2017/11/10

DOĞAL DİL İŞLEME

DERS 8: MESLEKİ BENZERLİK

ANA HATLARI

- **Sözcüksel ve Anlamsal Benzerlik**
- **Benzerlik**
  - Levenstein Mesafe
  - Jaccard Benzerliği
  - Kosinüs Benzerliği
- **Vektör Uzay Modeli**
  - İkili Ağırlık
  - Terim Frekansı (TF)
  - TFIDF

2017/11/10

SÖZLÜ VE SEMANTİK BENZERLİK

Sözcüksel benzerlik sadece kelimenin karakterini anlamsal benzerlik ortalama ile ilgilidir.

	Sözcüksel	Anlambilim
su-şu	0.5	0
su-bardak	0	0.7

MESLEKİ (KELİME) BENZERLİĞİ

Kelime benzerliği , kelimelerin form veya anlamına gelir.

- kelimesinin benzerliğin için bazı yöntemler;
  - Levenshtein Mesafesi,
  - Jaccard Endeksi,
  - Kosinüs Benzerliği.

Levenstein iki ipin mesafesini (farklılığını) ölçer. Yaparken mesafe artar, benzerlikleri azalır. Ancak Jaccard ve Kosinüs ölçüyor dizelerin benzerliği.

LEVENSHTEIN MESAFESİ

Levenshtein mesafe ölçümü için bir dize metrik iki dizi arasındaki fark.

Gayri resmi olarak, iki kelime arasındaki Levenshtein mesafesi minimum tek karakterli düzenleme (ekleme, silme veya yerine koyma) bir kelimeyi diğerine değiştirmek için gereklidir.

LEVENSHTEIN MESAFESİ

«İlgili» «fil»

(1) 'r' => ilgili
(2) 'p' => fil ekle
(3) 'v' => fil yerine 'h' koy

Levenstein mesafesi 3

LEVENSHTEIN MESAFESİ

Levenstein mesafesinin bazı kullanım alanları:

- OCR (Optik Karakter Tanıma) uygulamaları
- Yazım denetimi

Kelimelerin en yakın benzerliklerini bulabiliriz.

JAKAR ENDEKSİ

Jaccard endeksi, benzerliği karşılaştırmak için kullanılan bir istatistiktir ve örnek kümelerin çeşitliliği. Jaccard katsayısı benzerliği ölçer Sonlu örnek kümeleri arasında ve

kesişim örnek kümelerinin birleşiminin boyutuna bölünür

$$\text{Jakar Dizini} = \frac{\text{her iki kelimedeki karakterler}}{\text{her iki kelimedeki karakterler}} \times 100$$

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \times 100$$

JAKAR ENDEKSİ

Algoritması şu adımları içerir:

1. tarafından paylaşılan kesişen karakter sayısını sayın
   her iki kelime
2. Her iki gruptaki (paylaşılan ve
   paylaşılmamış)
3. Paylaşılan karakter sayısını toplam karakter sayısına bölün.
   karakterler
4. (3) 'te bulunan sayıyı 100 ile çarpın

KOSİN BENZERLİĞİ

Kosinüs benzerliği , sıfır olmayan iki arasındaki benzerliğin bir ölçüsüdür
iç kosinüsünün kosinüsünü ölçen vektörler
aralarında açı. 0 ° kosinüsü 1'dir ve herhangi biri için 1'den azdır.
diğer açı.

V1 ve V2: vektörler $\quad \text{Cos } V1, V2 = \frac{V1\ V2}{V1 * V2}$

Sayfa 6

2017/11/10

KOSİN BENZERLİĞİ

V1 ve V2: vektörler $\text{Cos V1, V2} = \frac{V1\ V2}{V1 * V2}$

Diyelim ki V1: [3 0 4] ve V2: [0 0 1]

$V1 = 3_{\ 2} + 0_{\ 2} + 4_{\ 2} = 5$

$V2 = 0_{\ 2} + 0_{\ 2} + 1_{\ 2} = 1$

$$\text{Cos V1, V2} = \frac{3 * 0 + 0 * 0 + 4 * 1}{5 * 1} = 0.8$$

VEKTÖR ALANI MODELİ

«Keliem beznerlii yötnemlerini örgendim.»

Cümleleri sayısal forma dönüştürmezsek, bilgisayar
üst cümleyi anladığımız için bunun ne anlama geldiğini anlayın.
Bu sorunu çözmek için yerine boyutsal vektörler kullanıyoruz
cümleler.

Sayfa 7

2017/11/10

VEKTÖR ALANI MODELİ

Bir dizi belgenin ortak bir vektörde vektörler olarak gösterilmesi
uzay vektör uzay modeli olarak bilinir ve puanlama için temeldir
sorgudaki belgeler.

Her benzersiz terim
belge temsil eder
vektör uzayda bir boyut.

