

BİL3102 METİN VE WEB MADENCİLİĞİNE GİRİŞ - Ödev

Türkçe Haber Metinleri için Belge Sınıflandırması

Ödevin Son Teslim Tarihi: 28 Nisan 2021 Çarşamba, saat: 20:00

(ek süre kesinlikle verilmeyecektir. Herhangi bir nedenle zamanında iletilmeyen ödevler, hiçbir mazeret kabul edilmeden 0 (sıfır) olarak notlandırılacaktır.)

Ödevin Teslim Şekli:

DEÜ Sakai sistemindeki ders sayfasında açılacak olan ödev yükleme (assignment) alanına; tüm dosyalar, rapor, vb. **zip / rar sıkıştırılmış tek bir dosya olarak yüklenecektir.**

Bu ödev tek kişiliktir. En ufak bir yardım, Internet'ten kod kopyalanması, vb eylemler kopya / intihal olarak değerlendirilecektir ve bunu yapan öğrenciler sınavdan ya da sınavın ilgili bölümünden **0 (sıfır)** alacaktır.

Veri kümesinin içeriği:

“Odev-veriler.rar” adlı sıkıştırılmış dosya içerisinde, **4 farklı haber sınıfına ait 230’ar, toplamda 920 haber metni** bulunmaktadır. Eğitim (train) ve test verileri iki ayrı dizinde olup, bunların her birisinin altında da 4 ayrı sınıfa (kategori) ait dizinler bulunmakta ve içerisinde de ilgili haber metinlerine ait dosyalar yer almaktadır.

Haber metinlerinin sınıfları:

ekonomi
magazin
sağlık
spor

Veri kümesinin olası kullanım alanı: Metin / belge sınıflandırma

Veri kümesindeki sınıf sayısı: 4

Veri kümesindeki eğitim (train) örnek sayısı: $150 \times 4 = 600$

Veri kümesindeki test örnek sayısı: $80 \times 4 = 320$

Ödevde Yapılacaklar ve İstenenler:

- Bu veri kümesini kullanarak, multi-class classification ile bu 4 sınıf için belge sınıflandırması yapılacaktır.
- Bu veri üzerinde **sözcüklerin ayrıştırılması (tokenization) gereklidir**. Bu şekilde sözcüklerden öznitelikler (features / attributes) oluşacaktır. Tokenization’da hangi karakterlerin ayrı olarak kullanılacağı (boşluk, virgül, nokta, vb) öğrencilere bırakılmıştır.
- Metinlerdeki büyük harflerin **küçük harfe çevrilmesi (lower-case), etkin olmayan sözcüklerin (stop-words) kullanılması da önerilir**. (Türkçe için stop-words dosyaları Moodle sisteminde önceki haftalarda ilgili hafta kısmında yüklü bulunmaktadır, onu kullanabilirsiniz).
- Sözcüklerin köklerine göre gruplanması (stemming) ve ilgili stemmer araçları da kullanılabilir. Türkçe için “Zemberek” uygulaması önerilmektedir.
- Özniteliklerin **seçimi / azaltılması (feature selection / reduction) yöntemlerinin de kullanılması özellikle önerilmektedir**. Sizlere derste anlatılan ve örnekleri verilen yöntemlerden bir veya birkaçını kullanabilirsiniz.
- Sözcüklerin metinlerde kaç kere geçtiğinin sayısal temsili, yani vektörel ve sayısal değerlere çevrilmesi de mutlaka gereklidir. **Binary vector, term frequency veya weighted / normalized tf-idf’den herhangi birisini seçip kullanabilirsiniz**.

- Sınıflandırma için hangi algoritma / algoritmaları (k-NN, Multinomial Naive Bayes, Rocchio) ve bunların ilgili hangi uzaklık / benzerlik metrikleri (Cosine, Pearson, Jaccard, Euclidean, vb) gene öğrencilerin tercihinine bırakılmıştır.
- Programınızda ilgili aşamada oluşturacağınız metin-sözcük verisini (son hale getirdiğiniz sözcüklerin de olduğu ve seçtiğiniz yöntemle göre her sözcüğün ilgili kayıttaki sayısal temsili değeri (tf-idf, binary vector, vb hangisini kullandıysanız) bulunan veriyi (train ve test hepsi bir arada) **ödev tesliminde ayrıca bir .txt dosya olarak (csv yani virgülle ayrılmış şekilde) teslim etmeniz zorunludur.** Aşağıda bir örneği verilmiştir (aşağıdaki örnekte tf-idf ile gösterilmiştir).

	s1	s2	s3	sn	Sınıf
1.txt	0	0	2.68	0	0	0	ekonomi
2.txt	0	1.24	0	0	3.567	0.88	ekonomi
..							...
..							...
Testdat 400.txt	0	1.78	0	0	0	0	spor

Bu veri dosyasını teslim etmezsiniz ödevinizden 100 **üzerinden 20 puan kırılabacaktır.**

- **Test sonucunda** elde edilen **performans ölçüm değerlerini de aşağıda gösterilen şekilde ayrı bir dosyada teslim etmeniz zorunludur.**

	ekonomi	magazin	saglik	spor	Ortalama
Precision					
Recall					
F-Score					

Bu sonuç dosyasını teslim etmezsiniz ödevinizden **100 üzerinden 25 puan kırılabacaktır.**

- Programınızda, hazır kütüphane / fonksiyon, vb kullanabilirsiniz, bu konuda bir kısıtlama yoktur.
- Ödevinizi C, C++, C#, .Net, Java, Python programlama dillerinden birisi ile yapabilirsiniz.