sadeleştirme

Bu adımda ana çizgeden spektral metin kümeleme sonucunda elde edilen iki alt çizgenin zayıf bağlantılarının standart sapma yöntemi ile elemine edilmesi anlatılmaktadır. Alt çizgelerde bulunan düğümler birer cümleyi temsil etmektedir. Düğümler arasında bulunan ayrıtlar ise, cümleler arasındaki ilişkinin gücünü temsil etmektedir. Öyle ki iki cümle arasındaki ilişki ne kadar yüksek olursa ayrıt ağırlığı aynı oranda yüksek olmaktadır aksi durumda cümleler arasındaki ayrıt ağırlıkları düşük olmaktadır. Bu çalışmada değerli cümlelerin ilişki olduğu cümlelerinde değerli olduğu ve özette bulunmaları gerektiği düşünülmektedir. İşte bu nedenle cümleler arasındaki bütün ilişkilerin değerlerinin standart sapma değerleri hesaplanmakta ve bu değerin altında kalan ilişkiler indirgenerek ortadan kaldırılmaktadır.

Standart sapma ile verilerin ne kadarının ortalamaya yakın olduğunu bulunur. Eğer standart sapma küçükse veriler ortalamaya yakın yerlerde dağılmışlardır. Bunun tersi olarak standart sapma büyükse veriler ortalamadan uzak yerlerde dağılmışlardır. Bütün değerler aynı olursa standart sapma sıfır olur.

Çizgelerde bulunan düğümlerin ayrıt ağırlıklarının standart sapması hesaplanmaktadır. V Düğümleri, E ayrıtları temsil etmek üzere $G = (V, E)$ olarak gösterilen bir çizgede ayrıt ağırlıkları $W = \{w_1, w_2, w_3, \ldots, w_n\}$ ve aritmetik ortalama $\overline{w}$ ile gösterilmektedir.

$$\overline{w} = \frac{\sum_{i=1}^{N} w_i}{N} \tag{5.5}$$

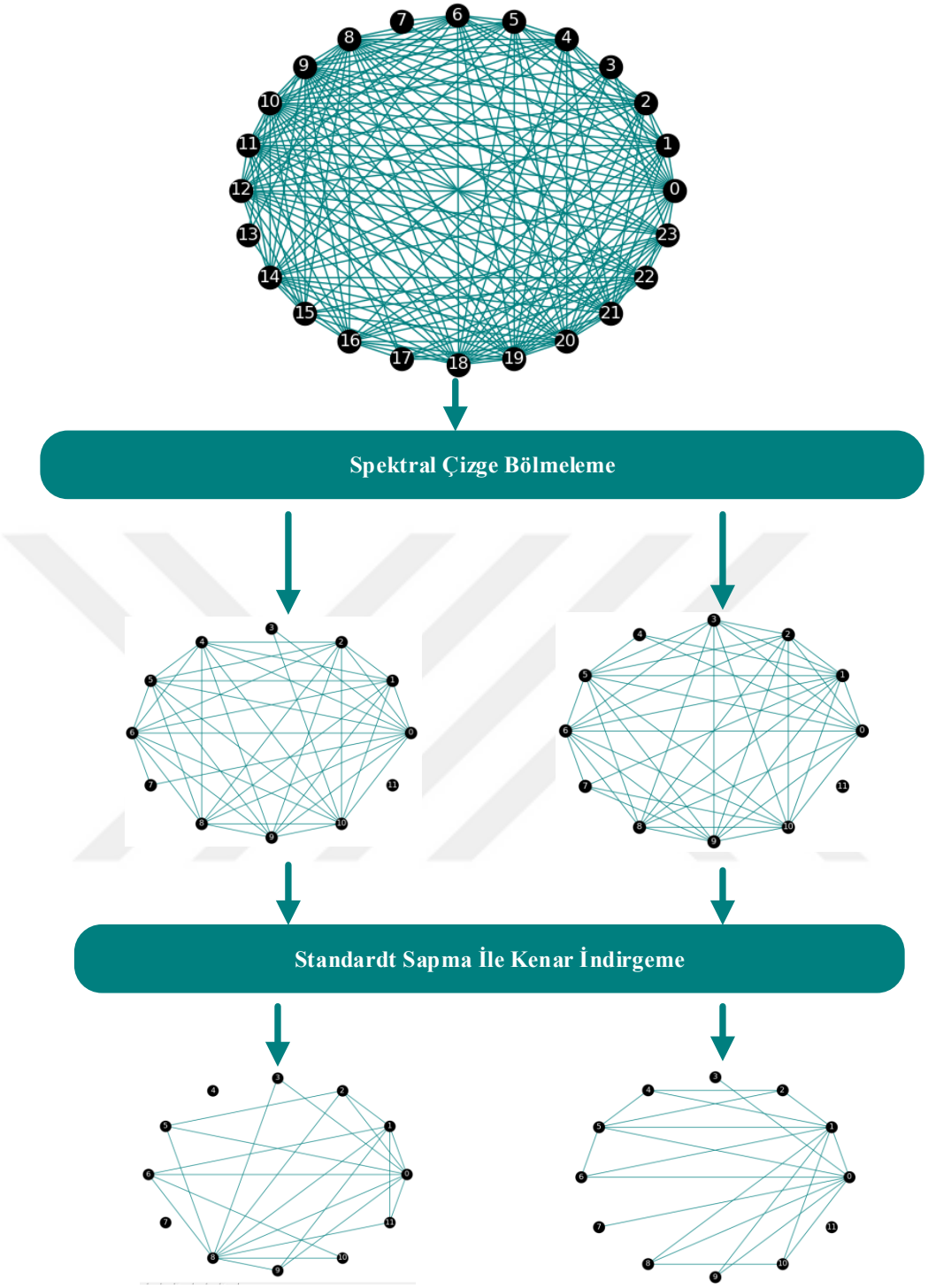
$$P(e_1) = \frac{w_1}{\sum_{i=1}^{n} w_i} \tag{5.6}$$

$$Var(W) = \sum_{i=1}^{n} (w_i - \overline{w})^2 * P(e_i) \tag{5.7}$$

$$\sigma = \sqrt{Var(W)} \tag{5.8}$$

Denklem 5.5, Denklem 5.6, Denklem 5.7 ve Denklem 5.8 de bulunan deklemler kullanılarak standart sapma σ değeri hesaplanmaktadır. Hesaplanan standart sapma değeri ile ihmal edilebilecek ayrıt ağırlıkları belirlenmektedir. Ayrıt ağırlıkları aritmetik ortalamanın standart sapma kadar üzerindeki mesafeye sahip ayrıt değerleri indirgenmektedir. Böylelikle Süper Ayrıt olarak isimlendirdiğimiz en değerli bağlantılar elde edilerek özette bulunacak cümlelerin daha güçlü cümlelerden oluşması sağlanmaktadır.

Bu çalışmada spektral metin kümeleme sonucunda elde edilen birbiri ile daha yakın ilişkili iki alt cümle gruplarının çizgeleri oluşturulmuştur. Oluşturulan çizgelerde cümlelerin içerdikleri ortak kelime sayıları dikkate alınarak ayrıt ağırlıkları belirlenmiştir. Belirlenen ayrıt ağırlıklarının standart sapma değerleri hesaplanarak ağırlıkların aritmetik ortalamalarından standart sapma miktarında yakın olan değerlere sahip ayrıtlar kaldırılmıştır. Böylelikle anlamsal olarak daha yakın ilişkili cümlelerin belirlenmesi sağlanmış daha zayıf ilişkili cümlelerin ilişkileri kesilmiştir. Şekil 5.3'te örnek bir metine ait alt çizgelerin standart sapma öncesinde ve sonrasındaki ayrıt bağlantıları gösterilmektedir.



Şekil 5.3. örnek bir belgeye ait çizge ,alt çizge ve sadeleştirilmiş çizgelerin bulunması

5.4.3. Deneysel Sonuçlar

Bu çalışmanın temel amacı, metin özeti çalışmalarını bir adım ileriye taşıyarak, çıkarılan metni özetlemek için denetimsiz ve çizge tabanlı bir süreç sunmaktır.

Önerilen CatSumm özetleme yönteminin performansını analiz etmek için, DUC-2002 ve DUC-2004 veri kümelerindeki metinlerin özetleri oluşturularak deneysel çalışmalar yapılmıştır. Bu çalışma kapsamında, CatSumm yönteminin performansını artırmak için KUSH adlı bir yazılım aracı kullanılmış ve metinler belirli dilsel süreçlerle belirli ön işlemlere tabi tutulmuştur. Sonrasında özetlerin başarısı, literatürde bulunan en yaygın kullanılan performans metrikleri hesaplanarak ölçülmüştür. Yaklaşımın performansı literaturde sıklıkla kullanılan ve performans karşılaştırmalarında kullanılan yedi farklı yöntemle karşılatırılarak tablo ve grafiklerle ayrıntılı analizler yapılmıştır.

Deneysel çalışmada, çalışmanın blok diyagramında gösterilen adımlar takip edildi. Öncelikle, istenmeyen ve temsili olmayan ifadeler ve karakterler olan durdurma sözcükleri veri kümesinden hariç tutulmuştur. Şekil 5.2'de gösterildiği gibi önişleme aşamasından hemen sonra ve Cümle Ayırıcı aşamasından önce kullanıldı. KUSH metin işleme yazılımı, yapılan deneysel sonuçlarında KUSH yazılımı ile gerçekleştirilen çalışmalar ile KUSH kullanılmadan gerçekleştirilen deneysel süreçler sonucunda performans olarak farklılıkların olduğu açıkça analiz edilmiş ve KUSH yazılımının yapılan çalışmanın doğruluğunda yüksek oranda artışa sebep olduğu gözlemlenmiştir.

Literatürde farklı Laplasian hesaplama yöntemleri bulunmaktadır. Önceki adımlarda da belirtildiği gibi bu çalışmamızda basit Laplacian hesaplama yöntemi kullanılarak sonuca gidilmiştir.

Oluşturulan Laplacian matrisleri kullanılarak elde edilen öz değer ve öz değer vektör çiftleri sıralanmıştır. Bu sıralama ile metinsel çizgeye ait olan ve çizgenin bağlantısallığı hakkında bilgi taşıyan Laplacian spektrumu oluşturulmuştur. İkinci en küçük öz değere karşılık gelen özdeğer vektörleri çizgenin cebirsel bağlılığını temsil etmektedir. Deneysel çalışmada çizgenin bölünmesinde bu vektör kullanılmıştır. Bu bağlamda çizgelerle temsil edilen metinlerin özetlenme performanslarını arttırmak adına SSC kullanılarak özeti temsil eden grafların bölmelenmesi ve spektral bölmeleme sonrasında elde edilen oranlar ölçüsünce özetler elde edilmektedir. Bu yenilikçi yaklaşım elde edilen performans metriklerine oldukça olumlu etkide bulunmaktadır. Geride bırakılan adımlara ilaveten bölütlenen çizgelerde yer alan en önemli düğümlerin elde edilmesi amacı ile belli bir standart sapma hesabı ile bir eşik değeri hesaplanmaktadır. Hesaplanan eşik değerinin altında değer alan ayrıtların çizgeden uzaklaştırılması yolu ile, ilgili ayrıt ve düğümleri temsil eden cümlelere özette yer verilmemektedir. Son aşamada elde edilen süper ayrıtlarla bağlantısı olan düğümlerin farklı düğüm merkezilik hesaplamaları ile merkezilik değerleri hesaplanmakta (Derece

Merkeziliği, Yakınlık Merkeziliği, Arasındalık Merkeziliği, Özvektör Merkeziliği) ve her bir yaklaşım için 100 (665 byte), 200 ve 400 kelimeden ibaret olmak üzere özetler elde edilmektedir.

Gerçekleştirilen çalışmaya ait özetleme performansını değerlendirmek için DUC-2002 ve DUC-2004 veri setlerinde yer alan metinlere ait model özet (100, 200 ve 400 kelimelik) ile CatSumm metodu tarafından elde edilen sistem özetleri (100, 200 ve 400 kelimelik) karşılaştırılmaktadır. Bu amaç ile özetleme sonuçlarının performansını ölçmek için ROUGE özetleme performans metrikleri kullanılmıştır. Deneysel çalışmada önerilen sistem özetinin özetleme başarısını ölçmek amacı ile Rouge-1, Rouge-2, Rouge-L, Rouge-W-1-2, Rouge-SU performans metrikleri hesaplanmıştır.

Önerilen CatSumm modeli öncelikle seçilen her bir düğüm merkeziyet değerleri için ayrı ayrı çalıştırıldı ve farklı özet değerleri elde edildi. Elde edilen 100, 200 ve 400 kelimelik özetlerin recall değerleri sırasıyla Çizelge 5.4, Çizelge 5.2 ve Çizelge 5.3'te gösterilmektedir. Sonuç tablolarından görülebileceği gibi, özvektör merkeziyet değeri 200 ve 100 kelimelik özetler için en iyi sonuçları ve 400 kelimelik özet için ikinci en iyi sonuçları vermiştir. Ayrıca Grafiksel olarak 200 kelimelik özetler için CatSumm modeli ile kullanılan düğüm merkezi değerlerinin ROUGE-1, ROUGE-2, ROUGE-L ROUGE-W-1.2 ve ROUGE-SU ölçümlerinin recall değerleri sırasıyla Şekil 5.4 ,Şekil 5.5 ve Şekil .56 'da gösterilmiştir. Şekil 5.4 'te açıkça görüldüğü gibi, Özvektör Merkezi ölçümleri dikkate alındığında, diğer merkezi ölçümlerden daha iyi sonuçlar verdiği açıkça görülmektedir. Benzer şekilde Şekil 5.5 'te Derece Merkeziyeti değerinin küçük bir farkla daha iyi sonuçlar verdiği ve 400 kelimeyle özetlendiğinde Özvektör değerinde çok rekabetçi sonuçlar verdiği gösterilmiştir. DUC-2004 veri kümesi kullanılarak yapılan 100 kelimelik özetler dikkate alındığında, Şekil 5.6'da Özvektör Merkezi değerinin daha iyi sonuçlar verdiği görülebilir. Bu çalışmalarda elde edilen sonuçlar Çizelge 5.2, Çizelge 5.3 ve Çizelge 5.4 dikkate alındığında, önerilen CatSumm modelinin seçilen tüm düğüm merkezi yöntemleriyle iyi sonuçlar verdiği, ancak en iyi sonucun Özvektör Merkezi ile elde edildiği görülmektedir.

Çizelge 5.3. 200 Kelimelik özetler için düğüm merkezilik değerleri ile Recall ölçüm sonuçları

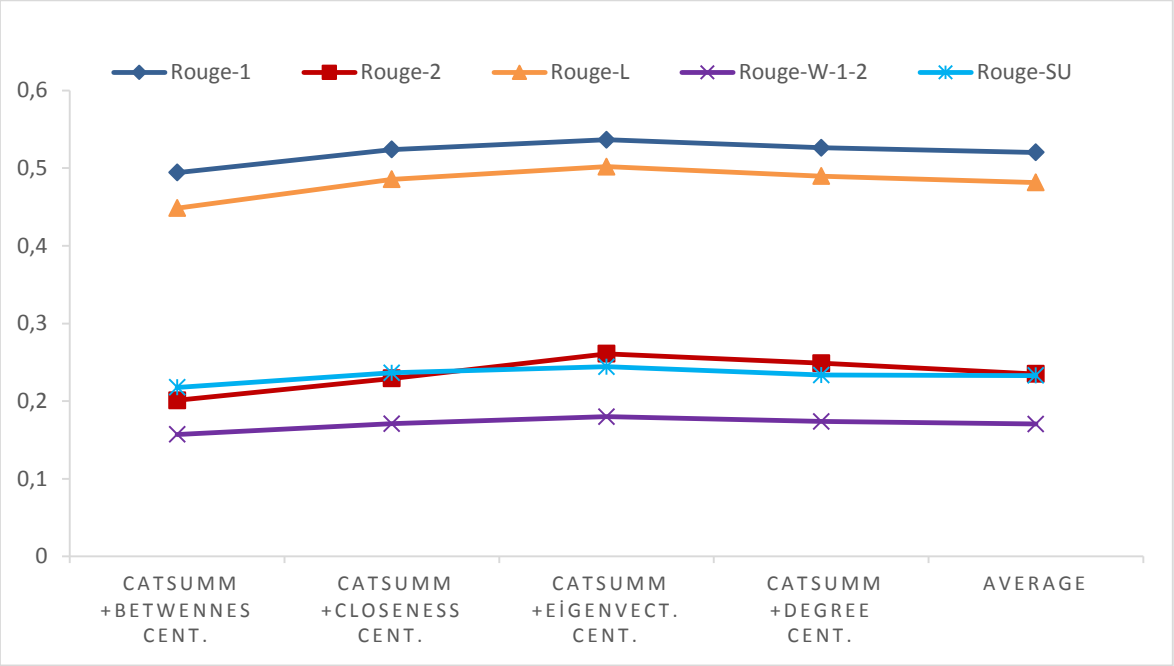
ROUGE değerlendirme ölçütü	CatSumm + Arasındalık Merkeziliği	CatSumm + Yakınlık Merkeziliği	CatSumm +Özvektör Merkeziliği	CatSumm + Derece Merkeziliği	Ortalama
ROUGE-1	0.49421	0.52407	**0.53657(1)**	0.52628	0.52028
ROUGE-2	0.20106	0.22930	**0.26097(2)**	0.24883	0.23504
ROUGE-L	0.44847	0.48566	**0.50195(1)**	0.48968	0.48144
ROUGE-W-1.2	0.15715	0.17091	**0.17990(1)**	0.17388	0.17046
ROUGE-SU	0.21778	0.23640	**0.24432(1)**	0.23347	0.23299

Çizelge 5.4. 400 Kelimelik özetler için düğüm merkezilik değerleri ile Recall ölçüm sonuçları

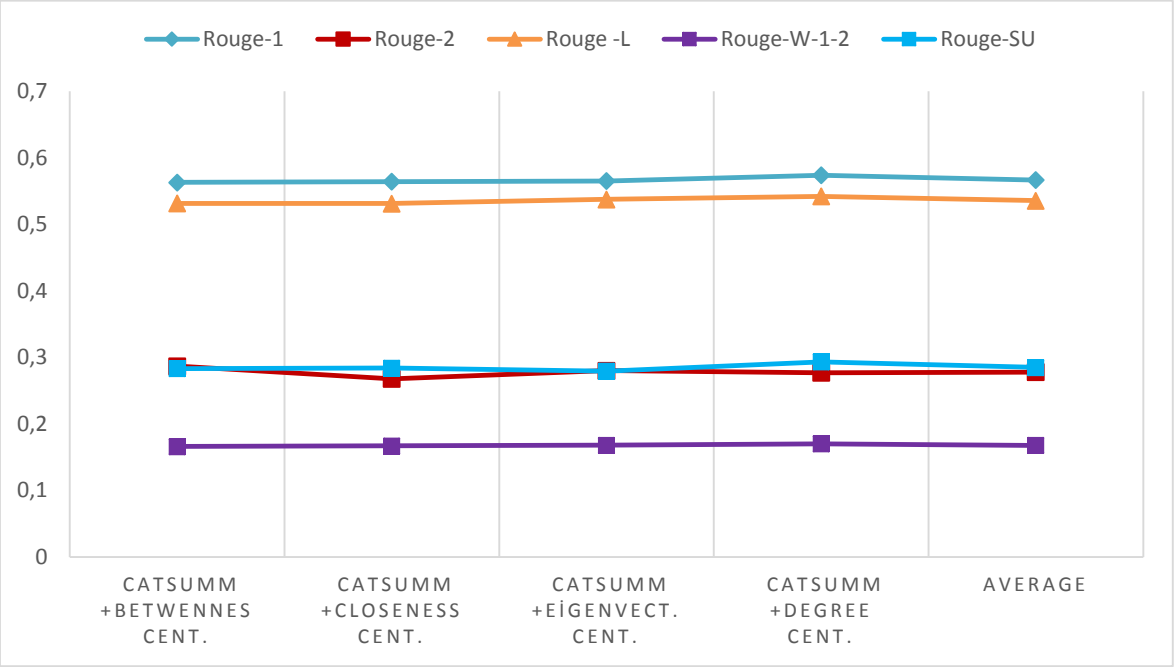
ROUGE değerlendirme ölçütü	CatSumm + Arasındalık Merkeziliği	CatSumm + Yakınlık Merkeziliği	CatSumm + Özvektör Merkeziliği	CatSumm + Derece Merkeziliği	Ortalama
ROUGE-1	0.56301	0.56414	0.56513 (2)	**0.57381(1)**	0.56652
ROUGE-2	**0.28698**	0.26789	**0.28009(1)**	0.27707(2)	0.27800
ROUGE-L	0.53151	0.53119	0.53749(2)	**0.54205(1)**	0.53556
ROUGE-W-1.2	0.16624	0.16702	0.16805(2)	**0.17033(1)**	0.16791
ROUGE-SU	0.28337	0.28396	0.27937(2)	**0.29344(1)**	0.28503

Çizelge 5.5. 100 Kelimelik özetler için düğüm merkezilik değerleri ile Recall ölçüm sonuçları

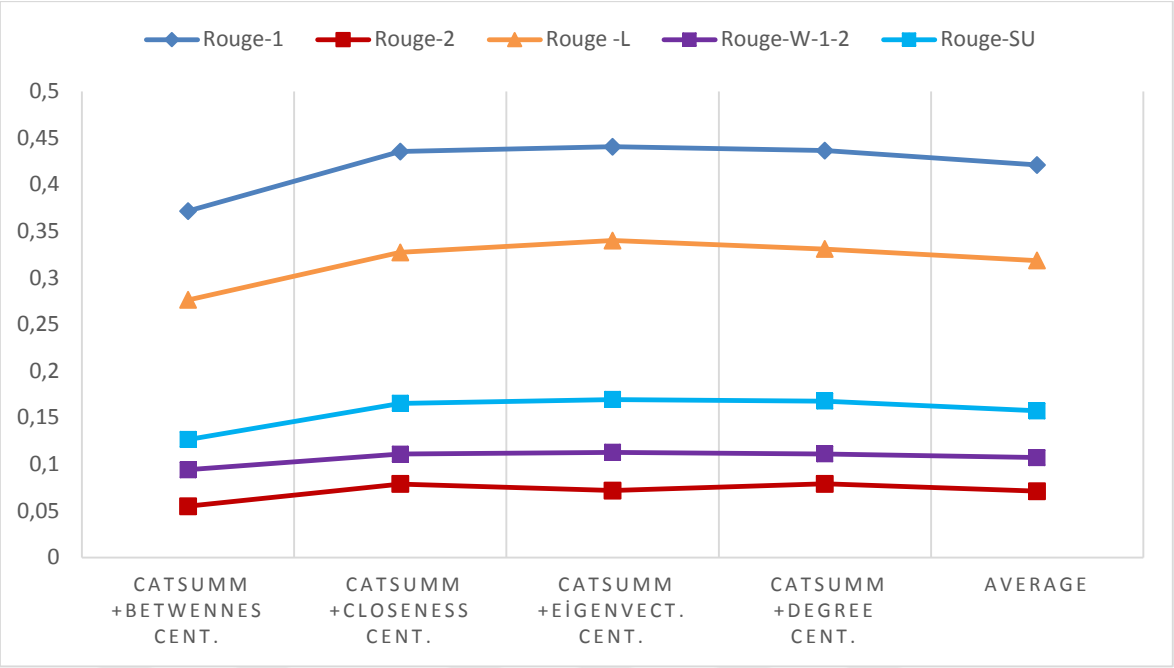
ROUGE Değerlendirme ölçütü	**CatSumm + Arasındalık Merkeziliği**	**CatSumm + Yakınlık Merkeziliği**	**CatSumm + Özvektör Merkeziliği**	**CatSumm + Derece Merkeziliği**	**Ortalama**
ROUGE-1	0.37167	0.43553	**0.44073 (1)**	0.43657	0.42112
ROUGE-2	0.05495	0.07876	0.07177 (2)	0.07894	0.07110
ROUGE-L	0.27632	0.32735	**0.34006 (1)**	0.33090	0.31865
ROUGE-W-1.2	0.09405	0.11086	**0.11279 (1)**	0.11110	0.1072
ROUGE-SU	0.12652	0.16541	**0.16939 (1)**	0.16783	0.15728



Şekil 5.4. 200 kelimelik özetlerin Recall değerlerinin karşılaştırılması



Şekil 5.5. 400 kelimelik özetlerin Recall değerlerinin karşılaştırılması



Şekil 5.6. 100 kelimelik özetlerin Recall değerlerinin karşılaştırılması

Çizelge 5.6 'te CatSum + Özvektör merkezilik değerinin DUC 2002 veriseti kullanılarak 200 kelimelik özetler için oluştrulması sonucu elde edilen sonuçlar gösterilmektedir. Tablolardan açıkça görüldüğü üzere 200 kelimelik özetler için önermiş olduğumuz modelin diğer modellerden daha iyi sonuç verdiği gözlemlenmektedir. Benzer şekilde 400 kelimelik özetler için elde edilen recall değeri sonuçları Çizelge 5.7'de ayrıntılı olarak verilmektedir. Buradan görüldüğü üzere 400 kelimelik özetlerde önerilen modellerin çoğundan daha iyi sonuçler elde edilmiş olup oldukça rekabetçi sonuçların elde edildiği gözlemlenmektedir.

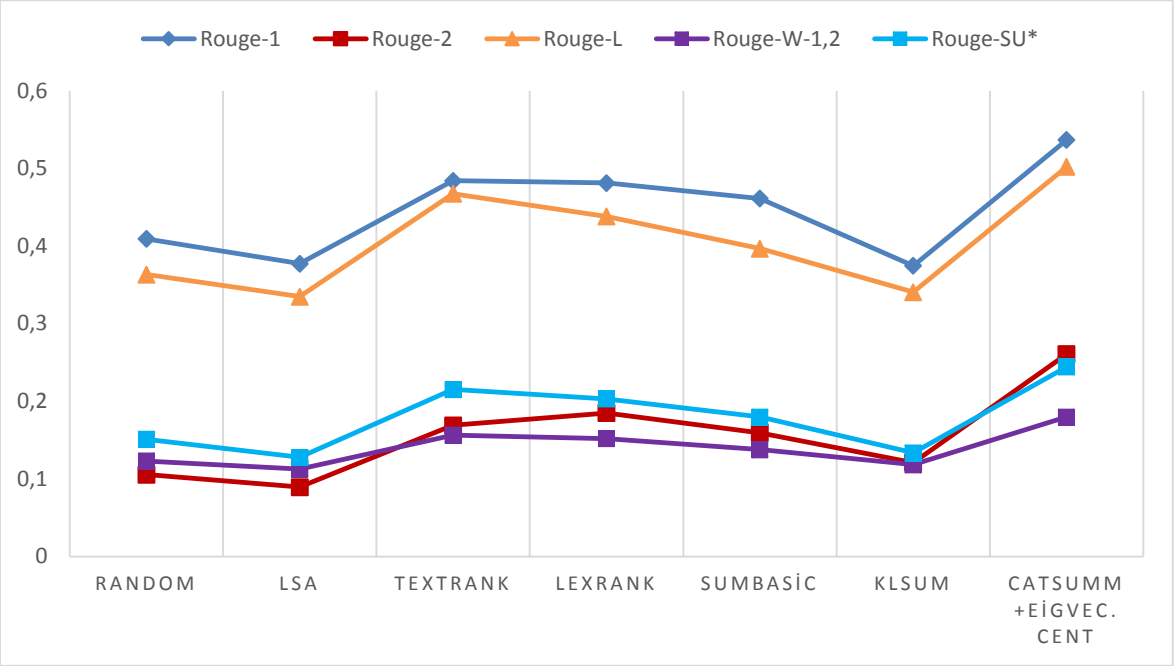
Elde edilen Ölçüm sonuçları ayrıca Şekil 5.7 ve Şekil 5.8'de grafiksel olarak gösterilmektedir. Şekil 5.7 de, Önerilen modelin DUC 2002 veriseti kullanılarak 200 kelimelik özetler için elde edilen ROUGE 1, ROUGE-2, ROUGE-L ROUGE-W-1.2 ve ROUGE-SU'nun Recall değerlerini açıkça göstermektedir. Ayrıca Şekil 5.8'de , DUC-2002 veri kümesi kullanılarak elde edilen 400 kelimelik özetler için ROUGE-1, ROUGE 2, ROUGE-L ROUGE-W-1.2 ve ROUGE-SU'nun Recall değerlerini ayrıntılı olarak göstermektedir.

Çizelge 5.6. 200 Kelimelik özetler için önerilen modelin diğer yöntemler ile karşılaştırılması

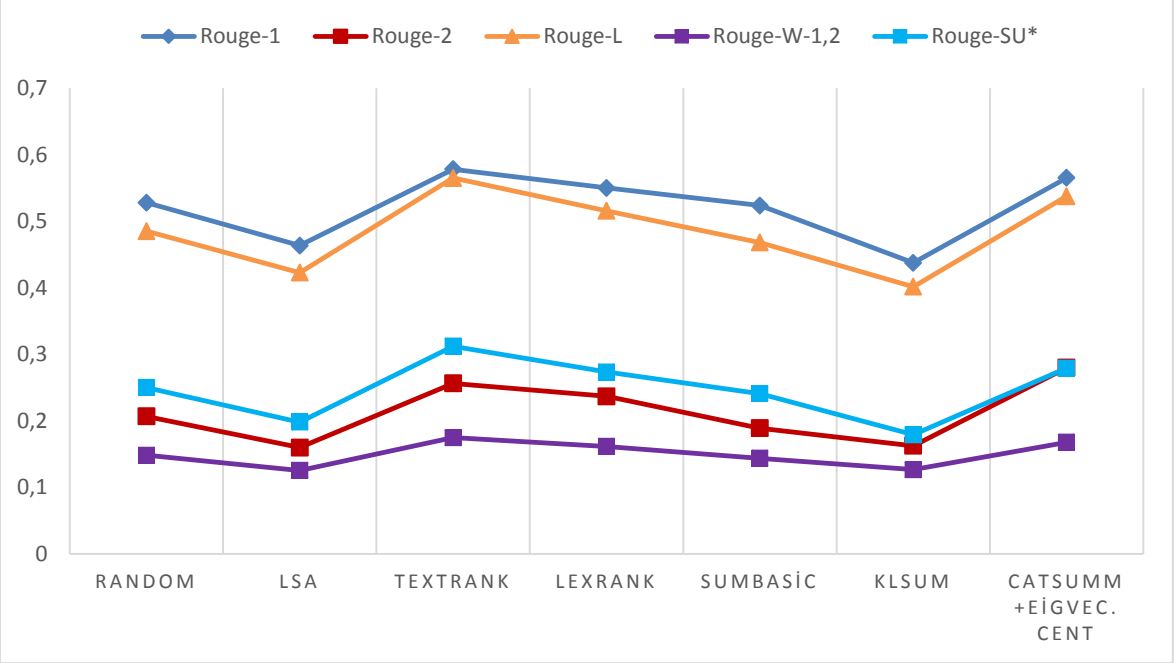
ROUGE	Random [13]	LSA [103]	TextRank [48], [105]	LexRank [80]	SumBasic [26]	KLSum [27]	CatSumm + Özvektör
ROUGE-1	0.40932	0.37723	0.48417	0.48124	0.46128	0.37464	**0.53657(1)**
ROUGE-2	0.10562	0.08950	0.16921	0.18507	0.15947	0.12113	**0.26097(1)**
ROUGE-L	0.36328	0.33477	0.46748	0.43813	0.39661	0.34066	**0.50195(1)**
ROUGE-W-1.2	0.12324	0.11295	0.15655	0.15201	0.13782	0.11839	**0.17990(1)**
ROUGE-SU*	0.15112	0.12783	0.21541	0.20321	0.17981	0.13361	**0.24432(1)**

Çizelge 5.7. 400 Kelimelik özetler için önerilen modelin diğer yöntemler ile karşılaştırılması

ROUGE	Random [13]	LSA [103]	TextRank [48], [105]	LexRank [80]	SumBasic [26]	KLSum [27]	CatSumm + Özvektör
ROUGE-1	0.52799	0.46361	0.57812	0.55018	0.52373	0.43746	0.56513(2)
ROUGE-2	0.20698	0.16001	0.25654	0.23706	0.18916	0.16256	**0.28009(1)**
ROUGE-L	0.48516	0.42283	0.56505	0.51566	0.46814	0.40158	0.53749(2)
ROUGE-W-1.2	0.14844	0.12558	0.17498	0.16162	0.14401	0.12675	0.16805(2)
ROUGE-SU*	0.25020	0.19823	0.31213	0.27351	0.24097	0.17922	0.27937(2)



Şekil 5.7. 200 kelimelik özetlerin Recal değerlerinin karşılaştırılması



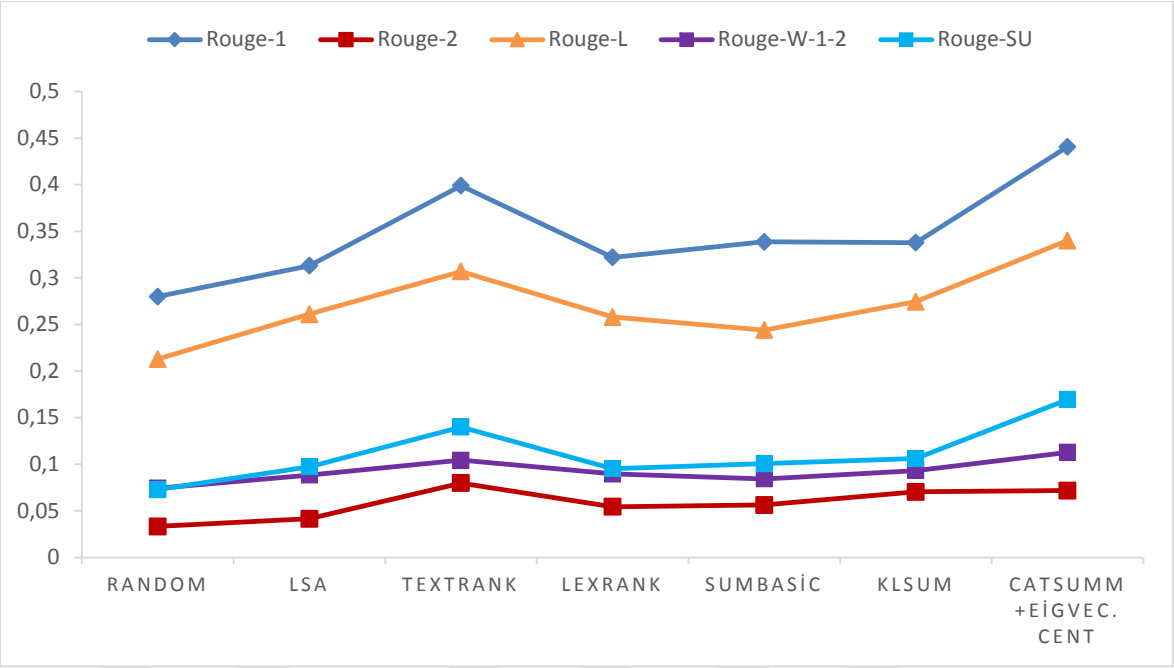
Şekil 5.8. 400 kelimelik özetlerin Recal değerlerinin karşılaştırılması

Çizelge 5.8, DUC-2004 veri kümesi kullanılarak elde edilen 100 kelimelik özetler için elde edilen ROUGE-1, ROUGE-2, ROUGE-L ROUGE-W-1.2 ve ROUGE-SU'nun Recall değerlerini açıkça göstermektedir. Çizelge 5.7'de gösterilen değerler dikkatlice incelendiğinde, önerilen CatSumm modeli ile Özvektör merkeziliği değerinin birleşimi sonucunda elde edilen Recall değerlerinin LexRank'in ROUGE-2 değeri hariç diğer

özetleme yöntemlerinden çok daha iyi sonuçlar verdiği görülmektedir. Önerilen modelin ROUGE-1 recall değerleri dikkate alındığında, Random modelinden yaklaşık % 37, LSA'dan % 57, LexRank'tan % 10 ve SumBasic ve KLSum yöntemlerinden yaklaşık olarak %29 daha iyi sonuçlar ürettiği görülmektedir. Önerilen Catsumm+Özvektör modeline ait elde edilen ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-W-1.2, and ROUGE-SU ölçüm metriklerinin Recall sonuçları ayrıca şekil 5.9'da grafiksel olarak gösterilmektedir.

Çizelge 5.8. 100 Kelimelik özetler için önerilen modelin diğer yöntemler ile karşılaştırılması

ROUGE Değerlendirme ölçütü	**Random** [13]	**LSA** [103]	**TextRank** [48], [105]	**LexRank** [80]	**SumBasic** [26]	**KLSum** [27]	**CatSumm + Özvektör**
ROUGE-1	0.32206	0.27995	0.31301	0.39893	0.33859	0.33778	**0.44073 (1)**
ROUGE-2	0.05439	0.03324	0.04145	0.**07977**	0.05623	0.07030	0.07177 (2)
ROUGE-L	0.25785	0.21295	0.26099	0.30685	0.24408	0.27448	**0.34006 (1)**
ROUGE-W-1.2	0.08957	0.07427	0.08836	0.10436	0.08427	0.09315	**0.11279 (1)**
ROUGE-SU*	0.09532	0.07285	0.09741	0.13998	0.10063	0.10609	**0.16939 (1)**

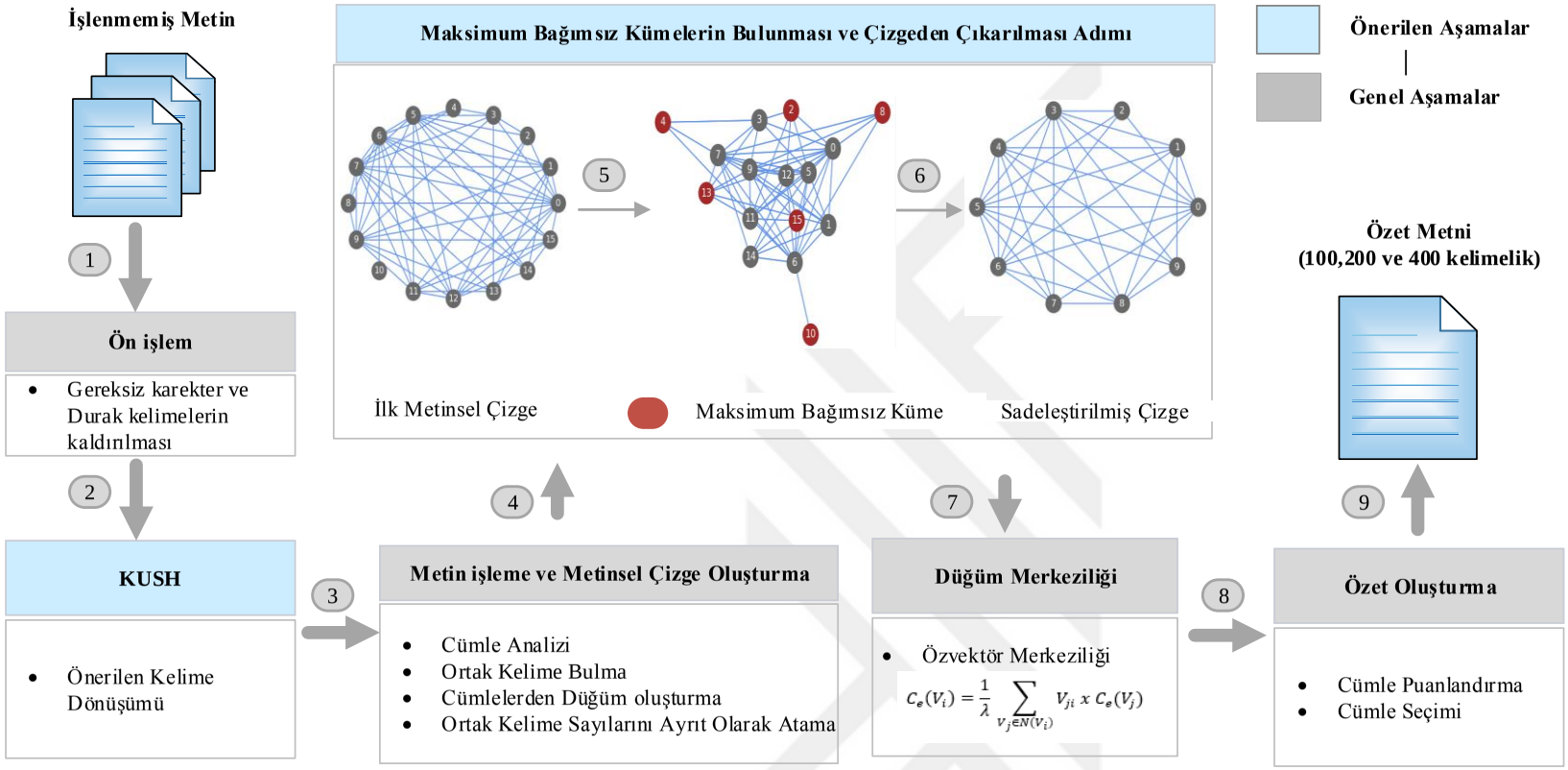


Şekil 5.9. 100 kelimelik özetlerin Recal değerlerinin karşılatırılması

5.5. Extractive Multi-Document Text Summarization Based On Graph Independent Sets(Maksimum Bağımsız Kümeye Dayalı Çoklu Doküman Özetleme)

Bu çalışmada, verilen metinlerden uygun özetler çıkarılması amacıyla çizge tabanlı genel, çıkarıcı ve çok belgeli bir özetleme yöntemi sunulmuştur. Önerilen metin özetleme yöntemi temelde üç ana aşamadan oluşmaktadır. İlk aşamada herhangi bir ayırt edicilik taşımayan durak kelimeler (zamirler, edatlar, bağlaçlar, vb..) veri setinden uzaklaştırılmıştır. Tarafımızdan geliştirilen KUSH metin ön işleme aracı kullanılarak birtakım önişlemler yapılmaktadır. Yazım şekilleri bakımından birbirlerinden farklı olan ancak anlamsal olarak ortak bir kelime kökünden türeyen yakın anlamlı kelimelerin çizgeler oluşturulurken farklı kelimeler gibi işleme alınmasının önüne geçilmektedir. İkinci aşamada kelime öbekleri arasındaki kelime ortaklıkları sırasıyla matematiksel ve çizgesel olarak temsil edilmektedir. Ek olarak bu aşamada Maksimum bağımsız kümeyi oluşturan düğümlerin tespiti ve bu düğümlere karşılık gelen cümlelerin ana çizgeden uzaklaştırılması adımlarıda yer almaktadır. Son aşama Özvektör düğüm merkeziliği yöntemi ile dokümanları oluşturan cümlelerin ağırlıklandırılması ve önemli cümlelerin seçimi aşamasıdır. Önerilen yöntemin bu aşamasında Top N kelime öbeği bir araya getirilerek 100, 200 ve 400 lük kelime boyutlarına sahip özetler ayrı ayrı oluşturulmaktadır. Elde edilen özetlere dair bir çok farklı ROUGE performans metriği açısından sunulan modelin başarısı detaylıca test edilmektedir.

Metin özetleme için önerilen modelin aşamalarına ait blok diyagramı Şekil 5.10'da gösterilmektedir.



Şekil 5.10. Önerilen modelin blok diyagramı

Tez kapsamında önerilen ikinci özetleme modeline ait blok diyagramı Şekil 5.10’ da gösterilmektedir. Diyagram dikkate alındığında özetleme sistemi temelde dört ana adımdan oluşmaktadır. Belirtilen bu adımların detaylandırılması ile birlikte toplamda sekiz adımdan oluşan bir işlem süreci gözlenmektedir. Özetleme sisteminde her bir adımda yapılan işlemler ayrıntılı olarak aşağıda maddeler halinde açıklanmaktadır.

1. **Metin ön işleme(Preproccesing)**

   Önerilen yöntemin ilk aşamasını oluşturmaktadır. Bu aşamada günlük konuşma dilinden oluşturulan ham metinler üzerinde temel metin ön işlemi uygulanmakta ve verilen metinde bulunan gereksiz kelimeler, gereksiz boşluklar ve ilgisiz karekterkerin metinden çıkarılarak metinin daha işlenebilir hale getirilmesi sağlanmakardır.

2. **KUSH Metin işleme**

   Bir önceki adımdan temizlenmiş olan metinler KUSH metin işleme aracına giriş metni olarak verilmektedir. Bu aşamada metin içerisinde bulunan kelimeler KUSH aracına özgü algoritmalar kullanılarak sadeleştirilerek cümleler arasında bulunan güçlü ilişkilerin tespit edilmesi sağlanmakta ve bir sonraki adımda oluşturulacak çizgenin doğrululuğu ve tutatlılığının artması sağlanmaktadır.

3. **Metinsel çizge oluşumu için ön hazırlık**

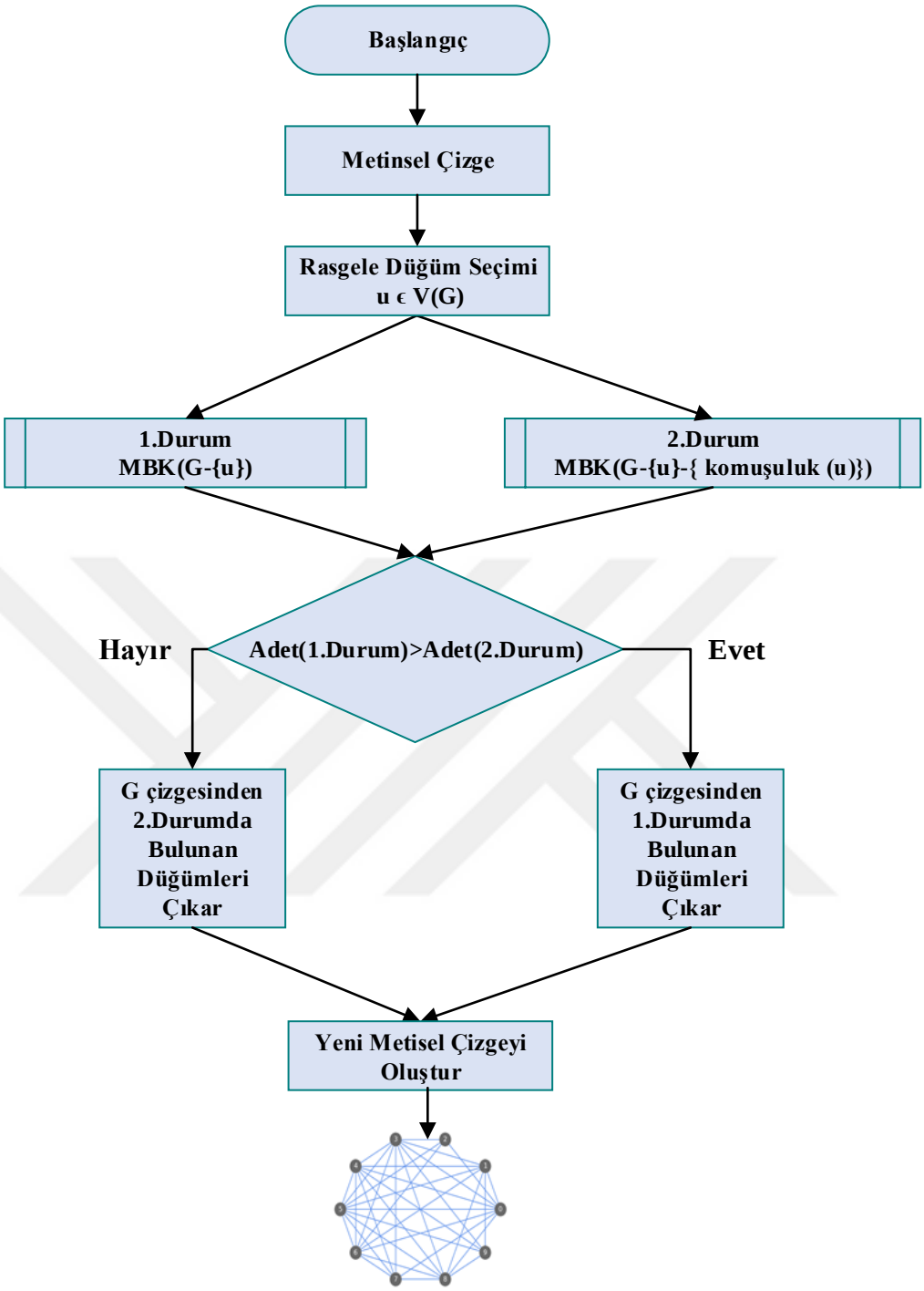
   Sisteme giriş olarak verilen ham metin üzerinde yapılan işlemler sonrasında elde edilen ve çizge oluşum adımına hazır metinler üzerinde işlemler yapılmaktadır. Bu işlemler metinlerin cümlelere ayırılarak cümle vektörünün , cümlelerin kelimelere ayırılarak kelime vektörlerinin ve her iki vektörün ilişilendirilerek cümle kelime matrisinin oluşturulması işlemidir.

4. **Metinsel çizge oluşumu(Textual Graph)**

   Bir önceki aşamadan elde edilen vektör ve matrisler dikkate alınarak metinsel çizge oluşturma aşamasına geçilmektedir. Burada herbir cümle bir düğümü temsil etmekte ve cümle kelime mattirisi baz alınarak cümleler arasındaki ortak kelime sayılarıda ayrıt ağırlıklarını oluşturmaktadır. Cümleler arasında herhangi bir ortak kelime bulunmaması durumunda ayrıt oluşturulmamaktadır.

