1906003132015

Doğal Dil İşleme

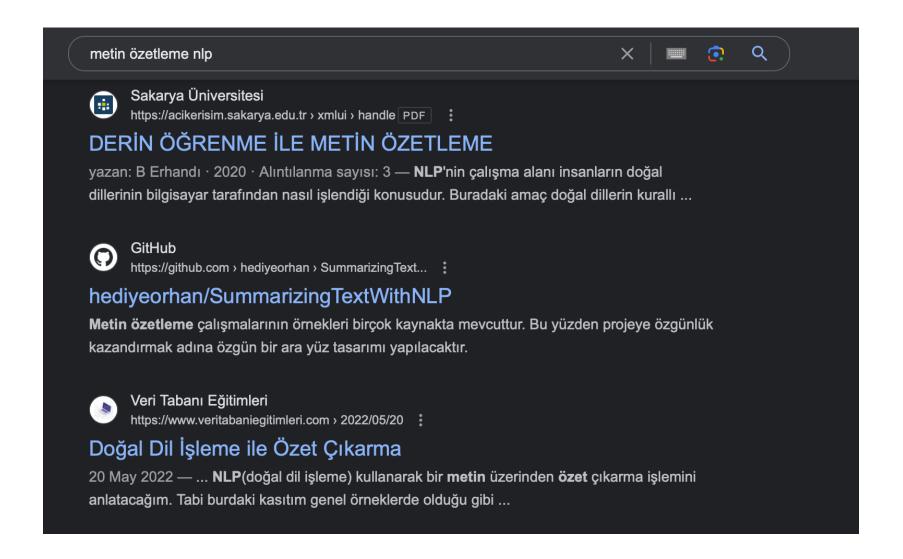
BAİBÜ Bilgisayar Müh.

Dr. Öğr. Üyesi İsmail Hakkı Parlak

ismail.parlak@ibu.edu.tr

Oda: 335

Metin Özetleme



Metin Özetleme

Metin özetlemenin 2 ana akımı vardır:

- Çıkarıcı (Extractive) özetleme: Metnin içerisinden metni özetlediğine karar verilen bazı cümleler olduğu gibi bir özet alt kümesine alınır.
- Soyutlayıcı (Abstractive) özetleme: Metin içerisinde bulunması zorunlu olmayan yeni cümleler oluşturularak metin özetlenir.
- Soyutlayıcı yöntem, çıkarıcı yöntemden daha etkili özetler oluşturabilir. Ancak, algoritmik düzeyde çıkarıcı yöntemlerden daha karmaşıktırlar.

Metin Özetleme (Çıkarıcı)

Örnek:

metin =
$$\{c_0, c_1, c_2, c_3, ..., c_{42}\}$$

özetleyici

özet = {
$$c_{12}$$
, c_4 , c_{35} , c_{27} , c_{40} }

c_i: i indeksindeki cümle

TF-IDF İle Çıkarıcı Özetleme

• $tfidf(t, d) = tf(t, d) \times idf(t)$, t: kelime, d: doküman

• tf(t, d) =
$$\frac{count(t,d)}{len(d)}$$
 idf(t) = $log(\frac{D}{D(t)})$

- TF-IDF ile dolgu sözcüklerinin doküman vektörlerindeki büyüklükleri algoritmik olarak küçültülür.
- Tüm dokümanlarda sıkça geçmeyen ancak dokümanda sıkça bulunan sözcükler önemli hale gelir.

TF-IDF İle Çıkarıcı Özetleme

- 1. Özeti çıkarılacak metin cümlelere ayrılır.
- Cümleler dolgu sözcüklerinden, noktalama işaretlerinden temizlenme, lemmatizasyon vb. işlemlerden geçirilerek temizlenir.
- 3. Cümleler TFIDF ile vektörleştirilir.
- 4. Her cümleye, TFIDF temsilleri kullanılarak bir puan verilir.
- 5. Puanı en yüksek cümlelerin bir kısmı puan sıralamasına göre seçilir.

TF-IDF İle Çıkarıcı Özetleme

cümle puanı = ortalama(vektor[vektor != 0])

Ör: vektor = [0, 0, 0.06, 0.1, 0, 0.2, 0, 0, 0]cümle puanı = (0.06 + 0.1 + 0.2) / 3 = 0.12

Özet için kaç cümle seçeceğiz?

- Top N
- Eşik değeri
- Top %N

Diğer Bazı Çıkarıcı Özetleme Yöntemleri

TextRank:

- https://web.eecs.umich.edu/~mihalcea/papers/mihalce a.emnlp04.pdf
- https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/
- Latent Semantic Analysis:
 - https://ieeexplore.ieee.org/document/7725296
 - https://iq.opengenus.org/latent-semantic-analysis-for-text-summarization/