VEKTÖRE GÖNDERME

Bir vektördeki her değer, karşılık gelen teriminin ağırlığıdır. Yapabiliriz aşağıdakilerden birini kullanarak ağırlıkları hesaplayın:

1. İkili,
2. Terim Frekansı (TF),
3. Terim Sıklığı - Ters Belge Sıklığı (TF-IDF).

İKİLİ AĞIRLIK

Bir belge bir terim içeriyorsa, o terimin
belge 1, aksi takdirde 0'dır.

Document1: Bir kalem ve bir defter aldım.
Doküman2: Bir kitap aldım.
VEKTÖR: [aldım bir defter kalem kitap ve]
(tüm belgelerdeki kelimelerin birleşmesi)
D1: [1 1 1 1 0 1]
D2: [1 1 0 0 1 0]

DÖNEM FREKANSI (TF) AĞIRLIĞI

Bir belge bir terim içeriyorsa, o terimin
belge TF'dir (terim sayısı), aksi takdirde 0'dır.

Document1: Bir kalem ve bir defter aldım.
Doküman2: Bir kitap aldım.
VEKTÖR: [aldım bir defter kalem kitap ve]
(tüm belgelerdeki kelimelerin birleşmesi)
D1: [1 2 1 1 0 1]
D2: [1 1 0 0 1 0]

TF-IDF AĞIRLIK

TF-IDF, bunun nasıl olduğunu yansıtmayı amaçlayan sayısal bir istatistiktir.
önemli bir kelime bir koleksiyon veya ceset bir belge için.

TF, bir belgede bir terimin kaç kez bulunduğunu temsil eder, IDF demek
bir terim kaç belgede var olduğunu.

$TF\text{-}IDF = TF_{normu} * IDF$

$TF_{normu} = TF / TF_{maks.}$

$$IDF = 1 + \ln \frac{\text{toplam belge sayısı}}{\text{gerçek kelimeyi içeren belge sayısı}}$$

TF-IDF AĞIRLIK

Document1: Bir kalem ve bir defter aldım.
Doküman2: Bir kitap aldım.

VEKTÖR: [aldım bir defter kalem kitap ve]
TF (D1): [1 2 1 1 0 1]
TF (D2): [1 1 0 0 1 0]
TF_{normu} (D1): [0.5 1 0.5 0.5 0 0.5]
TF_{normu} (D2): [0,5 0,5 0 0 0,5 0]

TF-IDF AĞIRLIK

$IDF_{aldım} = IDF_{bir} = 1 + \ln(2/2) = 1$
$IDF_{kalem} = IDF_{defter} = IDF_{kitap} = IDF_{ve} = 1 + \ln(2/1) = 1.693$
IDF: [1 1 1.693 1.693 1.693 1.693]

TF * IDF (D1): [0,5 1 0,85 0,85 0 0,85]
TF * IDF (D2): [0,5 0,5 0 0 0,85 0]

VEKTÖR UZAY MODELİNİN DEZAVANTAJI

Vektör uzay modelinin en büyük dezavantajlarından biri,
metin boyutu uzun olduğunda işe yaramaz hale gelir. Genel olarak,
bu özellik vektörlerinin bir sonucu olarak çok yüksek miktarda veri.

Örneğin, aşağıdakileri içeren bir web sitesinden 62.000 yorum aldık
IMDB film yorumlarının yorumları, az ya da çok 160.000 var
kelimeler. Yani bu, bir vektörün boyutunun 160.000 olduğu anlamına gelir. Bu
özellik vektörü neredeyse 37 GB (4B int vektör için) ve hiç olmayacak
bu verileri işlemek için bilgisayarda boş hafıza alanı.

VEKTÖR UZAY MODELİNİN DEZAVANTAJI

Bu boyut sorunu şu şekilde azaltılabilir:

- Metindeki **kelimelerin lemmalarını** kullanma .

Hem «ağaç» hem de «ağaçlar» için «ağaç» kullanabiliriz.

- Durma ve işlevsel kelimelerden kurtulun veya bazı özel anahtar kelimeler kullanın

SENTENCE BENZERLİĞİ

Kelime benzerliklerini birleştirerek cümle benzerlikleri buluyoruz.

«Cümle benzerliği» nerede kullanılabilir?

- Google arama motoru,
- Ad-sense,
- Bulma emsal dava.

11

2017/11/10

Sayfa 12

SENTENCE BENZERLİĞİ

Cümle benzerliği için basit adımlar:

- Durma ve işlevsel kelimelerden kurtulun
- Kelimeler yerine limon / sap kullanın
- Dizeleri Vector Space Modeline göre sayılara dönüştürme
- Benzerlik Fonksiyonunu kullanarak benzerliği ölçme

MİSAL

S: «Metin benzerliğini bulmak için TFIDF ve kosinüs benzerliğini kullanacağız. »

D1: «DDİ bilgisayar biliminin bir konusudur.»

D2: «Metin benzerliğini anlamak için DDİ yöntemlerini buluruz»

Stop Words : «bir», «için», «ve», «sayesinde», «bazı».

JAKARLI BENZERLİKLE ÇÖZÜM

S: «Metin benzerliğini bulmak için TFIDF ve kosinüs benzerliğini kullanacağız.»

D1: «DDİ bilgisayar biliminin bir konusudur.»

D2: «Metin benzerliğini anlamak için DDİ yöntemlerini buluruz»

Stop Words : «bir», «için», «ve», «sayesinde», «bazı».

paylaşılan kelime sayısı (Q, D1) = 0 J (Q, D1) = 0/11 = 0

paylaşılan kelime sayısı (Q, D2) = 2 J (Q, D2) = 2/11 = 0,18

Toplam kelime sayısı = 11

Jaccard Similarity'e göre sorgu ikinci belgeye benzer.
% 18 benzerlik oranı.

KOSİN BENZERLİĞİ İLE ÇÖZÜM

S: «Metin benzerliğini bulmak için TFIDF ve kosinüs benzerliğini kullanacağız. »

D1: «DDİ bilgisayar biliminin bir konusudur.»

D2: «Metin benzerliğini anlamak için DDİ yöntemlerini buluruz»

	DDI	Metin	Benzerlik	Bulmak	TFIDF				Kosinüs	Kullanmak	Yöntem
TF	0	0	0	0	1	2	1	1	1	1	0
IDF	1.41	2.1	2.1	2.1	1.41	1.41	1.41	2.1	2.1	2.1	1.1
TF $_{normu}$	0	0	0	0	0.5	1	0.5	0.5	0.5	0.5	0
TF * IDF	0	0	0	0	0.7	1.41	0.2	1.05	1.05	1.05	0

Sorgu (Q) cümlesi için

KOSİN BENZERLİĞİ İLE ÇÖZÜM

S: «Metin benzerliğini bulmak için TFIDF ve kosinüs benzerliğini kullanacağız. »
D1: «DDİ bilgisayar biliminin bir konusudur.»
D2: «Metin benzerliğini anlamak için DDİ yöntemlerini buluruz»

	DDI	Metin	Benzerlik	Bulmak	TFIDF				Kosinüs	Kullanmak	Yöntem
TF	1	1	1	1	0	0	0	0	0	0	0
IDF	1.41	2.1	2.1	2.1	1.41	1.41	1.41	2.1	2.1	2.1	1.1
TF $_{normu}$	0.5	0.5	0.5	0.5	0	0	0	0	0	0	0
TF * IDF	0.7	1	1	1	0	0	0	0	0	0	0

Belge 1 (D1) cümlesi için

KOSİN BENZERLİĞİ İLE ÇÖZÜM

S: «Metin benzerliğini bulmak için TFIDF ve kosinüs benzerliğini kullanacağız. »
D1: «DDİ bilgisayar biliminin bir konusudur.»
D2: «Metin benzerliğini anlamak için DDİ yöntemlerini buluruz»

	DDI	Metin	Benzerlik	Bulmak	TFIDF				Kosinüs	Kullanmak	Yöntem
TF	1	0	0	0	1	1	1	0	0	0	1
IDF	1.41	2.1	2.1	2.1	1.41	1.41	1.41	2.1	2.1	2.1	1.1
TF $_{normu}$	0.5	0	0	0	0.5	0.5	0.5	0	0	0	0.5
TF * IDF	0.5	0	0	0	0.7	0.7	0.7	0	0	0	0.5

Belge 2 (D2) cümlesi için

KOSİN BENZERLİĞİ İLE ÇÖZÜM

[DDİ Bilgisayar Bilim Konu Metin Benzerlik Bulmak TFIDF Kosinüs Kullanmak Yöntem]

S: [0 0 0 0 0.7 1.41 0.2 1.05 1.05 1.05 0]

D1: [0.7 1 1 1 0 0 0 0 0 0 0]

D2: [0,5 0 0 0 0 0,7 0,7 0,7 0 0 0 0,5]

Cos (Q, D1) = 0

COS (S, D2), 0.55 =

Kosinüs Benzerliğine göre, Sorgu (Q) ikinci belgeye benzer
% 55 değer.

LEVENSHTEIN HAKKINDA NELER VAR?

Levenshtein mesafesini kullanarak belge benzerliklerini ölçebilir miyiz?

Eğer öyleyse, bunu nasıl yapabiliriz?

https://www.twinword.com/api/text-similarity.php

Benzerliği karıştırmayın
bir soru ile
yanıtlama sistemi.

Burada ölçtük
«Sözcüksel Benzerlik».

PROJE

Botu SSS (Sık Sorulan Sorular) veritabanı için hazırlama.

16