5. Maksimum bağımsız kümenin bulunması

Özetleme sistemlerinin en önemli parametresi, oluşturulacak özet metine seçilecek cümlelerim maksimum kapsama değerine sahip olmasıdır. Yani özette bulunmaya değer cümlelerin tespitinin doğru yapılması gerekmektedir. Başka bir ifade ile özette bulunmaması gereken cümlelerin seçilerek metinden çıkarılması ve özete seçilme ihtimalinin ortadan kaldırılması gerekmektedir. Bu fikirden yola çıkarak önemsiz cümlelerin tespiti maksimum bağımsız kümenin bulunması yöntemi ile yapılmaktadır.Maksimum bağımsız kümeleri bulma fonksiyonu öz yinelemi olarak çalışmakta ve her defasında rastgele bir düğüm seçmektedir. Öncelikle seçilen düğümün komşuluk ilişkisi bulunan düğümler çıkarılarak maksimum bağımsız küme hesaplanmakta sonrasında komşuluğu bulunan düğümler arasında maksimum bağımsız küme hesaplanmaktadır. Her iki durumda hesaplanan kümelerde bulunan düğüm sayıları karşılaştırılmakta ve en fazla düğüme sahip olan küme maksimum bağımsız küme olarak belirlenmektedir. Bu işlem bütün düğümler için en fazla düğüm sayısına sahip küme bulunana kadar öz yinelemeli olarak çalıştırılmaktadır. Şekil 5.11' de bir adım için yapılan işlelerin grafiksel gösterimi ayrıntılı olarak ifade edilmektedir.



Şekil 5.11. Öz yinelemeli Maksimum Bağımsız küme bulma fonksiyonuna ait tek adım için grafiksel gösterim

6. **Yeni çizgenin oluşturulması**

   Elde edilen maksimum bağımsız kümelerin başlangıçta oluşturulan çizgeden çıkarılmasından sonra elde edilen yeni metinden yeni çizge oluşturulmaktadır.

7. **Düğüm ağırlıklandırma**

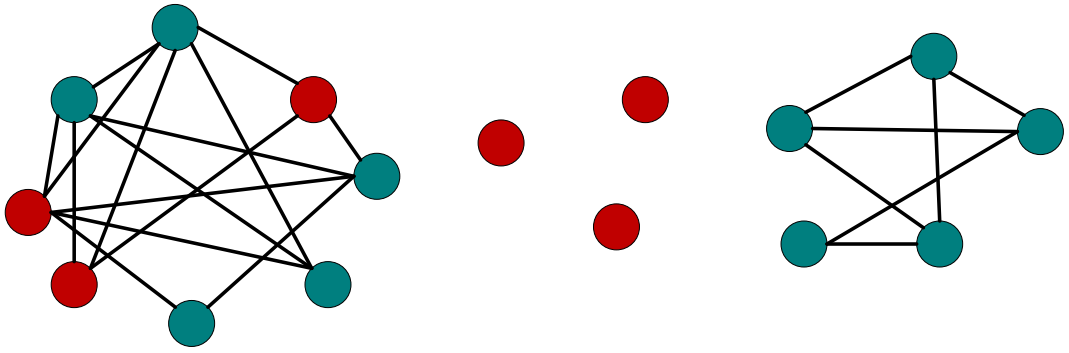
Bir önceki adımda elde edilen ve özette bulunmaması düşünülen cümlelerin çıkarıldığı çizge bulunan düğümler Özvektör merkezilik değerleri bulunmaktadır.

8. **Cümle Seçimi ve özet oluşturma**

Özvektör merkezilik değerleri ile ağırlıklandırılan düğümler ağırlık değerlerine göre büyükten küçüğü sıralanmakta ve istenilen özet boyutuna göre (100,200,400) en değerli düğümden başlanılarak özete koyulacak cümleler seçilmektedir.

5.5.1.1 Maksimum Bağımsız kümelerin bulunması

Bu tezde yapılan çalışmada özetleri elde edilecek dokümanlarda bulunan cümleler içerisinden özette bulunmaması düşünülen cümlelerin çıkarılması ve özet için cümle seçimi yapılmadan önce üzerinde çalışılacak cümle sayısının belli bir oranda azaltılması amaçlanmaktadır. Gereksiz ve anlam taşımayan cümlelerin ayrıştırılması işlemi çizgelere ait maksimım bağımsız düğüm kümeleri kullanılarak yapılmaktadır. Özeti oluşturulacak dökümanın öncelikle çizgeye dönüştürülme işlemi gerçekleştirilmekte ve sonrasında bu çizgeye ait maksimum bağımsız küme belirlenmektedir. Belirlenen bağımsız kümede bulunan düğümler ana dökümandan çıkarılarak geriye kalan cümleler üzerinde düğüm merkezilik değerleri kullanılarak özetleme adımına gidilmektedir.



Şekil 5.12 basit bir çizgeden bağımsız kümelerin bulunarak çizgeden çıkartılması

Şekil 5.12'de görüldüğü üzere küçük bir dokümanın çizgelere dönüşümü ardından bağımsız kümelerin bulunması ve dokümandan çıkarılma işlemi görülmektedir. Dokümanda bulunan cümle sayısı ve oluşturulan ilk çizgedeki düğüm sayısı 'n' olarak kabul edilmektedir. Ana çizgeden elde edilen bağımsız kümenin düğüm sayısı 'm' olmak üzere, ilk aşamada ana cümlede bulunan 'k' nıncı düğümün herhangi bir puanlandırma olmaksızın

özette bulunma ihtimali $P_k = (\frac{1}{n})$ iken bağımsız kümeler ana metinden çıkarıldıktan sonra $P_k = (\frac{1}{n-m})$ olmaktadır. Böylelikle düğüm merkeziliklerini hesaplamadan önce paragrafta anlamsal olarak güçlü kabul edilen cümlelerin özette bulunabilme olasılıkları belli bir oranda arttırılmaktadır. Öyle ki elde edilen başlangıç durumundaki çizgeden Özvektör merkezilik değerleri bulunurken çok daha fazla cümle dikkate alınarak işlem yapılmaktadır ve bu da özette bulunmaması gereken cümlelerin özette bulunma ihtimalini arttırmaktadır. Bağımsız kümede bulunan cümlelerin ana dökümandan çıkarılması ile Özvektör merkezilik değerleri yüksek olan fakat özette yer almaması gerektiği düşünülen cümlelerin tespiti yapılmakta ve böylelikle daha yüksek oranda doğrulukta özetler elde edilmektedir.

Önerilen yöntem kısaca şu şekilde açıklanabilir. Verilen metin (T) dosyasından G çizgesi elde edilir. $G = (V, E)$ bir çizge olmak üzere, bu çalışmada önerilen yöntemde bu çizgenin maksimum bağımsız kümesi bulunmaktadır. Bu küme $S \subset V$ olmak üzere maksimum bağımsız küme S olduğundan

$G2 = (V_2, E_2)$ şeklinde bir çizge tanımlanmaktadır.

T→G

$E_2 \subset E$ ve $S \subset V$.

$E_2 = \{(u,v) | u \in V\text{-}S$ ve $v \in V\text{-}S\ \}$ ve $V_2 = V\text{-}S$.

G_2→L matrisi (L matrisi G_2 çizgesinin Laplace matrisi).

L→Özvektör ve Özdeğer

Özvektör ve Özdeğer→Özetleme.

Maksimum bağımsız kümeyi bulmak yaklaşım gerektiren bir problem olup, kesin çözümü bulan ve etkili olan bir yöntem henüz geliştirilmemiştir. Bu çalışmada Maksimum bağımsız kümelerin tespitini yapabilmek için literatürde kullanılan maksimum bağımsız küme bulma algoritması kullanılmıştır.

Maksimum Bağımsız Küme Bulma Algoritması

```
Maximum_Independent_Set(G,S)
1   G=(V,E) and S=∅
2   if ∀v∈V and degree(v)=0 ise result S=V.
3   Other cases;
4     u∈V select node.
5     n1= Maximum_Independent_Set( (G-{u})
```

6 n2= Maximum_Independent_Set((G-{u}-{ neighbors (u)})

7 S=maximum(n1,n2)

8 Return S

Önerilen özetleme sisteminin algoritması

Input: KUSH metin ön işleme aracından elde edilen işlenmiş metin.

Output: Aday_özet

1 V : Düğümler Kümesi

2 E: Ayrıtlar Kümesi

3 G ← (V,E)

4 **I←G çizgesinde bulunan maksimum bağımsız kümeyi bul.**

5 **for each** I_i **in** I

6 Kaynak dökümandan $\mathbf{I}_i^{th}$ cümleyi çıkar.

7 **End For**

8 Yeni düğüm ve ayrıt ağırlıklarını hesapla

9 Yeni metinsel çizgeyi oluştur $G_{yeni} \leftarrow (V_{yeni}, E_{yeni})$

10 Düğüm merkeziliğini hesapla(Özvektör Merkeziliği)

11 Gyeni, de bulunan bütün düğümlerin hesaplanan merkezilik değerlerini baz alarak büyükten küçüğe sırala.

12 En yüksek değerli düğümleri kaynak dökümandan seçerek 100,200 ve 400 kelimelik aday özetler oluştur.

13 Return Aday_özet

5.5.2. Deneysel Sonuçlar

Çalışma kapsamında kullanılan DUC-2002 ve DUC-2004 veri setlerinde yer alan metinler .txt uzantısına sahip dosyalarda tutulmaktadır. Metin önişleme aşamasında kelime öbekleri arasındaki ilişkilerin metinsel çizgelere en doğru bir biçimde aktarılabilmesini sağlayabilmek için çalışma kapsamında önermiş olduğumuz KUSH adlı bir yazılım aracı kullanıldı. Basit ve kullanışlı KUSH aracı sayesinde değişikliğe uğrayan kelimelerin yerine, yine özetlenecek metin içinden atamalar gerçekleştirmiştir. Sonuç olarak metinlere ait yüksek temsil gücüne sahip çizgeler elde edilmiştir.

Çalışma kapsamında daha önce hiçbir özetleme çalışmasında kullanılmamış olan maksimum bağımsız kümeler kullanılmıştır. Maksimum bağımsız kümelerde yer alan düğümlere karşılık gelen cümlelerin özette yer almaması gerektiği öngörüsünden hareket edilerek çizgeler üzerinde bağımsız kümeleri oluşturan düğümler belirlenmiştir ve çizgeden çıkarılmıştır. Böylece düğümlerin genel çizge üzerindeki etkisi sayısal olarak ölçülmeden evvel özet üzerinde bir sınırlamaya gidilmiştir. Bu sınırlama ile oluşacak özette yer alacak kelime gruplarının tekrarı engellenerek çok daha kapsayıcı özetler elde edilebilmiştir. Ayrıca gerçekleştirilen deneysel süreçler, bir özetleme çalışmasında ilk defa kullanılan maksimum bağımsız kümelerin uzaklaştırılması yaklaşımının, ana metin içerisinde birbirleri ile kesişen minimum kelimeye sahip olan cümlelere özette yer verilmesini teşvik edici bir adım olarak etki etmektedir. Çalışmanın deneysel süreçleri boyunca rapor edilen değerler, bu yenilikçi yöntemin katkılarını açık bir biçimde ortaya koymaktadır. Son aşamada ise, elde edilen düğümlerin düğüm merkezilik yöntemleri ile merkezilik değerleri hesaplanmaktadır. Çalışma kapsamında, çizgeleri oluşturan düğümler ağırlıklandırılırken Özvektör merkezilik değeri kullanılmıştır. Sunulan özetleme sisteminin performansını etraflıca değerlendirmek için deneysel süreç 100, 200 ve 400 kelimeden oluşacak özetler için ayrı ayrı yürütülmüştür. Önerilen doküman özetleme modelinin performansı Çizelge 5.9' da yer verilen ROUGE performans metrikleri kullanılarak gösterilmektedir.

Çizelge 5.9 200 ve 400 kelimelik özetler için elde edilen ROUGE değerleri

ROUGE Değerlendirme ölçütü	**200 Kelimelik Özet**			**400 Kelimelik Özet**		
	Recall	**Precision**	**F-Score**	**Recall**	**Precision**	**F-Score**
Rouge-1	0.51954	0.47742	0.49726	0.59208	0.56983	0.58060
Rouge-2	0.24704	0.22714	0.23651	0.32471	0.31207	0.31819
Rouge-3	0.18677	0.17210	0.17900	0.26191	0.25137	0.25647
Rouge-4	0.17145	0.15794	0.16430	0.24083	0.23096	0.23574
Rouge-L	0.48306	0.44376	0.46227	0.56361	0.54225	0.55258
Rouge-W-1.2	0.17427	0.25233	0.20605	0.17856	0.27067	0.21510
Rouge-S*	0.23114	0.19577	0.21143	0.32032	0.29671	0.30775
Rouge-SU*	0.23388	0.19823	0.21403	0.32163	0.29798	0.30905

Değerlendirme sürecinde Rouge-1, Rouge-2, Rouge-3, Rouge-4, Rouge-L, Rouge-W-1-2, Rouge-S* ve Rouge-SU* metrikleri kullanılmıştır. Ayrıca her bir metrik türü Recall, Precision, F-Score odaklı ayrı ayrı rapor edilmiştir. 100,200 ve 400 kelimeden oluşan özetler

için değerler ayrı ayrı sunulmuştur. Çizelge 5.9'un ilk sütunu özetleme performans ölçümü için kullanılan ölçüm türlerini gösterirken kalan sütunlar 200 ve 400 kelimelik özetler için önerilen yöntem tarafından rapor edilen %95 güven aralığına sahip performans değerlendirme puanlarıdır.

Önerdiğimiz metin özetleme yöntemi ile elde ettiğimiz özetleme performans sonuçları önceki çalışmalar ile karşılaştırılmıştır. Daha önce gerçekleştirilen özetleme yaklaşımları ile uyumlu olması bakımından Çizelge 5.10 , Çizelge 5.11 ve Çizelge 5.12 'de görüldüğü gibi 200 , 400 ve 100 kelimelik özetler yedi farklı rekabetçi yöntem ile karşılaştırılmıştır.

Karşılaştırma için seçilen modellerin tercih edilmesindeki temel neden çalışma kapsamında sunulan özetleme yöntemi ile benzer şekilde matematiksel, istatistiksel veya çizge tabanlı teoriler kullanıyor olmalarıdır. Ayrıca seçilen yöntemlerin tamamı sunulan yöntemde olduğu gibi denetimsiz bir doküman özetleme yöntemidir.

Random, Luhn, LSA, TextRank, LexRank, SumBasic, KL-Sum ve bu çalışmada önerilen doküman özetleme yönteminin, DUC-2002 veri seti üzerinde rapor ettikleri 200 kelimelik özetlerine ait ROUGE performans metrik değerleri Çizelge 5.10'da sunulmaktadır. İlgili tablo üzerinde her satırda yer alan en yüksek değerler kalın rakamlarla vurgulanmıştır. Çalışma kapsamında sunulan yöntem istisnasız tüm rekabetçi yöntemleri geride bırakmaktadır. Çizelge 5.10 açıkça ortaya koymaktadır ki, 200 kelimelik özetler bazında yaklaşımımız rekabetçi yöntemlerle kıyaslandığında çok iyi sonuçlar vermiştir.

Sunulan yöntemin uygulanabilirliğini vurgulamak için tüm rekabetçi yöntemleri ve çalışma kapsamında önerilen metin özetleme yaklaşımının performanslarını daha detaylı biçimde incelemek için deneysel süreçler 400 kelimelik özetler için tekrarlanmıştır. Çizelge 5.11'de yer verilen sonuçlar elde edilmiştir. İlgili tabloda her satırda yer alan en yüksek değerler kalın rakamlarla vurgulanmıştır. Sunulan yöntem, Rouge-1 ve F-Score bazlı bir değerlendirme ile Random'dan %12, Luhn'dan %0.5, LSA'dan %24, TextRank'dan %5, LexRank'dan %7, SumBasic'den %12 ve KL-Summ özetleme yönteminden de %33 daha yüksek değerler rapor etmiştir. Açıkça görülmektedir ki 400 kelimelik özetler bazında da rekabetçi yöntemler birçok ölçüm değerinde geride bırakılmıştır.

Ayrıca farklı bir karakteristiğe sahip veri seti olan DUC-2004 veri kümesi kullanılarak 100 kelimelik özetlerle yapılan deneylerimizde, önerilen algoritmanın ROUGE-2 değerlerine dayanan tüm performans kriterleri için diğer rekabetçi yöntemleri aştığı görülebilir. ROUGE -1, ROUGE-L, ROUGE-SU ve ROUGE-W-1.2 hassas metrik türleri

için de en iyi değerler bildirilmiştir. Diğer metrik türleri için ise diğer yöntemlerin büyük bir çoğunluğunu geride bırakmakta ve oldukça rekabetçi sonuçlar sunmaktadır. Her satır için en yüksek değerler, sonuç tablolarında kalın yazı tipiyle vurgulanmıştır.

