*Sınıflar*

Öncelikle WordClass isimli bir sınıfımız var. Bu sınıf txt ve ya pdf belgelerinden text leri okuyup kelimelerine ayırdıktan sonra filtrelemeye yarıyor.

Bu sınıftan bir nesne oluşturulurken str tipinden bir dosya yolu girmemiz gerekiyor.

\_\_init\_\_ methodu içerisinde çağırılan ve alt sınıflarda override edilen ReadFile methodu çağırılarak doya yolundaki dosya okunuyor ve okunan veri split fonksiyonu çağırılarak kelimelerine ayrıştırılıp return ediliyor.

Sonrasında return edilen bu kelime listesi FilterWord isimli methoda aktarılıyor.

Bu methodda gelen liste şeklindeki her bir kelime filtreleme işleminden geçiyor. Bu işlem sırasında öncelikle her bir kelime, içerisindeki alfabetik olmayan karakterler çıkartılarak harfler küçü harflere çeviriliyor. Sonrasında oluşan yeni kelimenin uzuluğu yeterliyse ve BAN\_WORDS isimli terimsel olmayan kelimeler listesinde bulunmuyorsa, sınıfın words(self.words) isimli setine ekleniyor. Bu işlemle birlikte elimizde belgedeki terimsel sözcükler kalıyor.

WordClass isimli sınıftan kalıtım alan iki adet sınıfımız var, bunlar TextFile ve PDFFile. Bu sınıflarda ReadFile methodu override edilerek dosyalardan okuma işlemi sırasında dosyanın açılması ve text içeriğinin alınması işlemleri gerçekleştiriliyor.

Bu sınıflar dışında matematik, tıp ve tarih alanındaki alanları için Alan isimli sınıf bulunuyor. Bu sınıf oluşturulurken öncelikle alanın ismi, sonrasında da bu alanın terimsel verilerinin bulunduğu txt dosyasının dosya yolu veriliyor.

Son olarak Similarity sınıfında bir pdf dosyasının hangi alana daha yakın olduğu hesaplanıyor. Bunun için önce pdf dosyası, sonrasında da alanların listesi veriliyor. FindBestSimilarity sınıfında bu pdf dosyasının hangi alana daha yakın olduğu hesaplanıyor.



Benzerlik verisinin hesaplanmasında Jaccard Algoritması kullandık.

Örnek kod parçası:



Örnek kod çıktısı:

