

NLP Article type selector

Murat Bektaş



***Özetçe***

Bu projede Python dili ve kütüphaneleri kullanılarak girdi olarak verilen makalenin konusunun Matematik, Tıp, Tarih alanlarından hangisiyle ilgili olduğunu bulan bir yazılım geliştirilmiştir. Geliştirilen kod kendisine verilen farklı makale türlerinden yararlanarak bir veri tabanı oluşturur. Sonrasında test için girdi olarak verilen makalenin ise türünü belirler.

***Abstract***

In this project, a software has been developed that finds the subject of the article given as input by using the Python language and libraries, which is related to the fields of Mathematics, Medicine and History. The developed code creates a database by making use of the different article types given to it. Then it determines the type of the article given as input for the test.

***Anahtar Kelimeler***

**Besleme makalesi**: Doğru sonuçlar alabilmek için programa öğretilen matematik, tarih, tıp makaleleri

**Test makalesi**: Türünü öğrenmek için programa input olarak ismi verilen makale

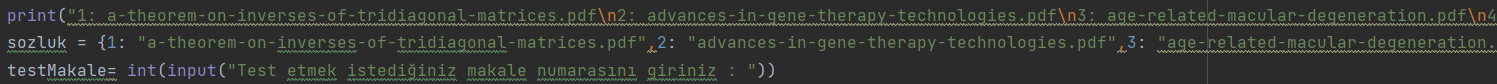
**Özel Kelimeler**: Makale türlerini birbirinden ayıran; edat, bağlaç, noktalama işaretleri gibi nitelik belirtmeyen karakterlerden arınmış kelimeler.

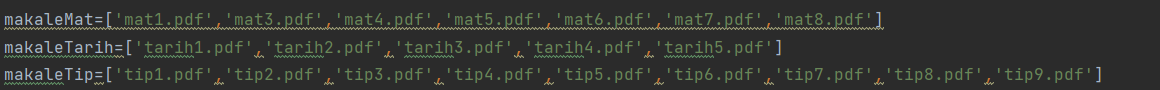
***Algoritma Ana Planı***

1. PDF okumak için gerekli kütüphaneleri import et.
2. Sözlük yapısı kullanarak test edilmek istenen makaleyi sor.
3. *Besleme makaleleri*ni for döngüsü kullanarak okut.
4. Gereksiz kelime (Stopwords) ve noktalama işaretlerini makalelerden kaldır.
5. *Besleme makaleleri*ndeki *özel kelimeleri* türlerine göre bir listeye at.
6. Seçilen *test makalesi*ndeki *özel kelimeleri* ayır.
7. *Test makalesi*nin *özel* kelimelerini bir listeye at.
8. *Özel kelimeleri* içeren listeleri karşılaştırma yapabilmek için set yapısına dönüştür.
9. *Test makalesi*nin kelimelerini içeren set’i *besleme makaleleri*nin içindeki kelimeleri içeren set’ler ile karşılaştır.
10. *Test makalesi* hangi tür makale tipiyle daha çok ortak kelime içeriyorsa *test makalesini* o tipe ata.
11. Sonucu ekrana bas.
12. (Teslim klasörü içerisinde algoritmaya dair .html formatında FlowChart dosyası bulunmaktadır)

***Kod Açıklaması***

* pdfminer kütüphanesi kullanılarak pdf’lerin okunması sağlanmıştır. Bu kütüphane sayesinde eldeki pdf’ler okunarak text haline gelmiştir. metin içeren bir resim

  Açıklama otomatik olarak oluşturuldu
* *Test makalesi* isimleri ekrana bastırılmıştır. Sözlük yapısı kullanılarak, kullanıcıdan hangi makaleyi test etmek istediği alınmıştır. 
* Makale içlerindeki nitelik belirtmeyen kelimeler(stopwords) bir .txt dosyasına tanımlanıp koda dahil edilerek çıkarılmak üzere okunmuştur. metin içeren bir resim

  Açıklama otomatik olarak oluşturuldu
* Gerekli for döngülerinde kullanabilmek üzere *besleme makaleleri*nin isimleri kendi türündeki liste değişkenlerine atılmıştır. 
* Üç adet for döngüsü kullanılarak tarih, tıp, matematik makaleleri okunmuş, gereksiz karakterlerden temizlenmiş ve nitelik belirten kelimeleri bir liste değişkenine eklenmiştir. metin içeren bir resim

  Açıklama otomatik olarak oluşturuldu metin içeren bir resim

  Açıklama otomatik olarak oluşturuldu metin içeren bir resim

  Açıklama otomatik olarak oluşturuldu
* Kullanıcıdan ismi istenen *test makalesi* için de aynı işlemler for döngüsü kullanılmadan tekrarlanmıştır. metin içeren bir resim

  Açıklama otomatik olarak oluşturuldu
* Özel kelimeler tekrar kullanılmak üzere liste değişkenlerinde saklanmıştır. metin içeren bir resim

  Açıklama otomatik olarak oluşturuldu
* Birbiriyle basitçe karşılaştırma yapabilmek için liste değişkenleri içindeki *özel kelimeler* set türündeki değişkenlere aktarılmıştır. metin içeren bir resim

  Açıklama otomatik olarak oluşturuldu
* *Test makalesi*nin hangi türde olduğunu bulabilmek için set değişkenleri üzerinde mantıksal işlemler yapılmıştır. Bulunan ortak *özel kelimeler* gerekli değişkenlere atılmıştır.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

* *Test makalesi*nin hangi tür ile daha çok ortak kelimeye sahip olduğunu bulabilmek için yukarıdaki değişkenlerin boyutlarını “len()” yapısı kullanılarak if-else içinde karşılaştırılmıştır. Çıkan sonuç “print()” ile ekrana basılmıştır. metin içeren bir resim

  Açıklama otomatik olarak oluşturuldu

***Kod Çıktısı***

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu