GİZLİ ANLAMBİLİMSEL DİZİNLEME YÖNTEMİNİN N-GRAM KELİMELERLE GELİŞTİRİLEREK, İLERİ DÜZEY DOKÜMAN KÜMELEMESİNDE KULLANIMI

Ahmet GÜVEN[1] , Ö. Özgür BOZKURT[2] ve Oya KALIPSIZ[3]

Bilgisayar Mühendisliği Bölümü, Yıldız Teknik Üniversitesi, Yıldız, İstanbul.
E-posta: [1]guven@ges.net.tr, [2]ozgurb@yildiz.edu.tr , [3]kalipsiz@yildiz.edu.tr

Anahtar sözcükler: *Demetleme, metin madenciliği, bilgiye erişim, kümeleme, yarı yapısal doküman madenciliği, Gizli Anlambilimsel Dizinleme*

ABSTRACT

Latent Semantic Indexing (LSI) is an automatic method that transforms the original textual data to a smaller semantic space and is one of the widely used feature extraction method for document clustering. LSI considers the words and their frequencies while acquiring the semantics. However, most of the time, phrases which are formed by more than a single word are required for determining the meaning. In this case, LSI may lead to unsuccessful clustering. We propose to enhance LSI, by applying the word based n-gram technique in 2-gram and 3-gram forms, to achieve better document clustering. Implementation results, demonstrating the usefulness of the proposed method, and comparison to original LSI are given.

1. GİRİŞ

Internet ve kişisel bilgisayarların yaygınlaşmasına bağlı olarak, gittikçe büyüyen hacme sahip doküman yığınları oluşmaktadır. Bu yığınlar içinde önemli bilgiler kaybolup giderken, değerli bilgilere ulaşmak için dokümanların içeriğinin belirlenmesi ve buna uygun sorgulanabilmesi ihtiyacı kendini hissettirmektedir [3,4].

Geleneksel bilgiye ulaşma yöntemleri, doküman yığınları içinden ana konu başlıklarına yönelik aramaları başarılı bir şekilde karşılayabilmektedir. Örneğin Google arama motoru, en başarılı araçlardan biri olarak, "müşteri ilişkileri yönetimi" gibi bir ana konuyla ilişkili dokümanları sorunsuz bir şekilde, sorgu sonucu olarak getirmektedir. Ancak, arama yapılan doküman sayısı artıkça, sorgu neticesinde gelen doküman sayısı binlerle ifade edilir hale gelebilmekte, bunun sonucu olarak da daha özel ihtiyaçları karşılayabilecek çözümlere ihtiyaç duyulmaktadır. Örneğin "müşteri ilişkileri yönetiminin işletme verimliliğine etkisi" gibi bir konunun sorgulanması ihtiyacı doğabilmektedir.

Google gibi sistemler, dokümanları, içindeki kelimelerle; dokümanların arasındaki ilişkileri de bu kelimelerin tüm doküman yığını içindeki istatistiki kullanım örüntüleri ile ifade etmektedirler [2]. Kelimeleri baz alan bu sistemler otomatik indeksleme yapabildiklerinden, dokümanların madenciliği açısından büyük kolaylıklar sağlamaktadır. Ancak bu sistemler belirli kısıtlara da sahip olmuşlardır.

İlk kısıt yazarların dokümanları oluştururken kullandığı kelimelerle, bu dokümanlar içinde arama yapmak isteyen kişilerin sorgu ifadelerinde kullandıkları kelimelerin birbirinden farklı olabilmesidir. Bunun nedeni kelimelerin çokanlamlı ve/veya eşanlamlı olabilmesidir [5]. Benzer şekilde farklı kültürlerde veya farklı disiplinlerde aynı kelimelerin farklı kavramları ifade edecek şekilde kullanılması durumu da bu kısıtın başka bir boyutudur. Bu nedenle, kelimeler her zaman aynı anlamı taşımayabilir. *Gizli Anlambilimsel Dizinleme (GAD)* yöntemi bu kısıtı ortadan kaldıran bir çözümdür.

İkinci kısıt kelimelerin tek başlarına taşıdığı anlamın dışında yan yana kullanımları ile bambaşka anlamlar ifade edebilmesi durumudur. Bu çalışma kapsamında geliştirilen *n-gram kelimelerle GAD* yöntemi ile, bu kısıt ortadan kaldırılarak, daha detay konularda arama yapmayı sağlamak mümkün hale getirilmiştir.

Üçüncü kısıt mevcut sistemlerin, kelimelerin anlama kattığı değeri hesaplamak için sadece ne sıklıkla kullanıldıklarına bakmasıdır, oysa kelimelerin dokümanların anlamını ifade etmede diğerlerinden daha etkili olduğu farklı durumlar da vardır. Örneğin Türkçe cümlelerde fiile yakın kelimeler anlam açısından daha önemlidir.

n-gram kelimelerle GAD yönteminin kullandığı kelime grubu oluşturma yaklaşımı ile bu kısıta dair de bir çözüm oluşturulmuştur.

Bu çalışmanın takip eden bölümünde *n-gram kelimelerle GAD* yöntemi açıklanmış, 3. bölümde önerilen yöntemin uygulaması anlatılmış, 4. bölümde yapılan uygulamanın detayları üzerinde durulmuş; oluşturulan çözüm, bu çalışma için derlenmiş bir Türkçe doküman seti ve reuters21578 İngilizce doküman seti üzerinde test edilerek bulgular 5. bölümde tartışılmıştır. 6. bölümde elde edilen sonuçların önemi ve gelecek çalışma konuları üzerinde durulmuştur.

2. N-GRAM KELİMELERLE GAD

GAD yöntemi dokümanları içerdikleri kelimelere göre değerlendirirken, aynı zamanda doküman kümesini de bir bütün olarak ele almaktadır. Bu şekilde, bir dokümanda yer alan kelimelerin başka hangi dokümanlarda yer aldığını ve diğer dokümanlarda bu kelimelerle birlikte kullanılan başka ortak kelimeleri de göz önünde bulundurur. Vektör uzay modelini kullandığından, ortak kelimeler içeren dokümanların mantıksal olarak benzeştiğini, ortak kelimeler içermeyen dokümanların konu olarak farklı olduğunu ortaya koyabilmektedir [2] Bu yöntem, insanların bir doküman yığınını incelerken dokümanlar ve aralarındaki ilişkiyi anlamak için kullandıkları yönteme çok yakındır; tek fark GAD algoritmasının bunu kelimelerin anlamını bilmeden yapmasıdır [8, 10, 11]

GAD yöntemi ile indekslenmiş bir veritabanında yapılan aramada, içeriği ifade etmek için seçilmiş tüm kelimelerin benzerlik değerlerine bakılır ve aranan anahtar kelime veya kelimelerin en yüksek benzerlik değerine sahip olduğu dokümanlar arama sonucu olarak döner. Herhangi iki doküman, ortak kelimelere sahip olmasalar da anlamsal olarak aynı olabileceğinden, GAD aranan anahtar kelimelerin indekslenmiş dokümanlarda birebir bulunup bulunmadığına bakmaz. Bu sayede daha gerçekçi bir arama işlemi yapılmış olur [5, 7].

Örneğin, matematik konusunda yazılmış dokümanlardan oluşan bir doküman kümesi GAD yöntemi ile indekslenmiş olsun. Bu kümeyi oluşturan dokümanlarda "matris", "lineer cebir", "doğrusal cebir" kelimeleri yeteri kadar dokümanda yer alıyorsa, bu kelimelerin anlamsal olarak birbirlerine yakın olduğu sonucuna ulaşılır. Matris anahtar kelimesini içeren dokümanları bulmak için başlatılan bir arama işlemi sonucunda, matris kelimesini içermeyen, ama, lineer cebir ve/veya doğrusal cebir kelimelerini içeren dokümanlar da yanıt olarak döner. Görüldüğü gibi arama mekanizması matematik konusunu bilmediği halde yeterli sayıda dokümanı inceleyerek matematik konusunda kullanılabilen kelimeleri öğrenerek arama işlemini buna göre yapmaktadır.

Yukarıdaki örnekten de görüldüğü gibi GAD yöntemi, aranmakta olan ifadelerin anlamlarını bilmeden anlamlarına göre aramayı sağlayacak bir yaklaşım sunmaktadır. Bu sayede, konu ve dilden bağımsız olarak dokümanlar indekslenebilir ve geleneksel arama yöntemlerinin ötesinde bir fayda sağlayacak şekilde dokümanlara erişim sağlanabilir.

