KOCAELİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLİŞİM SİSTEMLERİ MÜHENDİSLİĞİ
ANABİLİM DALI

YÜKSEK LİSANS TEZİ

TARANMIŞ GAZETE KOLEKSİYONU ÜZERİNDE TAM METİN ARAMA VE GÖRSELLEŞTİRME ARACI

HASAN BASRİ ŞAHİN

KOCAELİ 2022

KOCAELİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLİŞİM SİSTEMLERİ MÜHENDİSLİĞİ
ANABİLİM DALI

YÜKSEK LİSANS TEZİ

TARANMIŞ GAZETE KOLEKSİYONU ÜZERİNDE TAM METİN ARAMA VE GÖRSELLEŞTİRME ARACI

HASAN BASRİ ŞAHİN

Doç. Dr. Süleyman EKEN **Danışman, Kocaeli Üniv.**	……..……..
Dr. Öğr. Üyesi Alev MUTLU **Jüri Üyesi, Kocaeli Üniv.**	……..……..
Dr. Öğr. Üyesi Ekin EKİNCİ **Jüri Üyesi, Sakarya Uygulamalı Bilimler Univ.**	……..……..

Tezin Savunulduğu Tarih: 20.06.2022

ETİK BEYAN VE ARAŞTIRMA FONU DESTEĞI

Kocaeli Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım bu tez/proje çalışmasında,

- Bu tezin/projenin bana ait, özgün bir çalışma olduğunu,
- Çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı,
- Bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi,
- Bu çalışmanın Kocaeli Üniversitesi'nin abone olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş olduğu ölçütlere uygun olduğunu,
- Kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- Tezin/Projenin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez/proje çalışması olarak sunmadığımı, beyan ederim.

■ Bu tez/proje çalışmasının herhangi bir aşaması hiçbir kurum/kuruluş tarafından maddi/alt yapı desteği ile desteklenmemiştir.

☐ Bu tez/proje çalışması kapsamında üretilen veri ve bilgiler ..... tarafından ..... no'lu proje kapsamında maddi/alt yapı desteği alınarak gerçekleştirilmiştir.

Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

(İmza)

Hasan Basri Şahin

YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI

Fen Bilimleri Enstitüsü tarafından onaylanan lisansüstü tezimin/projemin tamamını veya herhangi bir kısmını, basılı ve elektronik formatta arşivleme ve aşağıda belirtilen koşullarla kullanıma açma izninin Kocaeli Üniversitesi'ne verdiğimi beyan ederim. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin/projemin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanımı bana ait olacaktır.

Tezin/projenin kendi özgün çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin/projenin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim kurulu tarafından yayınlanan **"Lisanüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"** kapsamında tezim aşağıda belirtilen koşullar haricinde YÖK Ulusal Tez Merkezi/ Kocaeli Üniversitesi Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

☐ Enstitü yönetim kurulu kararı ile tezimin/projemin erişime açılması mezuniyet tarihinden itibaren 2 yıl ertelenmiştir.

☐ Enstitü yönetim kurulu gerekçeli kararı ile tezimin/projemin erişime açılması mezuniyet tarihinden itibaren 6 ay ertelenmiştir.

■ Tezim/projem ile ilgili gizlilik kararı verilmemiştir.

(İmza)

Hasan Basri Şahin

ÖNSÖZ VE TEŞEKKÜR

Yüksek lisans eğitimimde benden desteklerini esirgemeyen hocam Doç. Dr. Süleyman Eken hocama teşekkürlerimi sunarım.

Ayrıca eğitim süresince yanımda olan eşime teşekkürlerimi sunarım.

Haziran - 2022 Hasan Basri ŞAHİN

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

Kısaltmalar

BLSTM : Bidirectional Long Short - Term Memory (Çift Yönlü Uzun Kısa-Dönem Hafıza)
DBLSTM : Deep Bidirectional Long Short - Term Memory (Derin Çift Yönlü Uzun Kısa-Dönem Hafıza)
DNA : Deoksiribo Nükleik Asit
HTTP : Hyper-Text Transfer Protocol (Hiper-Metin Transfer Protokolü)
HTML : Hyper-Text Markup Language (Hiper-Metin İşaretleme Dili)
JSON : JavaScript Object Notation (JavaScript Nesne Gösterimi)
LSPC : Locality Sensitive Pseudo Code (Yerelliğe duyarlı sözde kod)
LSTM-CRF : Long Short Term Memory with a Conditional Random Feild (Rastgele Alanlar ile Uzun Kısa-Dönem Hafıza)
NER : Name Entity Recognition (Varlık İsmi Tanıma)
NLP : Natural Language Processing (Doğal Dil İşleme)
NoSQL : Not Only SQL
OCR : Optic Character Recognition (Optik Karakter Tanıma)
PDF : Portable Document Format (Taşınabilir Dosya Formatı)
REST : Representational State Transfer (Temsili Durum Transferi)
RNA : Ribonükleik Asit
VİT : Varlık İsmi Tanıma
XML : Extensible Markup Language (Genişletilebilir İşaretleme Dili)

TARANMIŞ GAZETE KOLEKSİYONU ÜZERİNDE TAM METİN ARAMA VE GÖRSELLEŞTİRME ARACI

ÖZET

Gazete, 17. yüzyılın başlarında Avrupa'da ayrı bir kültürel form olarak ortaya çıktı. Tarihin erken modern dönemiyle bağlantılıdır. Tarih gazeteleri, milletler ve insanları için son derece önemlidir ve farklı disiplinlerden araştırmacılar, geçmişe dair anlayışımızı geliştirmek için, gazetelere güvenirler. Bu ihtiyacı karşılamak için İstanbul Üniversitesi Kütüphane ve Dokümantasyon Daire Başkanlığı, taranmış tarihi gazetelerden oluşan büyük bir veri tabanını, Gazetelerden Tarihe Bakış Projesi kapsamında erişime açmışlardır. Bunu bir adım daha ileri götürmek ve belgeleri daha erişilebilir kılmak için tüm veri tabanında optik karakter tanıma ve varlık ismi tanıma görevlerini çalıştırmamız ve sonuçları tam metin arama mekanizmasına izin verecek şekilde indekslememiz gerekmektedir. Varlık ismi tanıma; kişi, yer, kurum, tarih, formül ve para gibi varlık isimlerini dokümanlarda bulan doğal dil işleme ve metin madenciliğinde bilgi çıkarımı alanlarından biridir. Bu çalışmada, taranmış gazete dokümanlarındaki yer, kişi, organizasyon isimlerini etiketleyen bir sistem geliştirilmiştir. Orijinal web sitesindeki veri kümesini elde etmekten, arama sorgularını çalıştırmak için grafiksel bir kullanıcı ara yüzü sağlamaya kadar tüm bu hattı kapsayan bir sistem tasarladık ve bu beklentileri başarıyla gerçekleştiren bir algoritma geliştirdik. Ayrıca performans sonuçlarını doğru şekilde ölçümleyebilmek için bulut bilişim kullanılmış ve eşit kaynaklara sahip bilgisayarlar üzerinde geliştirilen yazılım çalıştırılmıştır. Önerilen sistem, kişi, kültür ve güvenlikle ilgili anahtar kelimeleri aramayı ve görselleştirmeyi sağlamaktadır.

Anahtar Kelimeler: Bulut Bilişim, Tam Metin Arama, Varlık İsmi Tanıma, Veri Gazeteciliği, Yapay Zeka.

FULL-TEXT SEARCH AND VISUALIZATION TOOL ON SCANNED NEWSPAPER COLLECTION

ABSTRACT

The newspaper emerged as a distinct cultural form in early 17th-century Europe. It is bound up with the early modern period of history. Historical newspapers are of utmost importance to nations and its people, and researchers from different disciplines rely on these papers to improve our understanding of the past. In pursuit of satisfying this need, Istanbul University Head Office of Library and Documentation provides access to a big database of scanned historical newspapers, Gazetelerden Tarihe Bakış. To take it another step further and make the documents more accessible, we need to run optical character recognition (OCR) and named entity recognition (NER) tasks on the whole database and index the results to allow for full-text search mechanism. Name Entity Recognition (NER); It is one of the fields of information extraction in natural language processing and text mining, which finds entity names such as person, place, institution, date, formula and money in documents. In this thesis, a system has been developed that labels the names of places, people and organizations in scanned newspaper documents. We design and implement a system encompassing the whole pipeline starting from scrapping the dataset from the original website to providing a graphical user interface to run search queries, and it manages to do that successfully. In addition, cloud computing was used to accurately measure performance results and the software developed on computers with equal resources was run. Proposed system provides to search people, culture and security-related keywords and to visualise them.

Keywords: Cloud Computing, Full Text Search, Name Entity Recognition, Data Journalism, Artificial Intelligence.

1. GİRİŞ

Dijitalleşen dünya ile birlikte üretilen bilgi her geçen gün artmakta dolayısıyla oluşan büyük miktardaki doküman koleksiyonlarını otomatik olarak organize etmek, analiz etmek, özetlemek, anlamak ve içerisinden istenilen bilgiyi elde etmek zorlu görevler olarak karşımıza çıkmaktadır. Bu zorlu görevleri gerçekleştirebilmek için ise dokümanları açıklayan terimlere ihtiyaç vardır. Doküman koleksiyonlarını açıklayan terimlere anahtar kelime denmektedir.

Anahtar kelime, dokümanın tamamını okumadan o doküman hakkında fikir sahibi olmamızı sağlayan kelimedir. Ancak dokümanların çoğu anahtar kelime içermemektedir. Bu durumda eğer doküman üzerinden bir görev gerçekleştirilmek istenirse koleksiyondaki tüm dokümanların okunması gerekmektedir. Tüm dokümanların bir ya da birkaç kişi tarafından okunması ve dokümanlar üzerinde çeşitli görevlerin gerçekleştirilmesi neredeyse imkansız olup oldukça fazla zaman gerektirmektedir. Aynı şekilde anahtar kelimelerin manuel olarak çıkartılması da neredeyse imkansızdır. Bir örnek ile açıklamak gerekirse elimizde on yıllık bir gazete koleksiyonu olsun. Gazetelerin her birinde pek çok yazı bulunmaktadır. Tüm yazılar içerisinden "yerel seçimler" ile ilgili olan kısımları bulmak istediğimizde on yıllık arşivdeki tüm yazıları okumamız mümkün değildir.

Ayrıca belgede adı geçen varlıklar da belge hakkında bizi bilgilendirir. Sonuç olarak, bu tarihi belgeleri araştırmak için üzerlerinde madencilik yapmamız gerekiyor.

1.1. Tezin Katkıları

Tezin literatüre katkılar şu şekilde sıralanabilir:

- Taranmış tarihi gazeteler üzerinde VİT ve optik karakter tanıma (OCR) işlemleri yapılmaktadır.
- Aranabilir tarihi gazetelere izin vermek için tam metin tabanlı bir mekanizma geliştirilmiştir.
- Örgütsel ve ulusal sorunları anlamak için tarih, kültür ve güvenlikle ilgili yapılan sorgular için tarihi gazetelere hizmet verilmektedir.
- Farklı indeksleme mekanizmalarının ölçekleme testleri yapılmıştır.

1.2. Tezin Organizasyonu

Tez dört bölümden oluşmaktadır. Giriş bölümünde tezin motivasyonu ve literatüre katkılar özetlenmiştir. 2. bölümde taranmış dokümanlar üzerinde çalışmalar ve varlık ismi tanıma üzerine yapılan araştırmalar sunulmuştur. 3. bölüm, önerilen sistemin detaylı açıklamasını sunmaktadır. 4. bölümde gerçekleştirilen testlere yer verilmektedir. Son bölümde elde edilen sonuçlar tartışılmış ve gelecekte neler yapılabileceğine değinilmiştir.

2. LİTERATÜR TARAMASI

Bu bölümde literatür taramasının detayları yer almaktadır. Taranmış dokümanlar ve varlık ismi tanıma üzerine yapılan çalışmalara yer verilmiştir.

2.1. Varlık İsmi Tanıma

Veri ve doküman madenciliği, bilgi çıkarımı ve bilgiye erişim, makina çevirisi, duygu analizi, sözdizimsel inceleme, multimedya indeksleme, soru-cevap sistemleri gibi birçok doğal dil işleme yöntemlerinin, önemli bir adımını oluşturmaktadır. Bir metinde yer alan varlıkların bulunarak önceden tanımlı kişi, yer, tarih, formül, yüzde, organizasyon ve para gibi sınıflardan bir tanesine atanması işlemi Varlık İsmi Tanıma (VİT) olarak ele alınmaktadır (Nadeau ve Sekine, 2007). Tüm bunlarla VİT bu veri tipleri ile sınırlı olmayıp farklı alanlardaki çalışmalarda ilgili alana özgü varlıkların tanınması ve işaretlenmesi için de kullanılmaktadır. E-posta adresleri (Minkov ve diğ., 2005), telefon numaraları, kitap başlıkları, proje isimleri, biyoinformatik ve kimya alanlarındaki metinlerde geçen genlerde bulunan protein isimleri (Tanabe ve diğ., 2005), RNA, DNA, hücre bilgileri, ilaç adları (Kim ve diğ., 2004), kimyasal adları (Eltyeb ve Salim, 2014) da varlık isimleri olarak çalışılan konulardandır. VİT işlemi ile genel olarak haber dokümanlarındaki varlık isimlerinin çıkarılmasına odaklanılmışken bankacılık, webte arama (Pasca, 2004), kimya, biyoloji gibi farklı alanlarda da kullanılmaktadır. Bu çalışmada, tarihi gazetelerde isimlendirilmiş varlıkları tanıdık ve ardından tam metin araması yapılmasını sağladık.

Ağırlıklı olarak İngilizce olmak üzere Arapça (Shaalan ve Raza, 2008), Rusça (Arkhipov ve Burtsev, 2017), Çince (Gao ve diğ., 2005), Fince ve Japonca (Isozaki, 2001) farklı dillerde VİT işlemi gerçekleştirilmiştir. Son yıllarda Türkçe VİT üzerine de birçok çalışma yapılmıştır. Özger ve Diri Türkçe dokümanlar için konudan bağımsız, kural tabanlı olarak kurum, yer, kişi isimleri ile tarih, para, saat varlık isimlerinin bulunması ve etiketlenmesini gerçekleştirmişlerdir (Özger ve Diri, 2012). Dalkılıç ve diğerleri Türkçe için dilin bazı gramatik kurallarına bağlı bir yöntem geliştirerek farklı dokümanlar üzerinde denemişlerdir (Dalkılıç ve diğ., 2010). Özkaya ve Diri resmi olmayan Türkçe e-postalar üzerinde bazı kurallar çıkarılarak, Şartlı Rastgele Alanlar kullanılıp kişi, kurum ve yer gibi üç farklı varlık isminin tanınmasını gerçekleştirmişlerdir (Özkaya ve Diri, 2011). Küçük ve diğerleri Türkçe için varlık isimleriyle işaretlenmiş haber metinlerinden oluşan

bir veri kümesi sunmuşlardır (Küçük ve diğ., 2016). Güngör ve diğerleri Türkçe varlık ismi tanıma görevi özyinelemeli sinir ağları kullanan bir yöntem geliştirmişlerdir (Güngör ve diğ., 2018). Güneş ve Tantuğ Türkçe VİT işlemleri için BLSTM ve DBLSTM yapay sinir ağı yapıları geliştirmişlerdir (Güneş ve Tantuğ, 2018). Eken tezinde Türkçe tweetler (kısa metinlerde) için varlık ismi tanıma sistemi geliştirmiştir (Eken, 2015). Küçük ve Arıcı Türkçe için önce Wikipedia tabanlı bir kişi ismi tanıma sistemi, sonrasında da yine Türkçe için Wikipedia tabanlı tam bir varlık ismi tanıma sistemi geliştirmişlerdir (Küçük ve Arıcı, 2016). Sarı ve Aktaş kapsamı tarih ve coğrafya alanları olarak belirlenen Türkçe ders metinleri için kural tabanlı bir VİT modeli geliştirmişlerdir (Sarı ve Aktaş, 2018). Arslan ve diğerleri büyük veri indeksleme ve arama yazılımı olan Apache Lucene kullanarak yarım milyar Web sayfası içinde en sık geçen e-posta, Web adresleri ve emoji varlıklarını tespit etmişler (Arslan ve diğ., 2018). Çekinel ve diğerleri Şartlı Rastgele Alanlar yöntemiyle haber metinlerinde varlık isimlerini tanınmıştır (Çekinel ve diğ., 2019). Akpınar ve diğerleri imge formatındaki banka talimatlarından yola çıkılarak varlık ve ilişki etiketleme işlemlerinin yarı-otomatik bir şekilde yapılmasına olanak sağlayan bir araç önermişlerdir (Akpınar ve diğ., 2019). Bütün bu çalışmalar modern (tarihi olmayan) dokümanlardaki Türkçe VİT görevleriyle ilgilidir.

2.2. Taranmış Dokümanlar Üzerinde Çalışmalar

Gazeteler, mektuplar, günlükler, tıbbi belgeler, mahkeme raporları vb. tarihi metinler, geçmişe ışık tutan önemli belgeler olup dijital ortama aktarılmakta, OCR yapılmakta, açıklama eklenmekte, saklanmakta ve internet ortamında herkesin ulaşabileceği şekilde yayınlanmaktadır. Sonuç olarak bu metinler özellikle araştırmacıların ilgisini çekmiş ve onlardan bilgi çıkarmanın yollarını aramışlardır. Bu nedenle geçmişten günümüze bu konuda birçok çalışma yapılmış ve halen devam etmektedir.

Literatür incelendiğinde, kullanıcıların gazetelerdeki bilgileri daha çok kişi, yer, kuruluş vb. adları kullanarak araştırdıkları görülmektedir. Bu durum göz önüne alındığında VİT'in bilgi aramada önemli bir işlevi olduğu rahatlıkla söylenebilir. Ayrıca literatürde geçen çalışmaların birçoğu direkt metin veya metinden oluşturulmuş dokümanlar üzerinde çalışmaktadır. Yaptığımız araştırmaya göre taranmış dokümanlar veya daha özelinde gazeteler üzerinde Türkçe bir VİT çalışmasına rastlanmamıştır.

Diğer dillerde yapılan çalışmalar ise şu şekildedir. Europeana Projesi (Willems ve Atanassova, 2015), yirmi üç Avrupa kütüphanesinden gelen tarihi gazete içeriğini düzelterek ve toplayarak sayısallaştırılmış tarihi gazetelere erişimi geliştirmeyi amaçlamaktadır. Gazete içeriği 1618 yılına kadar uzanan bu zengin koleksiyon, iki önemli kültürel mirası web sitesi aracılığıyla erişilebilir hale getirmiştir. Avrupa Kütüphanesi ve Europeana. Bu kapsamda Neudecker ve arkadaşlarının geliştirdiği VİT'in temel amacı, aranabilirliği artırmak için kişiler, konumlar ve kuruluşlar gibi varlıkları tam metinde tanımlamak ve sınıflandırmak ve daha sonra bunları çevrimiçi kaynak açıklamalarına ve yetki dosyalarına (authority) bağlamaktır (Neudecker ve diğ., 2014), (Neudecker, 2016), (Neudecker ve Antonacopoulos, 2016).

Kettunen ve diğerleri büyük ölçekli Fince dilindeki 1771-1910 arasını kapsayan tarihi gazete topluluğu üzerinde VİT gerçeklemişlerdir (Kettunen ve diğ., 2016). Ekbal ve diğerleri webte bulunan önde gelen Bengalce gazetesinin arşivinden geliştirilen kısmen varlık isimleri etiketli Bengalce News Corpus'un bir kısmı ile VİT gerçeklemişlerdir (Ekbal ve dig., 2008). Pirovani ve diğerleri özellikle kişi adları varlıklarını çıkarmak ve bir ad dizini ile kullanıcıya gazete sayfalarında ad bulmak için bir araç önermişlerdir (Pirovani ve diğ., 2018). Mac Kim ve Cassidy ise Avustralya Milli Kütüphanesi'ndeki 1803 yılına dayanan çok sayıda sayısallaştırılmış gazete Trove koleksiyonu ile üzerinde Stanford NER sistemini değerlendirmişlerdir (Mac Kim ve Cassidy, 2015).

Jones ve Crane, koleksiyonları aramaya uygun hale getirmek için 19. yüzyıldan kalma gazetelerden varlım isimlerini çıkardılar (Jones ve Crane, 2006). Yazarlar tarafından çalışma kapsamında çıkarılan varlık isimleri, kişi adları, yerleri, tarihleri, ürünleri, kuruluşları, sokakları, gazeteleri, gemileri, alayları ve demiryollarıdır. Bazı varlık isimlerinin çıkarılmasındaki başarı düşük olduğundan, eğitim verisi olarak iyi tanımlanmış kural setleri, eksiksiz bilgi kaynakları ve uzmanlar tarafından hazırlanan listelerin kullanılmasını önerdiler. Sadece 10 farklı varlık ismine odaklandılar ve bunları manuel olarak değerlendirdiler.

Biz ise bu tez kapsamında, üç ana varlık ismini çıkardık, indeksleme ve arama mekanizmaları çalıştırdık.

Borin ve arkadaşları İsveç edebi klasikleri, Litteraturbanken, araştırmasını geliştirmek için genel bir VİT sistemi uygulamıştır (Borin ve diğ., 2007). Genel VİT sistemi, klasik kural tabanlı VİT'e kelime benzerliğinin dahil edilmesi şeklinde çalışmakta ve %92,8'lik bir F-ölçüsü ile sekiz varlık ismi elde etme yeteneğine sahiptir. Sistemlerinde ayıklanan varlık isimleri için arama ve tarama arabirimleri yoktur.

Diğer çalışmalardan farklı olarak Labusch ve arkadaşları, aramada kullanılmak üzere hem tarihi Almanca metinler hem de çağdaş metinlerde VİT için derin öğrenme yöntemleri ile gömülü dilleri kullanmışlar ve iyi sonuçlar elde etmişlerdir (Labusch ve diğ., 2019). Sayısallaştırılmış belgelerin yeniden işlenmesi yoluyla kaynak OCR metinlerindeki gürültü seviyesiyle ilgilenmemişler. Ruokolainen ve Kettunen tarafından yapılan çalışmada, 19. ve 20. yüzyılın başlarında Fince yayınlanan tarihi gazeteler ve dergiler, aramayı iyileştirmek amacıyla varlık isimlerini çıkarmak için kullanılmıştır. (Ruokolainen ve Kettunen, 2020). Stanford NER ve LSTM-CRF modelleri ile OCR methodu ile işlenmiş verileri ve manuel olarak düzeltilmiş, kesinlik verileri üzerinde VİT gerçekleştirilmiştir. Stanford NER ve LSTM-CRF modelleri her iki veri seti için oldukça benzer bir başarıya sahipken, düzeltilmiş veri seti her iki yöntemle de varlık isimi çıkarmada en başarılı olan olarak gözükmektedir. Sistemlerinde ayıklanan varlık isimleri için arama ve tarama arabirimleri yoktur. Bir gazete koleksiyonunun, grafiksel kullanıcı arayüzündeki (GUI) isimlerin La Stampa tarzı kullanım şeklinin, kullanıcılar için daha bilgilendirici ve faydalı olduğuna inanmışlardır.

Apperley ve arkadaşları, tarihi Maori gazetelerinin Niupepa koleksiyonunda tam metin araması için Greenstone sistemini kullanmışlardır (Apperley ve diğ., 2001). İlk olarak, OCR kullanılarak gazeteler elektronik formata dönüştürülmüş; sayfa düzeyinde indeks kullanılarak sayfa sayfa arama sağlanmıştır. Niupepa koleksiyonu, her sayfa için ayrı bir dosyada tutulan, metin içeren, sayfa düzeyinde bir dizin içerir. Tez kapsamında biz de verileri indekslemek için Elasticsearch motorunu kullandık.

Gatos ve arkadaşları, eski tipte yazılmış belgelerden anahtar kelimeler çıkarmak için görüntü ön işleme, bölümleme, sentetik veri oluşturma, kelime belirleme ve kullanıcının geri bildirim teknolojilerini birleştirerek bölümlendirmeden bağımsız bir algoritma tasarlamıştır (Gatos ve diğ., 2005). Ahonen ve Hyvonen, kültürel nesnelerin anlamsal olarak

açıklandığı ortak bir ontoloji kullanarak Finlandiya Ulusal Kütüphanesi'nin tarihi metinleri için anlamsal bir arama sistemi oluşturmuştur (Ahonen ve Hyvonen, 2009). Çalışmalarında, açıklamalı malzemeye çok yönlü bir anlamsal arama arayüzü uygulanmıştır. Jerele ve arkadaşları 18. ve 19. yüzyıllardan kalma Sloven kitaplarında ve gazetelerinde tam metin aramayı basitleştirmek için OCR ve bir sözlük kullanmıştır (Jerele ve diğ., 2011). Çalışmalarında OCR metinleri sayısallaştırırken tarihsel kelimeleri çağdaş olanlarıyla ilişkilendirmiştir. Terasawa ve arkadaşları tarihi gazetelerin görüntü kalitesinin düşük olduğunu ve bu görüntülerde OCR kullanımının başarılı olmayacağını iddia etmiştir (Terasawa ve diğ., 2011). Bu nedenle, yazarlar tam metin araması için görünüm tabanlı algoritmaları özelleştirmişlerdir. Algoritma; tam metin arama sorununu, şekilleriyle karakterden karaktere eşleştirmeye dayalı sıralı eşleştirmeye indirgemiştir. Ayrıca, hesaplama yükünü azaltmak için, algoritmaya sözde kod ifadesi LSPC kabul edilmiştir. Sonuç olarak, arama ve hesaplama maliyeti açısından daha iyi sonuçlar elde edilmiştir. Yöntemi, farklı arama anahtar kelime uzunluğuna ilişkin arama doğruluğu açısından değerlendirmişlerdir.

Thompson ve arkadaşları, kullanıcısına British Medical Journal ve London Medical Officer of Health raporları olan tarihsel metinlerden, verimli bir arama sağlamak için metin madenciliği yöntemlerinden yararlanarak web tabanlı bir Tıp Tarihi aracı tasarlamışlardır (Thompson ve diğ., 2016). Araç, kullanıcı terimini, bibliyografik meta verilerini, varlığı, olayı ve adlandırılmış varlık tabanlı aramayı sunmuştur. Tarihsel tıp metinleri üzerine yapılan bir başka çalışmada da bu metinleri araştırma amacıyla hazır hale getirmek için kavramların, ilişkilerin ve varlık isimlerinin çıkarılması amaçlanmıştır (Thompson ve diğ., 2016). Bu görevi gerçekleştirmek için yazarlar VİT araçlarından ve tıbbi ontolojilerden yararlanılmıştır. Wilkinson ve arkadaşları, üç tarihsel metin veri setinde anlamsal arama için derin öğrenme yöntemlerinden yararlanmışlardır (Wilkinson ve diğ., 2018). Bu amaçla, Ctrl-F-Net olarak adlandırılan, segmentasyonsuz bir dizeye göre sorgulama ile kelime tespit modeli geliştirilmiştir. Modelin çalışması şu şekilde: Önce sayfada bölge önerileri elde edilmiş ve daha sonra bu öneriler aramanın yapıldığı kelime gömme alanına gömülmüştür.

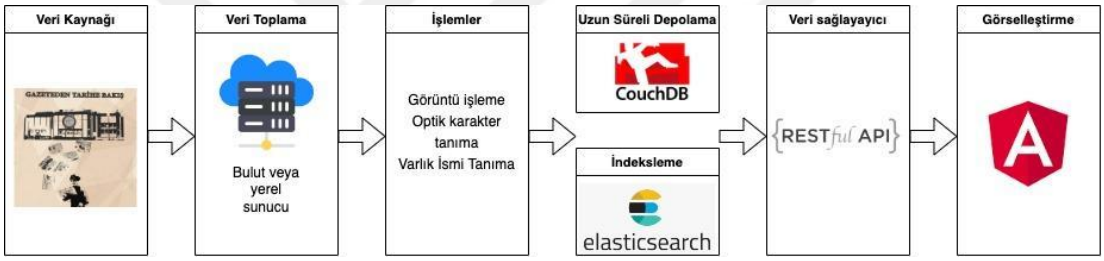
Ketunen ve arkadaşları aynı sayfadaki makaleleri ayrı ayrı elde etmek için PIVAJ makine öğrenimi tabanlı platformu kullanmışlar ve Fin dergisi Uusi Suometar'ın 1869–1898 yıllarındaki yayınları üzerinde bir deney yapmıştır (Ketunen ve diğ., 2019). Bunu yapmanın amacı, arama kalitesini iyileştirmek için makale makale arama sağlamak olduğunu söylemişlerdir. Bogaard ve arkadaşları, tarihsel gazete koleksiyonunda, işlevsel aramayı başarmak için meta veri kategorizasyonunu kullanarak, kullanıcı davranışını anlamayı amaçlanmıştır (Bogaard ve diğ., 2019). Eşzamanlı olarak bölge önerileri üretmiş ve bunları aramaların gerçekleştirildiği, bir kelime gömme alanına yerleştirmiştir. Atay ve arkadaşları, büyük miktarlarda taranan bu belgeler üzerinde içerik tabanlı şekil aramalarını mümkün kılan bir mimari geliştirmişlerdir (Atay ve diğ., 2018). Kullanıcı bazı anahtar kelimelerle arama yapabilir ve ilgili rakamları, dijital belgelerde başlıklarıyla görüntüleyebilmiştir.

Taranmış gazeteler üzerinde VİT işleminin veri gazeteciğine de katkı sunması beklenmektedir. Veri gazeteciliği etkin haberler üretebilmek amacıyla dijital verileri toplayıp işleyerek haber üretmektir. Öncelikle araştırmak ve haber yapmak istenilen konunun belirlenmesi gerekir. Daha sonra veri toplama, sorgulama, temizleme, işleme ve görselleştirme (Eken, 2020), analiz ve yorumlama ile habere dönüştürme adımları gelir (Gray ve diğ., 2012).

3. SİSTEM MİMARİSİ

Bu tez çalışmasında gazete verileri üzerinde arama işleminin yapılabilmesini sağlayan nihai veriyi elde edebilmek için çeşitli algoritmalar geliştirildi. Bu veriyi elde etmek için küçük görevler için adanmış ayrı ayrı yazılımlar entegre edildi. Bu yazılımların birbiriyle ortak bir şekilde çalışması sonucunda bir bütün olarak elimizde PDF olarak bulunun taranmış gazete koleksiyonu içerisinde arama yapılması sağlandı.

Geliştirilen yazılımın bu parçalarına kendisine ait başlıklarının altında detaylı bir şekilde değinilecektir. Bunlar; verinin toplanması, verinin ön işlemesi, optik karakter tanıma, varlık isimlerinin tanınması, verilerin CouchDB'de depolanması, verilerin Elasticsearch'te indekslenmesi, özelleştirmiş Restful API, web tabanlı GUI ve aramadan oluşmaktadır. Şekil 3.1.'de önerilen sistem mimarisi verilmiştir.



Şekil 3.1. Önerilen Sistem Mimarisi

3.1. Verilerin Toplanması

İstanbul Üniversitesi Gazetelerden Tarihe Bakış projesi kapsamında kütüphanesinde bulunan eski tarihli gazetelerinin bir kısmını proje sitesinde halka açık olarak taranmış şekilde paylaşmıştır. Bu gazeteler 1928-1942 yılları arasında yayınlanmış 688 cilt, 55 farklı ulusal ve yerel gazeteden oluşmaktadır. Toplamda 581106 sayfa gazete verisi ulaşılabilir şekilde sitesinde bulunmaktadır (URL-1).

Bu gazeteler Internetten veri kazıma yöntemi ile toplanmıştır. Bunu yapabilmek için Python'da Internet sitesin bize gazetelerin indirme linklerini, isimlerini ve basım tarihlerini bulan bir script yazılmıştır. Bu scriptte Requests (URL-2) kütüphanesi ile web sitenin HTML kodları çekilmiş sonrasında BeautifulSoup (URL-3) kütüphanesi kullanılarak bu HTML kod içerisinde ilgili HTML elementlerinin içerikleri ve özelliklerinden bu bilgiler

elde edilmiş. Requests ve BeautifulSoup birlikte, veri kümesinin web sitesini sorgulamamıza ve gazete adları, yıllar, aylar ve sayı numarası aracılığıyla ayrıştırmamıza izin vermiştir. Bir gazete bültenine ulaştığımızda, PDF'sini indirir ve gazete adı, tarihi ve orijinal URL'si gibi meta verilerini saklarız. Ardından indirme işlemini hızlandırmak için paralel bir indirme işi çalıştırıyoruz; ancak bu, bağlantının sunucu tarafından kapatılmasıyla sonuçlanır. Bu yüzden paralel işlerin sayısını maksimum 4'e düşürmenin yanı sıra istekler arasına 100 ms'lik bir gecikme eklemek zorunda kaldık. İndirilen sayfalara örnekler Şekil 3.2.'de verilmiştir.

(a) (b)

Şekil 3.2. Veri kümesinde taranan gazetelerin bazı bölümleri: (a) Acık Söz, 27 Ağustos 1936 ve (b) Jamanak, 1 Haziran 1936.

3.2. Veri Ön İşleme İşlemleri

Veriler elde edildikten sonra yapılacak olan karakter tanıma ve varlık ismi tanıma işleminin başarımını arttırmak için ön işleme adımlarına tabi tutulmuştur.

3.2.1. Gazete Sayfalarının Resimlere Çevrilmesi ve Temizlenmesi İçin PDF'i İşleme

Belge anlama genellikle taranan belgeler/görüntüler üzerinde yapılır. Lovegrove ve Braisford (1995) taranan PDF görüntülerinden birleşik satırları tanımlayan ilk çalışmayı yaptı. Popüler PDF formatı ile Anjewierden (2001), xpdf ile PDF'lerden metin ve grafik nesneleri çıkardı. Daha sonra Hadjar ve arkadaşları (2004) ayrıca PDF'den nesne çıkarma kütüphanelerinin performans sonuçlarını da karşılaştırdı. Chao ve Fan (2004), metin ve şekil nesnelerini doğrudan PDF kodundan (içerik analizi) ve çizgiler ve tablolar gibi vektör nesnelerini (düzen-tasarım analizi) bitmap görüntülerinden çıkardı. Hassan (2009)

görsel ilkelere göre bir aşağıdan yukarıya kümeleme algoritması önerdi.

OCR ile bazı karakterlerin yanlış tanınması mümkündür. Bu nedenle OCR ile NLP destekli bir düzeltme yapılması kaçınılmazdır. OCR düzeltme sonrası için makine öğrenimi, çoklu sistem çıktılarını birleştirme veya yüksek frekanslı sözcüklere dayalı farklı yaklaşımlar vardır. Niklas (2010) yanlış yazılmış kelimeleri en iyi şekilde düzeltmek için birkaç yöntemi birleştirdi. Bu yöntemler; kelime içerisindeki harflerin yerlerinin değiştirilmesi ile yeni kelimeler türetilmesi (Anagram Hash), karakterlerin şeklini temel alan yeni bir OCR metodu (OCR-Key) ve anlamsal bütünlük (Bigrams). Bu yaklaşım, 1785-1985 yılları arasında İngiliz gazetesinde yayınlanan London Times Gazetesi Arşivi'nde manuel olarak uygulanmış ve hata azaltma oranı %75'e yükseltilmiştir. 2007'de Hauser (2007), OCR düzeltme sonrası için bir düzenleme mesafesi geliştirdi. Généreux ve Spano, hızlı yaklaşık dizi eşleştirmesi sağlamak için CPMerge (Okazaki ve Tsujii, 2010) algoritmasını kullanan SimString kitaplığını (Généreux ve Spano, 2005) geliştirdi.

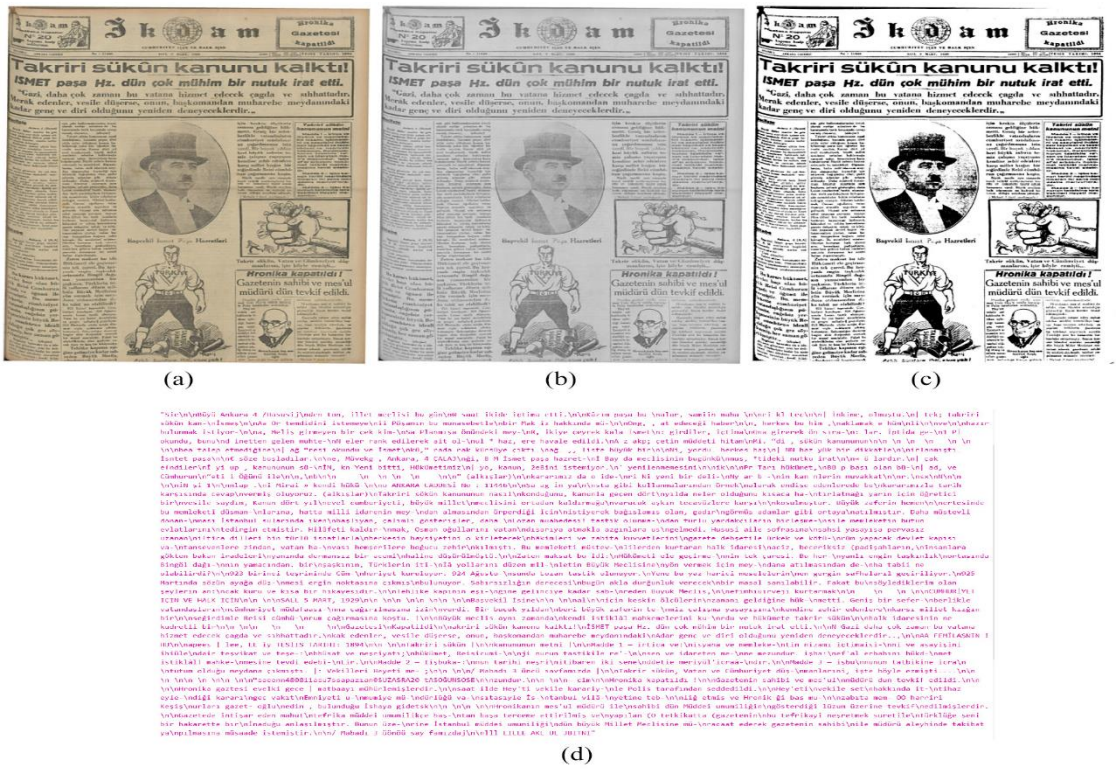
Gazete verileri pdf içerisinde gazetenin tüm içeriği sayfalar olarak bulunmaktadır. Bu pdfler içerisinde optik karakter tanıma işlemi için sayfalar ayrılarak her biri bir görüntü dosyası olacak şekilde ayrılmıştır. Bu işlem için python dili için geliştirilmiş pdf2image kütüphanesi kullanılmıştır.