Çizelge 5.10. 200 Kelimelik özetler için önerilen yöntemin diğer yöntemlerle karşılaştırılması

ROUGE Değerlendirme ölçütü		**200 Kelimelik Özet (DUC-2002)**							
		Metot							
		Random	**Luhn**	**LSA**	**TextRank**	**LexRank**	**SumBasic**	**KL-Sum**	**Ours**
Rouge-1	Rec.	0.40932	0.47503	0.37723	0.48417	0.48124	0.46128	0.37464	**0.51954**
	Prec.	0.38148	0.44489	0.37012	0.43685	0.45984	0.45114	0.36803	**0.47742**
	F-Sco.	0.39470	0.45924	0.37342	0.45868	0.46997	0.45597	0.37104	**0.49726**
Rouge-2	Rec.	0.10562	0.20820	0.08950	0.16921	0.18507	0.15947	0.12113	**0.24704**
	Prec.	0.09868	0.19558	0.08910	0.15195	0.17566	0.15584	0.11938	**0.22714**
	F-Sco.	0.10199	0.20160	0.08921	0.16000	0.18010	0.15757	0.12018	**0.23651**
Rouge-L	Rec.	0.36328	0.44490	0.33477	0.46748	0.43813	0.39661	0.34066	**0.48306**
	Prec.	0.33880	0.41687	0.32892	0.42183	0.41867	0.38768	0.33470	**0.44376**
	F-Sco.	0.35043	0.43022	0.33162	0.44289	0.42787	0.39194	0.33741	**0.46227**
Rouge-W-1.2	Rec.	0.12324	0.15858	0.11295	0.15655	0.15201	0.13782	0.11839	**0.17427**
	Prec.	0.18074	0.23417	0.17475	0.22316	0.22825	0.21231	0.18321	**0.25233**
	F-Sco.	0.14647	0.18902	0.13715	0.18382	0.18237	0.16710	0.14375	**0.20605**
Rouge-SU*	Rec.	0.15112	0.19901	0.12783	0.21541	0.20321	0.17981	0.13361	**0.23388**
	Prec.	0.13124	0.17444	0.12342	0.17468	0.18494	0.17190	0.12913	**0.19823**
	F-Sco.	0.14020	0.18556	0.12527	0.19200	0.19309	0.17548	0.13098	**0.21403**

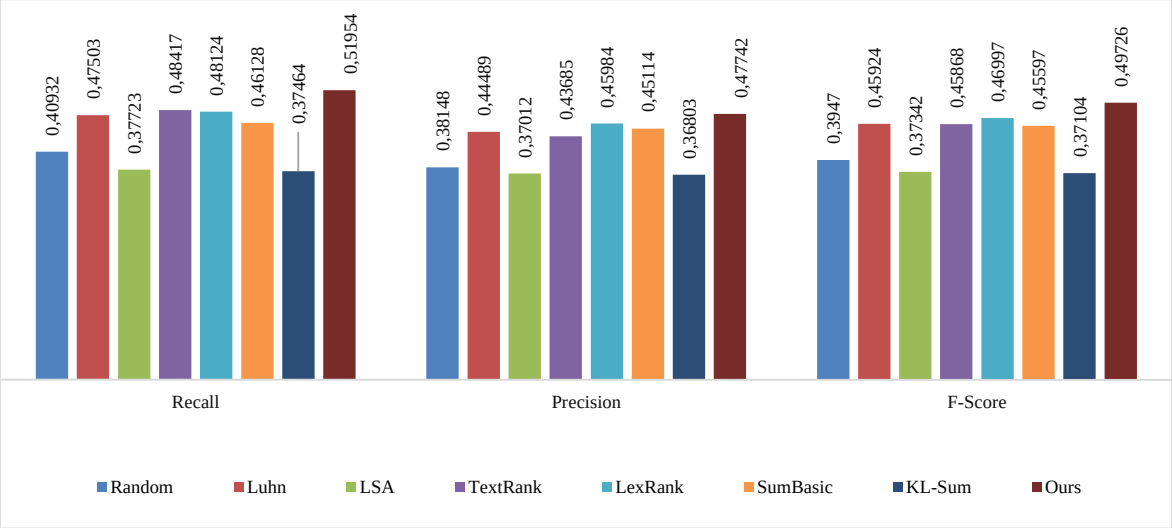
Çizelge 5.11. 400 Kelimelik özetler için önerilen yöntemin diğer yöntemlerle karşılaştırılması

ROUGE Değerlendirme ölçütü		**400 Kelimelik Özet (DUC-2002)**							
		Metot							
		Random	**Luhn**	**LSA**	**TextRank**	**LexRank**	**SumBasic**	**KL-Sum**	**Ours**
Rouge-1	Rec.	0.52799	0.57929	0.46361	0.57812	0.55018	0.52373	0.43746	**0.59208**
	Prec.	0.50893	**0.57584**	0.47293	0.52849	0.53431	0.50610	0.43254	0.56983
	F-Sco.	0.51819	0.57746	0.46804	0.55179	0.54188	0.51462	0.43490	**0.58060**
Rouge-2	Rec.	0.20698	**0.32775**	0.16001	0.25654	0.23706	0.18916	0.16256	0.32471
	Prec.	0.19967	**0.32546**	0.16297	0.23379	0.23088	0.18193	0.16017	0.31207
	F-Sco.	0.20322	**0.32654**	0.16141	0.24447	0.23382	0.18542	0.16132	0.31819
Rouge-L	Rec.	0.48516	0.55118	0.42283	0.56505	0.51566	0.46814	0.40158	**0.56361**
	Prec.	0.46755	**0.54784**	0.43129	0.51644	0.50089	0.45228	0.39708	0.54225
	F-Sco.	0.47611	0.54940	0.42685	0.53926	0.50793	0.45995	0.39924	**0.55258**
Rouge-W-1.2	Rec.	0.14844	0.17758	0.12558	0.17498	0.16162	0.14401	0.12675	**0.17856**
	Prec.	0.22572	**0.27834**	0.20148	0.25234	0.24770	0.21925	0.19767	0.27067
	F-Sco.	0.17905	**0.21675**	0.15464	0.20648	0.19550	0.17378	0.15442	0.21510
Rouge-SU*	Rec.	0.25020	0.29304	0.19823	0.31213	0.27351	0.24097	0.17922	**0.32163**
	Prec.	0.23189	0.28977	0.20537	0.26032	0.25871	0.22412	0.17485	**0.29798**
	F-Sco.	0.24052	0.29118	0.20140	0.28307	0.26541	0.23197	0.17688	**0.30905**

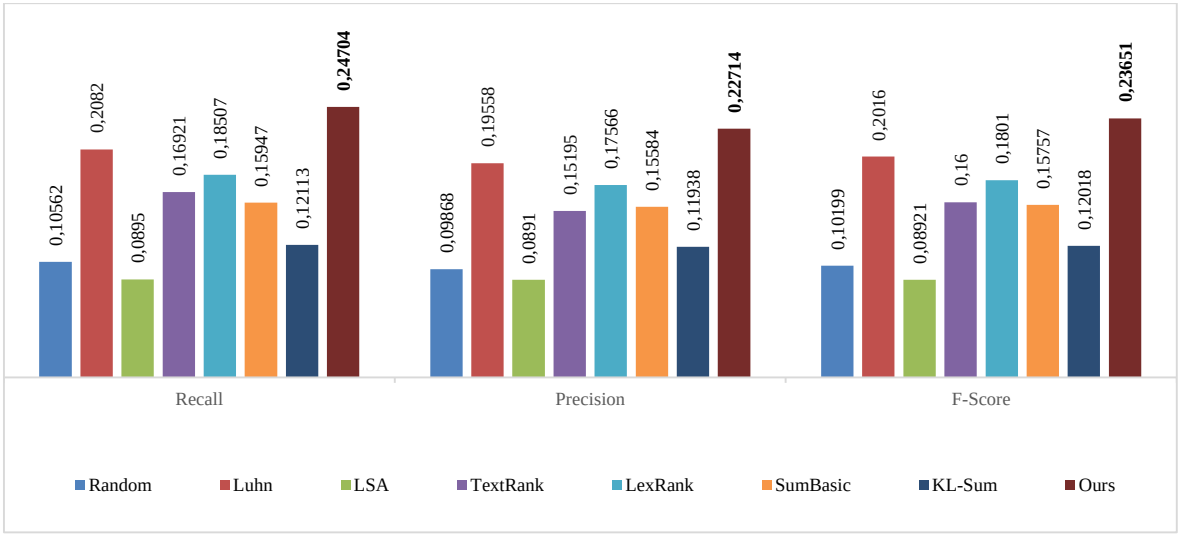
Çizelge 5.12. 100 Kelimelik özetler için öneriken yöntemin diğer yöntemlerle karşılaştırılması

ROUGE Değerlendirme ölçütü		**100 Kelimelik Özet (DUC-2004)**							
		Metot							
		Luhn	**LSA**	**Textrank**	**Lexrank**	**Random**	**SumBasic**	**KL-Sum**	**Ours**
Rouge-1	Rec.	0.35768	0.27995	0.31301	**0.39893**	0.32206	0.33859	0.33778	0.38072
	Prec.	0.29675	0.25740	0.28574	0.33462	0.30458	0.31882	0.32224	**0.34019**
	F-Sco.	0.32333	0.26733	0.29759	**0.36292**	0.31255	0.32808	0.32924	0.35879
Rouge-2	Rec.	0.05205	0.03324	0.04145	0.07977	0.05439	0.05623	0.07030	**0.08277**
	Prec.	0.04292	0.03068	0.03862	0.06831	0.05170	0.05226	0.06883	**0.07373**
	F-Sco.	0.04691	0.03182	0.03983	0.07338	0.05292	0.05413	0.06946	**0.07786**
Rouge-L	Rec.	0.27711	0.21295	0.26099	**0.30685**	0.25785	0.24408	0.27448	0.29037
	Prec.	0.22885	0.19611	0.23881	0.25868	0.24444	0.23038	0.26256	**0.25945**
	F-Sco.	0.24987	0.20354	0.24845	**0.27991**	0.25053	0.23680	0.26789	0.27364
Rouge-W-1.2	Rec.	0.09350	0.07427	0.08836	**0.10436**	0.08957	0.08427	0.09315	0.09908
	Prec.	0.14127	0.12498	0.14797	0.16113	0.15517	0.14533	0.16340	**0.16197**
	F-Sco.	0.11212	0.09286	0.11011	**0.12622**	0.11334	0.10656	0.11841	0.12278
Rouge-SU	Rec.	0.11866	0.07285	0.09741	**0.13998**	0.09532	0.10063	0.10609	0.13062
	Prec.	0.08109	0.06058	0.08277	0.09683	0.08421	0.08946	0.09795	**0.10372**
	F-Sco.	0.09516	0.06534	0.08820	0.11328	0.08884	0.09437	0.10117	**0.11493**

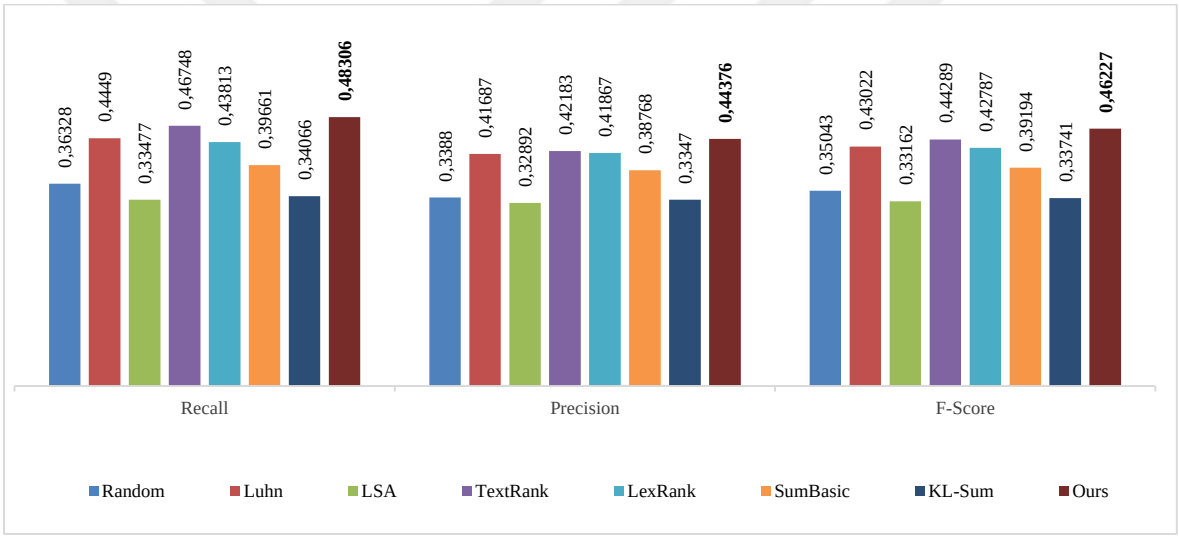
Yöntemimiz dahil olmak üzere sekiz sistemin tümü için % 95 güven aralıklarıyla ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-W-1.2 ölçüm metriklerinden alınan değerler sırasıyla Şekil 5.13, Şekil 5.14, Şekil 5.15 ve Şekil 5.16'da grafiksel olarak gösterilmektedir.. İlgili grafiklerde yer alan karşılaştırmaların tümü 200 kelimelik özetler için Recall, Precision ve F-Score değerleri bazında ayrı ayrı grafiksel olarak gösterilmiştir. Sunulan yöntemin, Rouge-1 ve F-Score bazlı bir değerlendirme ile Random'dan %25, Luhn'dan %8, LSA'dan %33, TextRank'dan %8, LexRank'dan %5, SumBasic'den %9 ve KL-Summ özetleme yönteminden de %34 daha yüksek değerler rapor etmiştir. İlgili tablo ve şekilde görüldüğü üzere benzer değerlendirme oranları diğer bütün ROUGE performans ölçüm yöntemleri içinde elde edilmiştir. Bu durum sunulan metin özetleme sisteminin son derece etkili ve verimli bir özetleme sistemi olduğunu doğrulamaktadır. Benzer Şekilde 400 kelimelik özetler için elde edilen sonuçlar Şekil 5.17, Şekil 5.18, Şekil 5.19 ve Şekil 5.20 'da, 100 kelimelik özetler için elde edilen sonuçlar ise Şekil 5.21, Şekil 5.22, Şekil 5.23 ve Şekil 5.24 de grafiksel olarak gösterilmektedir.



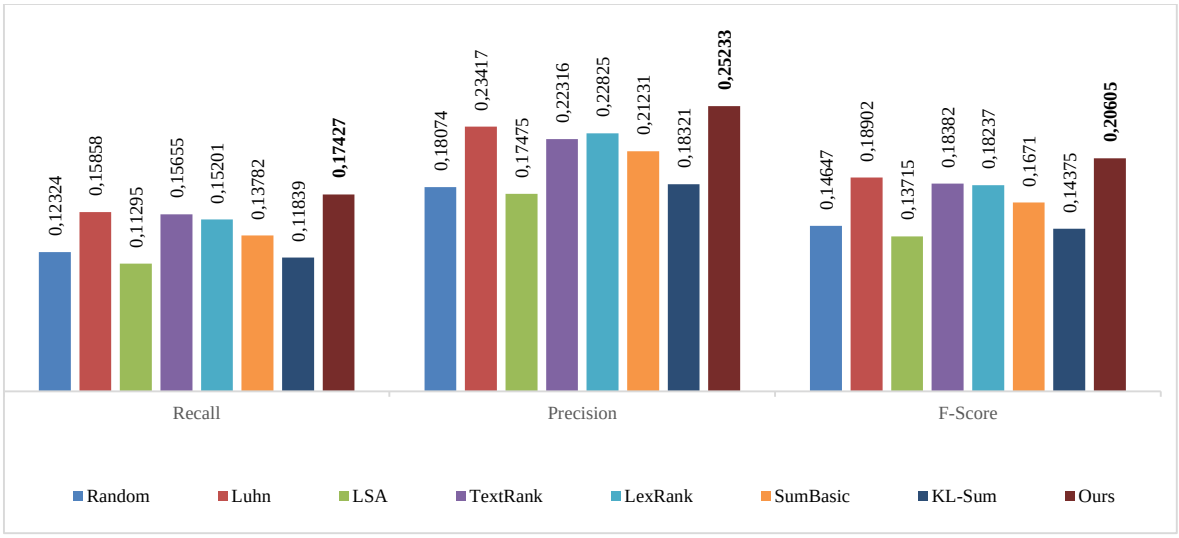
Şekil 5.13. 200 kelimelik özetler için ROUGE-1 değerinin karşılaştırılması



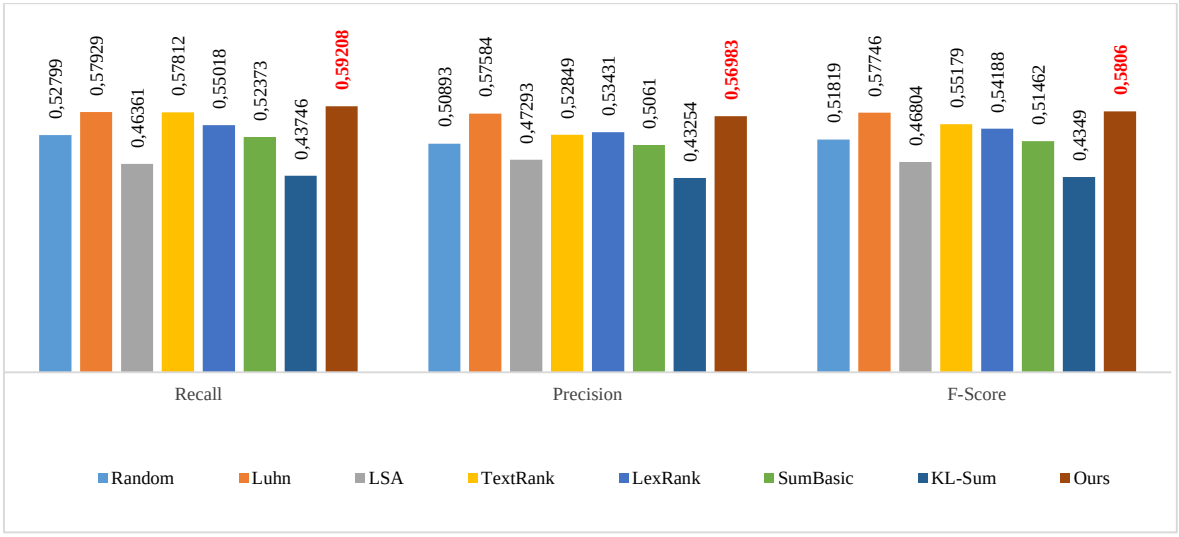
Şekil 5.14. 200 kelimelik özetler için ROUGE-2 değerinin karşılaştırılması



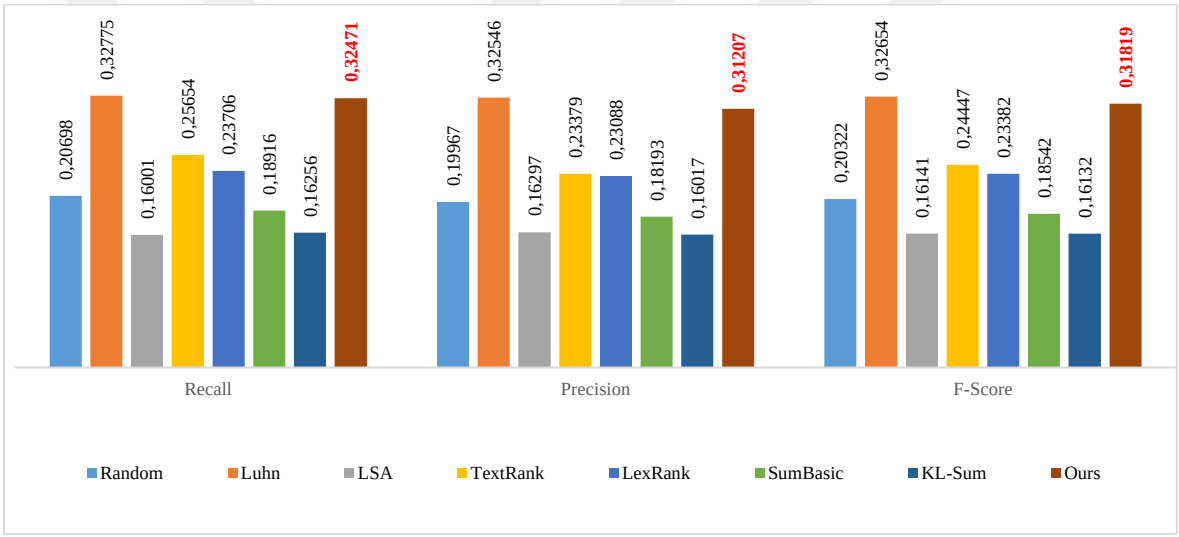
Şekil 5.15. 200 kelimelik özetler için ROUGE-L değerinin karşılaştırılması



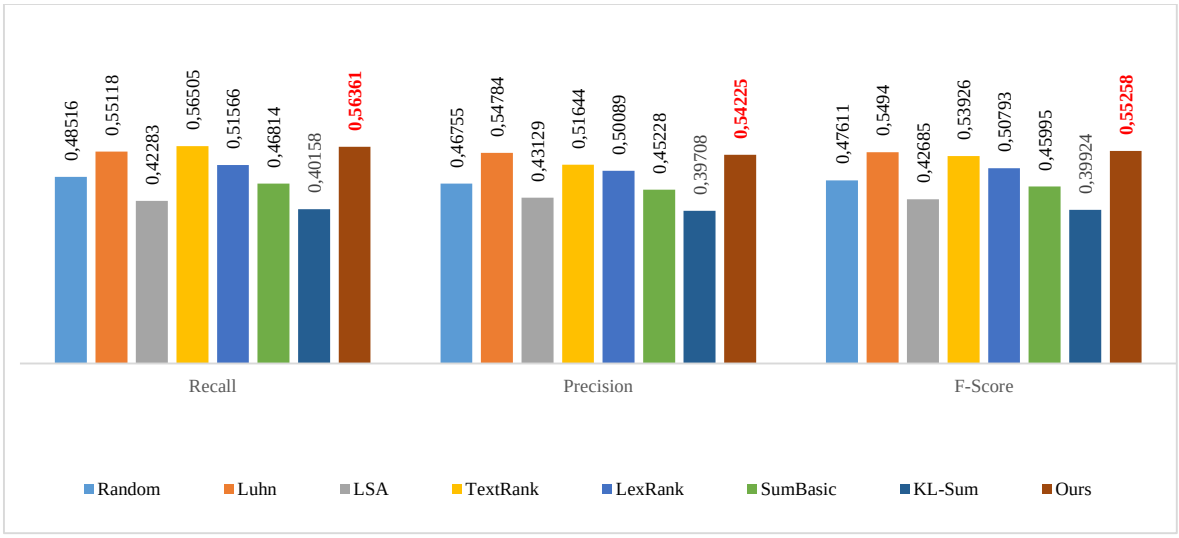
Şekil 5.16. 200 kelimelik özetler için ROUGE-W 1,2 değerinin karşılaştırılması



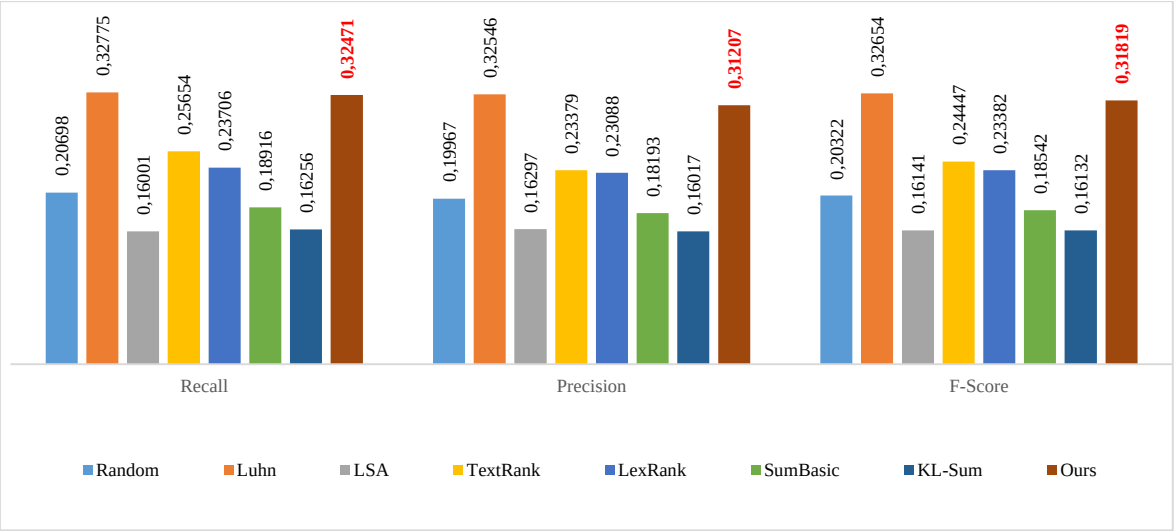
Şekil 5.17. 400 kelimelik özetler için ROUGE-1 değerinin karşılaştırılması



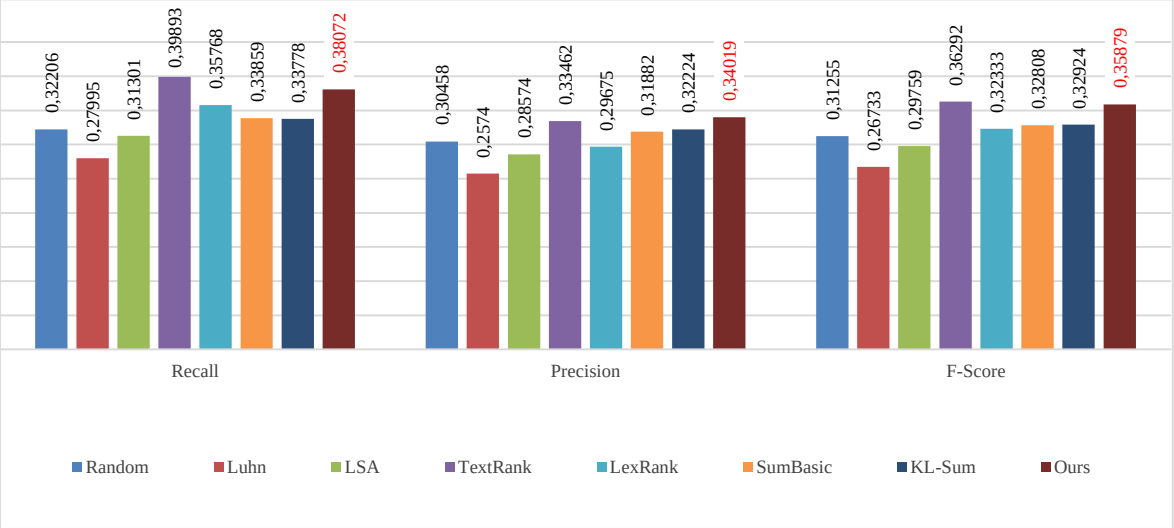
Şekil 5.18. 400 kelimelik özetler için ROGE-2 değerinin karşılaştırılması



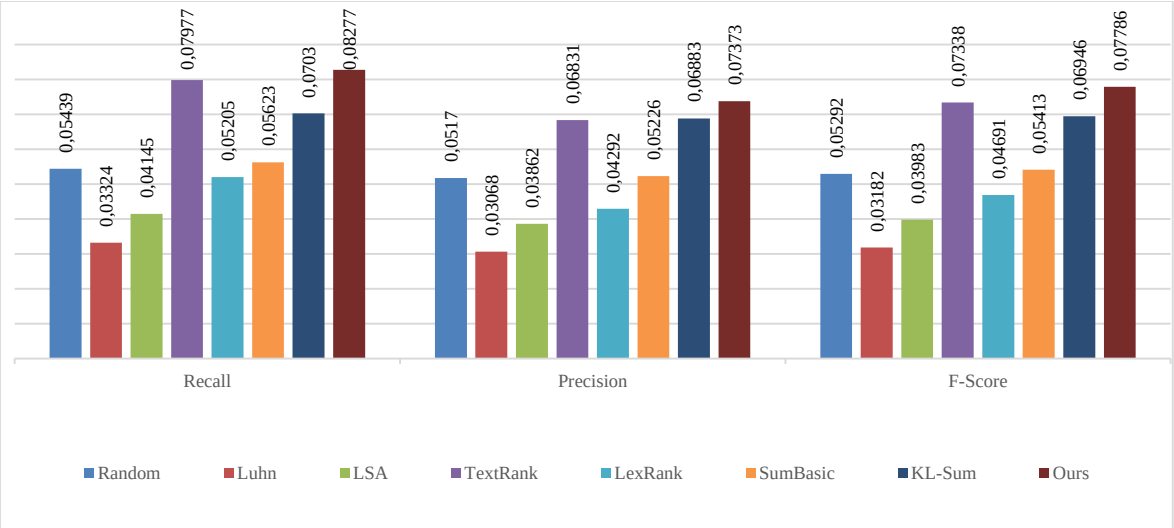
Şekil 5.19. 400 kelimelik özetler için ROUGE-L değerinin karşılaştırılması



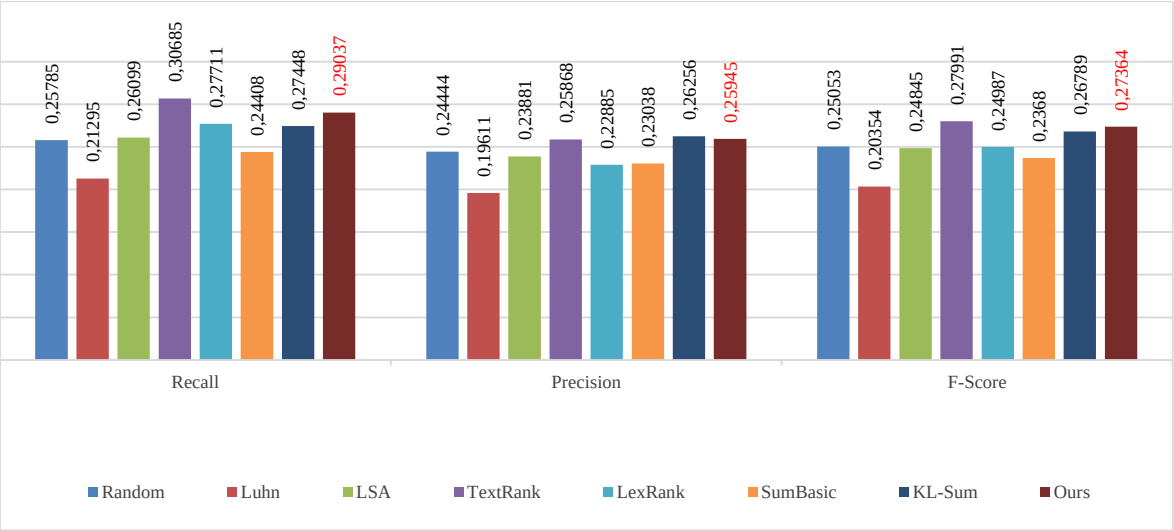
Şekil 5.20. 400 kelimelik özetlerin ROUGE-W 1.2 değerinin karşılaştırılması



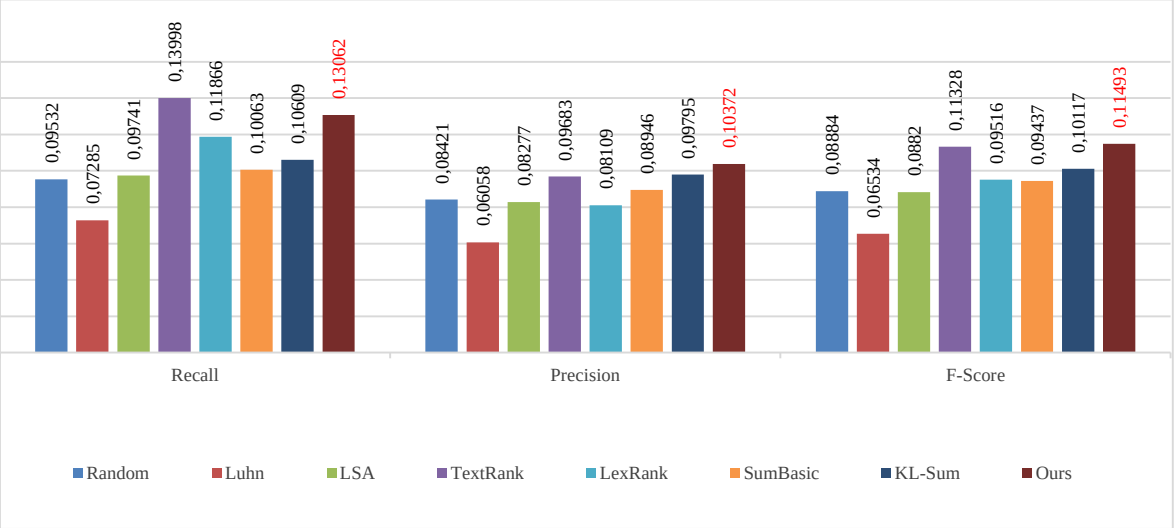
Şekil 5.21. 100 kelimelik özetlerin ROUGE-1 değerinin karşılaştırılması



Şekil 5.22. 100 kelimelik özetlerin ROUGE-2 değerinin karşılaştırılması



Şekil 5.23. 100 kelimelik özetlerin ROUGE-L değerinin karşılaştırılması



Şekil 5.24. 100 kelimelik özetlerin ROUGE-W 1.2 değerinin karşılaştırılması.

