METİN MADENCİLİĞİ: Temel Yöntemler ve Duygu Analizi

Yücel Saygın
Sabancı Üniversitesi,
İstanbul, TÜRKİYE

Büyük Veri

- 3V (Volume, Velocity, Veracity)
- 3Ç (Çok, Çabuk, Çeşitli)

Veri Madenciliği

- Operasyonel veri tabanları: Günlük veriler ve işlemler
- Veri toplamanın ve saklamanın kolaylaşması sonucu veri toplama yarışı
- Veri ambarları:
  - Geçmişe dönük veriler
  - Karar alma amacıyla uzun süren sorgular
- Veri Madenciliği

Veri Madenciliği

- Neden?
  - Büyük miktarlarda veri
  - Bunun incelenmesi ve faydalı bilgilere dönüştürülmesi
- Şu an Büyük Veri ile birlikte konuşulmakta
- Büyük Veri: Çağın Petrolü
- Büyük veri ile Veri Koruması daha fazla önem kazandı

Veri Madenciliği

- Büyük çaplı verilerden faydalı bilgi çıkarma
- Veri tabanları + İstatistik + Makine Öğrenmesi

Metin Veri Kaynakları

- Haber Grupları
- Bloglar
- Yorumlar (Gazete, Mekan, Sinema)
- Facebook
- Twitter
- ...

Metin Verilerin Farkı

- Düzensiz yapıda
- Dile bağlı
- Context (Bağlam) önemli

Bilgi Erişimi

- Büyük çaplı düzensiz metinlerin işlenip, indexlenip, sorgulanması
- Büyük verinin önemli itici güçleri: arama motorları

Veri Madenciliği Temel Yöntemler

- Supervised (Eğitim Şart)
- Un-supervised (Eğitmeden)

Eğitime Gerek Duymayan bir Yöntem

- Kümeleme (Clustering) :
  - Verilen büyük bir metin kümesini daha küçük kümelere ayırma.
  - Aynı kümedeki metinler birbirine yakın
  - Farklı kümelerdeki dokümanlar birbirinden uzak.
- Öncelikle metinler/dokumanlar arasında bir yakınlık ölçütü tanımlamak gerekiyor

Metin Kümeleme

- Metinleri kelime torbasına çevirme (her metin bir kelime kümesi)
- Jaccard Coefficient: İki kümenin ne kadar örtüştüğünü gösteren populer bir ölçüt.
- jaccard$(A,B) = |A \cap B| / |A \cup B|$
- jaccard$(A,A) = 1$
- jaccard$(A,B) = 0$ if $A \cap B = 0$
- A and B farklı boyutlarda olabilir.
- 0 ve 1 arası bir değer döndürür.

Jaccard skorunun problemleri

- Jaccard *kelimenin dokuman içindeki frekansına bakmıyor*
- Az geçen kelimeler daha fazla bilgi içerebilir. Jaccard buna da bakmıyor

Log-frekans ile ağırlıklandırma

- t teriminin d dokumanında geçme sıklığına bağlı

$$w_{t,d} = \begin{cases} 1 + \log_{10} \mathrm{tf}_{t,d}, & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0, & \text{otherwise} \end{cases}$$

- q ile d'nin benzerlik skoru

$$= \sum_{t \in q \cap d} (1 + \log \mathrm{tf}_{t,d})$$

idf ile ağırlıklandırma

- df_t : *t* terimini içeren dokumanların sayısı
  - df_t terimin önemi ile ters orantılıdır
  - $df_t \leq N$
- We define the idf (inverse document frequency) of *t* by

$$idf_t = \log_{10}(N/df_t)$$

tf-idf ağırlıklandırma

$$w_{t,d} = \log(1 + \mathrm{tf}_{t,d}) \times \log_{10}(N / \mathrm{df}_t)$$

- Şu ana kadar bilinen en iyi ağırlıklandırma yöntemi

Genel Yaklaşım

- Metin Verilerinde
  - Stop word çıkarılması
  - Kelimenin köklerine ayrılması
  - Vector Space Modele
  - Tf-Idf tabanlı ağırlıklandırma
- RAPIDMINER gibi veri madenciliği araçlarında yukarıdaki işlemler hazır olarak var.