Gazete sayfaları görüntü olarak elde edildikten sonra, görüntüler elimizde renkli olarak bulunmaktadır. Bu görüntüler önce gri olanlara dönüştürüldü. Bundan sonra, gri görüntüler Otsu eşikleme (1979) tekniği kullanılarak ikili biçimlere dönüştürülmüştür. Görüntüleri siyah beyaz formata çevirebilmek ve sonrasında temizleyebilmek için OpenCV kütüphanesinin Python için olan versiyonu opencv-python (URL-4) eklentisi kullanılmıştır. Siyah beyaz formatta görüntüler, siyahla beyaz arasındaki tonlarda renklerden oluşmaktadır. Görüntülerdeki yazıları ve arka planın birbirinden tamamen ayırabilmek için siyaha yakın olan tonlar tam siyah beyaza yakın olan tonlar tam beyaza çevrilmiş. Bu adımların görüntüler üzerindeki etkisi Şekil 3.3.'te gösterilmiştir.

3.2.2. Optik Karakter Tanıma ve Varlık İsmi Tanıma

Görüntüler ön işleme sonrasında optik karakter tanıma işlemine hazır hale gelmektedir. Bu işlem resimlerin içerisindeki karakter verilerini metin olarak elde etmemizi sağlayan

aşamadır. Bu işlem için Python dilinde pytesseract (URL-5) kütüphanesi kullanılmıştır. Temizlenmiş görüntüler ile çıkarılacak metinlerin dil bilgisini ilgili fonksiyonlara verildikten sonra bize metin verisini dönebilmektedir. Python-tesseract, Python için bir OCR aracıdır. Python-tesseract, Google'ın Tesseract-OCR Motoru için bir wrapper'dır. Tüm bu ön işlemler adımları çok iş parçacıklı (multi-threading) programlama ile gerçekleştirildi.



Şekil 3.3. 5 Mart 1929 tarihli İkdam'ın ön sayfasındaki veri ön işleme ve OCR sonucuna bir örnek: (a) ham görüntü, (b) gri görüntü, (c) ikili görüntü ve (d) OCR sonucu

OCR hatalarını en aza indirmek için bu kelimeler, Java ile geliştirilen ve Türkçe'ye özgü bir NLP kütüphanesi olan Zemberek-NLP (URL-6) kullanılarak normalize edilmiştir. Türkçe cümle normalleştirme yöntemi ile yazım hataları Zemberek-NLP'nin Türkçe morfolojisine göre düzeltilmiştir.

Zemberek-NLP, Türkçe için basit bir VİT modülü sağlar. Bu modül içerisinde eğitilmiş herhangi bir model bulunmamaktadır; bu nedenle, içerisindeki modeli eğitmemiz gerekiyor. Bir VİT modeli oluşturmak için önce bir eğitim setine, doğru örnekleri içeren bir dosyaya ihtiyacımız var. Eğitim verileri üç açıklama yöntemi kullanılarak hazırlanabilir: köşeli ayraç stili açıklama, OpenNLP stili açıklama ve Enamex stili açıklama.

Enamex sitili ile işaretlenmiş metin örneği: “<b_enamex TYPE="ORG">Enerji Verimliliği Merkezi<e_enamex> kurucu başkanı <b_enamex TYPE="PER">Bülent Yeşilata<e_enamex> , <b_enamex TYPE="LOC">Ankara'da<e_enamex> bir toplantıya katıldı.”

Zemberek kütüphanesinin içerisinde bulunan PerceptronNer yapay zeka modeli enamex veri seti ile eğitilmiştir. Bu model eğitildikten sonra Zemberek kütüphanesinin fonksiyonlarının kullanılabilmesi için Java ile geliştirilmiş kodların Python tarafında kullanılabilmesini sağlayan JPype1 (URL-7) kütüphanesi kullanılmıştır. Bu kütüphaneye Java runtime environment dosya yolu ve derlenmiş Java kodları olan jar uzantılı kütüphane dosyaları verilerek Zemberek kütüphanenin fonksiyonlarının çağrılabilmesi sağlanmıştır. Elde edilen PER (Person), ORG (Organisation) ve LOC (Location) varlık isimleri metin verisi ile birlikte CouchDB de depolanmıştır.

3.3. Mikroservis Mimarisindeki Servisler

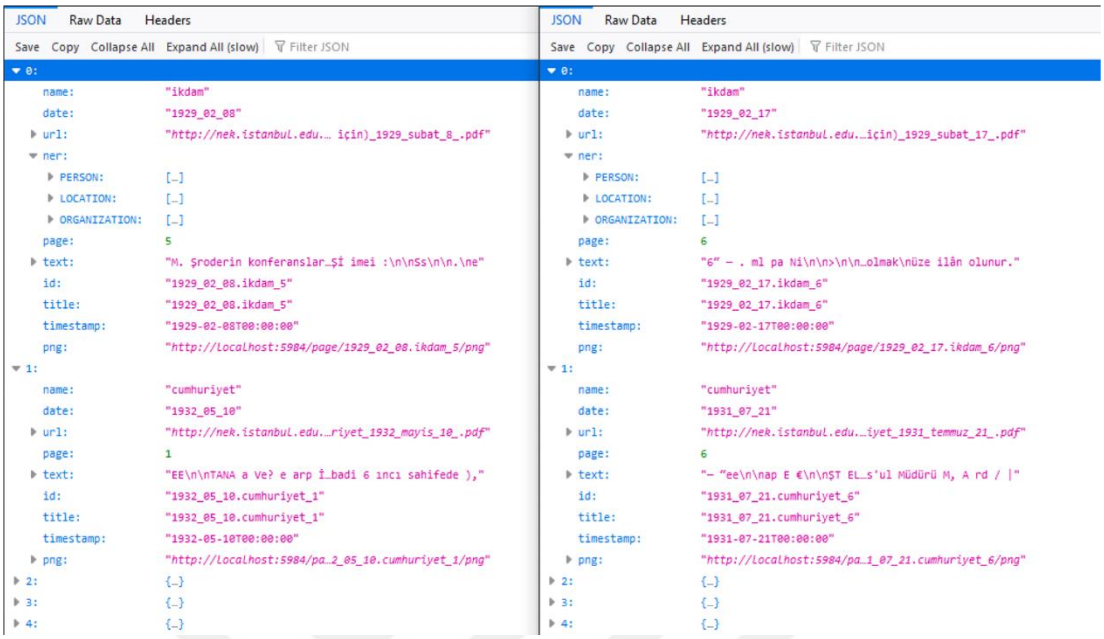
Mikroservis mimarisi, karmaşık bir sistemin, bağımsız olarak dağıtılabilen, hizmetler adı verilen, gevşek bir şekilde bağlı küçük parçalara bölündüğü bir mimari tarzıdır. Bu mimari, büyük, karmaşık uygulamaların hızlı, sık ve güvenilir bir şekilde teslim edilmesini sağlar.

Tüm veriler oluşturulduktan sonra arama yapmayı sağlayan kullanıcı arayüzü için her birinin kendine özel sorumlulukları olan adanmış servislerden oluşan bir mimari kurulmuştur. Bu mimaride konteyner teknolojisinden yararlanılmıştır.

Temsili Durum Transferi (REST) (Fielding, 2000) dağıtık sistemler tasarlamak için kullanılan bir mimari stildir. 2000 yılında Roy Fielding tarafından doktora tezinde tanıtılmış ve tanımlanmıştır. REST mimarisi HTTP protokolü üzerinden çalışır ve Genişletilebilir İşaretleme Dili (XML) ve JavaScript Nesne Gösterimi (JSON) verilerini taşıyarak istemci-sunucu mimarilerinde iletişim ve veri aktarımına izin verir.

GUI'miz ile elastik arama ve CouchDB örnekleri arasında oturmak için bir RESTful API kullanıyoruz, bu bize aşağıdaki biçimde bir GET isteği göndererek işlenmiş veri kümemizi sorgulamak için basit ve standart bir yol sunuyor: http://localhost:4000/query?keyword= <> start= <> end= <> . Ardından yanıt, Şekil 3.4.’de gösterildiği gibi bir

JSON belgesi biçiminde gelir.



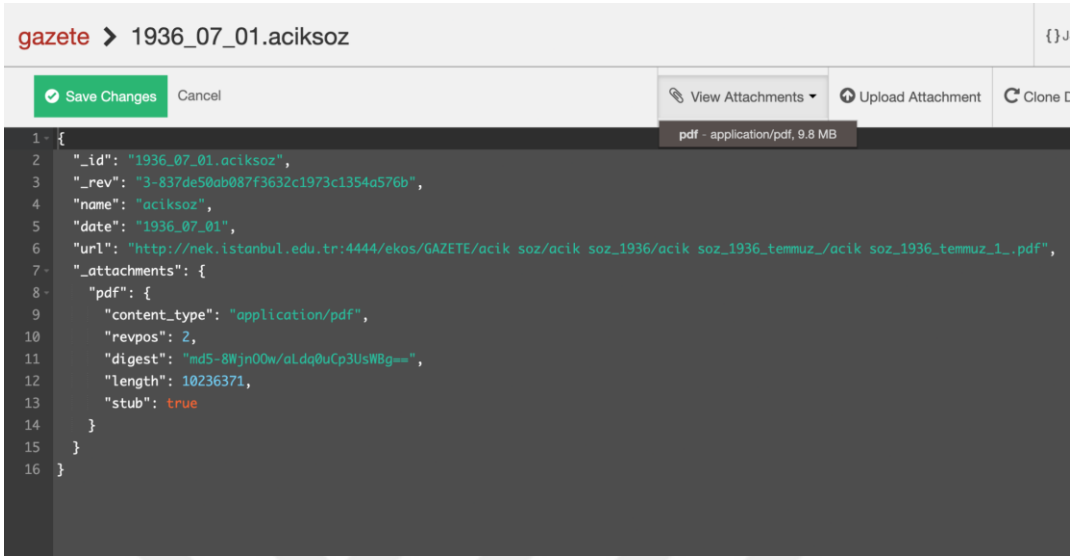
Şekil 3.4. JSON formatında "sansür" ve "sahte" anahtar kelimeleri için ilk sonuçları gösteren API yanıtı

3.3.1. Verilerin CouchDB'de Depolanması

Geleneksel olarak SQL veri tabanları herhangi bir veri depolama seçeneği için baskın seçim olmuştur; ancak üretilen verilerimizin muazzam büyümesi ve beraberinde getirdiği yeni uygulamaların yükselişi, SQL yerine yeni bir yaklaşıma olan ihtiyacı göstermiştir. NoSQL (Leavitt, 2010), SQL'deki katı veri şemasını daha esnek bir şemayla değiştirerek, dağıtılmış ve gerçek zamanlı uygulamalarda büyük yapılandırılmamış veri kümeleriyle çalışmamıza olanak tanır.

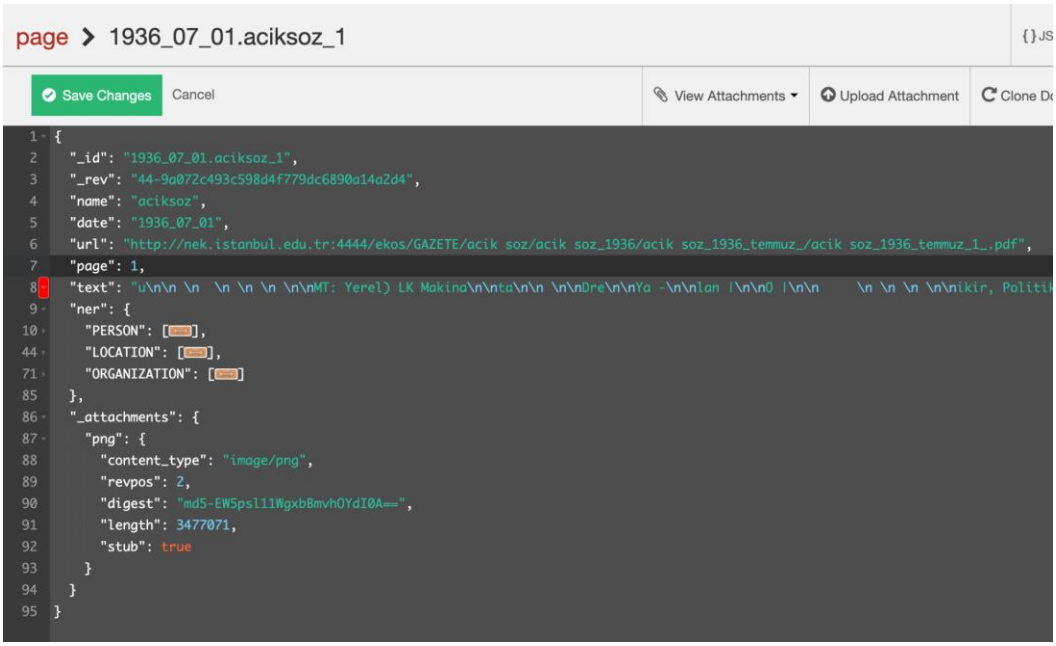
Veri kümemizdeki belgelerle uğraştığımızı düşünürsek, belge odaklı bir veri tabanı bunun için en uygun seçim olacaktır. Dolayısıyla CouchDB seçiminde bulunduk. Belge yönelimli bir veri tabanı, SQL'de olduğu gibi birden çok tabloya yaymak yerine, her nesneyi özel bir örnekte saklar. CouchDB doküman tabanlı NoSQL veri tabanıdır. 2005 yılında açık kaynak kodlu olarak Erlang diliyle geliştirilmeye başlanmıştır. CouchDB'de veriler JSON olarak tutulurlar ve bir şemaya bağlı değildir. Her bir JSON birbirinden farklı olabilir. Ayrıca bir JSON verisine ek olarak file dosyaları da saklayabilir. Bu çalışmada JSON içerisinde gazetenin adı, basım tarihi, indirme bağlantısı ile birlikte ek olarak ga-

zetenin PDF verisi ile birlikte tutulmaktadır. Bu bilgiler gazete adında bir tabloda tutulmaktadır. Gazete tablosuna örnek bir veri Şekil 3.5.'de gösterilmiştir.



Şekil 3.5. CouchDB Gazete tablosu örnek veri

Gazete PDF verilerinin sayfalara çevrilmiş şekli "page" adında ayrı bir tabloda tutulmuştur. Bu tabloda gazete sayfasının görüntü verisi haricinde görüntüde elde edilen metin verisine ek olarak varlık isimleri de JSON olarak tutulmuştur. Page tablosundaki örnek bir veri Şekil 3.6.'da gösterilmiştir.



Şekil 3.6. CouchDB Page Tablosu Örnek veri

3.3.2. Verinin Elasticsearch'te İndekslenmesi

Popülerliği ve kullanım kolaylığı nedeniyle, sistemin indeksleme ve arama motoru bölümünü uygulamak için Elasticsearch'ü (URL-8) kullanmaya karar verdik. Elasticsearch, Apache Lucene projesine (URL-9) dayanan bir tam metin arama ve analiz motorudur. Kibana ve Logstash ile birlikte, günlük kaydı, metin verilerini indeksleme ve tam metin arama için açık kaynaklı bir çözüm olan ELK (Elasticsearch, Logstash, Kibana) yığınını oluştururlar. Ölçeklenebilirlik, hızlı performans, çok dilli destek, belge yönelimli (JSON), otomatik tamamlama ve örnek arama ve şema ücretsiz gibi Elasticsearch'ü seçmenin birçok nedeni vardır.

Herhangi bir makinede veya yüzlerce düğüm içeren bir kümede mükemmel bir şekilde çalışır ve deneyim neredeyse aynıdır. Elasticsearch, dağıtılmış ters çevrilmiş dizinleri kullanarak en iyi eşleşmeleri hızla bulur. Elasticsearch, belgeler için serileştirme biçimi olarak JSON'u kullanır. JSON, NoSQL hareketi tarafından kullanılan standart biçim haline gelmiştir. Otomatik tamamlama, kullanıcıları yazarken alakalı sonuçlara yönlendiren ve arama hassasiyetini artıran bir gezinme özelliğidir. Elasticsearch, indeksleme işleminden önce indeks tipi ve alan tipi gibi bazı tanımlamalar gerektirmez. CouchDB'de depolanan gazete verilerinin isim, basım tarihi, metin verileri ve varlık isimleri Elasticsearch veri tabanında indekslenmiştir.

3.3.3. Web Tabanlı GUI ve Görselleştirme

Veriler toplanıp işlendikten ve üzerindeki işlemlerden sonra hazır hale geldiğinde bu işlenmiş veriler üzerinde arama yapmak ve görselleştirmek için web tabanlı bir arayüz geliştirdi.

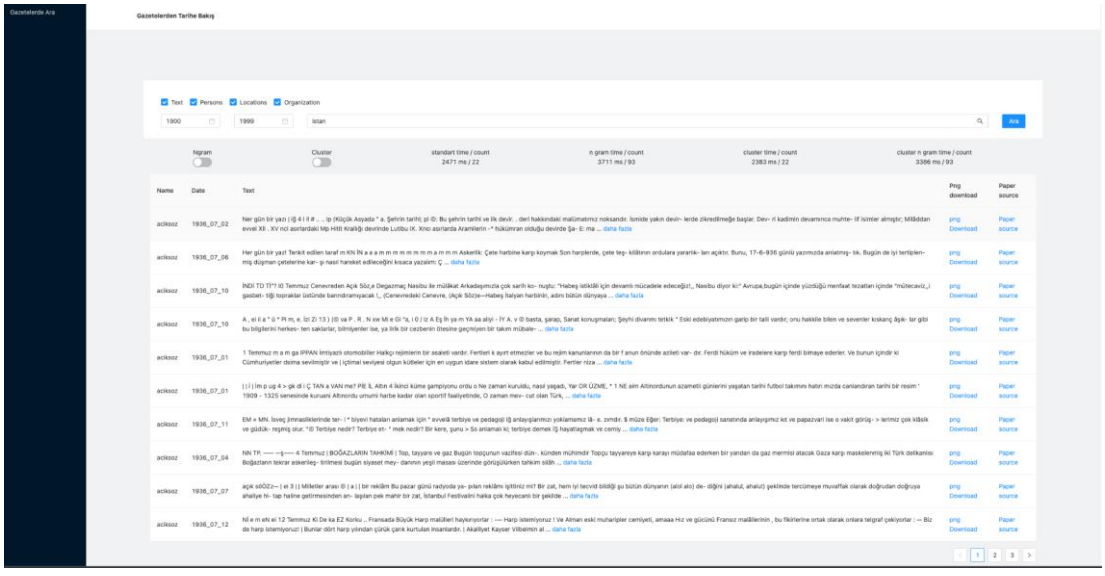
Bu arayüz için Python'da flask (URL-10) paketi kullanılarak bize belli parametreler alıp cevapları dönecek api geliştirildi. API verileri görselleştirecek arayüze veri tabanlarından bilgileri alıp aktarmak için yapıldı. API parametre olarak;

- gazeteler içinde arama yapılırken basım tarih aralığı (başlangıç ve bitiş tarihi olarak),
- hangi index türünde arama yapılacağı,
- gazetenin tam metni içerisinde arama yapılacak mı yoksa yapılmayacak mı,
- VİT sonucu bulunan insan isimleri içerisinde arama yapılacak mı yapılmayacak mı,

- VİT sonucu bulunan lokasyon bilgileri içerisinde arama yapılacak mı yapılmayacak mı,
- VİT sonucu bulunan organizasyon isimleri içerisinde arama yapılacak mı yapılmayacak mı,
- arama küme Elasticsearch sunucusunda mı yoksa tek düğüm olarak çalışan Elasticsearch sunucusunda mı yapılacak,

bilgilerini almaktadır. Cevap olarak da JSON formatta bir dizi içerisinde şartlara uyan gazete verilerini bize cevap olarak dönmektedir.

Yazılan API ile haberleşerek sonuçları gösteren uygulama Angular çerçevesi (URL-11) ile web ortamında çalışacak şekilde geliştirilmiştir. Angular Google tarafından desteklenen, typescript dili kullanarak nesneye dayalı programlama mimarisi ile web uygulamaları geliştirmemizi sağlayan çatı bir uygulamadır. Ara yüz bileşenleri içinde Angular uygulamasına entegre edebildiğimiz ng zorro kütüphanesi (URL-12) kullanılmıştır. Ng zorro kütüphanesinin girdi, tarih seçici, tablo gibi bileşenleri arayüz geliştirilirken kullanılmıştır. Ara yüz geliştirilen API'nin ihtiyaç duyduğu parametleri kullanıcıdan alıp dönen cevapları listelemektedir. Ara yüzün detaylı kullanımı ekte verilmiştir. Şekil 3.7.'de ara yüzün ekran görüntüsünün örneği verilmiştir.



Şekil 3.7. Arama sorguları göndermek için GUI

4. DENEYSEL TESTLER VE BULGULAR

Bu bölümde tez kapsamında önerilen sisteme ait alt bileşen testlerine ve ölçeklenebilirlik testlerine yer verilecektir.

4.1. Test Ortamının Hazırlanması

Bu proje Python 3'te uygulandı ve aşağıdaki özelliklere sahip yerel bir makinede test edildi: CPU: Intel Core i7-7700HQ, RAM: 12 GB DDR4-2400, GPU: Nvidia GeForce GTX 950M 2GB GDDR5, OS: Ubuntu 18.04 LTS.

4.2. VİT Testleri

Genel olarak, yalnızca birkaç fark edilebilir hata içeren VİT sonuçları aşağıda örnek olarak gösterilmektedir. Doğru versiyonlar parantez içinde verilmiştir.

4 Şubat 1929 tarihli Cumhuriyet gazetesinin birinci sayfası:

- PERSON: YUNUS NAD I _ Marehanesi (Idarehanesi), Haydarpaşa, Jermen La, Şozüttin, Emanullah Hazretlerin (Hazretleri), Habibulla (Habibullah), Ali Ahmet, Jermen La- (Jermen Labort), LawRens;
- LOCATION: Istanbul, Fırtına, Avrupa, Amerika, Amerikada, Türkiye, Berlin, Antalya, Karabiğa (Karabağ), Konya, Mersin, Izmir, Trabzonda, Londra, Moskova, Efganistanda, Atinaya;
- ORGANISATION: A Baş Muharriri, Liman Şiirketi, Rasathine (Rasathane).

4.3. İndeksleme Testleri

OCR yapılmış verileri indekslemek için, CouchDB'de depolanan verileri almak için sekiz paralel iş çalıştırıyoruz ve Elasticsearch'te indeksliyoruz, yaklaşık 80 doküman/sn'lik bir verim ölçtük. Bu, Elasticsearch'ün gerçek çıktı sınırından ziyade CouchDB sorguları ve ağ gecikmesi tarafından darboğazlıydı.

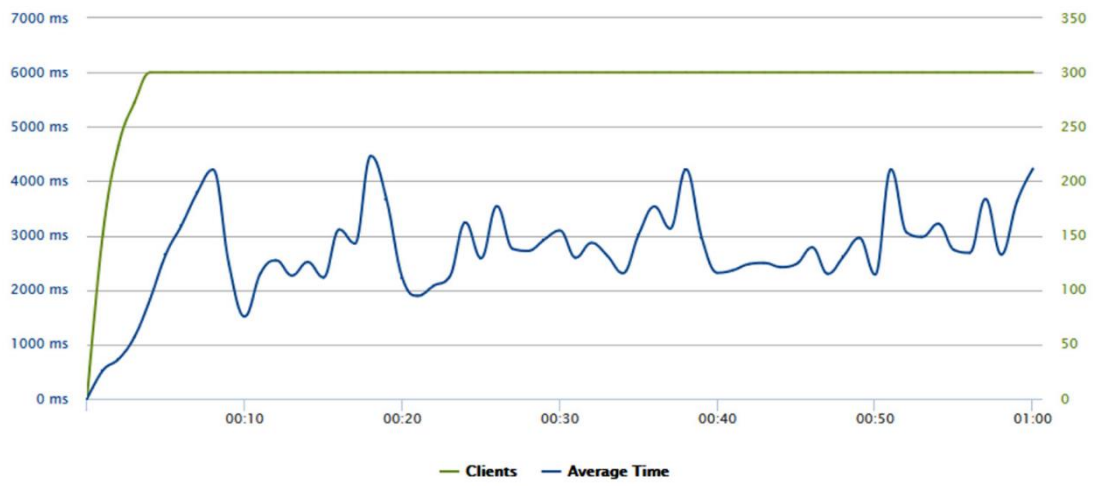
Tablo 4.1.'de gösterildiği gibi Elasticsearch indeksleme ve arama performansı hakkında daha iyi bir fikir edinmek için birkaç test daha yapıyoruz. Bu test sonuçlarında Elasticsearch'te n-gram belirteci kullanılmamıştır.

Tablo 4.1. Elasticsearch sonuçları

Sorgu	Sonuç sayısı	Zaman (ms)
q1	512	900
q2	350	371
q3	202	249
q4	41	175

4.4. Stres Testleri

Loader.io (URL-13) ve Postman (URL-14)'ün bir kombinasyonunu kullanarak API'mizi stres testiyle test ettik. Şekil 4.1.'de de gösterildiği gibi sonuçlar bir istemci için sorgu başına yanıt süresinin, 2500 ms tepki süresi ile yaklaşık 317 ms olduğunu ve 300 istemci/ms'lik bir yükü kaldırabileceğini gösteriyor.



Şekil 4.1. Arama sorguları için stres testleri

4.5. Ölçeklenebilirlik Testleri

Tüm geliştirmeler tamamlandıktan sonra Elasticsearch'te hem indeks farklılıklarının hem de küme ile tek düğüm olarak çalışmasının arama performansına etkisi incelenmiştir. İnceleme için eş bilgisayarların temininde Google Cloud servislerinin sanal sunucu imkanlarından faydalanılmıştır. Google Cloud servisleri bulut ortamda bulut hizmetlerinden faydalanmamızı sağlayan bir Google servisidir.

Elasticsearch'te tek düğüm ve çoklu düğüm karşılaştırması için Google Cloud servislerinden 4 adet aynı özelliklerde ve aynı bölgede çalışan sanal sunucu kiralanmıştır. Bu sanal sunuculardan bir tanesinde Elasticsearch tek düğüm olarak çalıştırılmıştır. 3 tanesinde Elasticsearch 3 düğüm olarak çalışacak şekilde konfigüre edilmiş ve çalıştırılmıştır. Sonrasında CouchDB'de depolanan bu elasticsearch veri tabanlarında indekslenmiştir.

N-gram indeksleme metin verilerini indekslerken n adet elamandan oluşan diziler şeklinde indeksler. Buda yine tam metin içerisinde hızlı arama yapmamızı sağlan bir metoddur. N sayısı değişkenlik gösterebilir. Bu verinin kaç elamandan oluşacak diziler şeklinde indeksleneceğini göstermektedir. Örneğin n değeri 3 ken 'merhaba' kelimesinin indekslemesi 'mer', 'erh', 'rha', 'hab', 'aba' şeklinde olacaktır. Bu indeksleme japonca ve çince gibi kelime sonu olmayan dillerde daha güçlüdür. Eğer Elasticsearch indeks ile karşılaştıracak olursak Elasticsearch'te merhaba kelimesinin aramak içim merhaba kelimesinin tamamının yazılması gerekirken n değeri 3 olan bir n-gram indekslemede 'mer' kelimesi aratıldığında merhaba kelimesi bulunacaktır.

Bu projede de elde edilen metin ve varlık ismi verileri n-gram indeksleme ile ayrıca indekslenmiş ve performans sonuçları Tablo 4.2 ve Tablo 4.3'te karşılaştırılmıştır. N-gram indekslemede n değeri 3 ile 10 arasındaki sayılar olacak şekilde konfigüre edilmiştir.

Tablo 4.2. Tek düğüm sonuçlar

Veri miktarı	Standart indeksleme zamanı (ms)	N-gram indeksleme zamanı (ms)
1 sayfa	286	370
10 sayfa	646	723
100 sayfa	620	645

Tablo 4.3. Küme bilgisayar sonuçlar

Veri miktarı	Standart indeks zamanı (ms)	N-gram indeksleme zamanı (ms)
1 sayfa	260	310
10 sayfa	537	539
100 sayfa	574	554

5. SONUÇLAR VE ÖNERİLER

Gazeteler çok çeşitli bilgiler içerir, bu nedenle çeşitli bilimsel disiplinler için mükemmel bir kaynaktır. Bunlar daha geniş halk için kolayca okunabilen materyallerdir, ancak yalnızca bu konular için değillerdir. İyi bilinen tarihi olaylarla bağlantılıdır ve okuyucuya tarihsel bir deneyim sunar. OCR ile işlenmiş gazete metinleri, kolektif ve işbirlikçi karakterleri göz önüne alındığında, bir topluluğun kelime dağarcığında metin madenciliği yapmak için oldukça kullanışlıdır. Bu da onları ilgili zaman diliminde kullanılan yazı dilinin iyi birer temsilcisi haline getirir. Kağıdın düşük kalitesi ve zaman zaman malzemeyi yavaş yavaş yok eden asidik mürekkebi nedeniyle tarihi matbaanın dijitalleştirilmesi gerekiyor. Gazeteler, etnik azınlıkların tarihi ve kültürü için belgeler olarak özellikle önemlidir.

Bu çalışmada, değerli Türk tarih gazeteleri veri setini aranabilir bir metin formatına dönüştürmek ve her alandan araştırmacının faydalanabilmesini sağlamayı kendimize görev edindik. Bu yüzden bazı insanlar, kültür ve güvenlik odaklı işler için kullanabilirler.

5.1. Tez Kısıtları

Optik karakter tanıma işlemini için kullandığımız Google'ın tesseract-ocr motorunun başarımı bizim elimizde değildir. Bu optik karakter tanıma başarımını kısıtlamaktadır. Ayrıca optik karakter tanıma uyguladığımız taranmış gazete verilerinde kağıdın eskimesine ve tarama işleminde oluşan hatalar resimlere yansımış ve optik karakter tanıma başarımını etkilemektedir.

Varlık ismi tanıma işleminde yapay zeka eğitimi için kullandığımız zemberek kütüphanesinin veri seti güncel Türkçe metinlerden oluşmaktadır. Oysa bizim gazete verilerimiz eski Türkçe metinlerde oluşmaktadır. Buda varlık ismi tanıma işlemi başarımı için bir kısıt oluşturmaktadır.

5.2. İleriki Çalışmalar

Optik karakter tanıma ile üretilmiş eski tarihli gazete verilerinde daha detaylı çıkarımlar yapılabilir. Bunlar gazetelerin özet verileri çıkarmak tarihlere göre gündem analizi yapılabilir. Ayrıca yapay zeka ile optik karakter tanıma başarımını olumsuz etkileyen tarama

hataları ve kağıt yıpranmasından kaynaklı hatalar giderilebilir.

Varlık ismi tanıma işleminde kullanılan yapay zeka güncel Türkçe metinlerden oluşan veriler ile eğitilmiştir. Buda eski tarihli gazete verilerinde varlık ismi tanıma işleminin başarımını düşürmektedir. Yapay zeka modeli eski tarihli metinlerden oluşacak veri seti ile eğitildiğinde varlık ismi tanıma işlemi başarımını arttıracaktır.

KAYNAKLAR

Ahonen, E., Hyvonen, E., (2009). Publishing Historical Texts On The Semantic Web-A Case Study, *IEEE International Conference On Semantic Computing*, Berkeley, CA, USA, 14–16 Eylül 2009.

Akpınar, M.Y., Oral, B., Engin, D., Emekligil, E., Arslan, S., Eryiğit, G., (2019). A Semi-Automatic Annotation Interface for Named Entity and Relation Annotation on Document Images, *4th International Conference on Computer Science and Engineering (UBMK)*, Samsun, Türkiye, 11-15 Eylül 2019.

Anjewierden, A., (2001). AIDAS: Incremental Logical Structure Discovery in Pdf Documents, *Sixth International Conference On Document Analysis And Recognition*, Seattle, WA, USA, 13 Eylül 2001.

Apperley, M., Cunningham, S.J., Keegan, T.T., (2001). Niupepa: A Historical Newspaper Collection, *Commun ACM 2001*, *44*(5), 86–87.

Anh, L.T., Arkhipov, M.Y., Burtsev, M.S., (2017). Application of A Hybrid Bi-LSTM-CRF Model to The Task Of Russian Named Entity Recognition, *Artificial Intelligence and Natural Language*, Petersburg, Rusya, 20-23 Eylül 2017.

Arslan, A., Alkılınç, A., Dinçer, B.T., (2018). Büyük Veri Setlerinde Varlık Tanıma: En Sık Geçen E-Posta, Web Adreslerinin ve Emojilerin Tespit Edilmesi, *Academic Perspective Procedia*, *1*(1), 399-406.

Atay, B., Sönmez, B.C., Eken, S., Sayar, A., (2018). DocDig: Dijitallestirilmis Dokümanlarda İcerik Tabanlı Figür Arama, *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, *6*(1), 68–78.

Bogaard, T., Hollink, L., Wielemaker, J., Ossenbruggen, J.V., (2019). Metadata Categorization For İdentifying Search Patterns in A Digital Library, *Journal of Document*, *75*(6), 87.

Borin, L., Kokkinakis, D., Olsson, L.J., (2007). Naming The Past: Named Entity And Animacy Recognition In 19th Century Swedish Literature, *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH)*, Prag, Çek Cumhuriyeti, 28 Haziran 2007.

Cekınel, R.F., Ağriman, M., Karagöz, P., Yilmaz, B., (2019). Named Entity Recognition with Conditional Random Fields on Turkish News Dataset: Revisiting the Features, *27th Signal Processing and Communications Applications Conference (SIU*), Sivas, Türkiye, 24-26 Nisan 2019.

Chao, H., Fan, J., (2004). Layout And Content Extraction For Pdf Documents, *International Workshop On Document Analysis Systems*, Floransa, İtalya, 8–10 Eylül 2004.

Dalkılıç, F.E., Gelişli, S., Diri, B., (2010). Named entity recognition from Turkish texts, *18th Signal Processing and Communications Applications Conference (SIU)*, Diyarbakır, Türkiye, 22-24 Nisan 2010.

Ekbal, A., Haque, R., Bandyopadhyay, S., (2008). Named Entity Recognition İn Bengali: A Conditional Random Field Approach, *Third International Joint Conference On Natural Language Processing: Volume-II*, Hydeberad, Hindistan, 7-12 Ocak 2008.

Eken, B., (2005). Kısa Metinlerde Varlık İsmi Tanıma, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, 389367.

Eken, S., (2020). Büyük Verinin İnteraktif Görselleştirilmesi: Tableau Üzerine Öğrenci Deneyimleri, *Avrupa Bilim ve Teknoloji Dergisi*, (18), 262-271.

Eltyeb, S., Salim, N., (2014). Chemical named entities recognition: a review on approaches and applications, *Journal of cheminformatics*, *6*(1), 17.

Fielding, R., (2000). Architectural styles and the design of network-based software architecture, Doktora Tezi, University of California, Irvine, Irvine, CA.

Gao, J., Li, M., Huang, C.N., Wu, A. (2005). Chinese word segmentation and named entity recognition: A pragmatic approach, *Computational Linguistics, 31*(4), 531-574.

Gatos, B., Konidaris, T., Ntzios, K., Pratikakis, I., Perantonis, S.J., (2005). A segmentation-free approach for keyword search in historical typewritten documents, *8th International Conference On Document Analysis And Recognition (ICDAR'05)*, Seoul, Güney Kore, 31 Ağustos–1 Eylül 2005.

Généreux, M., Spano, D., (2005). *NLP challenges in dealing with OCR-ed documents of derogated quality*, Replicability and Reproducibility in Natural Language Processing: Adaptive Methods, Resources and Software, https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdG-RvbWFpbnxhZGFwdGl2ZW5scDIwMTV8Z3g6Mzc1ZmNjNDg5ZDUzYTY5MA, (Ziyaret tarihi: 10 Şubat 2021)

Gray, J., Chambers, L., Bounegru, L., (2012) *The data journalism handbook: how journalists can use data to improve the news*, Sebastopol, California, O'Reilly Media, Inc.

Güneş, A., Tantuğ, A.C., (2018). Turkish named entity recognition with deep learning. *26th Signal Processing and Communications Applications Conference (SIU)*, İzmir, Türkiye, 2-5 Mayıs 2018.

Güngör, O., Üsküdarlı, S., Güngör, T., (2018). Recurrent neural networks for Turkish named entity recognition, *26th Signal Processing and Communications Applications Conference (SIU)*, İzmir, Türkiye, 2-5 Mayıs 2018.

Hadjar, K., Rigamonti, M., Lalanne, D., Ingold, R., (2004). Xed: a new tool for extracting hidden structures from electronic documents. *First International Workshop on Document Image Analysis For Libraries*, Palo Alto, CA, USA, 23–24 Ocak 2004.

Hassan, T., (2009). Object-level document analysis of pdf files. *9th ACM symposium on document engineering*, Münih, Almanya 15–18 Eylül 2009

Hauser, A.W., (2007). OCR-postcorrection of historical texts. Doktora Tezi, Ludwig-Maximilians-Universität, Münih, Almanya 2007.

Isozaki, H., (2001). Japanese named entity recognition based on a simple rule generator and decision tree learning, *39th Annual Meeting on Association for Computational Linguistics*, Toulousei Fransa, 4 Temmuz 2001.

Jerele, I., Erjavec, T., Pokorn, D., Kavčič-Čolić, A., (2011). Optical character recognition of historical texts: end-user focused research for Slovenian books and newspapers from the 18th and 19th century, *SEEDI conference*, Zagreb, Hırvatistan, 16–20 Mayıs 2011.

Jones, A., Crane, G., (2006). The challenge of virginia banks: an evaluation of named entity analysis in a 19th-century newspaper collection, *6th ACM/IEEECS joint conference on digital libraries (JCDL '06)*, Chapel Hill, NC, USA, 11–15 Haziran 2006.

Kettunen, K., Mäkelä, E., Ruokolainen, T., Kuokkala, J., Löfberg, L., (2016). *Old content and modern tools-searching named entities in a Finnish OCRed historical newspaper collection 1771-1910*, arXiv, https://arxiv.org/abs/1611.02839 (Ziyaret tarihi: 12 Mart 2021).

Kettunen, K., Ruokolainen, T., Liukkonen, E., Pierrick, T., Daniel, A., Thierry P., (2019). Detecting articles in a digitized Finnish historical newspaper collection 1771–1929: early results using the pivaj software, *3rd international conference on digital access to textual cultural heritage*, Bürüksel, Belçika, 8–10 Mayıs 2019.

Kim, J.D., Ohta, T., Tsuruoka, Y., Tateisi, Y., Collier, N., (2004). Introduction to the bio-entity recognition task at JNLPBA, *Natural language processing in biomedicine and its applications*, Cenevre, İsviçre, 28-29 Ağustos 2004.

Küçük, D., Arıcı, N., (2006). Türkçe için Wikipedia Tabanlı Varlık İsmi Tanıma Sistemi, *Politeknik Dergisi, 19*(3), 325-332.

Küçük, D., Arıcı N. (2016). A named entity recognition dataset for Turkish. *24th Signal Processing and Communication Application Conference (SIU),* Zonguldak, Türkiye, 16-19 Mayıs 2016.

Labusch, K., Kulturbesitz, P., Neudecker, C., Zu, S., (2019). *Bert for named entity recognition in contemporary and historical German*, Konvens, https://konvens.org/proceedings/2019/papers/KONVENS2019_paper_4.pdf (Ziyaret tarihi: 18 Nisan 2021)

Leavitt, N., (2010). Will NoSQL databases live up to their promise?, *IEEE Comput 2010*, *43*(2), 12–14.

Lovegrove, W.S., Brailsford, D.F., (1995). Document analysis of pdf files: methods, results and implications, *Electron Publish Originat Dissem Design 1995*, *8*(3), 207–220.

Kim, S.M., Cassidy, S. (2015). Finding names in trove: named entity recognition for Australian historical newspapers, *Australasian Language Technology Association Workshop 2015*, Parramatta, Australia, 8-9 Kasım 2015.

Minkov, E., Wang, R.C., Cohen, W., (2005). Extracting personal names from email: Applying named entity recognition to informal text, *Human language technology conference and conference on empirical methods in natural language processing*, British Columbia, Kanada, 6-8 Ekim 2005.

Nadeau, D., Sekine, S., (2007). A survey of named entity recognition and classification, *Lingvisticae Investigationes*, 30 (1), 3-26.

Neudecker, C., Antonacopoulos, A., (2016). Making Europe's Historical Newspapers Searchable, *12th IAPR Workshop on Document Analysis Systems (DAS)*, Santorini, Yunanistan, 11-14 Nisan 2016.

Neudecker, C., Wilms, L., Faber, W.J., Veen T. (2014). Large-scale refinement of digital historic newspapers with named entity recognition, *IFLA Newspapers/GENLOC Pre-Conference Satellite Meeting*, Cenevre, İsviçre, 13-14 Ağustos 2014

Neudecker, C., (2016). An open corpus for named entity recognition in historic newspapers, *10th International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenya, 23-28 Mayıs 2016.

Niklas, K., (2010). Unsupervised post-correction of OCR errors, Yüksek Lisans Tezi, Leibniz Universitat Hannover, Hannover.

Okazaki, N., Tsujii, J., (2010). Simple and efficient algorithm for approximate dictionary matching, *23rd international conference on computational linguistics*, Beijing, Çin, 23–27 Ağustos 2010.

Otsu, N., (1979) A threshold selection method from gray-level histograms, *IEEE T Syst Man Cybernet 1979*, *9*(1), 62–66.

Özger, Z.B., Diri, B., (2012). Türkçe dokümanlar için kural tabanlı varlık ismi tanıma, *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 5*(2), 2012.

Özkaya, S., Diri, B., (2011). Named entity recognition by conditional random fields from Turkish informal texts, *19th Signal Processing and Communications Applications Conference (SIU)*, Antalya, Türkiye, 20-22 Nisan 2011.

Pasca, M., (2004). Acquisition of categorized named entities for web search, *13th ACM international conference on Information and knowledge management*, Washington D.C., USA, 2-13 Kasım 2004.

Pirovani, J.P., Nogueira, M., Oliveira, E., (2018). Indexing Names of Persons in a Large Dataset of a Newspaper, *Computational Processing of the Portuguese Language*, Canela, RS, Brazilya, 24-26 Eylül 2018

Ruokolainen, T., Kettunen, K., (2020) *Name the name-named entity recognition in OCRed 19th and early 20th century Finnish news-paper and journal collection data*, CEUR Workshop Proceedings, http://ceur-ws.org/Vol-2612/paper10.pdf (Ziyaret tarihi: 21 Eylül 2021)

Sarı, Ö.C., Aktaş, Ö., (2018). Türkçe Ders Metinleri İçin Özelleştirilmiş Bir Varlık İsmi Tanıma Yapısı, *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 11*(2), 52-68.

Shaalan, K., Raza, H., (2008). Arabic named entity recognition from diverse text types, *Advances in Natural Language Processing*, Gothenburg, İsveç, 25-27 Ağustos 2008.

Tanabe, L., Xie, N., Thom, L.H., Matten, W., Wilbur, W.J., (2005). GENETAG: a tagged corpus for gene/protein named entity recognition, *BMC bioinformatics, 6*(S1), S3.

Terasawa, K., Shima, T., Kawashima, T. A., (2001). Fast appearance-based full-text search method for historical newspaper images, *International conference on document analysis and recognition*, Beijing, China, 18–21 September 2011.

Thompson, P., Batista-Navarro, R., Kontonatsios, G., Carter, J., Toom, E., McNaught, J., Timmermann, C., Worboys M., Ananiadou S., (2016). Text mining the history of medicine, *PLoS ONE 2016, 11*(1), e0144717, 10.1371/journal.pone.0144717

Thompson, P., Carter, J., McNaught, J., Ananiadou, S., (2015). Semantically enhanced search system for historical medical archives, *Digital Heritage*, Granada, İspanya, 28 Eylül–2 Ekim 2015.

URL-1: http://nek.istanbul.edu.tr:4444/ekos/GAZETE/ (Ziyaret Tarihi: 13 Şubat 2022)

URL-2: https://docs.python-requests.org/en/latest/ (Ziyaret Tarihi: 13 Şubat 2022)

URL-3: https://www.crummy.com/software/BeautifulSoup/ (Ziyaret Tarihi: 13 Şubat 2022)

URL-4: https://docs.opencv.org/4.x/index.html (Ziyaret Tarihi: 13 Şubat 2022)

URL-5: https://github.com/madmaze/pytesseract (Ziyaret Tarihi: 13 Şubat 2022)

URL-6: https://github.com/ahmetaa/zemberek-nlp (Ziyaret Tarihi: 13 Şubat 2022).

URL-7: https://jpype.readthedocs.io/en/latest/ (Ziyaret Tarihi: 13 Şubat 2022)

URL-8: https://www.elastic.co/elastic-stack (Ziyaret Tarihi: 13 Şubat 2022).

URL-9: https://lucene.apache.org/ (Ziyaret Tarihi: 13 Şubat 2022).

URL-10: https://flask.palletsprojects.com/ (Ziyaret Tarihi: 16 Mart 2022)

URL-11: https://angular.io/ (Ziyaret Tarihi: 16 Mart 2022)

URL-12: https://ng.ant.design/ (Ziyaret Tarihi: 16 Mart 2022)

URL-13: https://loader.io (Ziyaret Tarihi: 13 Şubat 2022).

URL-14: https://www.postman.com/ (Ziyaret Tarihi: 13 Şubat 2022)

Wilkinson, T., Lindström, J., Brun, A., (2018). *Neural word search in historical manuscript collections*, arXiv, https://arxiv.org/abs/1812.02771#:~:text=We%20address%20the%20problem%20of,to%20as%20%22word%20spotting%22 (Ziyaret Tarihi: 22 Haziran 2021).

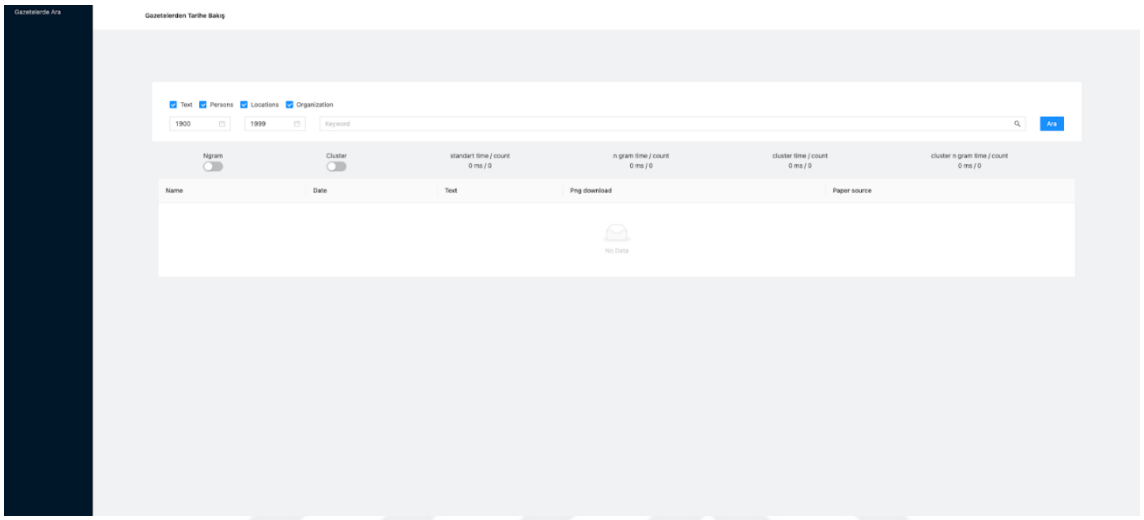
Willems, M., Atanassova, R., (2015). Europeana Newspapers: searching digitized historical newspapers from 23 European countries, *Insights, 28*(1), 51–56. DOI: 10.1629/uksg.218

EKLER

Ek-A

ARAYÜZ UYGULAMASI KULLANIM KILAVUZU

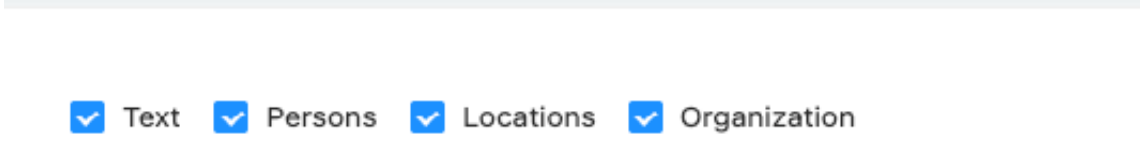
Uygulama ekranına girdiğinizde Şekil A.1.'deki gibi bir ekran sizi karşılayacaktır.



Şekil A.1. Açılış ekranı

ARAMA İLE İLGİLİ BÖLÜM

Bu ekranda en yukarıda Şekil A.2.'deki görüldüğü üzere text, person, locations, organization gibi chekboxlar bulunmaktadır. Bu checkboxlardan işaretlediğiniz veriler içerisinde arama yapılacaktır. Varsayılan olarak hepsi seçili gelmektedir.



Şekil A.2. Arama yapılacak varlık ismi türü ve tam metin seçimi

Altında Şekil A.3.'te görüldüğü gibi iki tane yıl seçim alanı bulunmaktadır. Burada seçtiğiniz yıllar arasında arama yapılacaktır.

Şekil A.3. Arama yapılacak yıl aralığı seçimi

Şekil A.4.'te görüldüğü gibi arama kelimesini yazabileceğiniz bir metin kutusu bulunmaktadır. Bu alana yazdığınız kelime gazete verileri içerisinde aranacaktır.

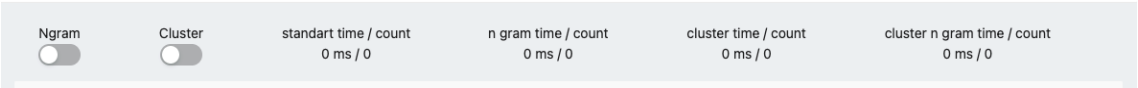
Şekil A.4. Aranacak metin girişi

Şekil A.5.'te görülen arama butonu, arama işleminin başlamasını sağlamaktadır.

Şekil A.5. Arama Butonu

SONUÇ LİSTELEME İLGİLİ AYARLAR VE BİLGİLER

Arama ile ilgili bölümün altında sonuçların listelenmesi ile ilgili ayarlar ve bilgiler bulunmaktadır. Şekil A.6.'da bu alan görülmektedir.



Şekil A.6. Listeleme sonucu verileri ve ayarları

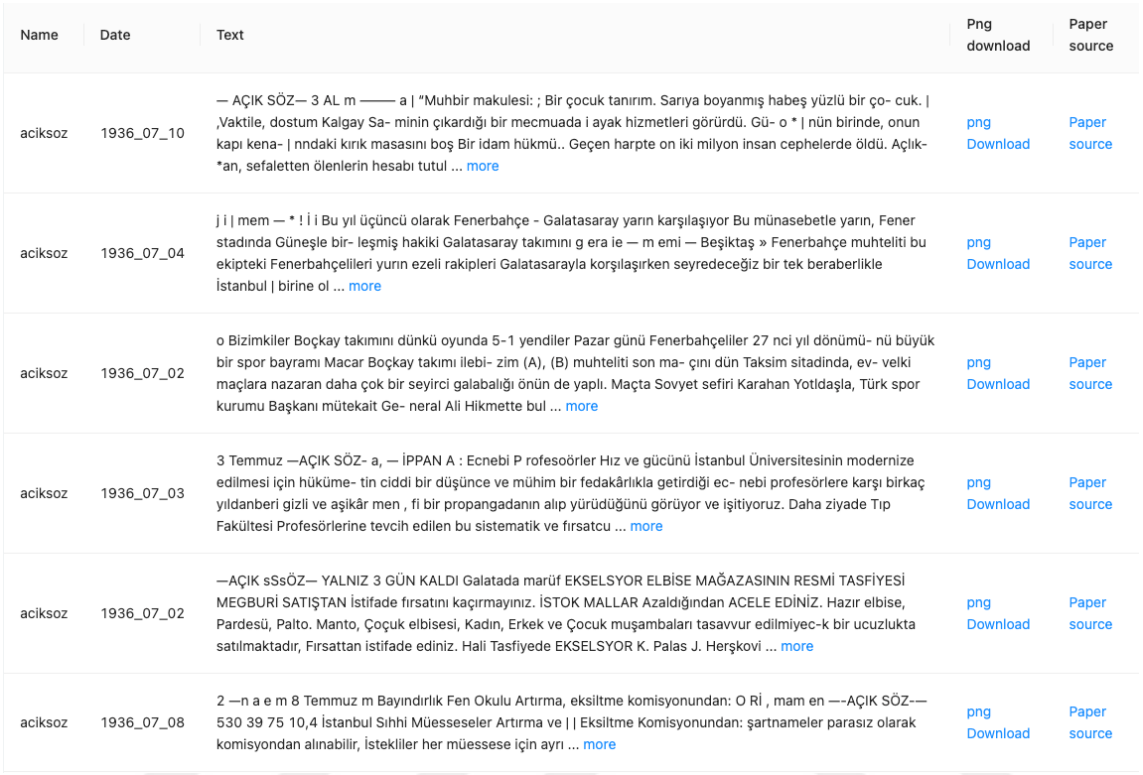
Ngram ve cluster seçimleri arama sonuçları listelenirken neye göre listeleneceği seçilmektedir. Ngram seçilirse ngram indeksine göre bulunan sonuçlar tabloda listelenecektir. Cluster seçimi seçilirse küme sunucudan gelen sonuçlar listelenecektir. Eğer n gram indeks seçilmezse standart indeks sonuçları, küme seçilmezse tek düğüm sonuçları listelenecektir.

Arama sonuçları ile ilgili bilgi içeren kısımlarda ise, arama sonuçlarının ne kadar sürede getirildiği ve arama sonuçlarında kaç adet kayıt bulunduğunu göstermektedir.

Not: Bu arama sürelerine, internet ortamının gecikmesi de dahildir.

SONUÇ LİSTESİ

Şekil A.7.'de 'İstanbul' kelimesi için örnek arama sonuçları gösterilmektedir.

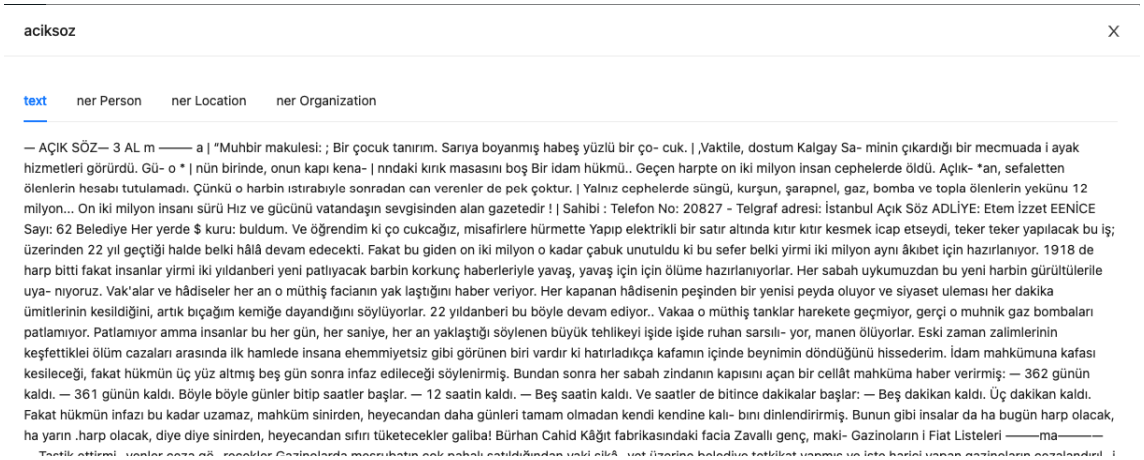


Name	Date	Text	Png download	Paper source
aciksoz	1936_07_10	— AÇIK SÖZ— 3 AL m —— a | "Muhbir makulesi: ; Bir çocuk tanırım. Sarıya boyanmış habeş yüzlü bir ço- cuk. | ,Vaktile, dostum Kalgay Sa- minin çıkardığı bir mecmuada i ayak hizmetleri görürdü. Gü- o * | nün birinde, onun kapı kena- | nndaki kırık masasını boş Bir idam hükmü.. Geçen harpte on iki milyon insan cephelerde öldü. Açlık- *an, sefaletten ölenlerin hesabı tutul ... more	png Download	Paper source
aciksoz	1936_07_04	j i | mem — * ! İ i Bu yıl üçüncü olarak Fenerbahçe - Galatasaray yarın karşılaşıyor Bu münasebetle yarın, Fener stadında Güneşle bir- leşmiş hakiki Galatasaray takımını g era ie — m emi — Beşiktaş » Fenerbahçe muhteliti bu ekipteki Fenerbahçelileri yurın ezeli rakipleri Galatasarayla korşılaşırken seyredeceğiz bir tek beraberlikle İstanbul | birine ol ... more	png Download	Paper source
aciksoz	1936_07_02	o Bizimkiler Boçkay takımını dünkü oyunda 5-1 yendiler Pazar günü Fenerbahçeliler 27 nci yıl dönümü- nü büyük bir spor bayramı Macar Boçkay takımı ilebi- zim (A), (B) muhteliti son ma- çını dün Taksim sitadinda, ev- velki maçlara nazaran daha çok bir seyirci galabalığı önün de yaplı. Maçta Sovyet sefiri Karahan Yotldaşla, Türk spor kurumu Başkanı mütekait Ge- neral Ali Hikmette bul ... more	png Download	Paper source
aciksoz	1936_07_03	3 Temmuz —AÇIK SÖZ- a, — İPPAN A : Ecnebi P rofesoörler Hız ve gücünü İstanbul Üniversitesinin modernize edilmesi için hüküme- tin ciddi bir düşünce ve mühim bir fedakârlıkla getirdiği ec- nebi profesörlere karşı birkaç yıldanberi gizli ve aşikâr men , fi bir propangadanın alıp yürüdüğünü görüyor ve işitiyoruz. Daha ziyade Tıp Fakültesi Profesörlerine tevcih edilen bu sistematik ve fırsatcu ... more	png Download	Paper source
aciksoz	1936_07_02	—AÇIK sSsÖZ— YALNIZ 3 GÜN KALDI Galatada marüf EKSELSYOR ELBİSE MAĞAZASININ RESMİ TASFİYESİ MEGBURİ SATIŞTAN İstifade fırsatını kaçırmayınız. İSTOK MALLAR Azaldığından ACELE EDİNİZ. Hazır elbise, Pardesü, Palto. Manto, Çocuk elbisesi, Kadın, Erkek ve Çocuk muşambaları tasavvur edilmiyec-k bir ucuzlukta satılmaktadır, Fırsattan istifade ediniz. Hali Tasfiyede EKSELSYOR K. Palas J. Herşkovi ... more	png Download	Paper source
aciksoz	1936_07_08	2 —n a e m 8 Temmuz m Bayındırlık Fen Okulu Artırma, eksiltme komisyonundan: O Rİ , mam en —-AÇIK SÖZ-— 530 39 75 10,4 İstanbul Sıhhi Müesseseler Artırma ve | | Eksiltme Komisyonundan: şartnameler parasız olarak komisyondan alınabilir, İstekliler her müessese için ayrı ... more	png Download	Paper source

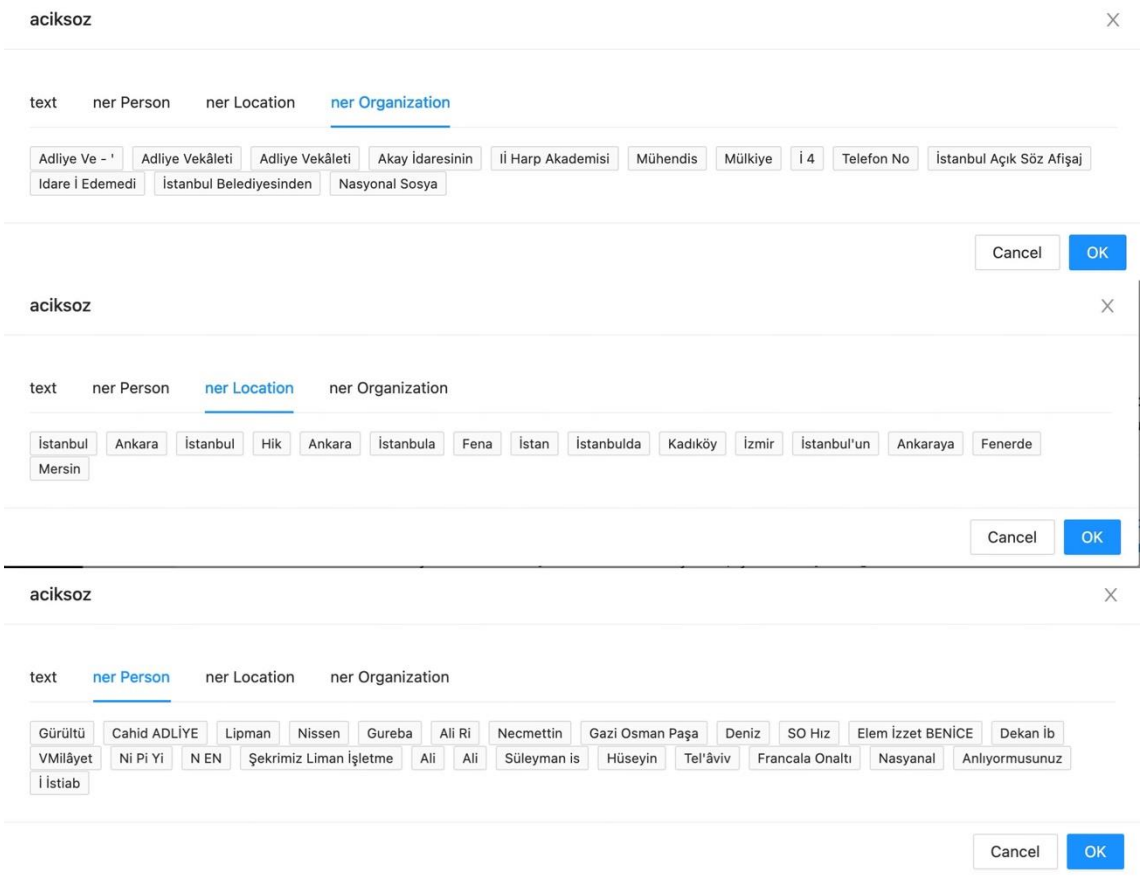
Şekil A.7. Arama sonucu bulunan gazete verilerinin listelendiği tablo

Listeleme tablosunda gazetenin adı, basım tarihi, optik karakter tanıma ile bulunmuş metin verisi, ilgili kelimenin geçtiği gazete sayfasının resim verisine erişebileceğiniz link ve gazetenin tam dokümanına pdf olarak erişebilecek link bulunmaktadır.

Text kolonunda bulunan daha fazla linkine tıklandığında, Şekil A.8. ve Şekil A.9.'daki gibi, o gazete sayfasının metin verisine ve hangi varlık isimlerinin bulunduğuna erişilebilmektedir.



Şekil A.8. Listelenen verilerin detaylı gözlem ekranı



Şekil A.9. Varlık isimlerinin detaylı listelendiği ekran

Png Download kolonundaki bağlantıya tıklandığında, Şekil A.10.'daki gibi, o gazete sayfasının resim verisine erişilmektedir.

Şekil A.10. Gazete sayfasının ham halinin görüntülendiği ekran

Paper source kolonu altındaki bağlantıya tıklandığında Şekil A.11.'daki gibi, gazete verisinin tamamına pdf olarak erişilmektedir.

Şekil A.11. Seçilen gazetenin tamamının görüntülendiği ekran

KİŞİSEL YAYINLAR VE ESERLER

Şahin H. B., Eken S., (2020) Taranmış Türkçe Gazete Dokümanları Üzerinde Varlık İsmi Tanıma, *3rd International Conference on Data Science and Applications (ICONDATA'20)*, Türkiye, 25-28 Haziran 2020.

Menhour H., **Şahin H. B.**, Sarıkaya R. N., Aktaş M., Sağlam R., Ekinci E., Eken S., (2021). Searchable Turkish OCRed historical newspaper collection 1928–1942, *Journal Of Information Science, 47*, 1-13. DOI:10.1177/01655515211000642

ÖZGEÇMİŞ

Sakarya'da Sakarya Anadolu Lisesi'nden mezun oldu. 2014 yılında Kocaeli Üniversitesi Bilgisayar Mühendisliği bölümüne girip, 2018 yılında mezun oldu. Mezuniyet sonrası çeşitli özel sektör şirketlerinde yazılım geliştirici olarak çalışmaya başladı. 2019 yılı içerisinde Kocaeli Üniversitesi Bilişim Sistemleri Mühendisliği anabilim dalında yüksek lisans yapmaya başladı.