Önerilen sistemin performansı, farklı veri setleri ve performans metrikleri kullanılarak ayrıntılı olarak test edildi. Tablolarda ve şekillerde önerilen yöntemin genellikle başarılı sonuçlar verdiği bildirilmiştir. Elde edilen yüksek performanslar kısmen önerilen metin işleme yazılımı (KUSH) kullanılarak zengin metin tablolarının oluşturulmasıyla da ilgilidir. Ancak, başarının ana nedeni; sezgisel olarak, düğüm ağırlıklandırma aşamasından önce bağımsız kümeleri oluşturan düğümlerin çıkarılmasıdır. Böylece, özete seçilecek uygun cümlelerin seçimi daha doğru bir şekilde yapılır. Bu, grafik tabanlı bir önişleme sağlar ve daha basit grafikler üzerinde çalışmayı sağlar. Bu da Özetleme çalışmalarında en önemli parametre olan özetin kapsayıcılık özelliğinin öne çıkmasını sağlamaktadır.

6. SONUÇ VE ÖNERİLER

Bu tez çalışmasında otomatik metin özetleme alanında yapılan çalışmaları daha öteye taşımak amacı ile çizge tabanlı çıkarıcı metin özetleme çalışmaları yapılmıştır. Yapılan çalışmalarda özetleme alanında en çok kullanılan verisetleri olan DUC-2002 ve DUC-2004 verisetleri kullanılmıştır. Otomatik metin özetleme işlemleri günlük hayatta kullanılan dilde ve yapılandırılmamış tamamen ham verilerden oluşan veri setleri üzerinde işlem yapmaktadır. Bu yüzden özetleme işlemine başlanılmadan önce kullanılacak verilerin detaylı ve dikkatli bir ön işlem aşamasında geçirilmesi zorunlu bir aşama olarak görülmektedir. Bu tez kapsamında öncelikle ön işlem aşamasının sağlıklı bir şekilde yürütülmesi cümleler arası ilişkilerin doğru bir şekilde korunabilmesi ve tespit edilmesi amacı ile KUSH adı verilen yeni bir ön işleme algoritmasi önerilmiş ve önişleme aracı geliştirilmiştir. Sonrasında geliştirilen ön işleme aracının kullanıldığı denetimsiz, çıkarıcı ve genel özetleme özetlliklerine sahip çizge tabanlı iki yeni metin otomatik özetleme yöntemi önerilmiştir.

Önerilen özetleme yaklaşımlarında konuşma dilinin beraberinde getirdiği kuralsızlıklar KUSH adını verdiğimiz yazılım aracı ile giderilmiştir ve yazılım aracı kendine özgü algoritması ile metinleri özetleme için hazırlanmıştır. Bu yazılım aracı cümleler arasında anlamsal olarak ayırt edilebilir ve ölçülebilir ilişkilerin elde edilebilmesi bakımından oldukça olumlu etkiler göstermiştir.

Özetleme kapsamında kullanılan metinler çizgeler ile ifade edilmektedir. Temsili çizgelerdeki ilk sonlu kümeyi cümleler temsil etmekte iken diğer sonlu kümeyi ise ortak kelime sayıları temsil etmiştir. Böylece metinlerin yönlendirilmemiş ve ağırlıklı çizgeler ile temsili sağlanmıştır.

Tez kapsamında geliştirilen ilk çalışma olan ve CatSumm adını verdiğimiz çalışma kapsamında özetleme performansına katkıda bulunan, spektral çizge bölmeleme olarak bilinen cebirsel çizge teorisindeki teknikler kullanılarak özeti temsil eden graflar oluşturulmuştur. Spektral bölmeleme sonrasında elde edilen oranlar ölçüsünce özetler elde edilmiştir. Ayrıca çizgelerin metinlere olan temsillerini arttırmak adına çizgelerden elde edilen bir standart sapma ile eşik değeri hesaplanmıştır. Eşik değerin altında değer alan ayrıtlar çizgeden çıkarılmıştır. Çalışmanın son aşamasında düğümlerle temsil edilen cümlelerin önemini, özetin oluşturulması için sayısallaştırabilmek amacı ile belirlenen düğüm merkezilik ölçüm değerleri hesaplanmıştır. Elde edilen puanlar kullanılarak her bir

düğüm merkeziliği yaklaşımı için ayrı ayrı özetler oluşturulmuştur. Bu bağlamda geride bırakılan adımların sonrasında sunulan yöntem bir düğümün derece merkezilik değeri (Degree Cntrality), diğer düğümlere olan yakınlığı (Closeness Centrality), diğer düğümlerle arasındaki en büyük mesafesi (Betweenness Centrality) ve önemli düğümlerin birbirlerine bağlanması (Eigenvector Centrality) bazında ayrı ayrı uygulanmıştır. CatSumm'ın bu farklı yöntemleri baz alarak uyarlanması adımından sonra özet için bir seçim stratejisi elde edilmiştir. Seçili merkezilik değerlerinden elde edilen sonuçlar incelendiğinde Önerilen CatSumm modelinin Özdeğer merkezilik yöntemi ile bir çok ölçüm değerinde daha iyi sonuçlar ürettiği gözlemlenmiştir. Bu sonuclar sonucunda Özvektor merkezilik değerinin önerilen modele daha uygun olduğu kabul edilmiş olup literatürde bulunan diğer modellerle CatSumm+ Özvektör modelinden elde edilen sonuçların karşılaştırılması yapılmaktadır. Tüm bu adımların sonunda kullanılan merkezilik ölçümlerininden en yüksek değerleri alan cümleler seçilerek boyutları 100,200 ve 400 kelime ile sınırlandırılmış özet dosyası oluşturulmaktadır. Spectral çizge bölmeleme yönteminin metin özetleme probleminde kullanılabileceğini gösteren deneyleri içeren çalışmamızda önerilen CatSumm modelinin performansı metin özetleme problemlerinde kullanılan ölçüm metrikleri ile incelenmiştir. Deneysel çalışma 100, 200 ve 400 kelimelik özetler elde edilecek şekilde ayrı ayrı gerçekleştirilmiştir. Çizelge 5.5'tde rapor edildiği gibi Özvektör merkeziliği bazlı CatSumm istisnasız bütün metrikler bazında Random, LSA, LexRank,TextRank, SumBasic ve KLSUM özetleme yöntemlerinden daha yüksek ROUGE değerleri elde etmiştir. Çizelge 5.6'da belirtildiği gibi 400 kelimelik özetler bazında da yüksek ROUGE değerleri rapor edilmektedir. İlgili tablo ve şekiller incelendiğinde 200 kelimelik özetlerden 400 kelimelik özetlere kıyasla daha yüksek metrik değerleri elde edildiği açıkça görülmektedir. Benzer şekilde Çizelge 5.7'de farklı bir veri seti olan DUC 2004 veri seti kullanılarak üretilen 100 kelimelik özetlere ait ROUGE değerleri rapor edilmektedir. Tablodan da açıkça görüldüğü üzere elde edilen sonuçlar ROUGE-2 Recall değeri dışında diğer bütün ölçüm değerlerinde karşılaştırılan yontemlerden daha iyi sonuçlara ulaşıldığı görülmektedir. Elde edilen bu bulguyu önerilen CatSumm yaklaşımının Spectral tabanlı bir süreç oluşu ile açıklamak mümkün olabilmektedir. Bilindiği üzere spektral çizge bölmeleme yöntemleri düğüm sayısı ile ters orantılı bir başarı grafiği ortaya koymaktadır ve dolayısı ile en iyi çözümü garanti edememektedir. Bu dezavantajı KUSH yazılım aracı ile kısmen gidermek mümkün olmaktadır. Bu çalışmada KUSH yazılım aracının kelimelerin aldıkları ekler dolayısı ile farklı tanınmasının önüne geçmesinden ötürü sadece metin özetleme, çalışmalarında değil metin sınıflandırma ve kümeleme çalışmalarında kullanılabileceği öngörülmektedir.

Çalışma kapsamında CatSumm yaklaşımının performansını geleneksel Radom, LSA, LexRank, TextRank, SumBasic ve KLSUM özetleme yöntemleri ile karşılaştırıldı. CatSumm yönteminin daha iyi sonuçlar verdiği gözlemlenmiştir. Açıkça ilgili tablo ve şekillerde ifade edildiği gibi kullanılan performans metrikleri açısından elde edilen ümit verici sonuçlar neticesinde, önerilen yöntemin metin özetlemede kullanışlı ve etkili bir araç olabileceği açıkça görülmektedir. Deneysel çalışmaların sonuçlarına dayanarak KUSH yazılımının araştırmacılar tarafından sıklıkla kullanılan sınıflandırma ve kümeleme yöntemleri öncesinde de kullanılabilecek bir araç olabileceğine inanıyoruz.

Tez kapsamında geliştirilen ikinci çalışmada bağımsız kümede yer alan düğümlere karşılık gelen cümlelerin özette yer almaması gerektiği öngörüsünden hareket edilmiştir. Sözü edilen öngörü ile çizgeler üzerinde bağımsız kümeyi oluşturan düğümler belirlenerek çizgeden çıkarılmıştır. Böylece düğümlerin genel çizge üzerindeki etkisi matematiksel olarak ölçülmeden evvel özet üzerinde bir sınırlamaya gidilmiştir. Çalışma kapsamında önerilen metin özetleme süreci gereği özetlenecek metinleri temsil eden çizgeler üzerinde bağımsız kümeyi oluşturan düğümler belirlenmiştir. Metinsel çizgeyi oluşturan düğümlere dair düğümlerin taşıdığı bilgi miktarları nicel olarak elde edilmeden önce düğümler arasında bir ön değerlendirme gerçekleştirilmiştir. Önerilen model ile daha önce hiçbir özetleme çalışmasında rastlanmayan ve matematiksel bir model ile kurgulanan altyapıya sahip bir özetleme yaklaşımı sunulmaktadır. Modelin başarısı Recall, Precision, F-Score tabanlı, Rouge-1, Rouge-2, Rouge-3, Rouge-4, Rouge-L, Rouge-W-1-2, Rouge-S*, Rouge-SU* performans metrikleri açısından ortaya konulmuştur. Ayrıca deneysel süreçler 100, 200 ve 400 kelimelik özetler için tekrarlanmıştır. Deneysel süreçler boyunca aşağıdaki sonuçlar elde edilmiştir. % 95 Güven aralığında 100 ve 200 kelimelik özetlere dair, çalışmaya konu olan istisnasız karşılaştırılan bütün metotlardan daha üstün sonuçlar rapor edilmiştir. % 95 güven aralığında 400 kelimelik özetlerde ise büyük oranda karşılaştırılan bütün metotlardan daha yüksek sonuçlar elde edilmiştir.

Düğümlerin genel çizge üzerindeki ağırlığı matematiksel olarak sıralanmadan evvel, özetleme çalışmaları kapsamında benzersiz bir yaklaşım sergilenerek bağımsiz kümeleri oluşturan düğümlerin ana çizgeden soyutlanması yolu ile özetlerin oluşturulması daha önce hiçbir döküman özetleme çalışmasında kullanılmayan bir yaklaşımdır. Bu sınırlama ile oluşacak özette yer alacak kelime gruplarının tekrarı engellenerek çok daha kapsayıcı özetler elde edilebilmiştir. Ayrıca gerçekleştirilen deneysel süreçler, bir özetleme çalışmasında ilk defa kullanılan bağımsız kümelerin uzaklaştırılması yaklaşımının, ana metin içerisinde

birbirleri ile ilişkisi maksimum olan cümlelere özette daha az yer almasını teşvik edici bir ön değerlendirme aşaması olarak etki etmektedir. Çalışmanın deneysel süreçleri boyunca rapor edilen değerler, bu yenilikçi yöntemin katkılarını ortaya koymaktadır. Sunulan metota ait sözü edilen aşama sunduğu matematiksel altyapısı sayesinde oldukça etkili ve tutarlıdır. Rapor edilen performans sonuçları bakımından sunulan modelin araştırmacılar açısından oldukça dikkat çekici olacağı öngörülmektedir.

7. KAYNAKLAR

[1] X. Mao, H. Yang, S. Huang, Y. Liu, and R. Li, *Extractive summarization using supervised and unsupervised learning*, **Expert Syst. Appl.**, (2019) 173–181.

[2] Osman Durmaz, *Metin Sınıflandırmada Boyut Azaltmanın Etkisi ve Özellik Seçimi*, (2011) **IEEE 19th Signal Processing and Communications Applications Conference (SIU 2011)**, (2011).

[3] C. Hark, A. Seyyarer, T. Uçkan, and A. Karci, *Doğal dil işleme yaklaşimlari ile yapisal olmayan dökümanlarin benzerliği*, **IDAP 2017 - International Artificial Intelligence and Data Processing Symposium**, (2017) 1–6.

[4] Ş. Canberk, G. , Sağıroğlu, *Bilgi ve Bilgisayar Güvenliği : Casus Yazılımlar ve Korunma Yöntemleri*. Ankara: Grafiker Yayıncılık, (2006).

[5] C. Hark, T. Uçkan, and A. Seyyarer, Abubekir Karci, *Metin Özetleme İçin Çizge Tabanlı Bir Öneri*, **IDAP 2018 - International Artificial Intelligence and Data Processing Symposium**, (2018).

[6] L. Ermakova, J. V. Cossu, and J. Mothe, *A survey on evaluation of summarization methods*, **Inf. Process. Manag.**, 56:5 (2019) 1794–1814.

[7] A. Joshi, E. Fidalgo, E. Alegre, and L. Fernández-Robles, *SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders*, **Expert Syst. Appl.**, (2019) 200–215.

[8] M. Gambhir and V. Gupta, *Recent automatic text summarization techniques: a survey*, **Artif. Intell. Rev.**, 47:1, (2017) 1–66.

[9] J. Tan, X. Wan, and J. Xiao, *Abstractive Document Summarization with a Graph-Based Attentional Neural Model*, **In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics**, (2017),1171–1181.

[10] R. Mihalcea and P. Tarau, *A Language Independent Algorithm for Single and Multiple Document Summarization*, **Proc. IJCNLP 2005, 2nd Int. Join Conf. Nat. Lang. Process.**, (2005) 19–24.

[11] K. Sarkar, K. Saraf, and A. Ghosh, *Improving graph based multidocument text summarization using an enhanced sentence similarity measure*, **IEEE 2nd Int. Conf. Recent Trends Inf. Syst. ReTIS 2015 - Proc.**, (2015) 359–365.

[12] H. Van Lierde and T. W. S. Chow, *Query-oriented text summarization based on hypergraph transversals*, **Inf. Process. Manag.**, 56:4 (2019) 1317–1338.

[13] S. Xiong and D. Ji, *Query-focused multi-document summarization using hypergraph-based ranking*, **Inf. Process. Manag.**, 52:4 (2016) 670–681.

[14] G. Erkan and D. R. Radev, *Lexrank: Graph-based lexical centrality as salience in text summarization*, **J. Artif. Intell. Res.**, (2004) 457–479.

[15] O. Kaynar, Y. Görmez, Y. E. Işık, and F. Demirkoparan, *Comparison of graph based document summarization method*, **International Conference on Computer Science and Engineering (UBMK)**, (2017) 598–603.

[16] S. Brin and L. Page, *The anatomy of a large-scale hypertextual web search engine*, **Comput. networks ISDN Syst.**, 30:1–7 (1998) 107–117.

[17] D. Parveen, H.-M. Ramsl, and M. Strube, *Topical coherence for graph-based extractive summarization*, **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, (2015) 1949–1954.

[18] E. Canhasi and I. Kononenko, *Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization*, **Expert Syst. Appl.**, 41:2 (2014) 535–543.

[19] G. Salton, A. Singhal, M. Mitra, and C. Buckley, *Automatic text structuring and summarization*, **Inf. Process. Manag.**, 33:2 (1997) 193–207.

[20] O. Medelyan, *Computing Lexical Chains with Graph Clustering*, **In Proceedings of the ACL 2007 student research workshop**, (2007) 85–90.

[21] Y.-N. Chen, Y. Huang, C.-F. Yeh, and L.-S. Lee, *Spoken Lecture Summarization by Random Walk over a Graph Constructed with Automatically Extracted Key Terms*, in **Twelfth Annual Conference of the International Speech Communication Association**, (2011).

[22] L. Plaza, M. Stevenson, and A. Díaz, *Resolving ambiguity in biomedical text to improve summarization*, **Inf. Process. Manag.**, 48:4 (2012) 755–766.

[23] T. K. Landauer, P. W. Foltz, and D. Laham, *An introduction to latent semantic analysis*, **Discourse Process.** 25:2 (1998) 259–284.