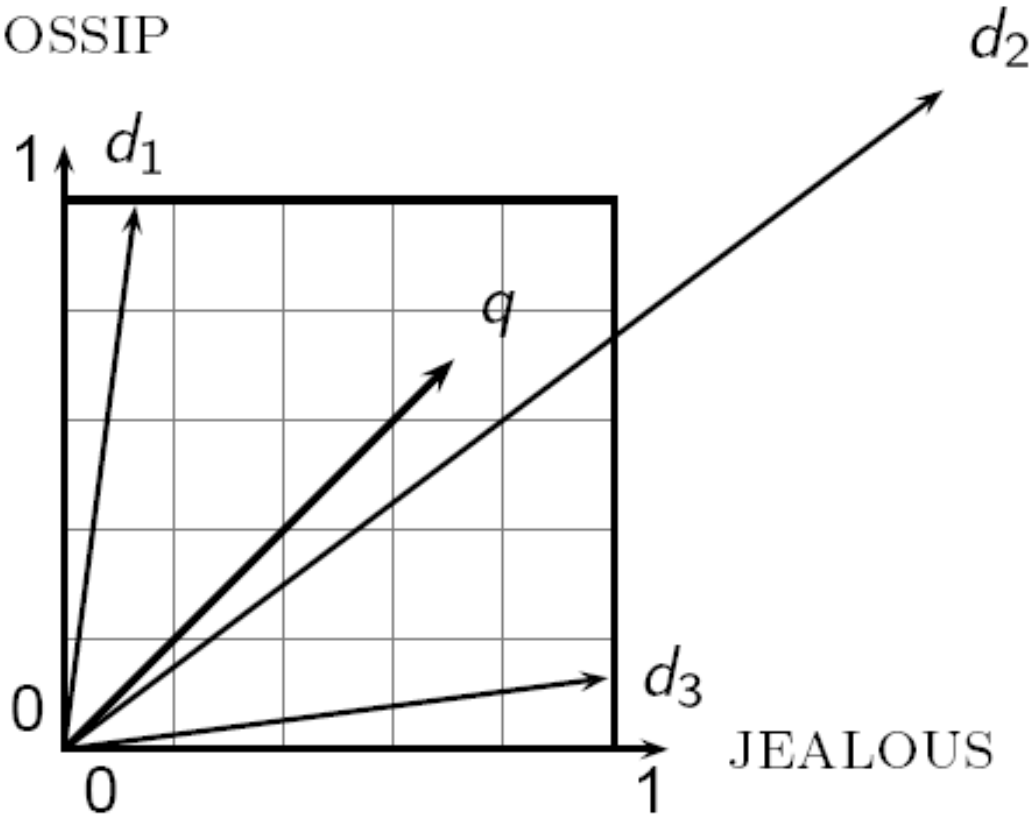
Dokümanların vektör ile temsili

- |V|-boyutlu vektör uzayı
- Terimler are farklı boyutları belirliyor
- Dokümanlar are vektör uzayında bir nokta
- Dokümanlar arasındaki uzaklık öklid uzaklığı olarak belirlenebilir

Öklid uzaklığı pek iyi bir fikir değil

q ve d2 dokumanlarının
Kelime dağılımı benzer
ama öklid uzaklıkları
çok fazla.

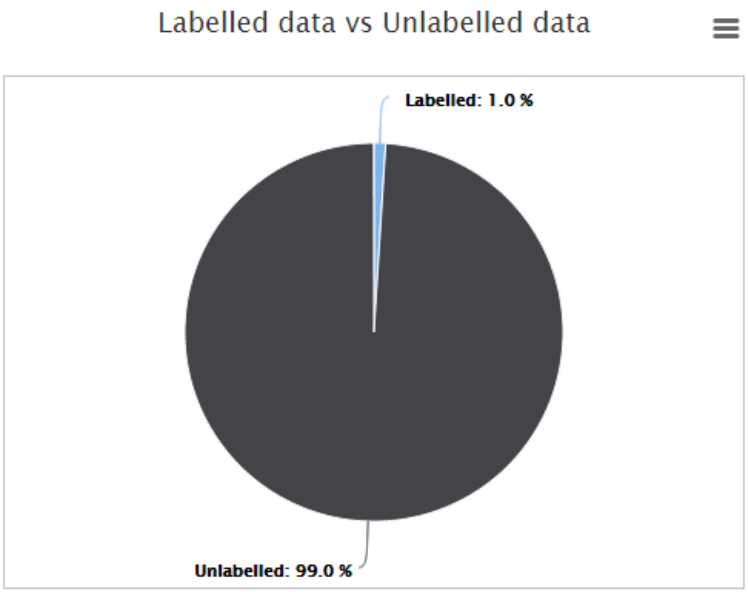
Bu durumda aralarındaki
açıya bakmak daha mantıklı



Tweetlerin Kümelenmesi

- Kelime torbası modeli çok uygun olmayabilir
  - Yanlış yazılan Tweetler
  - Kısaltmalar
  - Vs
- Standart algoritmalar çok uygun olmayabilir
  - K-means (ve varyansları)
  - Hiyerarşik kümeleme (çok pahalı)
  - Yoğunluk bazlı (Density based)

Tweetlerin işaretlenmesi



Tweets

Index	Sentiment	Sentiment	Sentiment	Reason	User	Tweet	Label
0	Positive	Objective	Negative		owner	RT @gtatlipinar: İHANET DEĞİLSE NEDİR BU KOBANİ EYLEMRİNDE ARABA YAKMAYA ÇALIŞAN PARALEL POLİS GÖREVDEN ALINDI	-2
1	Positive	Objective	Negative		owner	RT @YilmazGedik: Kobani Halk Meclisi Başkanı Ayşe Efendi: Türkiye IŞİD?in Yenilmesini Engelleyen Güçtür...	-2

Tweetlerin Kümelenmesi

- Berkay Dinçer MS tez çalışması
- İki Tweet'in yakınlığı için Longest Common Subsequence (LCS) uygun bir yöntem
  - Sıralı olacak şekilde karakter atlamaya musade ederek iki dizinin en buyuk ortak altdizisi
  - LCS'in uzunluğunu dizilerin uzunluğu ile normalize ettikten sonra yakınlık ölçütü olarak kullanabiliriz.

Temsili Tweetler ve eşik değeri kullanarak kümeleme

Algorithm 2 Clustering

```
1: for each c_i ∈ C do
2:   for each c_j ∈ C − c_i do
3:     if d(rep(c_i, c_j)) ≤ v then
4:       c_i ← c_i ∪ c_j
5:       C ← C − c_j
6:     end if
7:   end for
8: end for
```

- Temsili Tweetler eşik değerinden daha yakınsa birleşitiriyoruz.
- O (N2) Tweetlerin uzunluğu max 140 ama pratikte LCS hesaplaması O(nm).
- Is every tweet in a given cluster has less distance then the threshold?

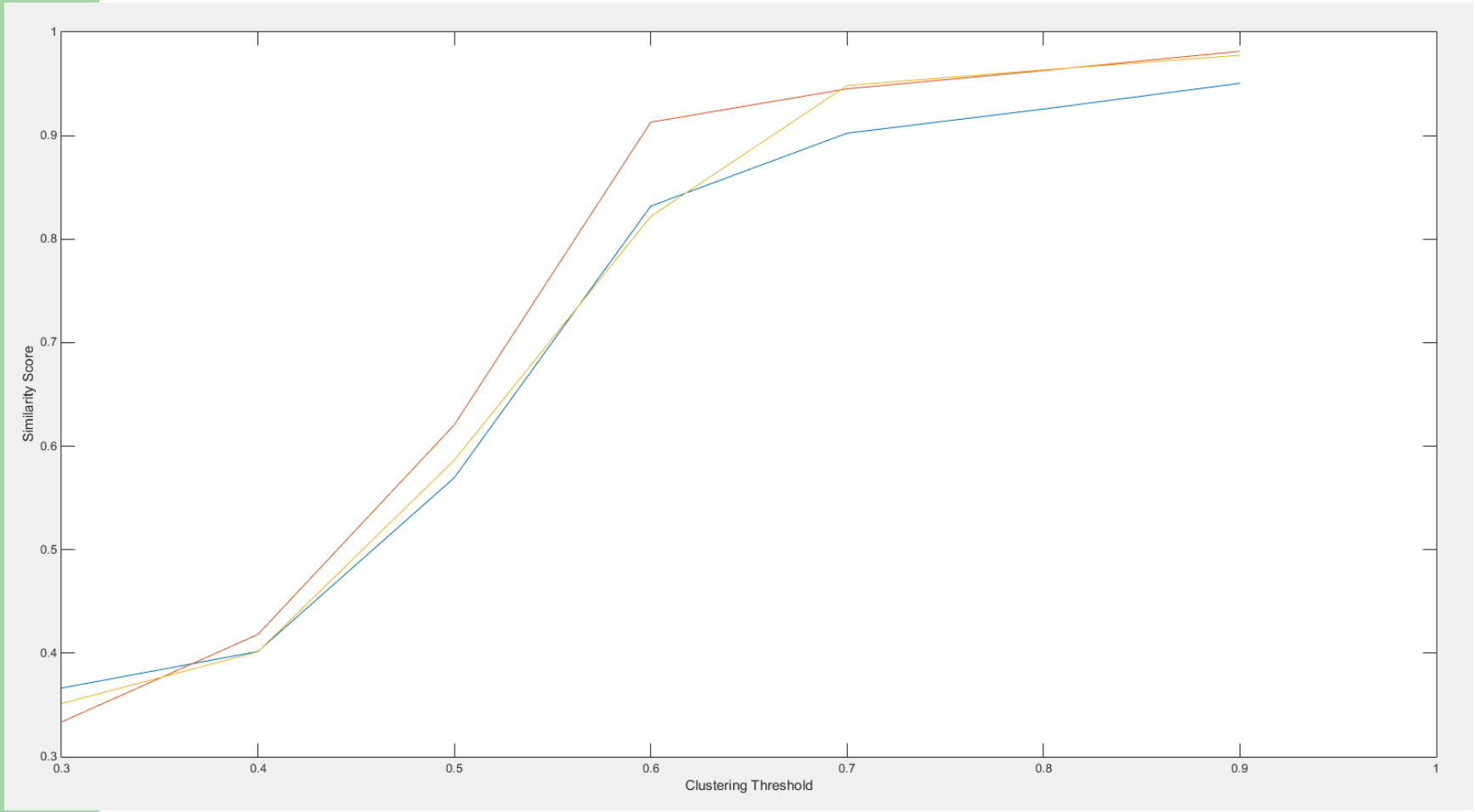
Tweetlerin Kümelenmesi

- Eşik değeri doğruluk ve hız arasında bir denge sağlıyor.
- $t_3 = rep(c_3)$ ve $t_1, t_2 \in c_3$
- $d(t_1, t_2)$ en fazla $2 - 2 \times threshold$ olabilir (triangular inequality)
- Deneylerle gerçek Tweetlere baktığımızda 0.6 eşik değeri içim en fazla 0.7 farklılık oluyor.

Deneysel Değerlendirme

- İki veri kümesi (Kobani ile ilgili tartışmalar ve sendeanlat hashtag ile toplanan Tweetler)
- Kümelerin saflığı ve eşik değeri
- Saflık: Kümeler içinde tüm ikililere bakarak bunların ortalama LCS uzaklıkları

Deneysel Değerlendirme



Deneysel Değerlendirme

Kobani direnişi 61. gününde	
0	Kobani direnişi 60.gününde.
1	RT @RamazanGuney2: Kobani direnişi 61. gününde
2	Kobani direnişi 61. gününde
3	RT @AjansaKurdi: Kobani direnişi 59. gününde
4	Kobani kuşatması 62. gününde
5	RT @imc_televizyonu: Kobani direnişi 60. gününde
6	?@imc_televizyonu: Kobani direnişi 61. gününde
7	Kobani direnişi 63. gününde

Deneysel Değerlendirme

turk takipci hilesi http://goo.gl/Jk82Qb http://t.co/VSKVNIFNLw #sendeanlat 754623

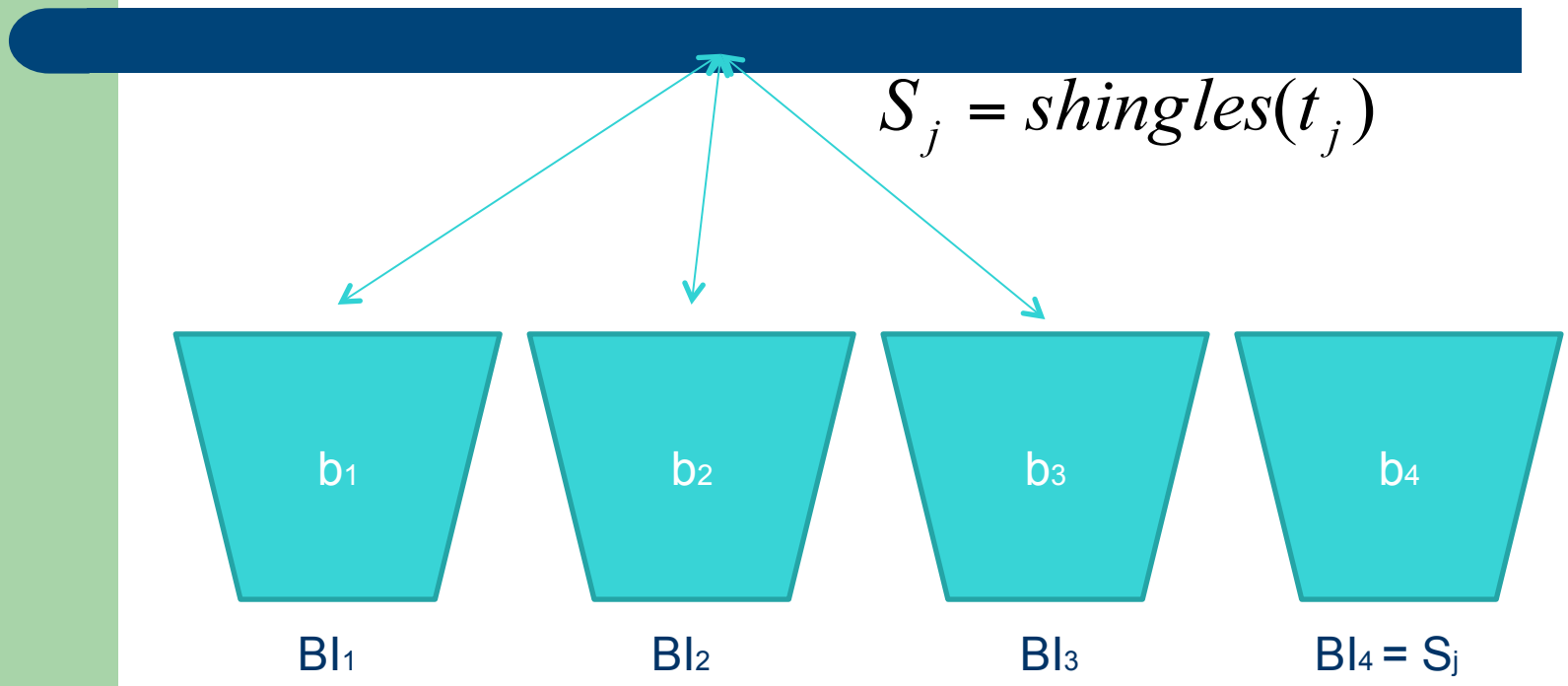
0	Twitter ucretsiz takipci http://goo.gl/Jk82Qb http://t.co/JOY11Zy4eL #sendeanlat 699927
1	En saglam takipci sitesi http://goo.gl/Jk82Qb http://t.co/Zg9g89lPqz #sendeanlat 481315
2	Twitter takipci arttirma sitesi http://tinyurl.com/kvhk6pv http://t.co/Cbx0XwBIoE - #sendeanlat 278646
3	En Iyi takipci sitesi http://goo.gl/Jk82Qb http://t.co/stRvM4Kgag #sendeanlat 211168
4	Twitter ucretsiz takipci http://goo.gl/Jk82Qb http://t.co/WpgH9Rs6JP #sendeanlat 361181
5	Twitter ucretsiz takipci http://goo.gl/Jk82Qb http://t.co/AcQAAex9MJ #sendeanlat 537808
6	Reklamsiz takipci arttir http://goo.gl/Jk82Qb http://t.co/E5CK1hgXu4 #sendeanlat 652800
7	turk takipci sitesi http://goo.gl/Jk82Qb http://t.co/CXy4pvv6mz #sendeanlat 782700
8	Super Takipci Sitesi http://goo.gl/Jk82Qb http://t.co/Y7JbOQwr4z #sendeanlat 447110
9	turk takipci hilesi http://goo.gl/Jk82Qb http://t.co/zh2ucoF2nl #sendeanlat 749089
10	Super Takipci Sitesi http://goo.gl/Jk82Qb http://t.co/Nq0gA2i2k6 #sendeanlat 219141

- Spam Tweetler tek bir clustera toplanıyor

W-Shingling

- W-shingling dokumanlar arasındaki uzaklığı hızlı hesaplamak için kullanılır.
- W pencere boyutu.
- Örnek:
  - “We are all born mad. Some remain so”
  - Shingle kümesi, w = 10 için:
    - S = {We are all, e are all, are all b, are all bo,....., remain so}
- Jaccard uzaklığı ile iki Tweetin Shingle kümesinin kesişimine bakılır

Clustering of Tweets



Büyük Veri ile İlgili Küçük Bir Deney

- Giray Havur Master Tezi Çalışmasından bir bölüm
- Twitter: 300 milyon kullanıcı
- Türkçe Tweet atanlar:10 milyondan fazla
- Twitterin %2'ye yakın hacmi Turkçe Tweetlerden oluşuyor

Büyük Veri ile İlgili Küçük Bir Deney

- Twitter kullanılarak politik partilerin oy oranları tahmin edilebilir mi?
- En fazla oy oranına sahip iki parti: AKP ve CHP
- Politik görüş bildiren ve bu partilerle ilgili olabilecek kelimeler/ terimler (AKP, CHP, AK Parti, Erdoğan, Ataturk, Kılıçdaroğlu, Sarıgül, ..)
- Bu kelimeleri ile Twitter API'ından 100.000 Tweet ile bir örneklem kümesi

Büyük Veri ile İlgili Küçük Bir Deney

- Bu terimleri kullananların destekleyen ya da muhalif olduğunu anlamak için Tweetlerin içeriğine bakmak gerekiyor ki bunu doğru yapabilmek çok fazla uğraş gerektiriyor
- Alternatif bir yaklaşım: Twitter profil resimleri (default yumurta)
- Resimlerden -> Anahtar kelimelere

Büyük Veri ile İlgili Küçük Bir Deney

- Önceden belirlediğimiz anahtar kelime kümeleri
  - AKP: (AKP, Recep Tayyip Erdoğan, ..., Osmanlı, Rabia)
  - CHP: (CHP, Kemal Kılıçdaroğlu, Ataturk, Mustafa Kemal, ...)
- AKP U CHP : 100.000 Tweet (2014)
- 65.000 farklı kullanıcı
- 65.000 profil fotografı
- 65.000 fotograf -> Google Fotograf Ters Arama -> 10957 terim
- 10.950 terimden 1274'u AKP U CHP içinde
- 1274 terim içinde
  - AKP terimlerine ait olan 781
  - CHP terimlerine ait olan 493 kelime
  - %61 - %39
  - 2014 yerel seçimlerde göreceli oy oranı %62 - %38

Bu deneyi nasıl genişletebiliriz?

- Anahtar kelimelerin yarı otomatik yöntemlerle belirlenmesi:
  - Twitterda anahtar kelimelerle beraber geçen diğer kelimelerin otomatik bulunması
  - Google'da anahtar kelimelerin aranmasıyla gelen sonuçlarda sıkca geçen diğer kelimeler
  - Mutual Information tarzı metriklerle eleme yapılabilir

Bu deneyi nasıl genişletebiliriz?

- Anahtar kelimelerin yarı otomatik yöntemlerle belirlenmesi:
  - Anlamsal veri kaynakları (Wordnet)
  - Wikipedia, Ekşisözlük gibi kaynaklar
- Sadece fotograf değil metin içerik analizi de yapılması

Büyük Veri ile İlgili Başka Bir Deney

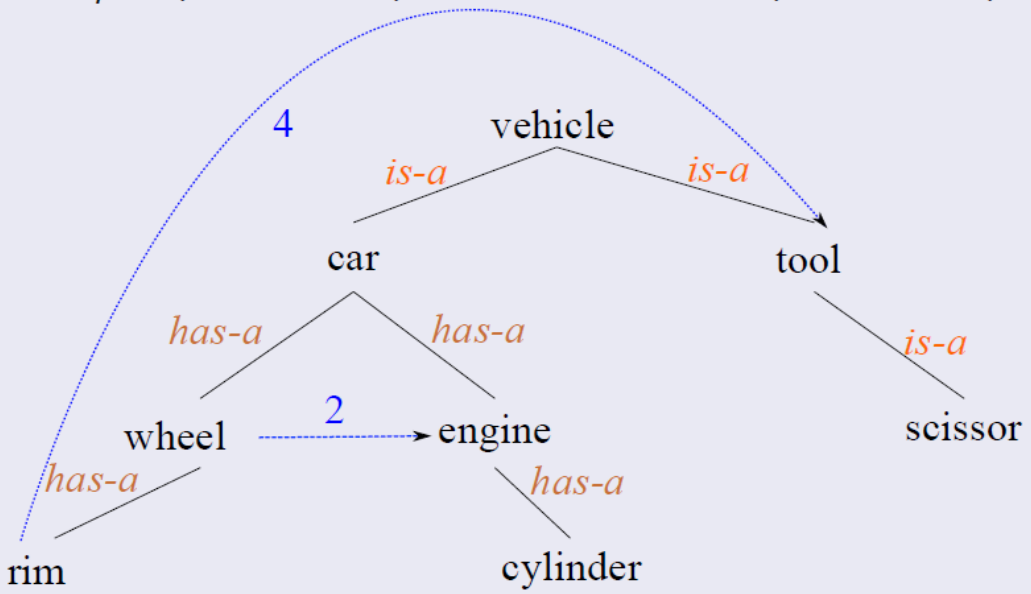
- Facebook kullanıcılarının ilgi alanlarını genel paylaşımlarından bulabilir miyiz?
  - Postlar
  - Likelar
  - Fotograflar
- İlgi alanı: Anahtar bazı kelimelerle tanımlı ama bunu genişletmek gerekiyor

Büyük Veri ile İlgili Başka Bir Deney

WordNet::Similarity Scoring

pathlen(engine,car) =#edges on the shortest path

sim_{path}(engine,car) = 1 - log*pathlen*(engine,car)



Daha zor bir problem: Duygu Analizi

- Yeni ipad ile ilgili insanlar ne düşünüyor?
- İnsanlar genelde beğenmiş mi?
- Ne gibi özelliklerini beğenmişler?
- Ne gibi özellikleri problemli?

Neden önemli?

- Başkalarının fikirleri bizim davranışımızı da etkileyebiliyor.
- Bir karar almadan önce başkalarının fikirlerine bakıyoruz (Otel, kitap değerlendirmeleri gibi).
- Eskiden anetlerle, pahalı çalışmalarla, danışmanlarla yapılanlar belki sosyal medya ve metinlerin incelenmesi ile de yapılabilir

Sabancı Üniversitesi

Duygu Analizi

Veri kümelerinde (Genellikle metin) ifade edilen duyguyu çıkarmayı hedefler:

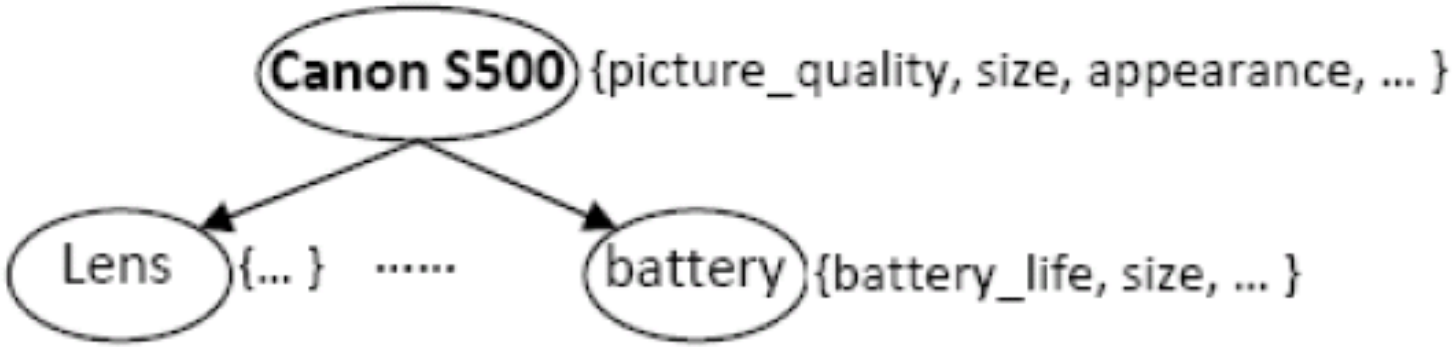
- **Otomatik ya da yarı-otomatik**
- **Polarite: poz./neg./nört.**

Farklı seviyelerde duygu analizi

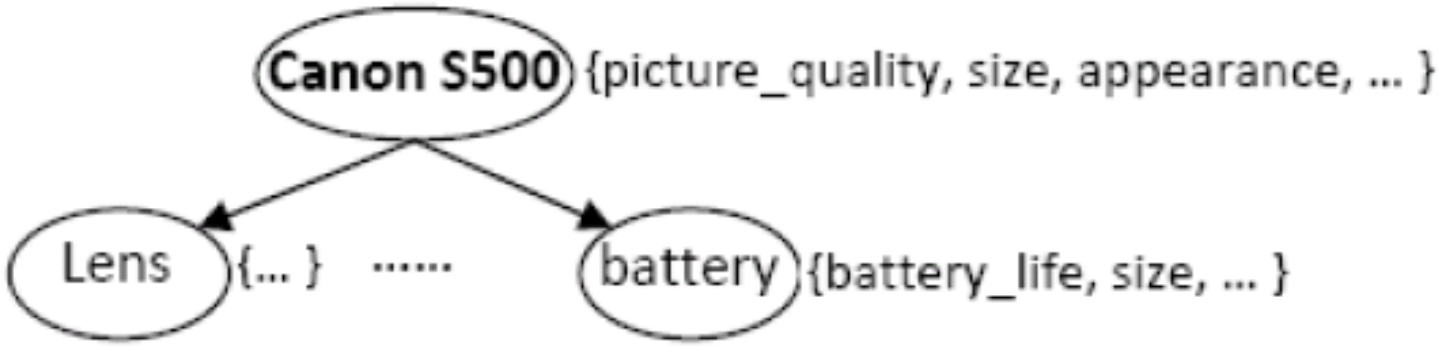
- One can look at this review/blog at the
  - Dokuman seviyesinde. (Tüm yorum genel olarak pozitif mi negatif mi?)
  - Cümle seviyesinde
  - Aspect seviyesinde

Entity (Varlık)

- Örnek: ürün, kişi, olay, ya da konu topic.
- Hiyerarşik bir yapıda olabilir.



Entity



- İnsanlar fikirlerini herhangi bir komponenti üzerinde beyan edebilir.
- ***Aspects*** tümünü temsil etsin.

Metin Sınıflandırma İşi

- Negatif, Pozitif, ve Nötr olarak
- Metinleri konulara ayırmaktan farkı
  - Metinleri konulara ayırırken konu ile ilgili kelimeler önemli
  - Duyguyu sınıflandırırken iyi, kötü, çirkin gibi kelimeler önemli.

Kural bazlı sınıflandırma (Turney, 2002)

- Sınıflandırılacak Veri: epinions.com sitesinden alınan yorumlar
- Adım 1:
  - Part-of-speech (POS) tagging
  - Yorumlardan sıralı kelime çiftlerini çıkarmak
  - Bunun için belli kurallar kullanılmış (İsim-sıfat, sıfat-isim gibi)

Kural bazlı sınıflandırma (Turney, 2002)

- Adım 2: PMI ölçütü kullanarak skorlama

$$PMI(word_1, word_2) = \log_2\left(\frac{P(word_1 \wedge word_2)}{P(word_1)P(word_2)}\right)$$

$$SO(phrase) = PMI(phrase, \text{“excellent”}) - PMI(phrase, \text{“poor”})$$

Kural bazlı sınıflandırma (Turney, 2002)

- Adım 3:
  - PMI kullanarak tüm grupları skorlama
  - Ortalama skorun pozitif, ya da negatif olmasına göre bir sınıf atama

SentiWordNet

- Word-Net tabanlı *domain independent* polarity lexicon
- Her kelimenin ne kadar pozitif, negatif, ve objektif olduğunu belirtiyor
- Her word-sense çiftne 3 değer atanıyor (positive, negative and objective)
- Örnek:

word-sense	negative pol.	objective pol.	positive pol.
"good"-adv	0.000000	0.812500	0.187500
"good"-adj	0.005952	0.386904	0.607142
"good"-noun	0.000000	0.468750	0.531750

- *Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06. pp. 417-422 (2006)).*

Dominant Polarity

Dominant Pol(t)

$$\begin{cases} pol^{+} & \text{if } pol^{+} \geqslant pol^{-} \\ pol^{-} & \text{if } pol^{-} > pol^{+} \\ 0 & \text{otherwise} \end{cases}$$

Problem

Otel Yorumu:

- "*The hotel had really **small rooms***" (-)

Fotograf Makinası Yorumu:

- "*This camera is great as it has a **small size***" (+)

- *pol ("small"-adj)* = 0.7250 (objective polarity).
- Domain-independent lexicon (SentiWordNet) *context (bağlam) bilgisini kaçırıyor*.

Word	POSTag	Neg-Pol	Obj-Pol	Pos-Pol
small	Adjective	0. 2625	0.7250	0.0125

SentiWordNet

- Gözlem: SentiWordNet word-sense ciftleri hep aynı polariteye sahip olduğunu farzediyor .
- Amaç: SentiWordNet polaritelerini farklı domainlere adapte edebilmek.

Yöntem

- Belli bir domain için işaretlenmiş yorumları kullanarak SentiWordNet gibi genel amaçlı kaynakları farklı domainlere adapte etmek.

Domaine Özel Kelimeleri Bulma

- Kelimelerin pozitif ya da negatif sınıflara ait olması durumunu anlamak için
  - Her kelimenin tf-idf skorlarını hesapladık.

$$tf.idf\ (w_i, +) = log_e(tf\ (w_i, +) + 1) * log_e\ (N\ /\ df(w_i))$$

$$tf.idf\ (w_i, -) = log_e(tf\ (w_i, -) + 1) * log_e\ (N\ /\ df(w_i))$$

Polarite Adaptasyonu İçin Ölçüt

- *(Δtf) idf*. Bir kelimenin polaritesini değiştirmek gerekir mi? Bunu anlamak için pozitif ya da negatif sınıfta görülme sıklığına bakıyoruz.

$$(\Delta tf)idf\ (w_i) = tf.idf\ (w_i, +) - tf.idf\ (w_i, -)$$

$$= [tf\ (w_i, +) - tf\ (w_i, -)]\ x\ idf\ (w_i)$$

- *(Δtf) idf skoru* ve *Pol (t) karşılaştırılarak gerekirse değiştirilir*.