N-gram metodu

N-gram yöntemi, bir metnin hangi dilde yazıldığını bilgisayar tarafından belirleyebilmek amacıyla kullanılır. Bunu kelimeleri oluşturan harflerin yan yana gelme örüntülerine bakarak yapar. Örneğin bilgisayar kelimesinin n-gramları:

2-gram

b bi il lg gi is sa ay ya ar r

3-gram

b bil ilg lgi gis isa say aya yar r

her dilin kelimelerinin 2-gram, 3-gram gibi n-gram kalıpları farklıdır. Bu yaklaşımla bir metnin hangi dille yazıldığı belirlenebilir. [9]

GAD ile N-gramın birleştirilmesi

Kelimelerin cümle içinde fiile yakınlığına bağlı olarak, anlama kattığı değer farklılaşmaktadır. Örneğin "Ahmet bugün camı kırdı" cümlesinde vurgu cam üzerinedir. "Ahmet camı bugün kırdı" dersek vurgu zaman üzerinde odaklanmaktadır. Aynı cümleyi "Bugün camı Ahmet kırdı" şeklinde kurduğumuzda ise anlam Ahmet üzerine yoğunlaşır. Görüldüğü gibi dört kelimeden oluşan basit bir cümlenin kelimelerinin yerlerini değiştirdiğimizde, vurgulanmak istenen farklılaşabilmektedir.

Diğer bir husus da kelime gruplarının, kendilerini oluşturan kelimelerden daha farklı bir anlam içermesinde ortaya çıkar. Mesela "kurt adam öldü" cümlesindeki "kurt adam" ile "kurt adam öldürdü" cümlesindeki kurt ve adam kelimeleri tamamen farklı anlamlarda kullanılmıştır.

N-gram metodunun kelimelere uygulanması ile GAD yönteminin bu eksikliklerini gidermek mümkündür. Bu sayede daha güçlü bir bilgiye ulaşma yöntemi ortaya çıkarılarak, giderek büyüyen doküman yığınlarından amaca uygun bilgi çıkarma sonucuna ulaşılmıştır.

3. UYGULAMA

1. Doküman havuzunun oluşturulması : Türkçe doküman seti olarak son beş yıla ait iş dünyası ve yöneticilikle ilgili değişik dergilerden 615 adet makale derlemiştir. Bu makaleler 13 kategoride toplanmıştır. İngilizce doküman seti olarak, doküman madenciliği çalışmalarında

uluslararası standart kabul edilmiş Reuters21578 doküman seti [14] kullanılmıştır.

2. Türkçe kelimelerin anlam kaybetmeyecekleri en yalın hallerine çevrilmeleri için Zemberek projesi [13] kapsamında geliştirilen sözlük ile Türkçe kelimelerin çözümlenmesi için modül oluşturulmuştur. İngilizce dokümanlar için standart PorterStemmer [15] algoritması kullanılmıştır.

3. Türkçe için, dokümanlara anlam katmayan kelimelerin tespiti yapılmıştır. İngilizce için stoplist adı altında var olan standart liste kullanılmıştır.

4. Her kelimenin, içinde geçtiği doküman için önem derecesini belirten matris değeri, bilgiye erişme yöntemlerinde sıkça kullanılan tfidf (term frequency – inverse document frequency) yöntemi yapılmıştır [17]. Elde edilen matrise Tekil Değer Ayrıştırması (TDA) uygulanmıştır [2]. TDA işlemi sonucu elde edilen daha düşük boyutlu matrisler üzerinde hiyerarşik kümeleme algoritması ile kümeleme yapılmıştır [16].

4.KÜMELEME ve DEĞERLENDİRME

Kümeleme sonuçlarının başarımının değerlendirilmesi için entropi ve f-measure ölçümleri kullanılır [17].

Her iki ölçümün de yapılabilmesi için kümelenecek dokümanların belirli sınıflara önceden atanmış olması gereklidir. Kümeleme işleminin başarısı elde edilen kümelerin, önceden belirlenmiş bu sınıflarla karşılaştırılması ile ölçülür.

Entropi, elde edilen bir kümenin ne kadar homojen olduğunun göstergesidir. Düşük entropi değeri kümenin daha homojen olduğunu ifade eder ve bu beklenir. Toplam entropinin belirlenmesi için, önce her kümenin entropisi ve daha sonra tüm kümelerin toplam entropisi hesaplanır. Bir kümenin entropi hesabı :

$$E_j = - \sum P(i,j) * \log P(i,j)$$

Burada P(i,j) *i* sınıfına ait dokümanların *j* kümesinde bulunma olasılığıdır. Toplam entropi hesabı:

$$E = \sum n_j / n * E_j$$

Burada n_j, j. kümenin boyutunu, n ise doküman seti içindeki toplam doküman sayısını ifade eder.

F-measure hesaplanırken bilgiye erişim sahasında kullanılan iki standart ölçümden yararlanılır.

Duyarlık (Precision) bir kümedeki dokümanların kaçının doğru kümede olduğunu gösterir.

Doğruluk (Recall) bir kümede olması gereken dokümanlardan kaçının o kümede olduğunu gösterir.

Toplam f-measure hesaplanırken önce her kümenin f-measure'ı hesaplanır

$$F(i,j) = \frac{2 * Recall(i,j) * Precision(i,j)}{Recall(i,j) + Precision(i,j)}$$

Daha sonra toplam F-measure hesabını yapmak için

$$F = \sum n_i / n * \max F(i,j)$$

Burada n_i, i. kümenin boyutunu, n ise doküman seti içinde ki toplam doküman sayısını ifade eder. [17]

5. BULGULAR

Bu çalışmadaki amaç, mevcut kümeleme tekniklerinden daha iyi bir kümeleme tekniği geliştirmektir. Bu amaçla GAD yöntemi n-gram yaklaşımı ile geliştirilmiştir. Geliştirilen yeni yöntemin de eğitmensiz bir sistem olması, incelenecek doküman sayısındaki artışa göre dışarıdan müdahale gerektirmeden ölçeklenebilmesi, bilgiye erişim alanına yönelik başarılı ve önemli bir katkı olacaktır. Önerilen yöntemin karşılaştırılabilir olması için GAD ve *n-gram kelimelerle GAD* ile elde edilen doküman dizinleri hiyerarşik kümeleme tekniği kullanılarak kümelenmiştir. Tablo 1 de kullanılan kümeleme yaklaşımları gösterilmiştir.

Tablo 1. Karşılaştırma için kullanılan algoritmalar

Algoritma	Açıklaması
Tfidf_GAD_HK	GAD tabanlı hiyerarşik kümeleme (Tfidf tabanlı vektör uzayı)
Tfidf_ngram_GAD_HK	n-gram destekli GAD tabanlı hiyerarşik kümeleme (Tfıdf tabanlı vektör uzayı)

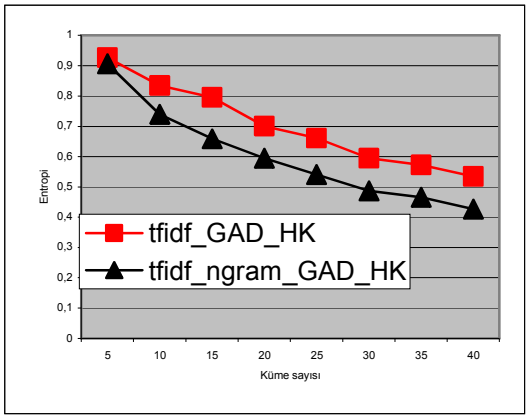
Tablo 2. Türkçe Doküman kümesi için entropi ölçümleri

Küme sayısı	Tfıdf_GAD_HK	Tfidf_ngram_GAD_HK
5	0,926381306	0,90608821
10	0,834725048	0,73937549
15	0,795165679	0,65872017
20	0,700502409	0,59382918
25	0,660731649	0,54026323
30	0,594019001	0,48636027
35	0,571697287	0,46587803
40	0,534692964	0,42687611

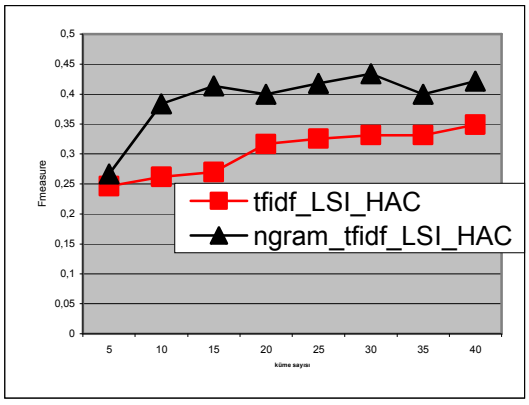
13 sınıfta kümelenmiş 615 Türkçe dokümana ait entropi ve f-measure ölçümleri Tablo 2 ve Tablo 3 de gösterilmektedir. Bu tablolarda verilen değerlerin Şekil 1 ve Şekil 2 de verilen grafiklerde yorumlanmıştır.

Tablo 3 Türkçe Doküman kümesi için f-measure ölçümleri

Küme sayısı	Tfidf_GAD_HK	Tfidf_ngram_GAD_HK
5	0,246469381	0,26671552
10	0,261534557	0,38382652
15	0,269808758	0,41365287
20	0,316816249	0,39988699
25	0,325498299	0,4177481
30	0,331417705	0,43376812
35	0,331566884	0,39960739
40	0,349225421	0,42172488



Şekil 1: Türkçe doküman kümesi için küme sayısına bağlı entropi değişim grafiği



Şekil 2: Türkçe doküman kümesi için küme sayısına bağlı f-measure değişim grafiği

Görüldüğü gibi *n-gram kelimelerle GAD* yöntemi hem kümelerin homojenliği (düşük entropi değeri) hem de kümeleme işlemenin kalitesi (yüksek f-measure değeri) açısından GAD yöntemine göre ciddi üstünlük sağlamıştır. Türkçe dokümanların normalde 13 sınıfta toplandığı belirtmişti. Bu dokümanların 13 kümeye dağıtımında n-gram destekli GAD daha başarılı olduğu gibi daha fazla kümeye ayrıştırıldığında da daha kaliteli ve homojen kümeler oluşmuştur. Bu; bir kümeye ait dokümanların da kendi içinde farklı konulara ayrıştırılabileceği ve önerilen *n-gram kelimelerle GAD* yönteminin bu ayrıştırma işlemini çok daha başarılı bir şekilde yaptığı anlamına gelmektedir.

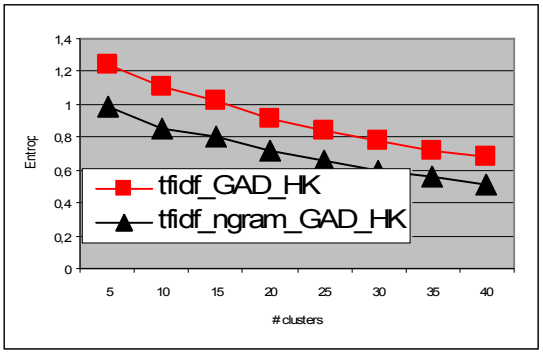
İngilizce doküman seti olarak Reuters21578 setinden rastgele seçilen 20 sınıfa, 1.680 doküman kullanılmıştır. Bu doküman kümesinde 1987 yılında Reuters yayınlarından derlenmiş 20.000 doküman bulunmaktadır [14]. Bu kümeye ait entropi ve f-measure ölçümleri Tablo 4 ve Tablo 5 de; bu tablolara ait grafikler Şekil 3 ve Şekil 4 de gösterilmiştir.

Tablo 4. İngilizce Doküman kümesi için entropi ölçümleri

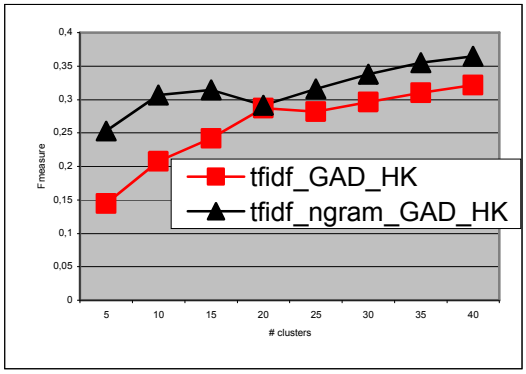
Küme sayısı	Tfidf_GAD_HK	Tfidf_ngram_GAD_HK
5	0,926381306	0,90608821
10	0,834725048	0,73937549
15	0,795165679	0,65872017
20	0,700502409	0,59382918
25	0,660731649	0,54026323
30	0,594019001	0,48636027
35	0,571697287	0,46587803
40	0,534692964	0,42687611

Tablo 5 İngilizce Doküman kümesi için f-measure ölçümleri

Küme sayısı	Tfidf_GAD_HK	Tfidf_ngram_GAD_HK
5	0,246469381	0,26671552
10	0,261534557	0,38382652
15	0,269808758	0,41365287
20	0,316816249	0,39988699
25	0,325498299	0,4177481
30	0,331417705	0,43376812
35	0,331566884	0,39960739
40	0,349225421	0,42172488



Şekil 3: İngilizce doküman kümesi için küme sayısına bağlı entropi değişim grafiği



Şekil 4: İngilizce doküman kümesi için küme sayısına bağlı f-measure değişim grafiği

İngilizce dokümanların kümelenmesi işleminde de önerdiğimiz *n-gram kelimelerle GAD* yöntemi, normal GAD yöntemine göre çok daha başarılı sonuçlar üretmiştir. Bu setteki dokümanların uzmanlar tarafından 20 grupta sınıflandırıldığı belirtilmişti. Normal GAD yöntemi sadece bu sayıda küme için *n-gram kelimelerle GAD* yöntemini yakalamıştır ancak bu küme sayısında dahi, kümelerin homojenliği açısından n-gram destekli GAD daha başarılı netice vermiştir.

6. SONUÇLAR

Gerek İngilizce gerek Türkçe dokümanların daha başarılı kümelenmesi için, dokümanlar içindeki kelimelerin yanında kelimelerin yan yana kullanımlarının sistematik olarak işleme katılmasının ortaya koyduğu başarı gösterilmiştir. Bunun anlamı, *n-gram kelimelerle GAD* yönteminin doküman içeriklerini daha iyi ortaya koymasıdır.

Bununla birlikte *n-gram kelimelerle GAD* yönteminin daha da iyileştirilebilmesi mümkündür. Bu yöntemi geliştirmek için bundan sonraki aşamada yapılacak planlananlar iyileştirmeler aşağıda verilen hususlar dikkate alınarak gerçekleştirlecektir :

- Dokümanların içeriğini yüklem ve yükleme yakın kelimeler şekillendirmektedir. Bu nedenle cümle içindeki yüklem ve yükleme yakın kelimelerin, doküman içeriğini ifade etmek için sahip oldukları değerin daha yüksek olması gerekir.

- Dokümanların içeriğini ifade etme gücü açısından kelime ve kelime gruplarını değerlendirirken, bunların dilbilgisi açısından yapılarının da önemi ayrıştırılmalıdır. Örneğin bir kelime nesne olarak kullanıldığındaki hali ile sıfat olarak kullanıldığı halinden farklı değerlere sahip olmalıdır.

7. KAYNAKLAR

[1] Bellot, P. and El-Beze, M., *A Clustering Method for Information Retrieval,* Technical Report IR-0199, Laboratoire d'Informatique d'Avignon,France, 1999.

[2] Berry, M. W., Drmac, Z. and Jessup E. R.: Matrices, Vector Spaces, and Information Retrieval, SIAM Review, v.41 n.2, p.335-362, June 1999

[3] Boley D., Principal direction divisive partitioning. Data Mining and Knowledge Discovery, 2(4), 1998.

[4] Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai "Class-based n-gram models of Natural Language", *Computational Linguistics*, vol. 18, pp. 467–479, 1992.

[5] Croft, W.B. and Xu, J.: Corpus-specific stemming using word form co-occurence. In Proceedings for the Fourth Annual Symposium on Document Analysis and Information Retrieval (pp. 147-159), Las Vegas, Nevada. 1995.

[6] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R.: (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407

[7] Duda, R. O., Hart, P. E., and Stork, D. G., Pattern Classification. Wiley, New York.2001

[8] Ekmekcioglu, F. C., Lynch, M. F. and Willett, P. (1996): Stemming and N-gram Matching For Term Conflation In Turkish Texts. Inf. Research, Vol. 2, No. 2.

[9] Kohonen, T., "The Self-Organizing Map," Proceedings of the IEEE, vol. 9, 1990, pp. 1464-1479.

[10] Lingpipe NLP Library http://www.aliasi.com/lingpipe

[11] Salton, G. and McGill, M. J.: Int. to modern information retrieval. McGraw-Hill

[12] Willet, P., Recent trends in hierarchical document clustering: a critical review. Information Processing and Management, vol. 24(5), pages 577-- 597, 1988.

[13] Zemberek Turkish NLP Library: https://zemberek.dev.java.net/

[14] "Reuters21578collection", http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

[15] Porterstemmer , http://www.tartarus.org/martin/PorterStemmer/

[16] Foundations of Statistical Natural Language Processing (Hardcover) by Christopher D. Manning, Hinrich Schütze

[17] Unsupervised Machine Learning Techniques For Text Document Clustering, Arzucan Özgür, Ethem Alpaydın