[24] T. K. Landauer and S. T. Dutnais, *A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge*, **Psychol. Rev.**, 104:2 (1997) 211–240.

[25] S. Xiong and D. Ji, *Query-focused multi-document summarization using hypergraph-based ranking*, **Inf. Process. Manag.**, 52:4 (2016) 670–681.

[26] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, *Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion*, **Inf.**

Process. Manag., 43:6 (2007) 1606–1618.

[27] A. Haghighi and L. Vanderwende, *Exploring content models for multi-document summarization*, (2009) 362.

[28] H. P. Luhn, *The Automatic Creation of Literature Abstracts*, **IBM J. Res. Dev.**,2:2 (1958) 159–165.

[29] R. Sinha and R. Mihalcea, *Unsupervised graph-based word sense disambiguation using measures of word semantic similarity*, **ICSC 2007 International Conference on Semantic Computing**, (2007) 363–369.

[30] M. A. Fattah and F. Ren, *GA, MR, FFNN, PNN and GMM based models for automatic text summarization*, (2008).

[31] C. Author, R. Shardanand Prasad, and U. Kulkarni, *Implementation and Evaluation of Evolutionary Connectionist Approaches to Automated Text Summarization*, **J. Comput. Sci.,** 6:11 (2010) 1366–1376.

[32] Y.-N. Chen, Y. Huang, C.-F. Yeh, and L.-S. Lee, *Spoken Lecture Summarization by Random Walk over a Graph Constructed with Automatically Extracted Key Terms*, **In Twelfth Annual Conference of the International Speech Communication Association**, (2011).

[33] A. Abuobieda, N. Salim, A. T. Albaham, A. H. Osman, and Y. J. Kumar, *Text summarization features selection method using pseudo genetic-based mode*l, **Proc. - 2012 Int. Conf. Inf. Retr. Knowl. Manag. CAMP'12**, (2012) 193–197.

[34] I. F. Moawad and M. Aref, *Semantic graph reduction approach for abstractive Text Summarization*, **Proc. ICCES 2012 2012 Int. Conf. Comput. Eng. Syst.**, (2012) 132–138.

[35] V. Gupta, *LNAI 8284 - Hybrid Algorithm for Multilingual Summarization of Hindi and Punjabi Documents*, (2013).

[36] H. L. Linqing Liu, Yao Lu, Min Yang, Qiang Qu, Jia Zhu, *Convolutional neural networks for sentence classification*, **EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.,** (2014)1746–1751.

[37] P. G. Student and D. M. COE, *A Comparative Study Of Hindi Text Summarization Techniques: Genetic Algorithm And Neural Network*, (2015).

[38] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gulçehre, and B. Xiang, *Abstractive text summarization using sequence-to-sequence RNNs and beyond*, **20th SIGNLL Conf. Comput. Nat. Lang. Learn. Proc**., (2016) 280–290.

[39] M. Nasr Azadani, N. Ghadiri, and E. Davoodijam, *Graph-based biomedical text*

summarization: An itemset mining and sentence clustering approach, **J. Biomed. Inform.**, (2018) 42–58.

[40] S. B. Enstit, S. Ana, B. Dal, L. Tezi, Y. Emre, and I. K. S. Ocak, *Otomatik doküman özetleme yöntemlerinin karşilaştirilmasi*, (2018).

[41] M. Tülek, *Türkçe için Metin Özetleme*, İstanbul Tecnical University, 2007.

[42] S. Alim, *Language Independent Multı Document Summarızatıon Usıng Latent Semantıc Indexıng/Clusterıng Technıques*, (2009).

[43] H. P. Lunh, *The Automatic Creation of Literature Abstracts*, ***IBM J. Res. Dev.,*** 2:2 (1958) 159–165.

[44] Ç. C. Birant, *Rule-Based Text Summarızatıon In Turkısh Rule-Based Text Summarızatıon In Turkısh*, (2015).

[45] M. Allahyari *et al., Text Summarization Techniques: A Brief Survey*, ***arXiv***, (2017).

[46] E. Akülker, *Extractive text summarization for turkish using tf-idf and pagerank algorithms*, (2019).

[47] C. Hark, T. Uçkan, E. Seyyarer, and A. Karci, *Graph-Based Suggestion For Text Summarization*, **2018 International Conference on Artificial Intelligence and Data Processing,** (2019).

[48] R. Mihalcea, *Language independent extractive summarization*, **Proc. ACL 2005 Interact. poster Demonstr. Sess. - ACL '05**, (2005) 49–52.

[49] G. Rossiello, *Neural abstractive text summarization*, **CEUR Workshop Proceedings,** (2016) 70–75.

[50] Y. Gong, *Generic Text Summarization Using Relevance Measure and.*

[51] H. Zha, *Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering*, **Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '02**, (2002) 13.

[52] R. M. Alguliev, R. M. Aliguliyev, and M. S. Hajirahimova, *GenDocSum + MCLR: Generic document summarization based on maximum coverage and less redundancy*, ***Expert Syst. Appl.,*** 39:16 (2012) 12460–12473.

[53] M. Gambhir and V. Gupta, *Recent automatic text summarization techniques: a survey*, ***Artif Intell Rev,*** (2017) 1–66.

[54] Y. Chali, S. A. Hasan, and Y. Chali, Language Engineering : *Query-focused multi-document summarization : automatic Query-focused multi-document summarization : automatic data annotations and supervised learning approaches*, (2015) 109–145.

[55] G. L. Alexanderson, *About the cover: Euler and Königsberg's bridges: A historical view*, **Bull. Am. Math. Soc.,** 43:4 (2006) 567–573.

[56] M. E. Berberler, *Çizgelerde Birleştirilmişliklik Sayısı Üzerine*, (2006).

[57] C. U. İleri, *Dıstrıbuted And Self-Stabılızıng Algorıthms For Capacıtated Graph Theory Problems*, (2018).

[58] A. Naji, The k-distance polynomial of some graph operations Asia Pacific Journal of Mathematics, 4: 1 (2017) 23-31.

[59] P. Biukaghazadeh, *Basics of Graph Theory Basic notions*, (2013) 1–13.

[60] Anonymous.(2019). https://en.wikipedia.org/wiki/Graph_(discrete_mathematics) (On-line access on 21 Dec,2019.

[61] Anonymous.(2019). https://www.geeksforgeeks.org/graph-types-and-applications/. (On-line access 21Dec,2019].

[62] L. Paul Chew, *There are planar graphs almost as good as the complete graph*, **J. Comput. Syst. Sci.**, 39:2 (1989) 205–219.

[63] P. Z. Gary Chartrand, *A First Course İn Graph Theory*. Dover Publıcatıons, Inc. Mineola, New York, 2003.

[64] A. Karci, *Çizge Algoritmaları ve Çizge Bölmeleme*, (1998).

[65] U. Von Luxburg, *A Tutorial on Spectral Clustering*, (2007).

[66] B. Slininger, Fiedler's Theory of Spectral Graph Partitioning, (1996).

[67] H. Qiu and E. R. Hancock, *Graph matching and clustering using spectral partitions*, ***Pattern Recognit.***, (2006) 22–34.

[68] H. Jia, S. Ding, and M. Du, *Self-Tuning p-Spectral Clustering Based on Shared Nearest Neighbors*, **Cognit. Comput.**, 7:5 (2015) 622–632.

[69] L. Macdonald, *Successive Approximations by the Rayleigh-Ritz Variation Methoti*, (1933).

[70] L. Hagen, A. K. *New spectral methods for ratio cut partitioning and clustering*. **IEEE transactions on computer-aided design of integrated circuits and systems**, 11:9 (1992) 1074-1085.

[71] S. Jianbo and M. Jitendra, *Normalized cuts and image segmentation*, **IEEE Trans. Pattern Anal**. **Mach. Intell.**, 22:8 (2000) 888–905.

[72] A. Ben Ayed, M. Ben Halima, and A. M. Alimi, *Adaptive fuzzy exponent cluster ensemble system based feature selection and spectral clustering*, **IEEE Int. Conf. Fuzzy Syst.** (2017).

[73] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, *A survey of kernel and spectral*

methods for clustering, **Pattern Recognit.**, 41:176–190 (2008).

[74] W. R. Casper and B. Nadiga, *A New Spectral Clustering Algorithm.*

[75] R. M. Alguliev, R. M. Aliguliyev, and M. S. Hajirahimova, *GenDocSum + MCLR: Generic document summarization based on maximum coverage and less redundancy*, **Expert Syst. Appl.**, 39:16 (2012) 12460–12473.

[76] M. Fiedler, *Algebraic connectivity of graphs*, **Czechoslov. Math. J.** 23:2 (1973) 298–305.

[77] F. R. K. Chung, *Lectures on Spectral Graph Theory*, (1996).

[78] A. Bavelas, *A Mathematical Model for Group Structures*, **Human Organization**, 7:3 (1948) 16–30.

[79] M. Kutlu, C. Ciğir, and I. Cicekli, *Generic text summarization for Turkish*, **Computer Journal**, 53:8 (2010) 1315–1323.

[80] G. Erkan and D. R. Radev, *LexRank: Graph-based lexical centrality as salience in text summarization*, **J. Artif. Intell. Res.**, 22:457–479 (2004).

[81] Linton C.Freeman, *Centrality in Social Networks Conceptual Clarification*, (1978).

[82] M. McPherson, L. Smith-Lovin, and J. M. Cook, *Birds of a Feather: Homophily in Social Networks*, **Annu. Rev. Sociol.**, 27:1 (2001) 415–444.

[83] B. N. Analysis, *Centrality and Hubs*,1979 (2016).

[84] F. Boudin .*A comparison of centrality measures for graph-based keyphrase extraction.* **In Proceedings of the sixth international joint conference on natural language processing**, (2013) *834-838.*

[85] Alex Kosorukoff, *Social Network Analysis Theory and Applications*. **Passmore, D. L,** (2011).

[86] A. See, P. J. Liu Google Brain, and C. D. Manning, *Get To The Point: Summarization with Pointer-Generator Networks*, (2000).

[87] M. E. B. Mehmet Gencer, *Genetik Algoritma İle Maksimum Bağımsız Küme Probleminin Çözümü*, (2017) 5–9.

[88] S. Oh, *Maxımal Independent Sets On A Grıd Graph*, (2017).

[89] M. J. Jou and G. J. Chang, *The number of maximum independent sets in graphs*, **Taiwan. J. Math.**, 4:4 (2000) 685–695.

[90] *S. Butenko, Maxımum Independent Set And Related Problems, Wıth Applıcatıons,* (2003).

[91] Ö. Arapoğlu and O. Dağdeviren,*Telsiz Duyarga Ağları için Enerji Etkin Dağıtık Öz Kararlı Maksi mal Bağımsız Küme Algoritmaları*, **Akademik Bilişim**, (2017) 1-6.

[92] S. M. Hedetniemi, S. T. Hedetniemi, D. P. Jacobs, and P. K. Srimani, *Self-stabilizing algorithms for minimal dominating sets and maximal independent sets*, **Comput. Math. with Appl.,** 46:5 (2003) 805–811.

[93] T. M. Chan and S. Har-Peled, *Approximation Algorithms for Maximum Independent Set of Pseudo-Disks* , **Discrete & Computational Geometry ,** (2012) 373–392.

[94] T. A. Feo, M. G. C. Resende, and S. H. Smith, *A Greedy Randomized Adaptive Search Procedure for Maximum Independent Set*, **Oper. Res**., 42:5 (2008) 860–878.

[95] N. Z. Shor, *Dual quadratic estimates in polynomial and Boolean programming*, **Ann. Oper. Res.,** 25:1 (1990) 163–168.

[96] S. Vijayarani, M. J. Ilamathi, M. Nithya, A. Professor, and M. P. Research Scholar, *Preprocessing Techniques for Text Mining -An Overview*, **Int. J. Comput. Sci. Commun. Networks,** 5:1,(2015) 7–16.

[97] M. Saad, *The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification*, Master Thesis, The Islamic University ,Palestine, (2010).

[98] M. J. Denny and A. Spirling, *Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It*, **Polit. Anal.,** 26:2 (2018) 168–189.

[99] Anonymous.(2002).https://www-nlpir.nist.gov/projects/duc/guidelines/2002.html. (On-line Access on 07 May,2019).

[100] Anonymous.(2019). https://duc.nist.gov/. (On-line Access on 24 Dec, 2019).

[101] C.-Y. Lin, *Rouge: A package for automatic evaluation of summaries*, **Text Summ. Branches Out,** (2004).

[102] C.Y. Lin and E. Hovy, *Automatic evaluation of summaries using n-gram co-occurrence statistics*. **In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics** ,(2003) 150-157.

[103] C. Republic, *Evaluatıon Measures For Text Summarızatıon Josef Steinberger , Karel Je˘zek,* **Comput. Informatics,** 28:2 (2009) 1001–1025.

[104] R. Mihalcea, *Language independent extractive summarization*, **Proc. ACL 2005 Interact. poster Demonstr. Sess.** (2005) 49–52.

[105] R. Mihalcea and P. Tarau, *Textrank: Bringing order into text.* **In Proceedings of the 2004 conference on empirical methods in natural language processing** .(2004) 404-411.

ÖZGEÇMİŞ

Ad Soyad : Taner UÇKAN

Doğum Yeri ve Tarihi : Van - 06/07/1984

Adres : Van Yüzüncü Yıl Üniversitesi

E-Posta : taneruckan@yyu.edu.tr

Lisans : Fırat Üniversitesi Bilgisayar Mühendisliği (2006)

Yüksek Lisans : Atatürk Üniversitesi Bilgisayar Mühendisliği (2015)

Mesleki Deneyimler : Ülker Data Teknik-İstanbul
Yazılım Uzmanı (2007-2008)

Acerpro Yazılım Ltd.Şti
Yazılım Uzmanı(2008-2009)

Verisoft A.Ş
Yazılım Uzmanı(2009-2010)

Hakkari Üniversitesi
Öğretim Görevlisi (2010-2012)

Van Yüzüncü Yıl Üniversitesi
Öğretim Görevlisi (2012-Devam)

TEZDEN TÜRETİLEN YAYINLAR

- **Uçkan T.,** A. Karcı., 2019: "Extractive multi-document text summarization based on graph independent sets," Egypt. Informatics J., no. xxxx, 2019, doi: 10.1016/j.eij.2019.12.002.

- **Uçkan T.,** Hark C., Seyyarer E.,Karcı A., 2019 : Ağırlıklandırılmış Çizgelerde Tf-Idf ve Eigen Ayrışımı Kullanarak Metin Sınıflandırma. Bitlis Eren Üniversitesi Fen Bilimleri Dergisi , 8 (4) , 1349-1362 . DOI: 10.17798/bitlisfen.531221

- Hark C., **Uçkan T.,** Seyyarer E.,Karcı A.,2019 : "Metin Özetlemesi için Düğüm Merkezliklerine Dayalı Denetimsiz Bir Yaklaşım". Bitlis Eren Üniversitesi Fen Bilimleri Dergisi , 8 (3) , 1109-1118 . DOI: 10.17798/bitlisfen.568883

- **Uçkan T.,** Hark C., Seyyarer E.,Karcı A.,2019 : A Extractive Text Summarization Via Graph Partitioning, International Conference on Data Science, Machine Learning and Statistics

- Hark C., **Uçkan T**., Seyyarer E.,Karcı A.,2019 : "Çizge Benzerliği Yöntemi ile Doküman Sınıflandırma," *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, Malatya, Turkey, 2018, pp. 1-4.doi: 10.1109/IDAP.2018.8620926

- Hark C., **Uçkan T.,** Seyyarer E.,Karcı A.,2019 : "Extractive Text Summarization via Graph Entropy," 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, pp. 1-5. doi: 10.1109/IDAP.2019.8875936

- Hark C., **Uçkan T.,** Seyyarer E.,Karcı A.,2018 : "Graph-Based Suggestion For Text Summarization," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, 2018, pp. 1-6. doi: 10.1109/IDAP.2018.8620738

- Hark C., **Uçkan T**., Seyyarer E.,Karcı A.,2017 : "Doğal dil İşleme yaklaşimlari ile yapisal olmayan dökümanlarin benzerliği," International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, 2017, pp. 1-6. doi: 10.1109/IDAP.2017.8090306

- **Uçkan T.,** Hark C., Seyyarer E.,Karcı A., 2016 : Fp Growth Algoritması Ve Big Data Uygulamaları," *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, Malatya, Turkey, pp. 1-4.doi: 10.1109/IDAP.2018.8620926

- Hark C., **Uçkan T.,** Karcı A.,2016 :"15 Temmuz Şehitler Köprüsü Kapsamında Hava Yol Durum Bilgileri Kullanılarak Trafik Kazalarının Nedenlerine Yönelik Birliktelik Çıkarımı," 2016 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya