Anlamsal Uzaylar

Mehmet Fatih AMASYALI
Hesaplamalı Anlambilim
Ders Notları
BÖLÜM 3

Yıldız Teknik Üniversitesi
Bilgisayar Mühendisliği Bölümü

İçerik

- İstatistiksel Dil Modelleri
- Kelimelerin Anlamsal Benzerliklerinin Ölçümü
  - Gizli Anlam İndeksleme (Latent Semantic Indexing - LSI)
  - Latent Dirichlet Allocation - LDA
  - Kavramsal hiyerarşilerin kullanımı
  - Birbiri yerine kullanılan kelimeler
  - Kelime kümeleme
  - Anlamsal Uzaylar
  - Kelime öbekleri

İstatistiksel Dil Modelleri*

- Bir cümlenin olasılığını bulmak
- Kullanım alanları
  - Makine Çevirisi
    - P(taze balık aldım)>P(yeni balık aldım)
  - Yazım düzeltimi
    - Ali soan yedi
    - P(ali soğan yedi)>P(ali sokan yedi)
  - Konuşmadan Metne
    - P(soğan yedim)>P(sol an yedim)
  - vb.

https://web.stanford.edu/~jurafsky/slp3/slides/LM_4.pdf

İstatistiksel Dil Modelleri

- Amaç 1: bir cümlenin olasılığını bulmak

  P(W)=P(w1,w2,…,wn)
- Amaç 2: bir kelime sekansından sonra gelecek kelimenin olasılığını bulmak

  P(w5|w1,w2,w3,w4)

  Bunları nasıl hesaplayacağız?

  Bayes Kuralı : P(A|B)=P(A,B)/P(B) → P(A,B)=P(B)*P(A|B)=P(A)*P(B|A)

  Zincir Kuralı: P(A,B,C,D)=P(A)*P(B|A)*P(C|A,B)*P(D|A,B,C)

İstatistiksel Dil Modelleri

- Zincir kuralının genel hali:
- $P(w_1,w_2,w_3,\ldots,w_n)=$
  $P(w_1)P(w_2|w_1)P(w_3|w_1,w_2)\ldots P(w_n|w_1,\ldots,w_{n-1})$
  Örnek: P(ali,okuldan,buraya,geldi)= P(ali) * P(okuldan|ali) * P(buraya|ali,okuldan) * P(geldi|ali,okuldan,buraya)
- P(geldi|ali,okuldan,buraya) yı nasıl bulacağız?
  Bir olasılık :

$$P(\text{geldi}|\text{ali,okuldan,buraya}) = \frac{\text{frekans(ali okuldan buraya geldi)}}{\text{frekans(ali okuldan buraya)}}$$

Ama bu yolla güvenilir değerler elde etmek için
yeterli büyüklükte bir derlemimiz elimizde olmaz.
Çoğu frekans 0 olacaktır. Çünkü dil, diskrit uzayda çok seyrek.
Peki ne yapalım?

İstatistiksel Dil Modelleri

- Markov imdada yetişir ☺
- Markov varsayımı: her olasılık kendinden önceki k bileşene bağlıdır.
- k=0 için her olasılık bağımsızdır
  - P(w1,w2,w3,w4)=P(w1)*P(w2)*P(w3)*P(w4)
  - P(w4|w1,w2,w3)=P(w4)
  - P(geldi|ali,okuldan,buraya)=P(geldi)
- k=1 için her olasılık sadece bir öncekine bağlıdır
  - P(w1,w2,w3,w4)=P(w2|w1)*P(w3|w2)*P(w4|w3)
  - P(w4|w1,w2,w3)=P(w4|w3)
  - P(geldi|ali,okuldan,buraya)=P(geldi|buraya)

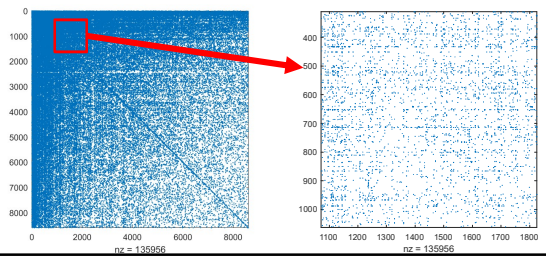
- k=2 için her olasılık sadece bir ve iki öncekine bağlıdır
  - P(w1,w2,w3,w4)=P(w3|w1,w2)*P(w4|w2,w3)
  - P(w4|w1,w2,w3)=P(w4|w2,w3)
  - P(geldi|ali,okuldan,buraya)=P(geldi|okuldan,buraya)
- k değerini istediğimiz gibi arttırabiliriz.

- Markov varsayımı ne yapıyor?
- Varsayım yokken
- $P(w_1,w_2,w_3,\ldots,w_n)=$
  $P(w_1)P(w_2|w_1)P(w_3|w_1,w_2)\ldots P(w_n|w_1,\ldots,w_{n-1})$
- $P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i|w_1, w_2, \ldots, w_{i-1})$
- Varsayım varken
- $P(w_i|w_1, w_2, \ldots, w_{i-1}) =$
  $P(w_i|w_1, w_2, \ldots, w_{i-k-1}, w_{i-k}, w_{i-k+1}, \ldots, w_{i-1}) \approx$
  $P(w_i|w_{i-k}, w_{i-k+1}, \ldots, w_{i-1})$

kırmızıları yok sayıyor

- Markov varsayımının avantajları:
  - daha az işlem
  - daha az seyrek veriler
- Dezavantajı:
  - Dil, uzak bağımlılıklar içerdiğinden çok yeterli bir model değil
    - Çine yaptığım ziyarette … çok uğraştım ama bir türlü Çinceyi öğrenemedim.
    - Burada Çinceyi kelimesi belki onlarca kelime öncesindeki Çine kelimesine bağlı.
- K seçimi burada bir ikilem

- Kod dil_modeli.m
- 1000 ekonomi haberi
- 5 ten az geçen kelimeleri sildikten sonra yaklaşık 8K*8K lık bir matris
- Çok seyrek, ~64M den sadece ~140K 0 değil



- ekonomi'den sonra en fazla geçen 5 kelime: $P(w_i|ekonomi)$ değeri en yüksek 5 kelime: bakanı, ve, ligi, bakanlığı, bankası
- para'dan sonra: politikası, cezası, ve, birimlerinin
- çok'dan sonra: önemli, daha, büyük, sayıda, iyi

- Bir cümlenin bigram modeline göre olasılığı:
- P(fuar otomotiv sektörüne önemli bir hareket getirdi) =P(otomotiv|fuar) * P(sektörüne|otomotiv) * P(önemli|sektörüne) * P(bir|önemli) * P(hareket|bir) * P(getirdi|hareket)
- $P(w_i|w_{i-1}) = \frac{fr(w_{i-1},w_i)}{fr(w_{i-1})}$
- P(otomotiv|fuar)=fre(fuar otomotiv)/fre(fuar)
- Log(P(fuar otomotiv sektörüne önemli bir hareket getirdi)) = -21.2067

w_i	$fr(w_{i-1})$	$fr(w_{i-1},w_i)$	$P(w_i\|w_{i-1})$
fuar	18	1	1/18
otomotiv	78	4	4/78
sektörüne	35	1	1/35
önemli	560	162	162/560
bir	3940	4	4/3940
hareket	42	1	1/42

0'larla başetmek Smoothing

- Log (P(fuar otomotiv sektörüne önemli bir hareket getirdi)) = log(P(otomotiv|fuar)) + log(P(sektörüne|otomotiv)) + log(P(önemli|sektörüne)) + log(P(bir|önemli)) + log(P(hareket|bir)) + log(P(getirdi|hareket)) = -21.2067
- Log (P(fuar otomotiv sektörüne önemli iki hareket getirdi)) = -INF
- **Add one estimation = Laplace smoothing**
- $P_{LS}(w_i|w_{i-1}) = \frac{fr(w_{i-1},w_i)+1}{fr(w_{i-1})+V}$
- Log (P_{LS}(fuar otomotiv sektörüne önemli bir hareket getirdi)) = -44
- Log (P_{LS}(fuar otomotiv sektörüne önemli iki hareket getirdi)) = -49

En olası dizilişi bulmak

- Kelimeler = önemli, hareket, bir
- Olası dizilişler
- P_{LS}(önemli bir hareket)= -11.8570
- P_{LS}(önemli hareket bir)= -18.1881
- P_{LS}(bir önemli hareket)= -17.4616
- P_{LS}(bir hareket önemli)= -16.8926
- P_{LS}(hareket önemli bir)= -13.0944
- P_{LS}(hareket bir önemli)= -17.4034
- Smoothing kullanmasaydık?
- Bir başka uygulama: Şu kelimelerle bir cümle kurun

Dil modellerini değerlendirmek

- Elimizde 2 dil modeli olsun (A, B). Hangisi daha iyi?
- Harici yöntem: her 2 modeli de farklı görevler için (Makine Çevirisi, Yazım düzeltimi, Konuşmadan Metne) kullan. O görevlerdeki performanslarına göre karşılaştır
- Dahili yöntem: Derlemi eğitim ve test kümesi olarak ayır. A ve B yi eğitim üzerinden oluştur.
  - Test kümesindeki cümlelere hangisi daha yüksek olasılık veriyorsa o daha iyidir.
  - Bir cümlenin başı verildiğinde sonunu hangisi doğru tahmin ediyorsa o daha iyidir.
  - Bu görevlerde Markov varsayımında k=0 seçimi nasıldır?

Dil modelleriyle cümle / dizilim üretimi

- Bir ilk kelime seçelim. Bundan sonra gelecek kelimeleri dil modeli ile belirleyelim.
- K=0 için her zaman en yüksek olasılığı seçersek hep aynı kelime tekrar eder
- K=1 için, her zaman en yüksek olasılığı seçersek bir kelimeden sonra hep aynı kelime gelir.
- K=2 için, ardışık 2 kelimeden sonra hep aynı 3. kelime gelir.
- Bunu aşmanın 2 yolu var
  - En yüksek olasılıklı t taneden birini rasgele seçmek
  - Hepsi içinden olasılıklarına göre seçim yapmak (Shannon Game)
- Ne zaman duracağız: İstediğimiz sayıda kelime / cümle ürettiğimizde

- Kod dil_modeli.m
- Bigram (K=1) modeli ile
- En yüksek olasılıklı t taneden birini rasgele seçer. t değeri azaldıkça metinlerde kopya çekme olasılığı artar. t=5
- 1000 ekonomi haberi ile
  - "fonların yüzde 5 de bu konuda son 10 ın da çok sayıda türk telekom gibi çok önemli olduğunu belirten bakan şimşek 2012 nin."
  - "mesleki eğitim ve ticaret bakanlığı ın piyasa ile ilgili olarak belirlendi bu nedenle bir önceki yıla da en fazla artış oranı ise lira oldu bu."
  - "akademik araştırmalar sonucunda yer aldığı bilgiye göre bir şekilde diyen ve yüzde 10 yılda bir süre içerisinde türkiye istatistik verilerine borsa seçimlerinin dikkati çekti türkiye."
- Burada first order (K=1) yapının problemi bariz görünüyor. "yer aldığı bilgiye" "yüzde 10 yılda"

- Kod dil_modeli_second.m
- trigram (K=2) modeli, t=5
- 1000 ekonomi haberiyle üretilen cümleler
  - "yaşanan sıkıntıları için devamlı ucuz bilet satıyoruz 100 90 ı aslında biz bu bitireceğiz inşallah yıl sonuna kadar türkiyede üretilen fazlasını belirterek geçen yıl kasım ayı ihracatını"
  - "birlikte fiyatlardaki üzerindeki etkileri takip edilecek gerek halinde düzenlemede birtakım yeni getirilecek talepler daha önceki bir açıklamasında 2012 yılının ilk 2 ihaleyi 500 adet olacak satış faaliyet"
- 1000 magazin haberiyle üretilen cümleler:
  - "arkadaşlarının da kendisini hiç söyleyen nilüfer kanser olduğunu tüm anlattı ilk itibaren dizinin kerem benim olduğu gibi bu hafta muhteşem neler olacak ali ikna eder ama şimdi"
  - "oyuncuların eğitimi yoktu ama hep başka şeyleri yapmak için burada bu arkadaşlar benim dönüş ve benim gibi tabii ki oldu ama devam çok bir adam alman sevgilisiyle"
- Ardışık 3 lüler daha anlamlı artık.

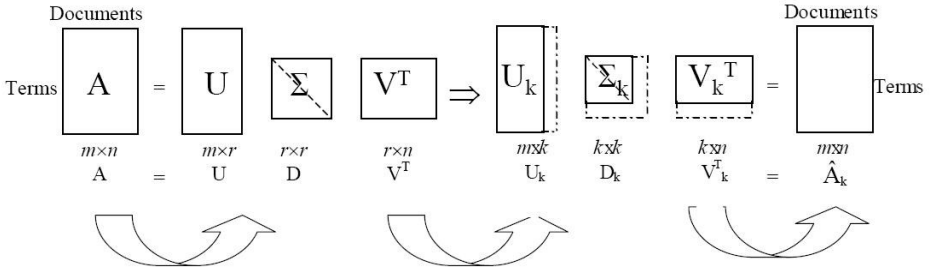
- Kod dil_modeli_term_doc.m
- Olasılık hesabında ardışık geçme yerine aynı dokümanda geçmeyi kullansak?
- Frekansı çok yüksek kelimeleri (stopWord) atmak gerekli. Yoksa her kelimeye en benzer kelimeler onlar olur.
- K=1 ve t=15 için üretilen cümle: "faiz oranı yüksek yönetim kurumsal yönetim kurulu genel 2 satış 5 enflasyon merkez cari kredi yapı nın türk türkiyenin ancak cari büyüme merkez bankası merkez cari."
- Cümle üretimi için iyi değil. Birlikte geçen kelimeleri sıralıyor.
- Faiz e en benzeyen kelimeler : "merkez" "kredi" "bankası" "oranı" "cari"
- Benzer anlama sahip kelimeleri bulmak için kullanılabilir. İleride göreceğiz.

Kelime benzerliği

- Karakter benzerliği (edit-distance)
  - Levenshtein vb. kod lev.m, lev_sample.m
- Anlam benzerliği
  - (Cimbom, Galatasaray,?)

Saklı Anlam İndeksleme
Latent Semantic Indexing

- Kelimelerin ardındaki saklı kavramları ortaya çıkarıp dokümanları bu kavramlar uzayında temsil etmek

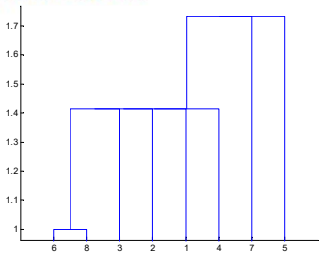


* https://liqiangguo.wordpress.com/2011/06/09/latent-semantic-analysis/

Saklı Anlam İndeksleme

```
% 2 sinifa ait dokumanlarımız olsun.
% 1. sınıf kedi 2. sınıf köpek
% kedi miyav süt fare köpek havla kemik hayvan tüy
D=[ 1   1     0     0     0     0      0     0     0 ; % kedi sinifi
    1   0     1     0     0     0      0     0     0 ; % kedi sinifi
    0   0     1     0     0     0      0     1     0 ; % kedi sinifi
    1   0     0     0     0     0      0     0     1 ; % kedi sinifi
    0   0     0     0     1     1      0     0     1 ; % köpek sinifi
    0   0     0     0     1     0      0     1     0 ; % köpek sinifi
    0   0     0     0     0     1      1     0     0 ; % köpek sinifi
    0   0     0     0     1     0      1     1     0 ];% köpek sinifi

D=D'; % matrisin kelime*dokuman olmasi gerek
[U,S,V] = svd(D);
%D = U*S*V'
%D 9*8 kelime*dokuman
%U 9*9 kelime*kavram
%S 9*8 singular values
%V 8*8 dokuman*kavram
```



Saklı Anlam İndeksleme

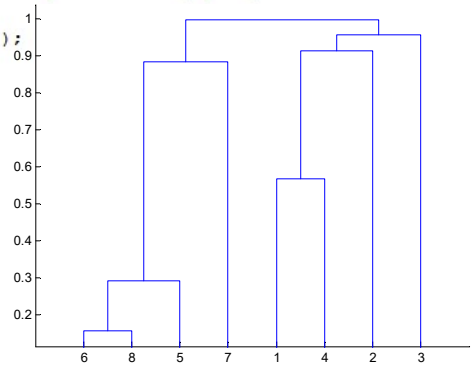
```
r=5; % dokumanlari kac gizli kavram boyutuna indirgeyecegimiz
V=V';
yeniD= U(:,1:r)*S(1:r,1:r)*V(1:r,:);
yeniD=yeniD(1:r,:);

figure;
Z = linkage(yeniD');
dendrogram(Z);
```

diag(S)

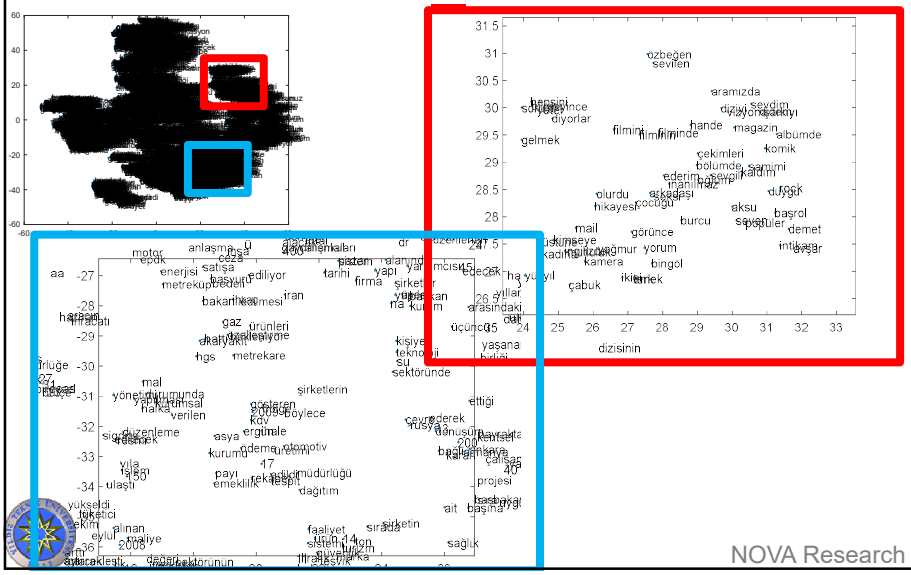
ans =

2.4893
2.0216
1.8073
1.3486
1.2349
0.8247
0.5943
0.2694



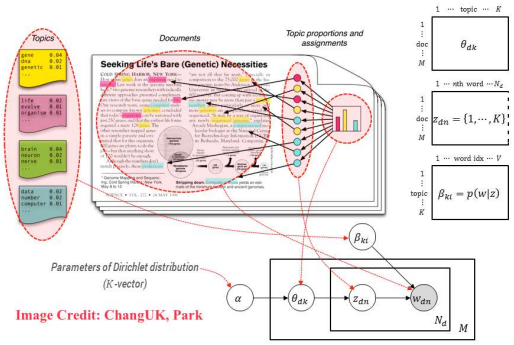
Büyük ölçekte LSI ,

kod lsi_modeli.m



Latent Dirichlet Allocation (LDA) model, Topic Modelling

- Konular kelimelerin karışımı
- Dokümanlar konuların karışımı
- http://littlesaiph.blogspot.com/2012/07/laymans-explanation-of-online-lda.html



kod lda_modeli.m

Kavramsal Ağaç (hiyerarşi) ile Anlamsal Benzerlik Bulma

- İki temel yaklaşım
  - Bağ sayma
    - Taksonomi (kavramsal ağaç) yeterli
  - Ortak/ Müşterek bilgi (Mutual Information)
    - Taksonomi ve derlem(corpus) gerekli

Leacock & Chodorow (1998)

$$sim_{LC}(c_1, c_2) = -\log \frac{len(c_1, c_2)}{2L}$$

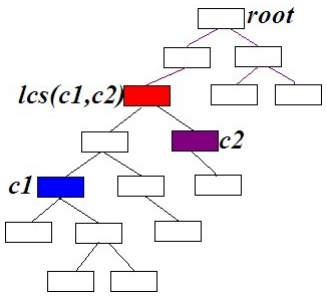
- ***len(c1,c2)*** iki synset arasındaki en kısa yolun uzunluğu. *(benzerlik değeriyle ters orantılı)*
- ***L***, tüm taksonominin derinliği

Wu & Palmer (1994)

$$sim_{Wu\&Palmer}(c_1, c_2) = \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3}$$

•***N1*** ve ***N2***: ***c1*** ve ***c2***'nin en yakın ortak üst synset'lerine (***lcs(c1,c2)***) IS-A bağlarıyla uzaklıkları (*benzerlik değeriyle ters orantılı*)

•***N3***, en yakın ortak üst synset'in kök synset'e IS-A bağlarıyla uzaklığı (*büyüklüğü ortak üst synset'in spesifikliğini gösterir, benzerlikle doğru orantılı*)

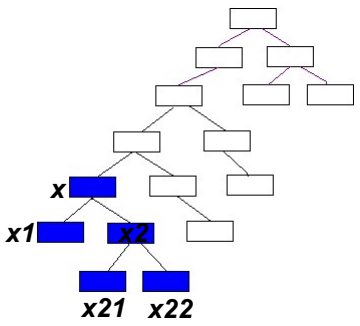


Jiang-Conrath (1997)- Lin (1998)

Bir kavramın bir korpus'taki olasılığı = corpus'ta geçme sayısı (frekansı)

P(concept) = freq(concept)/freq(root)

freq(x) → korpusta ***x*** synset'inin tepesinde bulunduğu
tüm synset'lerin frekanslarının toplamı
freq(x) = f(x)+f(x1)+f(x2)+f(x21)+f(x22)



Jiang-Conrath (1997)- Lin (1998)

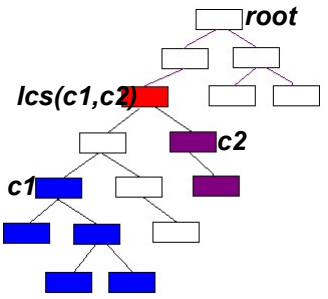
$$sim_{JC}(c_1,c_2) = \frac{2\log(p(lcs(c_1,c_2)))}{\log(p(c_1)) + \log(p(c_2))}$$

A B

$$sim_L(c_1,c_2) = \frac{2*\log(p(lcs(c_1,c_2)))}{\log(p(c_1)) + \log(p(c_2))}$$

A B

lcs(c1,c2) en yakın ortak üst synset



•A, her iki kavramı da içeren en spesifik kavramı kullanır
(İki kavramın beraber geçtiği doküman sayısına benzer)
•B, iki kavramdan herhangi birini içeren doküman sayısına benzer

Karşılaştırma

- Bütün metotlar İngilizce 38 kelime çiftine uygulanmış.
- Bulunan benzerlik değerlerinin, insan yargılarıyla olan korelasyonları yandaki tabloda.

Method	*Type*	*Correlation*
Wu & Palmer 1994	Edge Counting	0.74
Li 2003	Edge Counting	0.82
Leacock & Chodorow 1998	Edge Counting	0.82
Resnik 1999	Info. Content	0.79
Lin 1998	Info. Content	0.82
Lord 2003	Info. Content	0.79
Jiang & Conrath 1998	Info. Content	0.83
Tversky 1977	Feature Based	0.73
Adapted Lesk 2002	Feature Based	0.37*
Rodriguez 2003	Hybrid	0.71

Birbiri yerine kullanılan kelimeler Deniz Yüret, 2009

- W1 W2 X1 W3 W4
- W1 W2 X2 W3 W4
- X1 ve X2 birbiri yerine kullanılmış, aynı bağlamda kullanılmış, demek ki aralarında bir ilişki var.
- X1 ve X2 kaç kere aynı bağlamda kullanıldı ise aralarındaki bağ o kadar güçlü.

Birbiri yerine kullanılan kelimeler

İnci Düzenli'nin Bitirme Projesi'nde bulduğu örnekler:
Format: kelime1 kelime2 kaç farklı bağlamda birbirinin yerine kullanıldığı

dolar ytl 939
bulun olmak 619
avro euro 459
kanun yasa 350
avro dolar 344
oyna yap 309
dolar lira 251
dolar euro 249
milyar milyon 241
meclis tbmm 219
belir kaydet 194
futbol oyun 189
art azal 187
avro ytl 182
alan saha 177
gerek iste 176
lira ytl 173
haziran nisan 161

Kelime kümeleme

- Kelimeleri
  - terim*doküman matrisine göre
  - Birlikte geçme matrisine göre
  - Sınıflarda yer alma olasılıklarına göre

kümelersek aynı kümede yer alan kelimeler birbirine benzerdir diyebiliriz.

Anlamsal Uzay

- Kelimelerin anlamsal benzerliklerinin bir temsili
- Kavramların anlamsal yakınlıklarına göre yer aldıkları bir uzay
- Anlamsal yakınlığa bağlı koordinatlar nasıl oluşturulur?
  - Terim doküman matrislerini ayrıştırma (LSI)
  - Konu modelleri (LDA)
  - Birlikte geçme matrisleri (MDS, Glove)
  - Sinirsel dil modelleri (word2vec, fasttext)

Birlikte Geçme Matrisleri

- Harris der ki:
  - Birlikte kullanılan kavramlar birbirlerine anlamsal olarak benzerler.
- Birlikte kullanım: aynı cümlede, aynı metinde, sabit bir kelime penceresinde

Örnek çalışma-1
Kelimelerin Anlamsal Benzerlik Ölçümü, Amasyalı, 2006

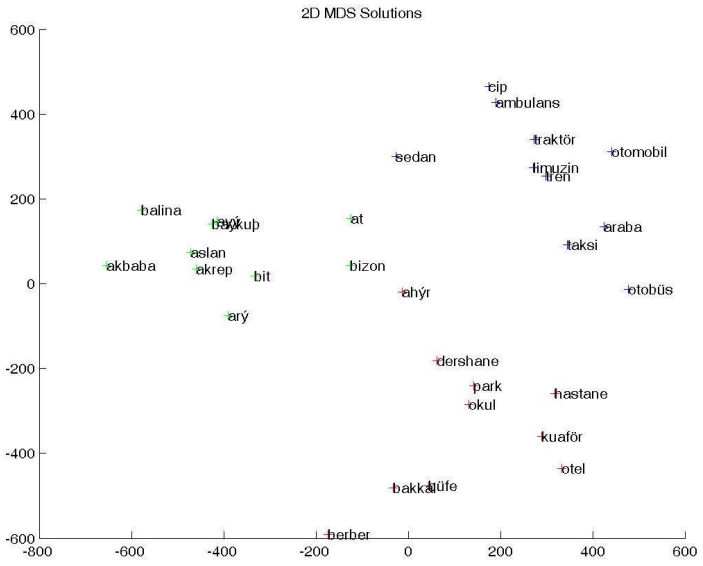
- İki kelimenin Internet'te yer alan sayfaların kaçında yan yana kullanıldıkları bulunarak belirlenmiştir.
- Bunun için arama motoruna

  "kelime1 kelime2" sorgusu gönderilerek gelen sonuç sayfasındaki sonuç sayısı alınmıştır.

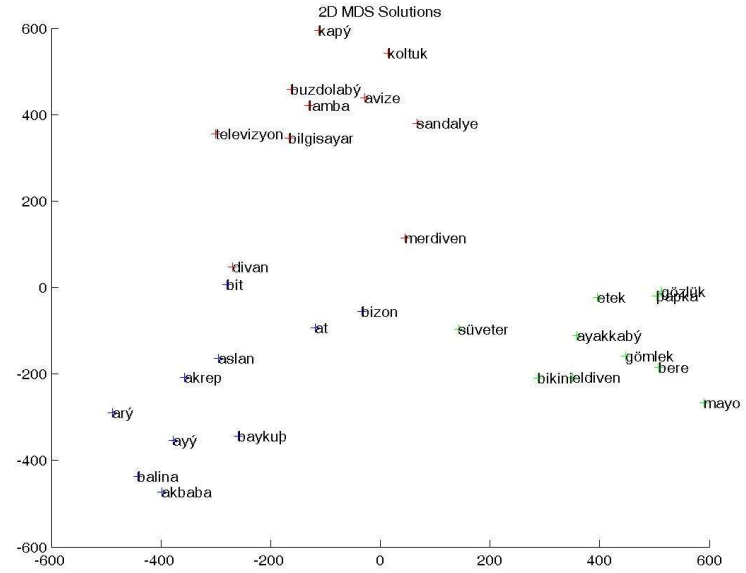
Kelimeler

Veri kümesi-I	*Yer* berber – bakkal- kuaför- hastane- otel- okul- dershane- büfe- ahır- park		*Hayvan* akbaba- arı- baykuş- balina- aslan- at – bizon – akrep - ayı- bit		*Taşıt* otobüs- cip- taksi- ambulans- araba- sedan- tren- limuzin- otomobil- traktör	
Veri kümesi -II	*Ev eşyası* divan - televizyon- avize- merdiven- bilgisayar- buzdolabı- kapı- sandalye- lamba- koltuk		*Hayvan* akbaba- arı- baykuş- balina- aslan- at – bizon – akrep - ayı- bit		*Giyecek* bikini- mayo- bere- süveter- etek- ayakkabı- eldiven- gömlek- şapka- gözlük	
Veri kümesi -III	*Yiyecek* tereyağı kaymak baharat peynir şeker tuz	*Giyecek* ayakkabı eldiven şapka gömlek gözlük etek	*Ev eşyası* avize sandalye lamba koltuk buzdolabı televizyon	*Hayvan* aslan arı at akrep ayı akbaba	*Yer* otel dershane park berber hastane okul	*Taşıt* taksi limuzin ambulans cip otomobil araba

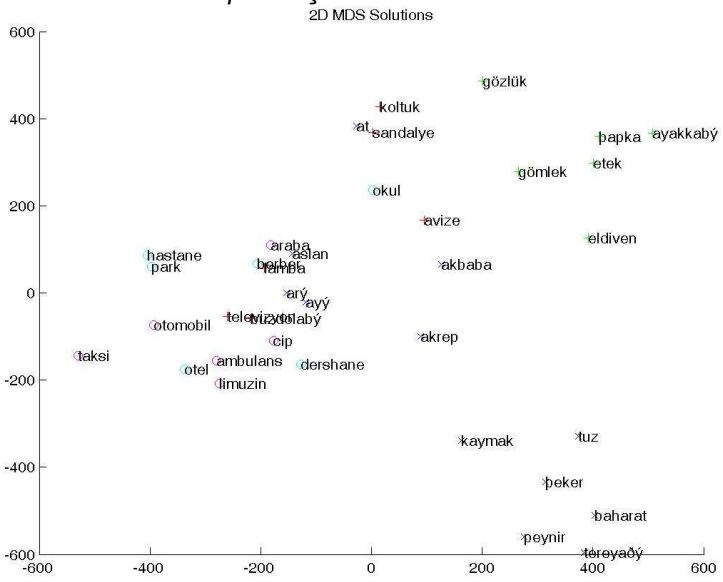
Veri kümesi–1 için Google ile bulunan benzerlik matrisinden hesaplanmış kelime koordinatları



Veri kümesi–2 için Google ile bulunan benzerlik matrisinden hesaplanmış kelime koordinatları



Veri kümesi–3 için Google ile bulunan benzerlik matrisinden hesaplanmış kelime koordinatları

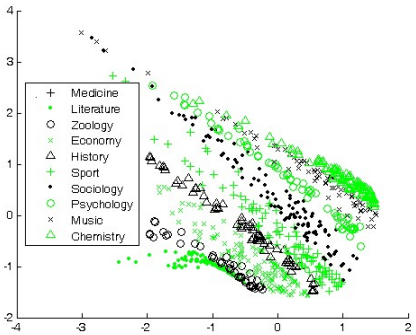


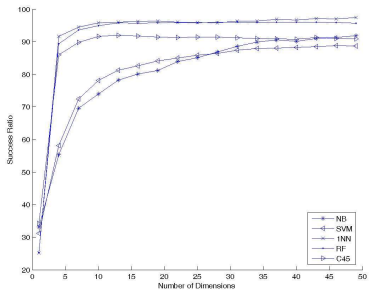
Kelimeleri Sınıflandırma Başarıları

	v1 10 boyut	v1 2 boyut	v2 10 boyut	v2 2 boyut	v3 10 boyut	v3 4 boyut
SVM	96,6	100	83,3	96,6	88,8	77,7
C4.5	83,3	83,3	90	90	77,7	75
RF	90	90	93,3	93,3	72,2	75
EM	66,6	96,6	66,6	96,6	66,6	83,3

Örnek çalışma-2
Teknik Terim sınıflandırma

- 967 teknik terimin koordinatları, 15 bin web sitesi (neden Google değil?) üzerinde ile





Örnek çalışma-3
Metin sınıflandırma
(Amasyalı, Beken, 2009)

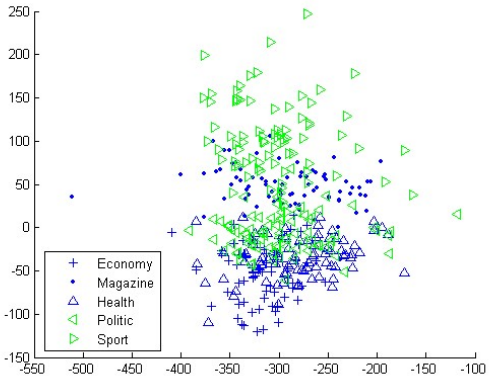
- 5 farklı haber sınıfına (ekonomi, magazin, sağlık, siyasi, spor) ait 230'ar metin
- Her sınıftan 150'şer haber metni eğitim, 80'er adedi test
- PCKimmo ile kelime kökleri
- farklı gövde sayısı yaklaşık 4500

Anlamsal Uzayların Metin Sınıflandırmada Kullanımı

- 4500 kelimenin sayısal karşılıklarını elde etmek için 15.000 web sitesinden oluşan bir külliyat kullanılmış.
- Kelimelerin birbirlerine anlamsal yakınlık matrisi bu külliyatta birlikte geçtikleri doküman sayıları
- uzaklık=1/yakınlık
- MDS (ÇBÖ)
- MDS'de ilk 100 ve ilk 10 boyut alınmış

Anlamsal Uzayların Metin Sınıflandırmada Kullanımı

- Metinler, içerdikleri kelimelerin koordinatlarının ortalamaları
- Metinler kelimelerle aynı boyutlu
- 1 ve 2. boyutlar



Anlamsal Uzayların Metin Sınıflandırmada Kullanımı

Sınıflandırma Algoritması	**100 Boyutlu Metinler**	**10 Boyutlu Metinler**
Lineer Regresyonla Sınıflandırma	**93.25**	89
Pace Regresyonla Sınıflandırma	92.75	88.5
En Yakın Komşu	70	79
Destek Vektör Makineleri	90	89.75
Rastgele Ormanlar (100 ağaçlı)	90.75	87.75

Sınıflandırma Algoritması	**Metinlerin Boyut Sayısı**	**Başarı yüzdesi (tf)**	**Başarı yüzdesi (tfidf)**
Klasik Naive Bayes	4500	87.25	
Diskrit Naive Bayes	4500	85.75	**89.25**
En Yakın Komşu	4500	34.25	43.5
Destek Vektör Makineleri	4500	87	86.5
Rastgele Ormanlar (100 ağaçlı)	4500	X	X
Lineer Regresyonla Sınıflandırma	4500	X	X
Pace Regresyonla Sınıflandırma	4500	X	X
C4.5	4500	74.75	23.5
Lineer Regresyonla Sınıflandırma	100	81.75	84.5
Pace Regresyonla Sınıflandırma	100	81.5	83.25
En Yakın Komşu	100	71.25	76.25
Destek Vektör Makineleri	100	80.25	87.75
Rastgele Ormanlar (100 ağaçlı)	100	85	84.5

Anlamsal Uzayların Metin Sınıflandırmada Kullanımı

- Daha yüksek başarı, Metin temsilinde daha az hafıza
- Benzerlik matrisi bulurken Google neden kullanılmadı?
- Yöntem dilden bağımsız mı?
- Eşsesli kelimelerin durumu?

Örnek çalışma-4

Kelime yörüngeleri (Amasyalı, Yener, Kaplan, 2012)

- *İnsanların düşünce süreçleri birbirine ne kadar benzemektedir?*
- *İnsanların düşünce süreçlerini gözlemlemek, direkt ölçmek mümkün değilse de onu, ürünlerinin bazılarıyla (konuşma ve yazılarıyla) dolaylı olarak gözlemlemek mümkündür.*

Kelime Yörüngeleri

- Yazıları, onları oluşturan kelimelerin zamana göre sıralanmış hali olarak düşünelim.
- Elimizde kelimelerin koordinatları olsa, bu koordinatları yazıdaki sıralarıyla birleştirerek bir yazıyı X boyutlu bir uzayda bir yörünge olarak ifade edebiliriz.

Hesaplamalı Anlambilim Ders Notları

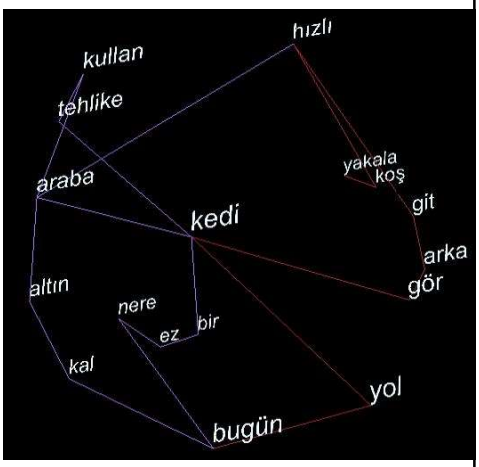
Kelime Yörüngeleri

- Aynı kişinin farklı yazılarının yörüngeleri birbirine benzer midir?
- Farklı kişilerin yörüngelerini birbirinden ayırmak mümkün müdür?
- Cevaplar için önce yörüngeleri oluşturmak gerekir. Bunun için ise önce kelime koordinatlarını bulmak gerekir.

NOVA Research Lab

Kelime Yörüngeleri

- Bugün yolda kedi gördüm. Arkasından gittim. Hızlı koşuyordu. Yakalayamadım.
- Hızlı araba kullanmak tehlikelidir. Kediler arabaların altında kalabilirler. Bugün neredeyse eziliyordu bir kedi.



Yörüngelerin Özellikleri

- 2 özellik:
- Kavramlar arası mesafeler,
- Kavramlar arası açılar.
- *(n, n+1, n+2, n+3)* arka arkaya gelen 4 koordinat olmak üzere *(n,n+1), (n,n+2), (n,n+3)* arası mesafelerin 10'luk histogram değerleri, frekansları ve *(n,n+1,n+2), (n,n+2,n+3)* arası açıların (PI cinsinden) 10'luk histogram değerleri, frekansları

Yörüngelerin Özellikleri

- *d* uzunluğundaki bir yörünge de *d-1* adet *(n, n+1)* arası mesafe ölçülmektedir. Bu *d-1* ölçümün eşit aralıklı 10 parçalık histogramı çıkarılmaktadır. Bu histogramın 10 adet değeri ve her değerin bir frekansı bulunmaktadır. Dolayısıyla yörüngenin *(n, n+1)* arası mesafelerini ifade eden 20 adet özellik çıkarılmaktadır. Bu işlem *(n,n+2), (n,n+3)* arası mesafelere ve *(n,n+1,n+2), (n,n+2,n+3)* arası açılara da uygulandığında bir metni ifade eden yörüngeye ait 100 özellik bulunmuş olmaktadır

Deneysel Sonuçlar

2 yazara ait 35'er yazı,
Bir yazı kime ait?
10'lu çapraz geçerleme sonuçları

Algoritma	Sınıflandırma Başarısı (%)
C4.5	97.15
Naive Bayes	97.15
En yakın komşu	95.71
Destek Vektör Makineleri	98.57

2 yazara ait 70 yazının (n, n+1) arası mesafelerinin histogramının ilk parçasının frekanslarının histogramı

Her yazara ait 35'er yazı,
Bir yazı kime ait?
10'lu çapraz geçerleme sonuçları

Algoritma	3 yazar	4 yazar	8 yazar
C4.5	91.43	90.71	53.21
Naive Bayes	95.24	92.14	62.5
En yakın komşu	80.95	62.14	50
Destek Vektör Makineleri	94.29	85.71	68.57
Zero Rule	33.33	25	12.5

Sonuç ve Tartışma

- 2 kişiye ait 35'er yazıdan oluşan veri kümesinde bir metnin yazarını tanıma başarısı %98
- Yazar sayısı arttıkça başarının düşüyor (8 yazar için %68)
- Metinde kullanılan kavramlardan bağımsız olması (sadece yörüngenin özelliklerinin kullanılıyor olması) aynı kişinin farklı konularda yazdıklarından elde edilen yörüngeler benzer

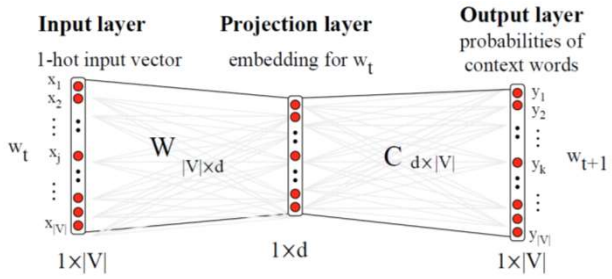
Olası kullanım alanları

- Psikolojik hastalıkların tespiti
- Psikolojik hastalıkların düşünce süreçleri üzerindeki etkilerinin araştırılması
- Cinsiyet, yaş, eğitim farklılıklarının düşünce süreçleri üzerindeki etkilerinin araştırılması

Sinirsel Dil Modelleri

- Word2vec*, fasttext**
- Tek gizli katmana sahip bir yapay sinir ağı
- C=2 için bağlam = (w_{t-2},w_{t-1},w_t,w_{t+1},w_{t+2})
- Skip-gram: giriş: w_t, çıkış: (w_{t-2},w_{t-1},w_{t+1},w_{t+2}) in herbiri
- C-bow: giriş: (w_{t-2},w_{t-1},w_{t+1},w_{t+2}) in herbiri, çıkış: w_t

Skip-gram modeli



* 2013, T. Mikolov ve ark. "Distributed representations of words and phrases and their compositionality"

** 2017, P. Bojanowski ve ark. "Enriching word vectors with subword information"

Sinirsel Dil Modelleri

- W matrisi kelime vektörleri
- Başlangıçta rasgele atanırlar. Eğitim bittiğinde kullanıma hazırlar.
- Amaç: bir kelimenin (w) genelde etrafında bulunan kelimelerle (c) benzer vektörlere, diğerleriyle (w_i) farklı vektörlere sahip olması
- 2 vektör benzer ise iç çarpımları yüksektir.
- Negatif örneklemeli Skip-gram için hata fonksiyonu:

$$\log \sigma(c \cdot w) + \sum_{i=1}^{\kappa} \mathbb{E}_{w_i \sim p(w)} \left[\log \sigma(-w_i \cdot w) \right]$$

$$\sigma(x) = \frac{1}{1+e^x}$$

https://medium.com/@Aj.Cheng/word2vec-3b2cc79d674
https://web.stanford.edu/~jurafsky/slp3/slides/vector2.pdf

Eğitim kümesi oluşturmak

- Text: w1 w2 w3 w4 w5 w6 w7 …
- Skip gram için eğitim örnekleri (giriş, çıkış) C=2 için:
  - benzer olmasını istediğimiz ikililer: (w3,w1), (w3,w2), (w3,w4), (w3,w5)
  - farklı olmasını istediğimiz ikililer: (w3, t1), (w3, t2), (w3, t3) … (w3, tK) burada t1…tK $p(w_i)$ ye göre rasgele seçilen kelimeler
- Sonra w4 ile devam …
- C-bow için ikilileri tersine çevir

Fasttext

- Birlikte bulunmaya ek olarak kelime içi n-gram benzerliklerini (sub-word similarity) de dikkate alır.
- Kelimeleri n-gram larına parçalar ve her bir n-gram içinde vektörler bulur.
- Bu sayede sözlükte olmayan (test örneklerinde çok karşılaşılan bir durum) kelimelerin vektörlerine bulur.
- Türkçe gibi ek alan diller için avantajlı (aynı kelimenin farklı ekler almış halleri, yanlış yazılmış halleri birbirine yakın koordinatlara sahip olur)
- Ama daha yavaş

kodlar

- word2vec_modeli_hazir.m (kullanım)
- word2vec_modeli_egitim.m (eğitim)
- emb42_ngramsiz.mat (kelime vektörleri)
- fast_emb.mat (kelime vektörleri)
- haber42bin.mat (derlem)

Hazır fasttext* vektörleri ile

ilk satırdaki kelimelere en yakın 19 kelime

"cimbom"	"terlik"	"virüs"	"elma"	"aslan"	"koş"	"bırak"
"cimbombom"	"terliksi"	"virüsun"	"elma,armut"	"aslaner"	"koşo"	"bıraktın"
"lik–"	"terlikler"	"virüslü"	"armut"	"aslana"	"koşoy"	"bırakın"
"penaltılarla"	"terlikleri"	"virüsle"	",armut"	"aslani"	"koşaç"	"bırakamam"
"#beşiktaş"	"noterlik"	"virüse"	"şeftali"	"aslankeser"	"koşa"	"bırakmam"
"rövanşta"	"kasiyerlik"	"virüstür"	"meyveş"	"aslanı"	"koşağaç"	"bıraz"
"lik/"	"önlük"	"virüsü"	"meyva"	"aslanın"	"koşovi"	"bıraksın"
"penaltısonuç"	"yeterlik"	"virüsler"	"elmaağacı"	"aslanyürek"	"koşun"	"bıraktım"
"beşiktaşlı"	"dülgerlik"	"virüsün"	"kayısı"	"aslanbaş"	"koşamaz"	"bırakayım"
"cimbilli"	"dizlik"	"virüsten"	"meyvelik"	"aslanhan"	"koşalı"	"bırakırsan"
"gülesin"	"göynek"	"virüsteki"	"ayva"	"özaslan"	"koşacı"	"bıra"
"galatasaray"	"yerlik"	"virüsleri"	"üzüm"	"aslanbay"	"koşaca"	"bırakırsam"
"iüsk"	"ayakkabıcılık"	"virüslere"	"elmakaya"	"aslanduz"	"koşköpür"	"bırakmayın"
"eledim"	"tonmaysterlik"	"mimivirüs"	"buğday,şeker"	"aslanpençesi"	"koşay"	"bıraktık"
",fenerbahçe"	"poşu"	"virütik"	"badem"	"aslantuğ"	"koşan"	"bırakmalı"
"haftabjk"	"köynek"	"virüslerde"	"arpa,buğday"	"aslanla"	"koşrabat"	"bırakalım"
"beşiktaş"	"yakıcılık"	"virüsleri#grup"	"armutu"	"aslanları"	"belıy"	"bırakalı"
"rövanşında"	"rençberlik"	"poliovirüs"	"şeftalinin"	"sarıaslan"	"koşanlar"	"bırakmadım"
"cim"	"terli"	"virüslerin"	"elmax"	"aslanlar"	"koşunca"	"bıraksam"
"kupadaki"	"üzerlik"	"virüslerle"	"armudu"	"yavuzaslan"	"koşkin"	"bırakma"

*Fasttext vektörleri 416 bin kelime içermektedir.
Ngram benzerliği de kullanır.

42 bin haber metniyle eğitilen vektörler*, ngram benzerliği yok ilk satırdaki kelimelere en yakın 19 kelime

"cimbom"	"istanbul"	"virüs"	"elma"	"aslan"	"ölmek"	"bırak"
"kırmızılılar"	"sergi"	"virüsü"	"meyve"	"aslanın"	"yaşamak"	"bırakın"
"beşiktа"	"gerçekleşecek"	"virüsün"	"meyveler"	"lakaplı"	"istemiyorum"	"dedikleri"
"wesley"	"sahipliğinde"	"bulaşan"	"lif"	"cimbom"	"asker"	"kardeşim"
"didier"	"katılımıyla"	"hiv"	"sebze"	"kırmızılı"	"erbakan"	"üzeri"
"galatasaraya"	"fashion"	"hastalık"	"limon"	"kırmızılılar"	"chavez"	"git"
"transfere"	"konser"	"yoluyla"	"çürük"	"serhat"	"lütfen"	"dediğini"
"kırmızılı"	"buluşacak"	"grip"	"salata"	"devler"	"bekleyen"	"hayattan"
"kırmızılıların"	"sarayında"	"aşı"	"pirinç"	"didier"	"kanserin"	"yahu"
"eboue"	"inci"	"zararlı"	"buğday"	"muhammet"	"ameliyata"	"genişletilmiş"
"cim"	"sergisi"	"yazılımı"	"zeytinyağı"	"toure"	"anın"	"memnunum"
"sarıkırmızılı"	"cnr"	"bilgisayar"	"patates"	"drogbayı"	"rio"	"aşkına"
"devler"	"seçkin"	"solunum"	"çiğ"	"lokomotiv"	"aileme"	"temsilcimiz"
"aysal"	"global"	"enfeksiyonları"	"muz"	"payını"	"ölümü"	"etkisine"
"ezeli"	"modern"	"kullanıcının"	"ürünlerde"	"ulaş"	"istemediğini"	"özetle"
"terimin"	"festival"	"belirtileri"	"fındık"	"oya"	"ağabeyim"	"dedim"
"ligindeki"	"louis"	"ağrısına"	"sarımsak"	"bulut"	"koydular"	"açıklamalarda"
"deplasmanında"	"dünyaca"	"enfeksiyon"	"portakal"	"yitirdi"	"özbeğenin"	"gidin"
"futbolcunun"	"konseri"	"bulaşıcı"	"hapı"	"ceren"	"yasak"	"yazdı"
"beraberlik"	"tarihleri"	"saldırganlar"	"ihracatı"	"ekrem"	"söylemiş"	"oynuyor"

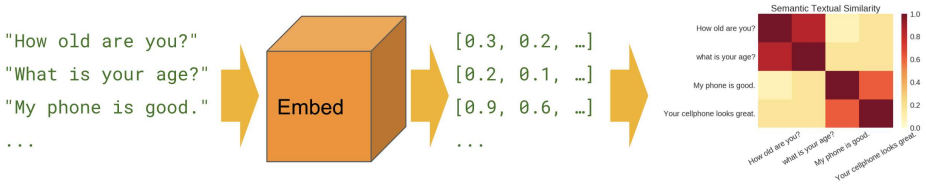
*Vektörler 19 bin kelime içermektedir.

Kelime vektörlerinin başarısını nasıl ölçeriz?

- Dahili / direkt yöntemler
  - İnsanların verdikleri cevaplarla korelasyon
  - Analoji veri kümeleri:
    - https://aclweb.org/aclwiki/Google_analogy_test_set_(State_of_the_art)
    - https://www.aclweb.org/anthology/C16-1332.pdf
- Harici / dolaylı yöntemler
  - Çeşitli NLP uygulamalarında kullanımları üzerinden

Cümle ve metin vektörleri

- En basit yol: içindeki kelimelerin ortalaması
- Universal sentence encoder https://arxiv.org/abs/1803.11175



Kelime Öbekleri

- 2, 3, 4 uzunluğunda genelde birlikte kullanılan kelime birlikleri içerdikleri her bir kelimeden ayrı bir bir kavramı ifade edebilirler.
- “Fatih Sultan Mehmet” ifadesi, “Fatih”, “Sultan” ve “Mehmet” in her birinden, farklı bir kavramı ifade eder.

Kelime Öbekleri

- Bu ifadeleri nasıl buluruz?
  - En basiti: frekansı en yüksek olanlar
  - Kelime öbeği olma olasılığı (ti):
    - ti nin frekansı/ti nin içerdiği tekil kelimelerin frekanslarının minimumu
    - log2(ti nin frekansı/ ti nin içerdiği tekil kelimelerin frekanslarının çarpımı)
- En basitini deneyelim.

Kod kelime_obekleri.m

- 42bin haber metni içindeki 1, 2, ve 3 uzunluğunda kelime ngramlarını bulalım
- Nadir kelimeleri silmezsek 11 milyon ngram çıkıyor ☹
- 50 den az geçen kelimeleri silersek
  - kelime silme oncesi tekil kelime sayısı= 356.680
  - kelime silme oncesi dokumanlardaki toplam kelime sayısı= 9.986.668
  - silinen tekil kelime sayısı= 336.019
  - kelime silme sonrasi dokumanlardaki toplam kelime sayısı= 8.320.579

```
Counts: [41992×8141753 double]
Vocabulary: [1×20661 string]
Ngrams: [8141753×3 string]
NgramLengths: [1 2 3]
NumNgrams: 8141753
NumDocuments: 41992
```

Ngram sayısı 8 milyon

Frekansı en yüksek 1 2 3 kelime ngramlar

Ngram			Count	NgramLength
"ve"	""	""	1.8454e+05	1
"bir"	""	""	1.4992e+05	1
"□"	""	""	1.2161e+05	1
"bu"	""	""	93130	1
"da"	""	""	73033	1
"de"	""	""	70376	1
"için"	""	""	57285	1
"ile"	""	""	54426	1
"çok"	""	""	39307	1
"olarak"	""	""	34066	1

Ngram			Count	NgramLength
"□"	"nin"	""	9411	2
"□"	"ın"	""	9061	2
"□"	"de"	""	8796	2
"□"	"da"	""	8793	2
"□"	"in"	""	7087	2
"□"	"nın"	""	6578	2
"türkiye"	"□"	""	6541	2
"diye"	"konuştu"	""	6450	2
"□"	"dedi"	""	6240	2
"ya"	"da"	""	5226	2

Ngram			Count	NgramLength
"türkiye"	"□"	"nin"	2187	3
"türkiye"	"□"	"de"	2171	3
"□"	"diye"	"konuştu"	2168	3
"spor"	"toto"	"süper"	1580	3
"şöyle"	"devam"	"etti"	1509	3
"başbakan"	"recep"	"tayyip"	1407	3
"türkiye"	"□"	"ye"	1017	3
"recep"	"tayyip"	"erdoğan"	996	3
"yönetim"	"kurulu"	"başkanı"	983	3
"bir"	"kez"	"daha"	909	3
"şunları"	"söyledi"	"□"	744	3
"i'stanbul"	"□"	"da"	728	3
"başta"	"olmak"	"üzere"	720	3
"toto"	"süper"	"ligde"	717	3
"genel"	"başkan"	"yardımcısı"	708	3
"çok"	"önemli"	"bir"	633	3
"erdoğan"	"□"	"in"	627	3
"şöyle"	"konuştu"	"□"	576	3
"olduğunu"	"ifade"	"eden"	561	3
"aa"	"muhabirine"	"yaptığı"	559	3
"bir"	"süre"	"sonra"	556	3
"muhabirine"	"yaptığı"	"açıklamada"	555	3
"bir"	"an"	"önce"	551	3
"olay"	"yerine"	"gelen"	526	3

- 8 milyon ngramdan sadece 36 bin'i 50 den fazla kez geçiyor. Sadece bunlarla ilerleyelim. Tekil kelime sayısı 20 bin.
- Ngramları kendi içlerindeki kelimelerden ayrı kendi başlarına anlamları / koordinatları olsun diye tek kelimelere çevirelim.

"bitirdikten sonra"	"bitirdiktensonra"
"işlemlerinde"	"işlemlerinde"
"ısrarlı"	"ısrarlı"
"dağıldı"	"dağıldı"
"fırtınası"	"fırtınası"
"yer alırken"	"yeralırken"
"700 milyon"	"700milyon"
"sırada yer alırken"	"sıradayeralırken"
"telefonundan"	"telefonundan"
"yi˙ne"	"yi˙ne"
"kız kardeşi"	"kızkardeşi"
"cep telefonundan"	"ceptelefonundan"
"i˙nternet üzerinden"	"i˙nternetüzerinden"
"başbakan da"	"başbakanda"
"ve şöyle"	"veşöyle"
"bir tane"	"birtane"

Haberlerin yeni hali

- (1,1) 111 tokens: ortak vizyonumuz var dışişleribakanı davutoğlu yunanistan iletürkiye arasındaki farlılıkların ortak vizyon ile çözülebileceğini söyledi dışişleribakanıahmet davutoğluve yunan mevkidaşı dimitris avramopulos başbaşa ve heyetler arası görüşmelerinardından ortakbasıntoplantısı düzenledi davutoğlu yunanistanda kendisine gösterilen sıcak karşılama içinteşekkür ederek ziyaretinin türkyunan dostluğunun bir nişanesi olduğunudile getirdi ikiülke arasında görüş ayrılıkları vefarklı yaklaşımlar olabileceğini amaaynı vizyonun paylaşılmasıyla aradaki sorunların problemlerin daharahat çözülebileceğini ifadeeden bakan davutoğlu samaras ve avramopulos ile görüşmelerinden eldeettiği sonucun dabu yönde olduğunu söyledi davutoğlu ikiyılönce ydiki bu anlayışla kurduklarını yunanistanda istikrarlıbir hükümet kurulur kurulmaz da toplantının ikincisini gerçekleştirmeyi planladıklarını vebugün ön hazırlıkları önemliölçüde tamamladıklarını söyleyerek birçok alanda ortak mutabakatların ilk tohumları atıldı ...

42 bin haber metniyle eğitilen kelime ngramı vektörler*, karakter ngram benzerliği yok ilk satırdaki kelimelere en yakın 19 kelime

"cimbom"	"istanbul"	"virüs"	"elma"	"aslan"	"ölmek"	"bırak"
"galatasaray"	"sergi"	"virüsün"	"sebze"	"aslanın"	"giymek"	"git"
"g"	"gerçekleşecek"	"virüsü"	"soğan"	"hakan"	"direniş"	"dedim"
"saray"	"festival"	"bulaşan"	"meyveve"	"pelin"	"benim"	"başkabirşey"
"galatasarayın"	"fashion"	"bilgisayara"	"domates"	"irem"	"okumak"	"gel"
"sarıkırmızılılar"	"istanbul"	"yazılımı"	"patates"	"sevgilisiyle"	"cenazetöreni"	"bul"
"drogba"	"filarmoni"	"bulaşıcı"	"çürük"	"yalçınkaya"	"üzereyken"	"yahu"
"terim"	"konser"	"hastalık"	"salata"	"mehmet"	"benim"	"demesiüzerine"
"fenerbahçe"	"fuarı"	"bilgisayar"	"portakal"	"kaptı"	"ölüm"	"hiç"
"sarıkırmızılı"	"lütfikırdar"	"bakteri"	"tavuk"	"yaşar"	"şiiri"	"diyecek"
"realmadrid"	"festivali"	"enfeksiyon"	"meyve"	"naz"	"yaşamak"	"kızım"
"sneijder"	"solo"	"bağışıklık"	"tadı"	"neslihan"	"öldürmekle"	"geridönen"
"f"	"orkestrası"	"kullanıcıların"	"kabak"	"yusuf"	"hastalığınedeniyle"	"mahsun"
"fildişili"	"sanatseverlerle"	"antibiyotik"	"ürünlerin"	"abdurrahman"	"istemiyorum"	"görmüyorum"
"sarıkırmızılıların"	"konseri"	"şebeke"	"muz"	"bom"	"sormak"	"ama"
"galatasarayda"	"bilkent"	"tedavisinde"	"üzüm"	"hasan"	"geldin"	"müddetçe"
"beşiktaş"	"grand"	"telefonlar"	"şekerli"	"erol"	"bana"	"bilmiyorum"
"ligi"	"prixsi"	"dolandırıcılık"	"sebzeve"	"2523"	"ogün"	"sana"
"cim"	"caz"	"grip"	"kiraz"	"çini"	"çocuğumu"	"sapan"
"şampiyonlar"	"konserler"	"porno"	"meyveler"	"shanghai"	"evlenmek"	"hiçkimse"

*Vektörler 26 bin kelime/ngram içermektedir.

Faydalı bağlantılar

- Dil modellerini değerlendirmek: https://thegradient.pub/understanding-evaluation-metrics-for-language-models/
- Klasik dil modelleme: https://github.com/vrann/next-word-prediction
- Generalization through Memorization: Nearest Neighbor Language Models: https://arxiv.org/abs/1911.00172