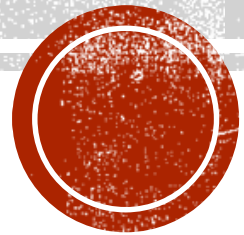
TF-IDF

Term Frequency–inverse Document Frequency



Musa ŞANA – musana.net

NEDİR ?

- Bir döküman/dataset içerisinde geçen kelimelerin önemini, istatistiksel yöntemlerle hesaplayarak bunları birer vektör haline getiren bir yaklaşımdır.

- TF-IDF yöntemi iki ana parçadan oluşmaktadır.

→ Term Frequency(TF) ve inverse Document Frequency(IDF)

TF NASIL HESAPLANIR ?

- TF (**Term Frequency**) : Bir kelimenin/terimin dökümanın (herhangi bir sayfasında) içerisisinde geçme sıklığını ifade eder.

➢ TF aşağıdaki formul ile hesaplanır.

✓ Terimin dökümanda geçme sıklığı(sayısı) / Dökümandaki toplam terim sayısı

- Uzun dökümanlarda ilgili terimin kullanım sayısı artacağından dolayı, terim, dökümandaki toplam terim sayısına bölünerak normalizasyon işlemi yapılmaktadır.

BİR ÖRNEK

- Örnek Dataset:

Boşa çıksın reislerin, kahinlerin, şairlerin kuvveti.

Güler yüzlü olmak **neydi** onu hatırlayın.

Neydi söğüt gölgesinde gülümsemek

Ağız dolusu gülmeden taşlıkta

Tokenize İşlemi:

list_ = [‘bosa’, ‘ciksin’, ‘reislerin’, ‘kahinlerin’, ‘sairlerin’, ‘kuvveti’, ‘guler’, ‘yuzlu’, ‘olmak’ ‘neydi’, ‘onu’, ‘hatirlayin’, ‘neydi’, ‘sogut’, ‘golgesinde’, ‘gulumsemek’, ‘agiz’, ‘dolusu’, ‘gulmeden’, ‘taslikta’]

Tf Hesaplama:

- «neydi» teriminin tf hesaplanması/değeri:

  Terim Sayısı / Toplam Terim → 2 / 20 = **0,1**

- «gülümsemek» teriminin tf hesaplanması/değeri:

  Terim Sayısı / Toplam Terim → 1 / 20 = **0,05**

- Dökümandaki her kelime için tf değeri çıkartılarak bir matris oluşturulur.

- Eğer dökümanda kullanılan ama terim olmayan ifadeler varsa (Stop Word), bunların terim olarak sayılmaması için ilgili metoda bu parametrenin değeri verilebilir.

→Örneğin ve, veya, yada gibi bağlaçlar

TF değeri hepsalanıp matris oluşturulduktan sonra IDF değeri hesaplanarak ikinci bir matris daha oluşturma işlemine geçilir.

IDF NASIL HESAPLANIR ?

- IDF(**inverse Document Frequency**), tf ile hesaplanan terimlerin toplam dökümanın(**corpus**) kaç yerinde en az bir kere geçtiğini hesaplar. Burda terimin geçtiği dökümanlardaki sayısı önemli değildir. Kaç dökümanda bu terime (en az bire) rastlanılmıştır ona bakılır.

➢ Aşağıdaki formül ile hesaplanır.

✓ **LOG(**Terimin geçtiği döküman sayısı / toplam döküman sayısı**)**

Bu işlem ile terimin bütün dökümanlar(corpus) içerisindeki sıklığına bakılarak önemi/değeri hesaplanmaya çalışır.

BİR ÖRNEK

- Örnek Corpus

- D1 = Bu bir **Yusuf** masalıdır de
  D2 = Bunu söyle ve fakat
  D3 = Şunu da sor
  D4 = **Yusuf**'un masalı neden
  D5 = **Yusuf**'la başlamıyor?
  D6 = Bir varmış bir yokmuşla başlıyor bütün masallar gibi
  D7 = Bir Şivekâr varmış, bir gençkız
  D8 = **Yusuf** yokmuş, cinler
  D9 = Kaçırmış, yazgı saklamış onu
  D10 = **Yusuf** bir ayna mıdır?

- **IDF Hesaplama:**

→**LOG(**Terimin corpusta bulunan dökümanların kaç tanesinde geçtiği / toplam döküman sayısı**)**

❖**«Yusuf»** kelimesinin idf hesaplanması/değeri:

LOG(5/10) **= 0,301029995663981**

TF-IDF SONUÇ

- Bir terime ait hesaplanan TF ve IDF değeri çarpma işlemine tabii tutulduktan sonra nihai bir matris oluşmaktadır. Oluşan matrisdeki numeric değerler ile makine öğrenmesinin çıkış sutunu olan değerler arasında bir matematiksel model oluşturmak için makine öğrenmesi algoritmaları kullanılır. Böylece bir kelimenin geçme sıklığı yüksek ise o kelime iyi/good şeklinde sınıflandırılırken az geçen kelimeler bad/kötü olarak etiketlenmektedir.

TEŞEKKÜRLER...