TOJDE DERGİSİ ÜZERİNDE LDA İLE KONU MODELLEME
Yüksek Lisans Tezi

Yusuf KARTAL

Eskişehir, 2017

TOJDE DERGİSİ ÜZERİNDE LDA İLE KONU MODELLEME

Yusuf KARTAL

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Doç. Dr. Özgür YILMAZEL

Eskişehir
Anadolu Üniversitesi
Fen Bilimleri Enstitüsü
Mayıs, 2017

JÜRİ VE ENSTİTÜ ONAYI

Yusuf KARTAL'ın "TOJDE dergisi üzerinde LDA ile Konu Modelleme" başlıklı tezi 02/05/2017 tarihinde aşağıdaki jüri tarafından değerlendirilerek "AnadoluÜniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği"nin ilgili maddeleri uyarınca, Bilgisayar Mühendisliği Anabilim dalında Yüksek Lisans tezi olarak kabul edilmiştir.

	Unvanı Adı Soyadı	**İmza**
Üye (Tez Danışmanı)	**Doç.Dr. Özgür YILMAZEL**	**.....**
Üye	**Prof.Dr. Rıfat EDİZKAN**	**.....**
Üye	**Yrd.Doç.Dr. Ahmet ARSLAN**	**.....**

.....

Enstitü Müdürü

ÖZET

TOJDE DERGİSİ ÜZERİNDE LDA İLE KONU MODELLEME

Yusuf KARTAL

Bilgisayar Mühendisliği Anabilim Dalı
Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Mayıs, 2017

Danışman: Doç. Dr. Özgür YILMAZEL

Çeşitli bilgilerin kayıt altına alınması hususunda bilgisayar sistemlerinin güvenlik, maliyet, erişilebilirlik gibi konularda sağladığı avantajlar ile birlikte içinde bulunulan bilgi çağında hızla büyüyen verilere erişimin sağlanması, bu veriler içerisinden aranılan bilginin çıkarılması konusu, üzerinde çalışılması güç problemler doğurmuştur. Latent Dirichlet Allocation gibi konu modelleme algoritmaları ve bu algoritmalar üzerine geliştirilmiş konu modelleme araçları binlerce kayıt arasında sıklıkla bahsedilen konuların saptanmasını sağlayabilmektedir. Bu tez kapsamında yapılan çalışma, The Turkish Online Journal of Distance Education (TOJDE) dergisi tarafından kayıt altına alınmış makalelerin araştırılabilir biçime çevrilmesi ve bu metin veriler üzerinde Latent Dirichlet Allocation algoritması ile konuların algılanmasını amaçlamaktadır. Konuların algılanması ile birlikte kullanıcı tarafından anlaşılır görsel analiz sonuçları sunan bir sistem ortaya koyarak yıllara göre konu dağılımlarını gösteren grafiklere ulaşılmıştır. Çalışmanın gerçekleştirilmesi için elde edilen metin veriler Latent Dirichlet Allocation algoritması ile analiz edilmeden önce makale arşivinde yer alan metinlerde geçen kelimeler, kök bulma gibi işlemlerle sadeleştirilerek konu algılama işleminin başarısının arttırılması sağlanmıştır.

Anahtar Sözcükler: Latent Dirichlet allocation, Konu modelleme, Konu algılama, Metin analitiği.

ABSTRACT

TOPIC MODELLING OF TOJDE JOURNAL WITH LDA

Yusuf KARTAL

Department of Computer Engineering

Anadolu University, Graduate School of Science, May, 2017

Supervisor:Assoc. Prof. Dr. Özgür YILMAZEL

In order to record various information, advantages of computer systems such as security, cost, accessibility as well as the provision of access to rapidly growing data in the information age, the issue of extracting the information sought from these datahas caused difficulties. Topic modeling algorithms such as Latent Dirichlet Allocation and topic modeling tools developed on these algorithms is often used to determine the topics mentioned between thousands of documents. In this thesis, the study aims to generate searchable texts from the articles registered by The Turkish Online Journal of Distance Education (TOJDE) journaland perceive the topics with using Latent Dirichlet Allocation algorithm on these searchable texts. Along with the detection of the topics, a system that presents visual user-friendly analysis charts developedand reached graphical outputs which show the distribution of the topics according to years. The words in the article archive have been simplified with operations such as stemming before analyzing texts with the Latent Dirichlet Allocation, thereby increasing the success of the topic modelling process.

Keywords: Latent Dirichlet allocation, Topic modelling, Topic detection, Text analytics.

TEŞEKKÜR

Tez çalışmamın gerçekleştirilmesi ve tez metninin hazırlanması sırasında yol gösteren ve her aşamasında destek olan danışmanım Doç. Dr. Özgür YILMAZEL'e teşekkürlerimi sunarım.

Desteğini hiçbir zaman esirgemeyen sevgili eşim Tuğba KARTAL'a ve varlığıyla bana güç veren sevgili oğlum Batu KARTAL'a teşekkürü bir borç bilirim.

Yusuf KARTAL

Mayıs, 2017

02/05/2017

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Bu tezin 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 edilemeyen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmanın Anadolu Üniversitesi tarafından kullanılan "bilimsel intihal tespit programı"yla tarandığını ve hiçbir şekilde "intihal içermediğini" beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçlara razı olduğumu bildiririm.

Yusuf KARTAL

İÇİNDEKİLER

Sayfa

TABLOLAR DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

SİMGE ve KISALTMALAR DİZİNİ

LDA	**L**atent **D**irichlet **A**llocation (Saklı Dirichlet Ataması)
MALLET	**MA**chine **L**earning for **L**anguag**ET**oolkit
TOJDE	The **T**urkish **O**nline **J**ournal of **D**istance **E**ducation
HTML	**H**yper **T**ext **M**arkup **L**anguage
CSS	**C**ascading **S**tyle **S**heets
SVG	**S**calable **V**ector **G**raphics
DOM	**D**ocument **O**bject **M**odel
PDF	**P**ortable **D**ocument **F**ormat

1 GİRİŞ

Çeşitli bilgilerin kayıt altına alınması hususunda bilgisayar sistemlerinin güvenlik, maliyet, erişilebilirlik gibi konularda sağladığı avantajlar ile birlikte içinde bulunulan bilgi çağında hızla büyüyen verilere erişimin sağlanması, bu veriler içerisinden aranılan bilginin çıkarılması konusu üzerinde çalışılması güç problemler doğurmuştur. Bunun yanında büyük veri sistemleri ve bilgi erişim sistemleri üzerine yapılan çalışmalar gün geçtikçe ilerlemekte ve böylece artan elektronik kayıtların analiz edilme yöntemleri değişirken bu kayıtlardan daha fazla bilgi sağlamak da olanaklı hale gelmektedir (Baeza-Yates& Ribeiro-Neto, 2012).

Konu modelleme algoritmaları ve bu algoritmalar üzerine geliştirilmiş konu modelleme araçları binlerce kayıt arasında sıklıkla bahsedilen konuların saptanmasını sağlayabilmektedir(Srivastava & Sahani, 2009, s. 71-72). Farklı bilimsel alanlardan üzerinde sıklıkla çalışılan konular ya da nadiren çalışılan konular konu modelleme araçları sayesinde binlerce çalışma arasından özetlenerek; araştırmacılar için kendi alanında çalışmalarına başlamadan önce bilgi sağlamaktadır. Yıllara göre hangi araştırma konularının önemini yitirdiği ya da hangi konuların gündeme geldiği bilgisine erişim, konu modelleme araçları sayesinde olabilmektedir.

Büyük metin veriler üzerinde çalışılırken bir metin içinde işlenen konunun saptanmasını sağlamaya yönelik konu modelleme algoritmaları arasında *Latent Dirichlet Allocation (LDA), Topic Model Visualization Engine, Collaborative modeling for recommendation, Dynamic topic models and the influence model* gibi algoritmalar yer almaktadır. Bu algoritmalar benzer gibi görünse de birisi zamana göre değişen konuları ve bu konuların nasıl değiştiğini belirlemeyi amaçlamışken bir diğeri konu sayısını saptamayı amaçlamıştır. Bununla birlikte konuların hiyerarşik bir düzen içerisinde çıkarılmasını amaçlayan algoritmalar da mevcuttur[1].

Bu tez kapsamında yapılan çalışma, kayıt altına alınmış metin verilerin araştırılabilir biçime çevrilmesi ve üzerinde yapılan analizler ile yazılım tarafından tanınan ve kullanıcı tarafından anlaşılır analiz sonuçlarını sunan bir sistem ortaya koymak amacını gütmektedir. The Turkish Online Journal of Distance Education (TOJDE) dergisi makale arşivi üzerinde çalışılarak, makale metinlerinden temiz ve net

[1]David M. Blei, https://www.cs.princeton.edu/~blei/topicmodeling.html (Erişim Tarihi: 05/11/2016)

bir şekilde ayırt edilebilir sonuçların çıkarılması için doğru yöntemlerin saptanması sağlanacaktır.

TOJDE dergisi makale arşivi İngilizce makaleler içermektedir[2]. Her bir İngilizce makaleden elde edilecek düz metin veri üzerinde algılamanın iyileştirilmesi için metin içeriğini bozmadan veri temizliği ve düzenlemesi yapılarak her yıl hangi konular üzerinde makalelerin yayınlandığı çıkarılacaktır. Böylece araştırmacıların yıllara göre hangi konular üzerinde araştırma yapmaktan vazgeçerken hangi konuların daha çok ilgi gördüğü ile ilgili bilgiye erişilmiş olunacaktır.

Tezin ana hatları şu şekildedir; arama ve konu modelleme hakkında genel bilgi, yapılacak konu modelleme deneylerinde kullanılacak LDA algoritması ve kullanım alanları ile konu modelleme üzerinde yapılan geçmiş araştırmalar Bölüm 2'de ele alınmıştır. Deneylerde kullanılacak verilerin elde edilmesi, analize hazır hale getirilmesi ve üzerinde konu modellemesinin nasıl yapılacağına dair detaylı bilgiye Bölüm 3'te yer verilmektedir. Bölüm 4 ile elde edilen bulguların ortaya konması ve detaylandırılmasısağlandıktan sonra Bölüm 5'te yapılan çalışma ve elde edilen sonuçlar üzerinde yapılan değerlendirmelere ve gelecek için sunulan önerilereyer verilecektir.

[2]http://tojde.anadolu.edu.tr/ (Erişim Tarihi: 05/11/2016)

2 ALT YAPI

Bilgi erişim sistemleri üzerine yapılan araştırmalar ile aranan belgelere hızlı ve doğru erişimi amaçlayan çalışmalar zamanla çok büyük verilerin kategorize edilip indekslenmesi ihtiyacına karşılık konu modelleme yani büyük metin veriler içinden konu çıkartma algoritmalarının ortaya çıkmasına sebep olmuştur (Rosell, 2006). Böylece bilgi erişimi farklı bir boyut kazanmış ve sadece bir arama kelimesine karşılık doğru belgeyi bulmak ile kalmayıp LDA gibi konu modelleme algoritmaları ile tüm belgelerde bahsedilen konuların saptanıp analizlerde kullanılmasına da imkan sağlanmaya çalışılmıştır (Chakraborty vd., 2013).

2.1 Bilgi Erişim Sistemleri

Bilgiye erişme konusunda ilk başlarda makaleler gibi metin içerikli belgelerde yapılan kelime odaklı aramalar daha sonraları geliştirilen algoritmalar ile sadece kelimenin içinde bulunduğu değil gerçekten aranan kelime ya da cümlelerin alakalı olduğu dokümanların bulunmasının sağlanarak kullanıcıya sunulması ile geliştirilmiştir. Burada, doğru bilginin sunulduğu; bazen kullanıcılardan gelen geri beslemeler ile onaylanırken bazen denetçiler yardımı ile hatalı ya da gereksiz bilgilerin temizlenmesi sağlanarak kullanıcılara sunulan sonuçların kalitesi arttırılmaya çalışılmıştır (Yom-Tov vd., 2005, s. 512-519). Bazen kullanıcıdan alınan arama kriteri doğrudan uygulanmak yerine zorluk derecesinin tespit edilerek parçalara ayrılması ve böylece her parça için tespitte bulunulması ve ortak sonuçların gruplanarak sunulması sonuçlardaki kaliteyi yükseltmiştir (Arguello vd., 2009, s. 315-322). Arama motorlarında amaç kullanıcıya en doğru bilgiyi en hızlı şekilde göstermek olduğundan bazı durumlarda arama kelimeleri üzerinde yapılan analizler ile aranan bilginin hangi alana ya da kategoriye ait olduğu bilgisine erişilebilmesi ile aranacak belge sayısı ciddi şekilde azaltılabilmiştir (Arguello vd., 2009, s. 315-322).

Birçok benzer kelime ya da öbekler ile birlikte kullanıcıların gerçekten ne aradığının bulunması sadece parametre olarak görülen arama kelimesi ile mümkün olmamaktadır. İlerleyen çalışmalar ile kullanıcının tarayıcılar ya da arama motorları ile olan etkileşimleri izlenerek kullanıcının gezinme izleri vasıtasıyla daha kaliteli sonuçların sunulabilmesi sağlanmıştır (White vd., 2007, s.159-166). Gelişen algoritmalar bazen arama kriteri olarak kullanıcı tarafından girilen anahtarlar üzerinde bazen de kullanıcıların arama yöntemleri üzerinden gerçekten kullanıcı tarafından

aranan bilgiye erişim imkanı sağlamaya odaklanmıştır. Ancak uzun çalışmalar ile bazen basit seviyedeki kullanıcıların dahi ne aradıklarını tam bilmedikleri durumların olduğu tespit edilmiştir ve bu gibi durumlarda da kullanıcılara destek olunabilmesi için aynı bilgiye erişmeye çalışan daha uzman kullanıcıların deneyimlerinin uzman olmayan kullanıcılara aktarılması ile daha iyi bir arama deneyimi oluşturulmaya çalışılmıştır(Pickensy vd., 2008, s.315-322). Arama motorlarında girilen bir arama kelimesinin arama motoru algoritmaları tarafından farklı çeşitlerdeki kelime öbekleri ile olan ifadelerinin önerilmesi bu duruma örnek gösterilebilir.

Metin içerikli belgelerde yapılan basit aramalar ile başlayan bilgi erişim sistemleri doğası gereği kullanıcı davranışlarını içeren karmaşık algoritmaların ortaya çıkması ile birçok parametreyi göz önünde bulundurarak paralel çalışan çok büyük sistemlerin ortaya çıkmasına sebep olmuştur. Metin içerikli dokümanlarda yapılan aramaların yanı sıra imaj veya video gibi metin içermeyen içeriklerde bile daha farklı algoritmaların geliştirilmesi ile arama yapılabilmesine imkan verilmiştir (Arguello vd., 2009, s. 315-322). Geliştirilen akıllı algoritmalar çeşitli alanlarda hatta basit ev kullanıcılarının kendi bilgisayarlarında kullanımına sunularak kişisel kullanıma açılmış ve bu şekilde geri beslemeler ile büyük veri yığınlarındaki arama algoritmaları daha da iyileştirilmiştir (Sanderson& Croft, 2012, s. 1444-1451).

Günümüzde insanların kullandıkları teknolojik cihazlar arttıkça bu cihazlar ile uyumlu olarak çalışan yazılımlar da gelişerek çeşitlenmektedir. Akıllı telefonlar ya da tablet bilgisayarların gündelik hayatın içine iyice yerleşmesi ile insanların telefonlarını kullanım amaçları zaten bildikleri bilgiye doğrudan erişmekten çok, planlanan şeyin tahmin edilmesi ve bu şekilde kullanıcı ile interaktif olarak etkileşimde bulunularak aranacak bilginin kullanıcıya önerilmesi şeklini almıştır (Maarten vd., 2014, s. 681-686). Akıllı telefonların sürekli çevrimiçi olarak kullanımı ile bilgi erişim sistemleri ilk ortaya çıktığı şekliyle yapılan aramaya en iyi sonucu bulma hedefinin yanında kullanıcıyı doğru uygulama, görev ya da içeriğe yönlendirme imkanı da vermektedir (Yuasa vd., 2011). Sabit bilgisayarlardan yapılan aramalara ek olarak mobil cihazlarda donanımsal olarak eklenmiş konum bulma servisleri, bilgi erişim sistemlerine parametre olarak konum bilgisini de sağlayarak konuma özel sorgu sonuçları dönme imkanı sağlayabilmektedir (Masahide vd., 2011).

Paralel olarak gelişen dil tercüme yazılımları ile entegre çalışan bilgi erişim sistemleri herhangi bir dilde yapılan aramayla alakalı olabilecek ve belki daha kabul

edilebilir bilgiye sahip farklı bir dildeki belgeyi kullanıcıya sunabilmektedir (Berger& Lafferty, 1999, s. 222-229). Öncelikli olarak arama yapılan dilin göz önünde bulundurulması ya da sonuç bulunamadığı durumlarda başka dillerdeki sonuçların kullanıcıya önerilmesi gelişen arama motorlarında karşılaşılan durumlardır (Pirkola vd., 2001, s. 209-230).

Bilgi erişiminin merkezinde erişim modelleri bulunmaktadır. Bir çok erişim modeli tanımlanmış ve bunlar üzerinde çeşitli çalışmalar yapılmış olmasına rağmen en iyi performansı hep sezgisel olarak uyarlanmış erişim modelleri sergilemektedir. Bu erişim modellerinin altında belli belirsiz bazı yaklaşımların olduğu görülmekle birlikte küçük farklılıklar ile daha başarılı sonuç elde etmek amaçlanmıştır. Erişim performansının iyileştirilmesi için bazı yöntemlerin kullanıldığı görülmektedir ancak burada ortaya çıkan soru "bunu matematiksel olarak nasıl ifade edebiliriz?"'dir. Bunun için bu kısıtların hep birlikte sağlanabildiği bir formül üretebilmek adına bazı deneysel çalışmalar yapılarak elde edilebilecek en iyi sonuç için hangi yöntemlerin nasıl modifikasyonlardan geçmesi gerektiği üzerinde tartışılmaktadır (Fang vd., 2004).

2.2 Konu Modelleme

Günümüzde kağıt ortamında yayınlanmış ve saklanan binlerce doküman taranarak sayısal ortama aktarılmıştır. Bu şekilde oluşmuş büyük doküman arşivleri çevrimiçi olarak ulaşılabilir olarak yayınlanmaktadır ancak bu verilerin otomatik olarak analiz edilerek kategorize edilmesi ve indekslenmesi gerekmektedir (Blei & Lafferty, 2007, s. 17-35). Büyük veri kümelerini içeren arşivlerin otomatik olarak indekslenmesi, kategorize edilmesi, aramaların yapılması için geliştirilen yeni metotlar istatistiksel modelleme için yeni imkanlar da sunmuştur. Makine öğrenme ve istatistik alanlarındaki yeni çalışmalar ile birlikte doküman kümeleri üzerinde kelime kalıplarının saptanması konusunda hiyerarşik sezgisel modellerin kullanılmasına yönelik yeni teknikler geliştirilmiştir (Blei & Lafferty, 2006, s. 113-120).Bununla birlikte makale arşivleri kategorize edilmiş olsa bile bazı kategorilerde çakışmalar ortaya çıkabilecektir ve her geçen gün artan metin veriler üzerinde bu şekilde aranan bilgiye ulaşılabilmesi çok mümkün olmamaktadır. Bunun sonucu olarak makine öğrenme üzerine araştırma yapan araştırmacılar sezgisel konu modelleme tekniğini geliştirmişlerdir. Konu modelleme algoritmaları orijinal metin içerisindeki kelimeleri analiz ederek ilişkili konuların

saptanmasını sağlarken konuların birbirleriyle olan ilişkilerini ve zaman içerisindeki değişimlerini de ortaya koymaktadır (Blei, 2012). Konu modelleme algoritmaları, insan gücüyle çok zor olacak analiz ve indeksleme işlemini herhangi bir insan kaynağı olmadan doğrudan metin veriler üzerinden yapabilmeye imkan vermektedir.

2.3 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation, bir çok konu çıkarma modelinin temelini oluşturan, metinsel verilere uygulanan ve yaygın olarak kullanılan bir konu modelleme (topic modelling) algoritmasıdır (Li & McCallum, 2006, s. 577-584). Konu modelleme yöntemleri, bir doküman arşivinde yer alan dokümanları incelerken hem bir dokümanda geçen kelimeleri, hem de farklı dokümanlarda geçen kelimelerin birlikte kullanıldığı kelimeleri inceleyerek her belgenin bir veya birden fazla konuya ait olabileceği sonucunu veren modeli çıkarır. Konu sayısı verildikten sonra LDA tarafından her belge için saptanan konuların olasılık dağılımı ile daha anlamlı sonuçlara ulaşılır. LDA algoritması ile dokümanlar, tüm arşivdeki metin verilerin bir özeti niteliğinde ve en belirleyici anahtar kelimeleri içeren konular ile ifade edilebilmektedir (Öztürk vd., 2014).

LDA algoritmasına göre tüm kelimeler bir konuyu belli oranda temsil etmektedirler. Bütün belgeler de belli oranda bu konuları içermektedir yani her bir belge birden fazla konunun karışımı olarak ifade edilebilmektedir. LDA,bir konunun bir belgede olma olasılığını hesaplayarak bütün arşiv üzerindeki konuların saptanmasını sağlamaktadır.

Daha ayrıntılı olarak, LDA belgelerin şu şekilde üretildiğini varsayar;[3]

- Belge içerisindeki kelime sayısı(N) belirlenir.
- Dirichlet dağılımındaki $k$tane konudan bir kaç tanesi seçilir.
- Belgeye bir konuya aitw_i kelimesi eklenir.
    - İlk olarak olasılık dağılımına göre bir konu seçilir.
    - Eklenecek kelime seçilen konununmultinomial dağılımına göre seçilir.

[3]http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/ (Erişim Tarihi: 02/05/2017)

Bu şekilde bir dizi belgenin elde edildiğinivarsayarak *k* tane konuyu ve bu konulara ait kelimeleri LDA algoritması ile saptamak için şu adımlar izlenir;

1. Her bir belge için belgedeki bir kelime rastgele bir konuya atanır.
2. Bu rastgele atama, hem tüm belgelerin hem de tüm konuların kelime dağılımlarını verir.
3. Her bir belge için atamalar iyileştirilir.
    a. *d* belgesindeki her bir *w* kelimesi ele alınır.
        i. Her bir konu için iki oran hesaplanır;
            1. Belgede o anda konuya atanmış sözcüklerin oranı
            2. *w* kelimesinden gelen konunun tüm belgeler içindeki oranı
        ii. *w* kelimesi olasılık dağılımı hesaplanan(oranlar çarpımı)yeni bir konuya atanır.
4. Üçüncü adım defalarca tekrarlandığında konu atamalarının istikrarlı olduğu bir duruma ulaşılır. Böylece bu atamalar ile her bir belgenin içerdiği konu olasılıkları ve her konuya ait kelimeler, tekrarlanma sayıları ile birlikte elde edilir.

Adım 3.a'da o anda ele alınan *w* kelimesi hariç diğer tüm kelimelerin doğru atandığı varsayılmaktadır. Sadece ele alınan kelime için güncelleme yapılmaktadır.

LDA algoritması tarafından belgeler üzerinde işletilen bu adımları örnekleyerek daha açık hale getirmek gerekirse şu cümlelere göz atalım[4];

- Balık ve sebze yerim.
- *Balıklar evcil hayvanlardır*
- *Kedim* balık yer.

Burada LDA, altı çizili kelimeleri Y konusu altında toplar. Y konusunu Yiyecek olarak etiketleyebiliriz. LDA, aynı şekilde italik kelimeleri de H konusu altında toplar. H konusunu da Hayvanlar olarak etiketleyebiliriz.

LDA'in kelime seviyesinde işlem yapması iki açıdan önemlidir.

1. Her bir cümlenin içeriği kelime sayısı ile çıkarılabilir.

    Birinci cümle: %100 Y konusu

    İkinci cümle: %100 H konusu

[4]https://algobeans.com/2015/06/21/laymans-explanation-of-topic-modeling-with-lda-2/ (Erişim Tarihi: 02/05/2017)

Üçüncü cümle: %33 H konusu ve %66 Y konusu

2. Her kelimenin ilgili konudaki oranı çıkarılabilir. Örneğin Y konusu %40 balık, %40 yemek, %20 sebze içermektedir.

LDA bu sonuca daha önce bahsedilen ve aşağıda örneklenecek olanüç adımda ulaşmaktadır. Bu adımların cümlelerde değilde belgeler üzerinde işletildiğini düşünelim.

1. Adım: Toplam kaç konu olduğu algoritmaya girdi olarak verilmesi gerekmektedir. Bunun için bir tahmin yapmak gerekir. Bu aşamada konu sayısını verilen örnek için cümleleri gözden geçirerek tespit edebiliyor olsak da çok sayıda belge için Perplexity hesaplanması ile konu sayısı tahmini yapılabilir.

2. Adım: Algoritma her bir kelimeyi geçici bir konuya atar. Bu atamalar 3. adımda güncellenmek üzere geçici olarak yapılmıştır. Bu aşamada bir kelime birden fazla defa geçiyorsa bunlar ayrı konulara atanmış olabilirler. Bu atamalar yapılırken bağlaç(ve, ile) gibi yardımcı kelimeler dikkate alınmazlar.

3. Adım: Algoritma her belgedeki her kelime için konu atamalarını kontrol eder ve yeniden düzenler. Bu yeniden atama işlemi için iki kriter gözönünde bulundurulur;

1. Konular içinde bu kelime ne kadar geçiyor
2. Belge içerisinde konular ne kadar geçiyor

Bu iki kriterin şu şekilde kontrol edilir;

Tablo 2.1*LDA için belge örnekleri*

	Belge 1		**Belge 2**
Y	Balık	?	Balık
Y	Balık	Y	Balık
Y	Yemek	Y	Süt
Y	Yemek	H	Kedi
Y	Sebze	H	Kedi

Tablo 2.1'de Balık kelimesi için birinci kriterin değerlendirildiğinidüşünürsek bu kelimenin %100 Y konusuna ait olduğu görülüyor. Rastgele seçilen bir balık kelimesi hep Y konusunda olacaktır. Balık kelimesi ikinci kriter için

değerlendirildiğinde ise Belge 2'de kelimelerin yarı yarıya iki konu arasında dağıldığı görülüyor. Bu iki kriter birlikte ele alındığında Belge 2'deki Balık kelimesi Y konusuna atanır. Bu şekilde tüm kelimeler için bu adım defalarca tekrarlanır. Bu şekilde nihai sonuca ulaşılır.

Bu çalışmada TOJDE dergisi arşivi üzerinde LDA algoritmasının saptayacağı konu olasılıklarının hesaplanmasında MAchine Learning for LanguagE Toolkit (MALLET) kütüphanesi kullanılmıştır. MALLET tüm dokümanları içeren bir dizini girdi olarak alır ve çıktı olarak konu anahtar sözcükleri ve bunların frekanslarını, olasılığı en yüksek anahtar sözcükleri ve her belgenin konular üzerinde dağılımını gösteren olasılıkları verir.[5]

2.4 Benzer Çalışmalar

Uzun (2011), Türkçe için Kavram Çıkarma sistemi Geliştirilmesi adlı tez çalışması sonucunda beklenenden fazla sayıda kavram üretilmiş olmasına rağmen dilin karmaşıklığı ve konuların birebir geçmeme ihtimalleri göz önününde bulundurulduğunda sonuç başarılı olarak kabul edilmiştir. Bu çalışmada konu modelleme için kümeleme yöntemi kullanılmıştır.

Hugo Liu ve Munindar P. Singh tarafından konu modelleme yapmak için sağduyu bilgi veritabanı olan ConceptNet[6] kullanılmıştır (Liu & Singh, 2004, s. 211-226).

Abdelattif Rahmoun ve arkadaşları tarafından WordNet[7] kullanılarak yani sözlük anlamları üzerinden eşler oluşturularak konu modelleme çalışması yapılmıştır (Elberrichi vd., 2008, s. 16-24).

Loulwah AlSumait ve arkadaşları tarafından LDA algoritmasının Matlab Topic Modeling Toolbox aracı kullanılarak uygulanması ile konu modelleme çalışması yapılmıştır (AlSumait vd., 2009, s. 67-82).

[5]http://mallet.cs.umass.edu/ (Erişim Tarihi: 05/11/2016)
[6]http://conceptnet.io(Erişim Tarihi: 05/11/2016)
[7]https://wordnet.princeton.edu/(Erişim Tarihi: 05/11/2016)

2.5 Yardımcı Uygulamalar ve Algoritmalar

Makale arşivinin analiz edilebilmesi için yapılan ön hazırlıklar ve sonrasında elde edilen verilerin görsel çıktılara dönüştürülmesi için kullanılan ek bileşenler ve algoritmalara bu başlık altında yer verilecektir.

2.5.1 Kök bulma (Stemming) algoritması

Bir kök bulma algoritması verilen bir kelimenin dilbilimsel olarak normalleştirilerek kökünün bulunmasını sağlayan bir algoritmadır. Orijinal "Stemming Algorithm" makalesi, büyük bir projenin bir parçası olarak 1979 yılında Cambridge'de yazılmıştır (Rijsbergen vd., 1980). Rijsbergen'in teşvikiyle de 1980 yılında M.F. Porter tarafından yayınlanmıştır (Porter, 1980, s. 130-137). Daha sonra 1997 yılında Karen Sparck Jones ve Peter Willet tarafından yeniden yazılmıştır (Jones & Willet, 1997).

Kök bulma algoritması, Veri Madenciliği uygulamalarında doğal dil işleme fonksiyonları kadar önemli olan bir ön işlemdir. Bir çok bilgi erişim sistemi için de çok önemli bir algoritmadır. Kök bulma algoritmasının hedefinin; çekim eki ve bazen yapım eki almış kelimeleri bu eklerden kurtararak üzerinde çalışılan kelime çeşitliliğini azaltmak olduğunu söyleyebiliriz (Jivani vd., 2011, s. 1930-1938).

Kök bulma algoritması uygulanırken dikkat edilmesi gereken hususlardan bazılarını şu şekilde sıralayabiliriz[8].

1. Konuşma dili için değil yazışma dili için uygun çıktılar üretecektir.
2. Her dil için kelime eklerinin saptanma kuralları değişeceğinden bir dil üzerinde çalışmak için o dile hakim olmak gerekmektedir.
3. Kök bulma algoritması örneğin Japonca, Çince gibi bir dil için uygulanabilir değildir.
4. Genel olarak 2 tip hata ile karşılaşılabilir;
    a. Farklı köke ait 2 kelimenin aşırı kırpılması ile aynı köke ulaşılması.
    b. Aynı köke ait 2 kelimenin yetersiz kırpılması ile farklı 2 köke ulaşılması.

Kök bulma algoritmasının bir kelime üzerinde uygulanma örneği Tablo 2.2'de görülmektedir.

[8]https://xapian.org/docs/stemming.html (Erişim Tarihi: 05/11/2016)

Tablo 2.2 *Kök bulma işlemi*

Kelime	Kök
Stemmer Stemming Stemmed Stems	Stem

2.5.2 MALLET

MALLET, doğal dil işleyen, belge sınıflandıran, konu algılama yapan ve diğer makine öğrenme araçlarını da içeren Java tabanlı bir uygulama ve Common Public License[9] ile korunan açık kaynaklı bir kütüphanedir. Belge sınıflandırma konusunda gelişmiş araçları bünyesinde barındırır. Etiketlenmemiş metin veriler içeren çok büyük çaplı arşivlerin analizi için konu modelleri çok kullanışlıdır ve MALLET konu modelleme aracı;Pachinko Allocation, Latent Dirichlet Allocation ve Hierarchical LDA algoritmalarını verimli ve örnek temelli olarak gerçekleştirir. MALLET içerisinde bulunan konu modelleme paketi, Gibbs Sampling[10]'i çok hızlı ve ölçeklenebilir bir biçimde gerçekleştirebilmektedir.[11]

2.5.3 Perplexity

Dil modellemesinde sıklıkla kullanılan perplexity, test verilerinin olasılığında monoton olarak azalır ve kelime olasılığı için geometrik ortalamanın tersine matematiksel olarak eşittir. Daha düşük hesaplanan bir perplexity değeri daha iyi genelleme ve tahmin performansı sergiler. Blei ve arkadaşları tarafından yapılan araştırmalar da sonuç olarak LDA algoritmasının çok daha düşük perplexity değerleri ortaya koyduğunu göstermektedir(Blei vd., 2003, s.993-1022). Matematiksel olarak *K* tane belge içeren bir veri kümesi için perplexity hesaplayan formul2.1'de gösterilmektedir.

$$perplexity(D_{test}) = exp\left\{-\frac{\sum_{d=1}^{K} \log p(w_d)}{\sum_{d=1}^{K} N_d}\right\} \quad (2.1)$$

[9]http://www.opensource.org/licenses/cpl1.0.php (Erişim Tarihi: 16/11/2016)
[10]http://www.mit.edu/~ilkery/papers/GibbsSampling.pdf (Erişim Tarihi: 16/11/2016)
11http://mallet.cs.umass.edu/ (Erişim Tarihi: 16/11/2016)

Formul 2.1'de; K test verileri içinde bulunan doküman sayısıdır. w_d, d dokümanındaki kelimeleri, N_d ise d dokümanındaki kelime sayısını belirtmektedir.

2.5.4 D3 (Data-Driven Documents)

D3.js verilere dayalı belgeleri işlemeyi sağlayan bir JavaScript kütüphanesidir. D3; HTML, SVG ve CSS kullanarak verileri görsel olarak çeşitli grafiklere yansıtır. Web standartlarına tam uyum sağlayarak ve modern tarayıcıların yeteneklerini tam olarak kullanarak veri odaklı yaklaşım ile güçlü görselleştirme bileşenleri sunar[12].

D3, verinin bir Doküman Nesne Modeli(DOM) ile ilişkilendirilmesini sağlayarak veri odaklı dönüşümü uygular. Örneğin; D3 kütüphanesini kullanarak bir sayı dizisinden bir HTML tablo bileşeni oluşturulabilir veya aynı veri kullanılarak kullanıcı etkileşimine sahip bir çizgi grafik oluşturulabilir.

D3 tek başına bir framework değildir. Verilerin verimli bir şekilde dönüşümlerinin sağlanması için çözüm sunar. Web standartları olan HTML, SVG ve CSS'in tüm yeteneklerini kullanarak esneklik sağlar. Bunlarla birlikte D3, büyük veri kümeleri üzerinde olağanüstü hızlı gösterimler ile kullanıcı etkileşimi ve animasyon desteği sunarak dinamik davranışlara imkan verir.[13]

[12]https://github.com/d3/d3/wiki (Erişim Tarihi: 05/11/2016)
[13]http://d3js.org (Erişim Tarihi: 05/11/2016)

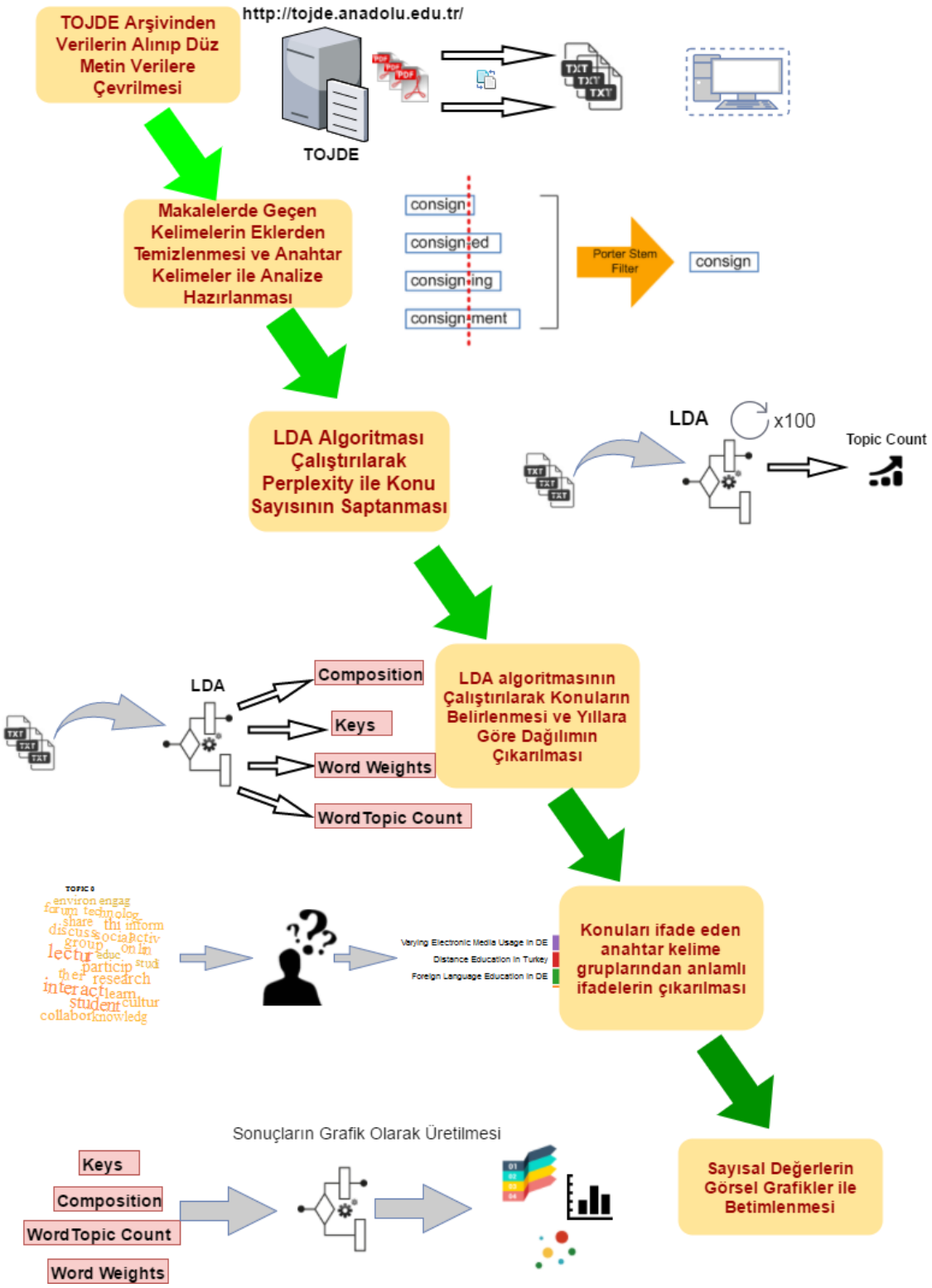
3 YÖNTEM

TOJDE dergisine ait makale arşivinde yer alan her bir makalenin metni alınarak verilerin anlamlandırılması ve benzerliklerin daha net ortaya konması adına kök bulma algoritması ile benzer anlamlı kelimelerin algoritma tarafından bir konu olarak algılanması sağlanmaya çalışılmıştır. Ayrıca makale yazarları tarafından belirtilmiş anahtar kelimeler parçalanmayacak şekilde organize edilerek bağlaç gibi kelimeler ile ortaya çıkan birden fazla kelime içeren benzersiz konuların da analize katılmaları sağlanmaya çalışılmıştır.

Yazılım[14] tarafından, hazırlanan veriler üzerinde LDA algoritmasını uygulayan MALLET uygulaması yardımıyla perplexity hesaplanmış ve üzerinde çalışılan makaleler için uygun konu sayısı saptanmıştır. Belirlenen konu sayısı üzerinden yine LDA algoritması ile analizler yapılarak, çeşitli yıllarda farklı sayılarda yayınlanmış makalelerin hangi yıllarda hangi konulardan bahsedildiğini D3.js JavaScript kütüphanesi yardımıyla görsel olarak ortaya koyarak sonuçlandırılmıştır. Bu sonuç geçerliliğini yitirmiş çalışma alanları ile birlikte akademik çevrelerce sıklıkla bahsedilmeye başlanılan konuların da rahatlıkla saptanabilmesini sağlamıştır.

Uygulanan adımlar ve detaylarıŞekil 3.1'de görsel olarak ortaya konulmuştur. Her adımın kendi içindeki mini adımlara yer verilerek;ilgili adım için girdi olan veriler ile o adımın işletilmesi sonucunda oluşan çıktılar belirtilmiştir.

[14]https://github.com/ykartal/topic-modelling, (Erişim Tarihi: 05/11/2016)



Şekil 3.1*Sistemin işleyişi*

İlk adım olarak; Java programlama dili kullanılarak makalelerin her biriTOJDE internet adresinden PDFbiçiminde indirilmiş ardından düz metin verilere çevrilerek ve yıl, ay, makale adı bilgileri ile birlikte yerel diske kaydedilmiştir. Bu çalışma

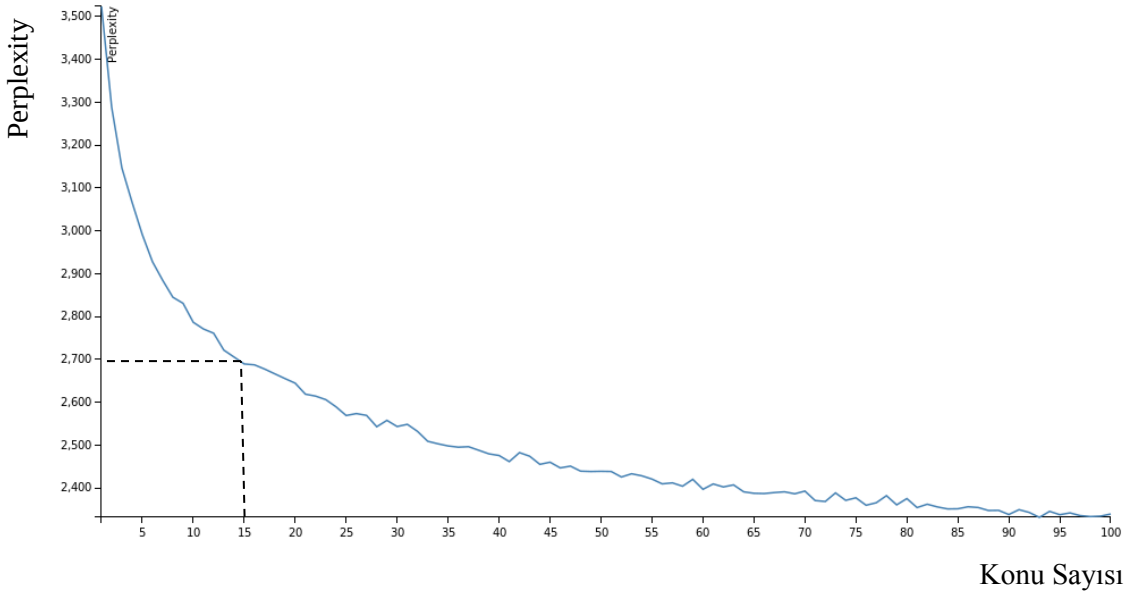
kapsamında önemli görülen ve kullanılacak olan yıl bilgisidir. Bu aşamada her bir makale için yazarları tarafından belirtilen anahtar kelimeler de toplanarak arşiv genelinde anahtar kelime listesi oluşturulmuştur. Bu şekilde birden fazla kelime birlikte kullanıldığında anlam ifade ediyorsa anahtar kelimeler yardımıyla daha iyi sonuçlar üretilmesi amaçlanmıştır. Geliştirilen yazılım[15] tarafından yerel diske kaydedilen makalelerin, birden fazla kelime içeren anahtar kelimelerdeki kelimeler arası boşlukların, '_' karakteri ile değiştirilmesi sağlanmıştır. Böylece tek başına anlam ifade etmeyen ve anlamlı olmayan konu başlıklarının LDA algoritması tarafından tespit edilmesine sebep olacak kelimeler en aza indirgenmeye çalışılmıştır. Örneğin; Distance Education tüm makalelerde birlikte kullanıldığında anlam ifade ederken analize tek tek girmeleri Education kelimesini bir konu olarak belirlerken Distance kelimesinin önemsiz kalmasına sebep olabilecektir. Tüm makale metinleri içerisinde Distance Education ibarelerinin Distance_Education olarak düzeltilmesi daha anlamlı sonuçların elde edilmesini sağlayacaktır.

İkinci adım olarak; Martin Porter tarafından Java programlama dili ile geliştirilmiş Stemmer sınıfı[16],geliştirilen yazılıma entegre edilerek tüm makale metin içerikleri ve anahtar kelimelerin eklerinden temizlenmiş kopyaları oluşturulmuştur. Böylece makale metinleri içerisinde geçen Educated, Educates, Education gibi kelimelerin ayrı ayrı konular olarak saptanması yerine Educate şeklinde ortak kelime kökünün analiz sonucunda daha doğru sonuçlar vermesi amaçlanmıştır.

*Üçüncü adım olarak;*LDA algoritması analize girdi olarak verilen metin dosyaları üzerinde işletilirken kaç konu başlığı arandığı bilgisini de girdi olarak almaktadır. Perplexity hesaplamak için bir konu sayısı belirleyip metin dosyaları ile birlikte LDA algoritmasına parametre vermek gerekmektedir. LDA algoritmasının işletilmesinin ardından üretilen çıktılar ile perplexity hesabı yapılır ve bu değer kayıt altına alınır. Bu çalışma kapsamında konu sayısı 1'den 100'e kadar olacak şekilde her konu sayısı için perplexity değeri hesaplanmıştır. Perplexity hesaplama süresi belirtilen konu sayısı ve üzerinde çalıştırılan bilgisayarın hızına göre değişiklik göstermekle birlikte burada 1 konu sayısından 100 konu sayısına kadar her bir konu sayısı için hesaplama yapılmış ve tamamlanma süresi toplam 22 saat sürmüştür. Hesaplamalar

[15]https://github.com/ykartal/topic-modelling, (Erişim Tarihi: 05/11/2016)
[16]https://tartarus.org/martin/PorterStemmer/java.txt, (Erişim Tarihi: 05/11/2016)

sonucunda çıktılar değerlendirilirkenŞekil 3.2'de görülen perplexity değerinin değişkenliğini yitirmeye başladığı 15 konu sayısı, üzerinde çalışılan makaleler için uygun konu sayısı olarak kabul edilmiştir.



Şekil 3.2*Her Konu sayısı için bulunan perplexity değerleri grafiği*

Son adım olarak; belirlenen konu sayısı ile birlikte veri analizi için hazırlanan metin girdiler işleme alınarak çıktı olarak istatiksel sonuçları içeren veri dosyalarının oluşturulması sağlanmıştır. LDA algoritmasının işletilmesinin ardından üretilen çıktıları şu şekilde detaylandırabiliriz;

- ✓ Keys: Her bir konu için saptanan anahtar kelimeleri içeren çıktı dosyasıdır.
- ✓ Composition: Her bir makalenin tespit edilen konuların her biri ile ne oranda ilgili olduğu bilgisini sayısal olarak ifade eden çıktıları içeren dosyadır.
- ✓ Word Topic Count: Konulara ait anahtar kelimelerin her birinin hangi konu için kaç adet tekrarlandığı bilgisini içeren dosyadır.
- ✓ Word Weights: Her konuya ait anahtar kelimelerin o konu için ağırlıklarının belirtildiği çıktı dosyasıdır.

LDA algoritmasının işletilmesinin ardından metin halinde oluşturulan çıktılar üzerinde görsel olarak ifade edilebilmesi amacıyla hesaplar yapılarak grafiklere yansıtılabilecek sayısal veriler üretilmiştir.Yıllara göre değişen verilerin hesaplanması için; ilgili makalenin ait olduğu yıl bilgileri kullanılarak bu makale için LDA ile hesaplanan konuyu barındırma olasılığı yıllar bazında gruplanarak ve bu olasılık değerleri toplanarak ilgili yılda ilgili konudan bahsedilme olasılığı hesaplanmıştır.Tablo

3.1'de sadeleştirilmiş olarak örneklenmiş çıktı kesiti göz önünde bulundurulduğunda 2000 yılı için 1. konunun 1.62, 0.36, 0.06 değerlerinin toplamı kadar yüzde oranına sahip olduğu bilgisi ortaya çıkmaktadır.Her bir konu için o sütundaki tüm değerler toplamı 100 değerini vermektedir.

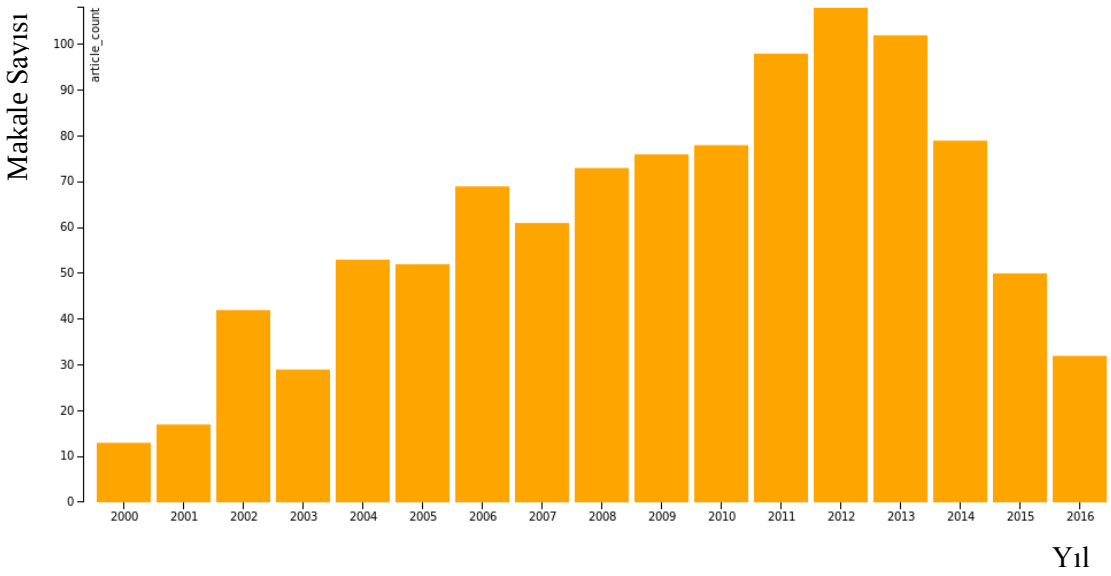
Tablo 3.1 *LDA algoritması sonucu composition çıktısı örnek kesit*

Makale	Konu 1	Konu 2	Konu 3	Konu 4
'2000-1-DISTANCE_****__THE_RO.txt	1.6299918500407498E-4	0.009942950285248574	0.015321923390383048	6.519967400162999E-4
'2000-1-Driving_****ly_Cns.txt	0.3614070691696263	0.05445628276678505	0.08895653644512093	1.6911889058007779E-4
'2000-1-The_Scho*lo__1989_1.txt	0.06361887853232727	0.053410267791093355	0.17058736499482172	0.054741825713863
'2001-2-Desc*******cating_C.txt	0.08744082428292956	0.05987190197716514	1.392369813422445E-4	0.15636313004734056
'2001-2-Succ***port_Project.txt	0.01554907677356657	0.1205053449951409	0.6098847702346244	0.0055532417048452035
'2001-2-Telev*elevion_at_Op.txt	0.0015642108556233381	0.0010949475989363367	0.007195369935867355	0.0015642108556233381
'2002-3-A_C******u_Universi.txt	7.473841554559043E-4	0.031016442451420O3	0.018124065769805683	0.009155455904334827

Tüm konular için hesaplanan değerler uygun çıktılar halinde diske kaydedilmiş ve böylece kelime bulutu, çizgi, sütun ve yığın grafikler için veriler hazır hale getirilmiştir. Yazılım en nihayetinde her bir veri dosyası ile ilgili grafiği üreten HTML/JavaScript kodlarını entegre ederek grafikler oluşturmuş ve nihai çıktılara ulaşılması sağlanmıştır.

4 BULGULAR

Her yıl için TOJDE makale arşivlerinde yer alan makale sayısının yıllara göre dağılımını gösteren grafikŞekil 4.1'de görülmektedir.Makale arşivinde yıllar içerisinde yayınlanan makale sayısının genel olarak artış gösterdiği ve en fazla makalenin 2012 yılında yayınlandığı görülmektedir.



Şekil 4.1 *Yıllara göre makale sayısı*

Şekil 4.2'de konulara gösterilen ilginin yıllara göre dağılımını gösteren çizgi grafik yer almaktadır. Bu grafikteki iniş çıkışlar o konu hakkında yıldan yıla değişen popülariteyi göstermektedir.

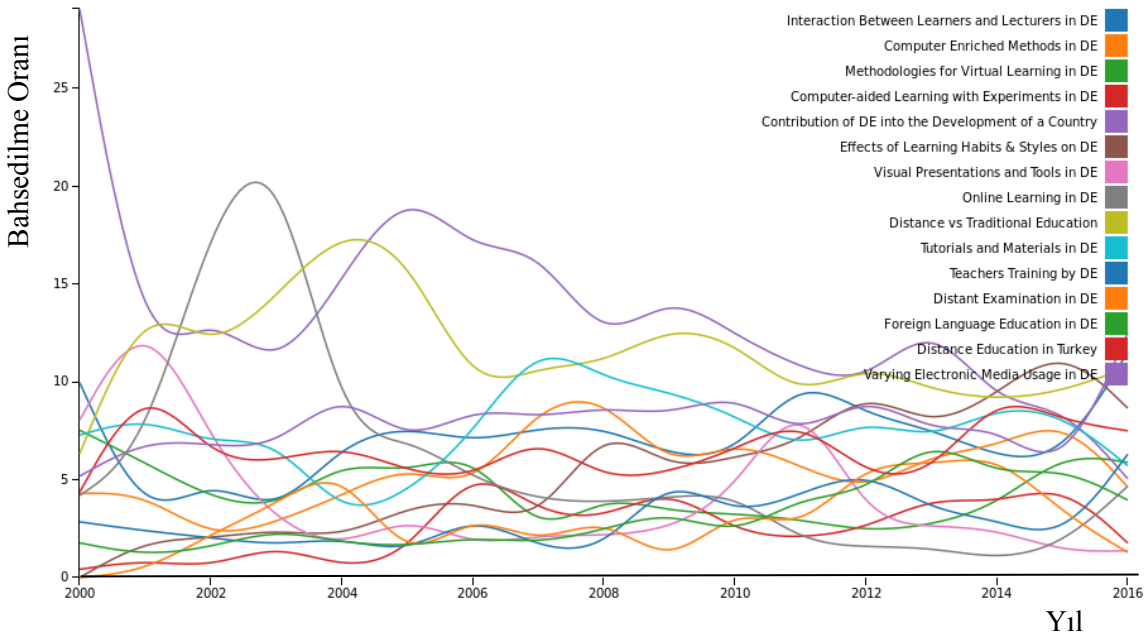
Şekil 4.2 *Yıllara göre konuların dağılımı*

Her konu için anahtar kelimelerin oranlarının görsel olarak ifade edilmesini sağlayan kelime bulutu grafikleri Şekil 4.3'te belirtildiği gibi görülmektedir. Üretilen bu kelime bulutları, algılanan konuların anlamlandırılmasında kolaylık sağlamaktadır.

Şekil 4.3 *Tespit edilen konulara ait kelime bulutları*

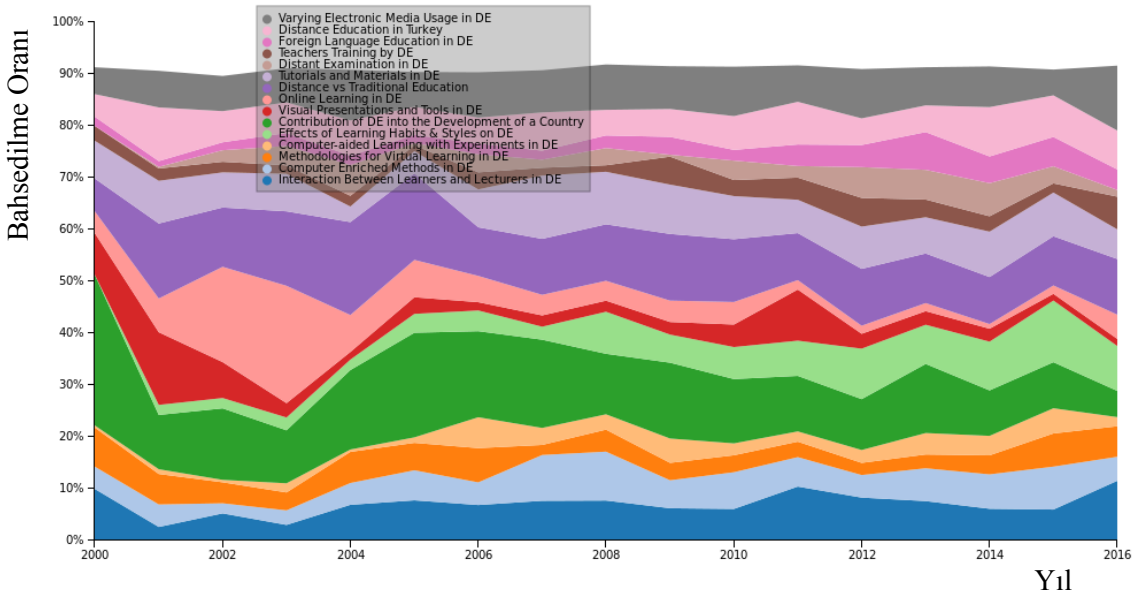
Öncelikle elde edilen ve konuları ifade eden anahtar kelime gruplarından anlamlı ifadeler elde edilmeye çalışılmıştır. Bu amaçla kelime grupları içerisinde en fazla tekrarlanan anahtar kelimelerden yardım alınarak konu tahminleri yapılmıştır. Dağılımın homojen olduğu veya anahtar kelimelerden tam olarak anlamlı ifadelerin çıkarılamadığı konular için geliştirilen yazılımın ürettiği sonuçlar yardımıyla ilgili konunun sıklıkla geçtiği makaleler incelenerek tahminler yapılmıştır. Yani üretilen grafiklerin anlamlı veriler ile bütünleştirilmesi için LDA algoritması sonucunda üretilen anahtar kelime bulutları ve bu anahtar kelimeleri içeren makaleler incelenerek anlamlı konu başlıkları haline getirilmiştir. LDA algoritmasının uygulanması sonucunda üretilen kelime öbeklerine karşılık gelen konu başlıkları grafik üzerine yerleştirildikten sonra grafik

Şekil 4.4'teki halini almıştır.



Şekil 4.4 *Yıllara göre konuların dağılımı (Anlamlandırılmış konular)*

Şekil 4.5'te konulara gösterilen ilginin yıllara dağılımını gösteren yığılmış[17] grafik yer almaktadır. Yığılmış grafikler ile konular arasında daha kolay karşılaştırma yapılabilinmektedir.Şekil 4.5'te yer alan grafik incelendiğinde bazı konuların tüm yıllarda popülaritesini koruduğu görülmektedir. Bu konulara*Varying Electronic Media Usage in Distance Education*ve *Distance Education in Turkey*örnek olarak verilebilir. Bununla birlikte bazı konulardan 2000 yılında hiç bahsedilmiyorken ilerleyen yıllarda artan bir ilgi ile karşılaştığı görülmektedir. Bu tip konulara ise *Foreign Language Education in Distance Education*, *Distant Examination in Distance Education*, *Effects of LearningHabbits & Styles in Distance Education* ve *Computer-aided Learning with Experiments in Distance Education* örnek verilebilir.Bu durumun tersi olarak *Visual Presentations and Tools in Distance Education* ve *Contribution of Distance Education into the Development of a Country* konularında ise daha önceki yıllara göre daha az araştırma yayınlandığı görülmektedir.*Online Learning in Distance Education* gibi bazı konuların popülaritesinin dalgalı bir şekilde yükselip düştüğü de tespit edilen bulgular arasında gösterilebilir.



Şekil 4.5 *Yıllara göre konuların dağılımı yığılmış grafik*

[17] Yığılmış grafikler aynı X değerine ait birden fazla başlığın Y değerine göre üstüste yığılmış halde görüntülenmesini sağlar. Bir başlık için Y değeri yükseldikçe grafikte o başlık daha kalın bir aralık kaplar.

5 SONUÇ VE ÖNERİLER

Bu araştırma kapsamında LDA algoritmasının kullanılarak konu modellemenin başarılması ile birlikte daha iyi ve net sonuçlar üretilebilmesi için yapılabilecek sadeleştirme çalışmaları üzerinde durduk. Martin Porter'ın geliştirmiş olduğu kök bulma algoritması ve metin arşivinde her bir makale için ilgili makaleyi yazan araştırmacılar tarafından belirlenmiş birden fazla kelime içeren anahtarların tek kelimeymiş gibi davranmasının sağlanması ile benzer konuların aynı çatı altında toplanmasını sağladık. Sonuç olarak; bir dizi belgeyi girdi olarak alan ve LDA konu modelleme algoritmasınıkullanarak üretilen ara çıktılar üzerinde yapılan işlemler ile konuların yıl bilgisi ile birlikte çıkarılmasını ve görsel grafikler ile betimlenmesini sağlayan bir yazılım geliştirdik. Herhangi bir metin arşivi üzerinde, alan bilgisine sahip olmadan, ilgili yıllardaki makalelere ait olasılık değerlerinin kullanılması ile hangi konuların yıllar içerisinde popülerliğinin artıp hangi konuların popülerliğini yitirdiği bilgisine ulaşılabileceğini ve geliştirilen ek yöntemler ile sonuçların daha net olarak çıkarılabileceğini gözlemledik.

Geliştirilen yazılım ile TOJDE dergisinde yer alan binden fazla makalenin içeriklerininsadeleştirilmesinin ardından LDA algoritması ile analiz edilmesi ve elde edilen analizler üzerinde yapılan hesaplamalarsonucunda makale arşivinde araştırılan konuların kapsamı hakkında bilgi sahibi olurken araştırmacıların yıllara göre hangi konular üzerinde araştırma yapmaktan vazgeçerken hangi konuların daha çok ilgi gördüğü ile ilgili bilgiye ulaştık. Bir araştırmacının, yıllara göre hangi konular üzerinde araştırma yapılmaktan vazgeçilirken hangi konuların daha çok ilgi gördüğü ile ilgili bilgiye erişmesine imkan sağlamış olduk.Bununla birlikte tespit edilen konulara ait kelime bulutunun çıkarılması sonucundabir konu ile ilişkili alt konular hakkında da bir bilgi üretmiş olduk. Örneğin; yapılan çalışma sonucunda TOJDE dergisi makale arşivinde yer alan hiç bir makale hakkında bilgi sahibi olmamamıza rağmen TOJDE dergisinde yer alan makalelerin genel olarak uzaktan eğitim üzerinde durduğunu ayrıca uzaktan eğitimin alt kırılımları olarak Türkiye'de uzaktan eğitim, ölçme değerlendirme ve kullanılan teknolojiler konularının bu alandaki ana temaları oluşturduğunuve zamanla hangi konuya ilginin artıp azaldığını tespit ederken görsel sunum ve araçların ilk başlarda üzerinde konuşulan bir konu olmasına rağmen yıllar içerisinde azalarak ilginin bu konuda gitgide kaybolduğunu gözlemledik.

Geliştirilen sistemin, farklı makale arşivleri ya da dergiler gibi metin veriler içeren arşivler üzerinde de uygulanabilecek bir kaynak olduğu görülmektedir. TOJDE dergisi gibi İngilizce arşivler üzerinde başarılı sonuç üreten çalışmanın ilerleyen dönemlerde dil yetenekleri geliştirilerek Türkçe arşivler için de çalışmalar yapılması planlanmaktadır. Bununla birlikte örnek olarak sağlıksektöründe geçmiş verilerin analizi ile karar destek mekanizmalarına yardımcı olabilecek çıktılar üretecek sistemler ortaya konulabilecektir.

KAYNAKÇA

AlSumait, L. Barbará, D. Gentle, J. and Domeniconi, C. (2009). Topic Significance Ranking of LDA Generative Models, Machine Learning and Knowledge Discovery in Databases, s. 67-82

Arguello, J. Díaz, F. and Callan, J. (2009). Sources of Evidence for Vertical Selection, SIGIR'09, s. 315-322

Baeza-Yates, R. and Ribeiro-Neto, B. (2012). Modern Information Retrieval, Addison Wesley Longman Publishing Co. Inc.,

Berger, A. and Lafferty, J.(1999). Information Retrieval as Statistical Translation, SIGIR'99, s. 222-229

Blei, D. Carin, L. Dunson, D. (2010). Probabilistic Topic Models, IEEE Signal Processing Magazine, vol. 27, no. 6, s. 55-65,

Blei, D. Ng, A. and Jordan, M. (2003). Latent Dirichlet Allocation, Journal of Machine Learning Research, s. 993-1022

Blei, D. and Lafferty, J. (2006). Dynamic Topic Models, International Conference on Machine Learning'06, s. 113-120

Blei, D. and Lafferty, J. (2007). A Correlated Topic Model of Science, The Annals of Applied Statistics(1), s. 17-35

Chakraborty, G. Pagolu, M. and Garla, S. (2013). Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS, SAS Institute

Elberrichi, Z. Rahmoun, A. and Bentaallah, M. A. (2008). Using WordNet for Text Categorization, The International Arab Journal of Information Technology, s. 16-24

Fang, H. Tao, T. and Xiang, C. (2004). A Formal Study of Information Retrieval Heuristics, SIGIR'04, s. 49-56

Jivani, A. G. Shingala, A. and Virparia, P. (2011), The Multi-Liaison Algorithm, International Journal of Advanced Computer Science and Applications, Vol 2 (6), s. 1930-1938

Jones, K. S. and Willet, P. (1997). Readings in Information Retrieval, San Francisco: Morgan Kaufmann, ISBN 1-55860-454-4.

Li, W. and McCallum, A. (2006). Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations, ICML '06 Proceedings of the 23rd international conference on Machine learning, s. 577-584

Liu, H. and Singh, P. (2004). ConceptNet - A Practical Commonsense Reasoning Tool-Kit, BT Technology Journal, s. 211-226

Öztürk, S., Sankur, B., Güngör, T. and Yılmaz, M. B. (2014), Turkish Labeled Text Corpus, 2014 22nd Signal Processing and Communications Applications Conference (SIU), Trabzon, 2014, s. 1395-1398

Pickensy, J. Golovchinskyy, G. Shahz, C. Qvarfordty, P. and Backy, M. (2008). Algorithmic Mediation for Collaborative Exploratory Search, SIGIR'08, s. 315-322

Pirkola, A. Hedlund, T. Keskustalo, H. andJärvelin, K.(2001). Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings, Information Retrieval, s. 209-230

Porter, M. F. (1980). An algorithm for suffix stripping, Program, 14(3) s. 130−137

Rijsbergen, C. J. Robertson, S. E. and Porter, M. F. (1980). New models in probabilistic information retrieval. London: British Library. (British Library Research and Development Report, no. 5587)

Rosell, M. (2006). Introduction to Information Retrieval and Text Clustering, KTH CSC

Sanderson, M. and Croft, W. B. (2012). The History of Information Retrieval Research, Proceedings of the IEEE, s. 1444-1451

Srivastava, A. N. and Sahami, M.(2009). Text Mining: Classification, Clustering, and Applications, CRC Press, s. 71-72

Uzun, M. (2011). Türkçe için Kavram Çıkarma Sistemi Geliştirilmesi, İstanbul

White, R. W. Bilenko, M. Cucerzan, S. (2007). Studying the Use of Popular Destinations to Enhance Web Search Interaction, SIGIR'07, s. 159-166

Yom-Tov, E. Fine, S. Carmel, D. and Darlow, A. (2005). Learning to Estimate Query Difficulty, SIGIR'05, s. 512-519

Yuasa, M. Nishida, T. Ohyama, M. and Sera, T. (2011). Information retrieval system using location and transportation by GPS traces as search criteria, TENCON 2011 - 2011 IEEE Region 10 Conference, Bali, 2011, s. 221-225