Metin Sınıflandırma

Mehmet Fatih AMASYALI

Kemik

Akış

- Görev
- Eğiticili Eğiticisiz Öğrenme
- Metin Özellikleri
- Metin Kümeleme
- Özellik Belirleme
  - Çok Boyutlu Verilerle Çalışmak
  - Özellik Belirleme Metotları
    - Stop - Functional words
    - Ağırlıklandırma
    - Gövdeleme
    - Filtreler (Information Gain, S2N vs.) (*Zaten Görmüştük!*)
    - Kelime Grupları
    - Kelime Koordinatları
    - Projeksiyonlar (LSI, PCA, LDA)
    - Metin resimleri
- Metin Sınıflandırmada bir Metot: Naive Bayes

Kemik

Görev

- Verilen: bir metin kümesi
- İstenen: metinlerin kategorilere ayrılması

- Örnekler:
  - Haber metinleri: POLİTİK, SPOR, SAĞLIK, MAGAZİN vs. haber başlıklarına ayırmak
  - Web siteleri: EĞİTİM, EĞLENCE, BİLİM vs. türlerine ayırmak, bir sayfaya benzeyen diğer sayfaların bulunması (Arama motorlarındaki gibi)
  - E-mailler: İSTENEN, İSTENMEYEN şeklinde ayırmak
  - Bir metnin yazarını/dilini bulmak

Kemik

EĞİTİCİLİ- EĞİTİCİSİZ

- Elimizdeki örneklerin etiketleri varsa eğiticili, yoksa eğiticisiz metotlar kullanılır.
- Eğiticili→ sınıflandırma
- Eğiticisiz → kümeleme

Kemik

Metin Özellikleri

- Metinleri ifade etmek için kullanılan özellikler:
  - Kelimeler
  - Kelime türleri
  - Ngramlar
  - Ekler (Morfolojik analiz, Zemberek)
  - Ek türleri
  - ... ?

Kemik

Yazar belirlemede kullanılan özellikler

ID	Style Markers	ID	Style Markers
1	Num. of sentences	12	Avg. Num. of pronoun in a sentence
2	Num. of words	13	Avg. Num. of conjunctions in a sentence
3	Avg. Num. of words in a sentence	14	Avg. Num. of exclamations in a sentence
4	Avg. word length	15	Num. of points
5	Num. of different words	16	Num. of commas
6	Word richness	17	Num. of colons
7	Avg. Num. of nouns in a sentence	18	Num. of semicolons marks
8	Avg. Num. of verbs in a sentence	19	Num. of question marks
9	Avg. Num. of adj. in a sentence	20	Num. of exclamation marks
10	Avg. Num. of adverb in a sentence	21	Num. of inverted / Num. of all sentences
11	Avg. Num. of particle in a sentence	22	Num. of incomplete / Num. of all sentences

Kemik

Metinlerin Kelime Frekanslarıyla İfadesi

Örnek metin

Manchester United won 2 – 1 against Chelsea , Barcelona tied Madrid 1 – 1 , and Bayern München won 4 – 2 against Nürnberg

Metnin kelime sayılarıyla ifadesi

Manchester	1	0.04
United	1	0.04
won	2	0.08
2	2	0.08
–	3	0.12
1	3	0.12
against	2	0.08
Chelsea	1	0.04
,	2	0.08
Barcelona	1	0.04
tied	1	0.04
Madrid	1	0.04
and	1	0.04
Bayern	1	0.04
München	1	0.04
4	1	0.04
Nürnberg	1	0.04

Kemik

Her metinde aynı kelimeler yer almaz

- dokümanlar * kelimeler

$$\begin{pmatrix} & T_1 & T_2 & \dots & T_t \\ D_1 & d_{11} & d_{12} & \dots & d_{1t} \\ D_2 & d_{21} & d_{22} & \dots & d_{2t} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ D_n & d_{n1} & d_{n2} & \dots & d_{nt} \end{pmatrix}$$

Kemik

Metinlerin N-gram'larla İfadesi

- Kelime
- Karakter

D1: "Army troops searched for nuclear weapons."

D2: "Military personnel investigated reports of dirty bombs"

	D1	D2
army troops	1	0
dirty bombs	0	1
for nuclear	1	0
investigated reports	0	1
military personnel	0	1
nuclear weapons	1	0
of dirty	0	1
personnel investigated	0	1
reports of	0	1
searched for	1	0
troops searched	1	0

İki metnin kelime bigramları ile ifadesi

Kemik

Metin Kümeleme

- Hiyerarşik kümeleme
- K-means
- SOM

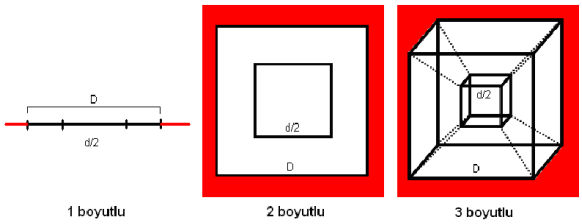
Kemik

Özellik Belirleme

- Metinleri özellikle kelime frekanslarıyla ifade edildiğinde verisetimizin boyut sayısı çok yüksek olur. (binler)
- Çok yüksek boyutta işlem yapmak iyi değil
- Neden?
- Bir sebep işlem hızı
- Başka?

Kemik

Çok Boyutlu Verilerle Çalışmak-1



1 boyutlu 2 boyutlu 3 boyutlu

Boyut Sayısı	Merkeze daha yakın noktaların oranı (%)
1	50
2	25
3	12,50
…	…
P	$(1/2)^P$

Boyut sayısı arttığında verilerin çok büyük bir kısmı sınıfları ayıran sınırlara çok yakın yerlerde bulunacağından sınıflandırma yapmak zorlaşmaktadır.

Kemik

Çok Boyutlu Verilerle Çalışmak-2

- Tek boyutlu uzayda [0,1] aralığı temsil eden 10 nokta
- Rastgele bir noktanın, uzayı temsil eden noktalardan en yakın olanına ortalama uzaklığı = 0.5
- İki boyutlu uzayda rasgele bir noktanın en yakın noktaya olan ortalama uzaklığının düşey ya da dikey (manhattan) 0.5 olması için gerekli temsilci nokta sayısı = 100

Boyut Sayısı	Gerekli temsil eden nokta sayısı
1	10
2	100
3	1000
…	…
p	10^p

Doğru sınıflandırma yapmak için gereken örnek sayısı artıyor.

Kemik

Özellik Belirleme Metotları

- Stop - Functional words
- Ağırlıklandırma
- Gövdeleme
- Filtreler (Information Gain, S2N vs.)
- Özellik alt küme seçicileri (Wrappers)
- Projeksiyonlar (LSI, PCA, LDA)
- Kelime Grupları
- Kelime Koordinatları

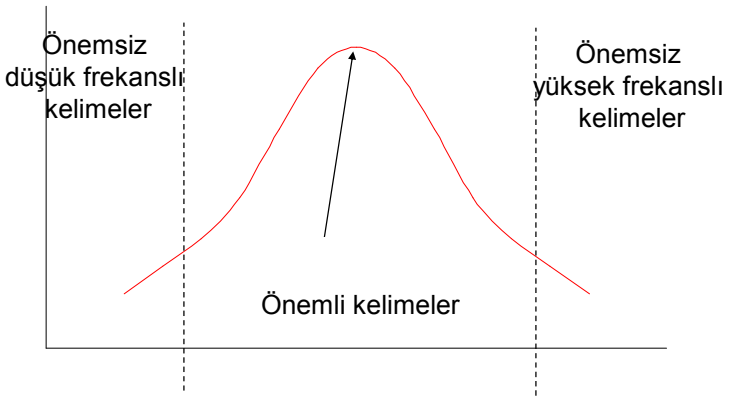
Kemik

Stop - Functional words

- Metinlerde geçen bütün kelimeleri kullanmak yerine bir kısmını silsek:
  - "bir, ben, o, ve" gibi frekansı çok yüksek, ancak bir anlam ifade etmeyen (?) kelimeler (Stop-word elimination)
  - Bütün dokümanlarda sadece 1-3 kere geçen düşük frekanslı kelimeler (Document frequency thresholding)

Kemik

Stop - Functional words



Kemik

Ağırlıklandırma

- TF*IDF = kelime frekansı * ters doküman frekansı
- t_k kelimesinin d_j dokümanı için ağırlığı

$$tfidf(t_k, d_j) = \#(t_k, d_j) \cdot \log \frac{|Tr|}{|Tr(t_k)|}$$

  - $\#(t_k, d_j)$: t_k kelimesinin d_j dokümanında geçme sayısı
  - Tr: tüm dokümanların sayısı
  - $Tr(t_k)$: içinde en az bir kere t_k kelimesi geçen doküman sayısı

Kemik

Ağırlıklandırma ama Neden?

Bütün dokümanlarda geçen kelimelerin önemini azaltmak için.

$$tfidf(t_k, d_j) = \#(t_k, d_j) \cdot \log \frac{|Tr|}{|Tr(t_k)|}$$

B A

A'nın payı ve paydası birbirine eşit/yakın olursa 1'e yaklaşır

B'nin içi 1'e yaklaşırsa B de 0'a yaklaşır ve terimin ağırlığı azalır.

Kemik

Gövdemele (Stemming)

- Özellikle Türkçe gibi eklemeli diller için gerekli
- Ağaçlarımı = ağaçlarını = ağaç

Kemik

Kelime gruplama

- Her kelime için aynı türden dokümanlardaki geçme sayılarının ortalamasını alırsak;
- Her kelime sınıf sayısı boyutunda bir vektörle ifade edilir.
- Bu kelimeleri kümeleme metotlarıyla kümelersek, X adet küme elde ederiz.
- Gerçek veri setimizdeki kelimeler yerine bu kümeleri alırsak, özellik sayımız kelime sayısı yerine küme sayısına düşer.
- Yeni veri seti oluşturulurken, küme içindeki kelimelerin toplam geçiş sayısı, kümenin geçiş sayısı olacaktır.

Kemik

Kelime gruplama-Örnek

- Veri seti (6 boyutlu 6 metin)

kelimeler	metin1	metin2	metin3	metin4	metin5	metin6
balık	1	6	2	1	6	5
kedi	2	8	1	0	6	6
aslan	0	9	0	2	8	6
araba	5	1	2	8	1	0
limuzin	7	0	9	8	1	0
tren	8	0	7	4	0	1
metin sınıf	taşımacılık	hayvanlar	taşımacılık	taşımacılık	hayvanlar	hayvanlar

Kemik

Kelime gruplama-Örnek

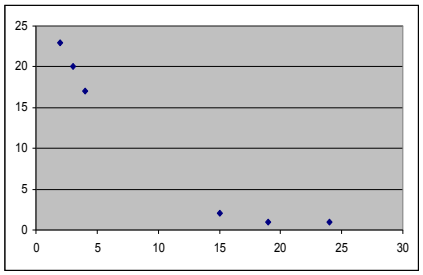
Herbir kelimenin doküman türlerinde kaç defa geçtiği

	taşımacılık	hayvanlar
balık	4	17
kedi	3	20
aslan	2	23
araba	15	2
limuzin	24	1
tren	19	1

Kemik

Kelime gruplama-Örnek

- Kelimeleri gruplama

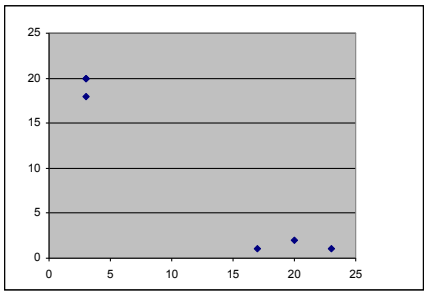


	taşımacılık	hayvanlar
balık	4	17
kedi	3	20
aslan	2	23
araba	15	2
limuzin	24	1
tren	19	1

Kemik

Kelime gruplama-Örnek

- Metinlerin yeni boyutlarda ifadesi (2 boyutlu 6 metin)



	metin1	metin2	metin3	metin4	metin5	metin6
küme1	3	23	3	3	20	17
küme2	20	1	18	20	2	1
metin sınıf	taşımacılık	hayvanlar	taşımacılık	taşımacılık	hayvanlar	hayvanlar

Kemik

Kelime Koordinatları

- Öncelikle metinlerin içinde geçen kelimelerin koordinatları bulunur.
- Metinler içinde geçen kelimelerin koordinatlarının ortalamasıyla ifade edilir.

Kemik

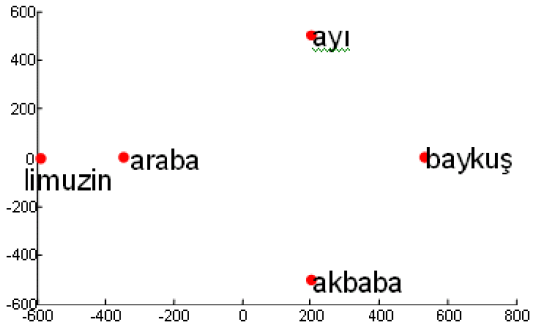
Kelime Koordinatları-Örnek

- Kelimelerin koordinatlarını bulmak için birlikte geçtikleri doküman sayılarından yararlanılır.
- Birlikte geçtikleri doküman sayılarını nasıl bulabiliriz?
  - Kendi veri kümemizle
  - Ya da?

	akbaba	ayı	baykuş	araba	limuzin
akbaba		0	20	1	0
ayı	0		33	4	0
baykuş	20	33		0	0
araba	1	4	0		38
limuzin	0	0	0	38	

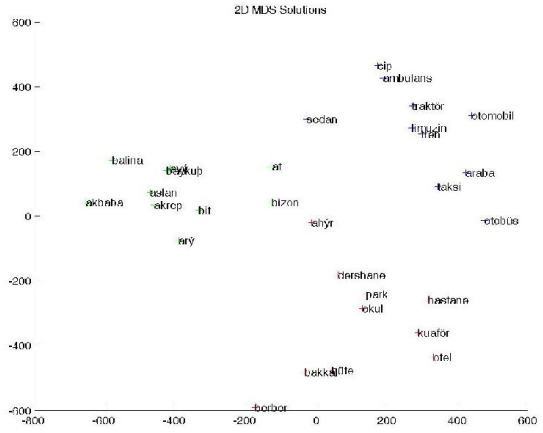
Kemik

Benzerlik Matrisinden "Çok Boyutlu Ölçekleme(ÇBÖ)"/ "Multidimensional Scaling(MDS)" ile elde edilen kelime koordinatları



Kemik

Yer	*Havyan*	*Taşıt*
berber – bakkal- kuaför- hastane- otel- okul- dershane- büfe- ahır- park	akbaba- arı- baykuş- balina- aslan- at – bizon – akrep - ayı- bit	otobüs- cip- taksi- ambulans- araba- sedan- tren- limuzin- otomobil- traktör



Kemik

Projeksiyonlar

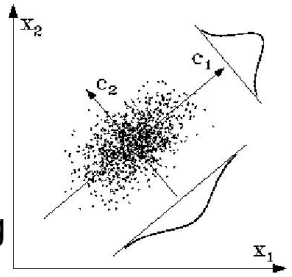
- Verileri n boyuttan r boyuta indirgeyen bir projeksiyon matrisi bulunur. Tüm veriler bu projeksiyon matrisiyle çarpılarak boyutları indirgenmiş olur.
- Gerçek veri= k örnek * n boyut
- Projeksiyon matrisi = n * r boyut
- Yeni veri seti = gerçek veri * projeksiyon matrisi

  = (k*n) * (n*r) = k*r boyut

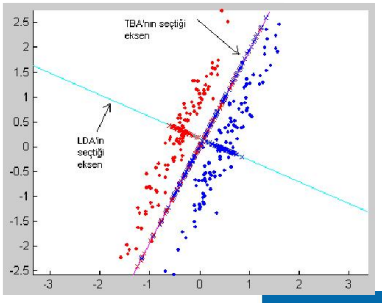
Kemik

Projeksiyonlar

- PCA
- LDA
- Latent Semantic Indexing



$$D_{l\times n} = U_{l\times r} \begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{pmatrix}_{r\times r} V^T_{r\times n}$$



Kemik

Naïve Bayes

Eğitim:

- Sınıf olasılıklarını bul.
  - Toplam K adet doküman varsa, Y sınıfından Z adet doküman varsa → Y sınıfının olasılığı (Z/K)
- Her kelimenin her doküman sınıfında yer alma olasılığını bul. İki yaklaşım:
  - Y sınıfındaki K adet dokümanın Z tanesinde yer almışsa → (1+Z)/K
  - Y sınıfındaki dokümanlarda K adet kelime varsa ve kelimemiz bu dokümanlarda Z defa geçmişse → (1+Z)/K

Kemik

Naïve Bayes

Dökümanın c sınıfına ait olma olasılığı dokümandaki her kelimenin c sınıfına ait olma olasılıklarının çarpımının c sınıfının olasılığı ile çarpımına eşittir.

Test:
Sınıfı istenen doküman *X*
n →X dokümanındaki kelime sayısı
X'in sınıfı:

$$\underset{c_i \in C}{\operatorname{argmax}} P(c_i) \prod_{i=1}^{n} P(a_i \mid c_i)$$

Herbir sınıf için bu olasılık bulunur ve doküman en yüksek olasılığa sahip sınıfa dahil edilir.

ai kelimesinin cj sınıfında yer alma olasılığı (önceki slayt)

Bir dokümanın cj sınıfından olma olasılığı

Kemik

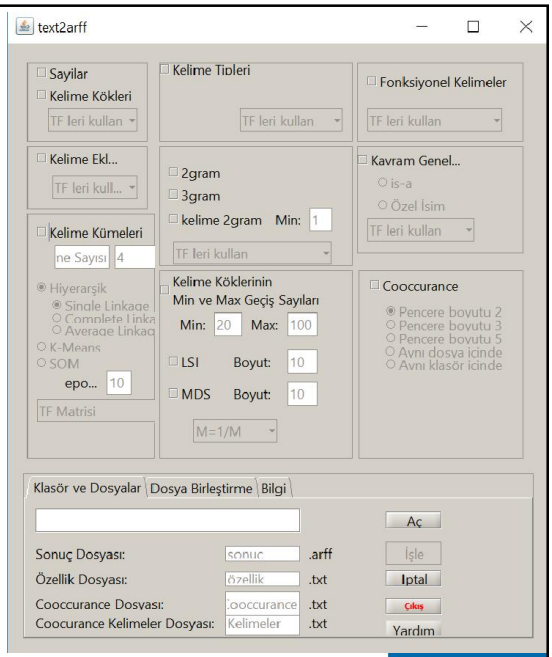
Sonuç

- Metinlerin makine öğrenmesi metotlarıyla işlenebilmesi için öncelikle sayısallaştırılmaları gerekir.
- Bu derste bunun için birçok metot gördük.
- Artık metinler sayı olduğuna göre onlar üzerinde kümeleme, sınıflandırma işlemlerini gerçekleştirebiliriz.

Kemik

Text2arff

- Gördüğümüz yöntemleri içeren bir doğal dil işleme kütüphanesi
- Metin girdilerini arff dosyasına çevirir



Kemik

Kaynaklar

- Alpaydın E. (2004) “Introduction to Machine Learning”, The MIT Press
- Helena Ahonen-Myka, Processing of large document collections
- Philipp Koehn, Data Intensive Linguistics — Lecture 12, Text Classification and Clustering
- SUNY Learning Network, Text Classification
- Christopher Manning, Opportunities in Natural Language Processing
- M.Fatih Amasyalı, Arama Motorları Kullanarak Bulunan Anlamsal Benzerlik Ölçütüne Dayalı Kelime Sınıflandırma

Kemik