Sabancı Üniversitesi

Kelime Seviyesinde Duygu Analizi

- **Değişken olmayan polarite:**

  Güzel

- **Farklı kültürlerde değişebilen:**

  Atatürk, cami, tanrı

- **Domain-bağımlı polarite:**

  (Büyük) : Room size (Hotel domain)

  (Büyük) : Battery size (Camera domain)

Sabancı Üniversitesi

Aspect Seviyesinde

Oyunculuk bence iyi

ama efektleri pek sevmedim

$$P(a_j) = \sum_{\forall n_k \in NG, s.t.\ a_j \in n_k} \frac{\sum_{t_i \in n_k} pos(t_i)}{|n_k|} \qquad (1)$$

$$N(a_j) = \sum_{\forall n_k \in NG, s.t.\ a_j \in n_k} \frac{\sum_{t_i \in n_k} neg(t_i)}{|n_k|} \qquad (2)$$

Binary classification accuracy: 79%
Ternary classification accuracy: 70%

Sabancı Üniversitesi

Cümle ve Doküman Seviyesinde

1) kendini kabul ettiren çok eğlenceli bir romantik komedi.
2) aralarındaki samimiyet oldukça hoş ve gerçekten insanı doğru insan tanımı yapmaya zorluyor:
3)"kendimi onunla yaşlanırken hayal edebileceğim birisi olmalı" derken.
4) fakat asıl dikkat çeken acı bir yön ise çapkın bir playboy olan drew hanım kızın sevgiliciğinin onu garson olmasına rağmen yanında tutma gerekçesi.
5) çünkü garson olması aslında onun için sorun ama işte.
6) kendi açıklıyor bu kesim için oyunun kuralları böyle değil mi gerçekte de zaten?
7) izlenesi bir film iyi eğlenceler.

Sabancı Üniversitesi

Yoğunlaşma

Iyi (good)
Çok iyi
Bayağı iyi
Gerçekten iyi

Kötü (bad)
Biraz kötü
azcık kötü

Polariteleri arttırmak ya da azaltmak gerekir

Negatifleşme

pişman(neg) değilim (neg->pos)

izlettğim onca insandan sevmedim (neg) diyen çıkmadı (neg->pos)

Polarite değerini düşürmek gerekir

Emoticonlar

Vakit geçirmek için fena değil ☺
ama daha fazlası yok ☹

Sabancı Üniversitesi

Logistic Regression
WEKA
5-fold Cross-validation

f_1: *average positive score of words in S using STN*
f_2: *average negative score of words in S using STN*
f_3: *average positive score of words in S using SN*
f_4: *average negative score of words in S using SN*
f_5: *number of positive words in S using PWS*
f_6: *number of negative words in S using PWS*
f_7: *occurrence of positive emoticons in S*
f_8: *occurrence of negative emoticons in S*
f_9: *number of adjectives and adverbs in S*
f_{10}: *number of (first letter) capitalized words in S*
f_{11}: *number of domain-specific indicative words in S*
f_{12}: *length of sentence (number of tokens in S)*
f_{13}: *is S a conditional sentence?*
f_{14}: *is S an interrogative sentence?*
f_{15}: *is S a negated sentence?*
f_{16}: *is S an exclamative sentence?*

Sabancı Üniversitesi

Cümlelerle İlgili Öznitelikler

Group Name	Feature	Name
Basic	F_1	Average review polarity
	F_2	Review purity
Occurrence of Subjective Words	F_3	Freq. of subjective words
	F_4	Avg. polarity of subj. words
	F_5	Std. of polarities of subj. words
$\Delta TF * IDF$	F_6	Weighted avg. polarity of subj. words
	F_7	Scores of subj. words
Punctuation	F_8	# of Exclamation marks
	F_9	# of Question marks
Sentence Level	F_{10}	Avg. First Line Polarity
	F_{11}	Avg. Last Line Polarity
	F_{12}	First Line Purity
	F_{13}	Last Line Purity
	F_{14}	Avg. pol. of subj. sentences
	F_{15}	Avg. pol. of pure sentences
	F_{16}	Avg. pol. of non-irrealis sentences
	F_{17}	$\Delta TF * IDF$ weighted polarity of 1st line
	F_{18}	$\Delta TF * IDF$ scores of subj. words in the 1st line
	F_{19}	Number of sentences in review

Metin Madenciliği Uygulama Alanları

- Politika: Seçim tahminleri
- Politikaların belirlenmesi: Farklı politikalara toplumun tepkisi
- Sosyal ağ analizi ile etkili insanların bulunması ve bunların insanların fikirlerine etkisi

Metin Madenciliği Uygulama Alanları

- Çağrı merkezi verilerinin incelenmesi
  - Şikayet konuları
  - Şikayet sebepleri
  - İletişim silsilesi sonucunda memnuniyet
  - Churn durumlarının önceden belirlenmesi, Churn'e sebep olan durumlar

İlerisi için araştırma alanları

- Association'dan Causality'e
- Active Learning
- Crowd sourcing