BİL3102 METİN VE WEB MADENCİLİĞİNE GİRİŞ : Ödev

Türkçe Tweet Kayıtlarından Oluşan Veri Kümesi için Metin Sınıflandırması

Ödevin Son Teslim Tarihi: 22 Mart 2020 Pazar, saat: 22:00 (TSİ)

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

Ödevin Teslim Şekli:

CSC ÖBS (Moodle) sistemindeki ders sayfasında açılacak olan ödev yükleme (assignment) alanına; tüm program kaynak kodu, dizinler, kütüphneler, dosyalar, vb. **zip / rar sıkıştırılmış tek bir dosya olarak yüklenecektir**. Eğer Moodle'a yüklemede sorun olursa, öğrenci github, dropbox vb ödevine erişim / indirme adresini bir dosyaya yazıp Moodle'a ödev olarak o dosyayı yükleyecektir.

Bu ödev **tek** kişiliktir. **En ufak bir yardım, vb kopya / intihal olarak** değerlendirilecektir ve **yardım veren / alan** öğrencilerin hepsi bu ödevden **0 (sıfır) alacaktır.**

Veri kümesinin içeriği:

"Odev-veriler.rar" adlı sıkıştırılmış dosya içerisinde, "Raw_texts" klasörü ve bu klasör altındaki 3 farklı kalasörde, 3 farklı sınıfına ait toplamda 3000 ayrı tweet kaydı bulunmaktadır. Bu kayıtlar, Internet'ten alınmıştır ve çeşitli kişilere ait çeşitli farklı konularda gerçek kayıtlardır. Tweet'ler Türkçe yazılmıştır ama yabancı kökenli sözcükler, ayrıca Türkçe dil kurallarında tanımlı olmayan kısaltmalar, vb. içerebilir.

Metin sınıflandırması, 3 ayrı sınıf şeklinde yapılacaktır ve ilgili tweet sınıfları aşağıdaki gibidir:

1- Olumlu Tweet ler (toplam 756 kayıt)
2- Olumsuz Tweet ler (toplam 1287 kayıt)
3- Nötr Tweet ler (toplam 957 kayıt)

Her **sınıfa** ait tweet metinleri, **ayrı bir klasörde** yer almaktadır.
Veri kümesindeki sınıf sayısı: 3
Veri kümesindeki toplam kayıt sayısı: 3000

Ödevde Yapılacaklar ve İstenenler:

- Bu veri kümesini kullanarak, multi-class classification ile bu 3 sınıf için belge sınıflandırması yapılacaktır. Metinler, yani tweet kayıtları ham veridir.
- Bu veri üzerinde **sözcüklerin ayrıştırılması (tokenization) gereklidir.** Bu şekilde sözcüklerden öznitelikler (features / attributes) oluşacaktır. Tokenization'da hangi karakterlerin ayırı 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) ve bunlara ilişkin hangi uzaklık / benzerlik metrikleri (Cosine, Pearson, Jaccard, Euclidean, vb) gene öğrencilerin tercihine bırakılmıştır.

- Programınızda ilgili aşamada oluşturacağınız metin-sözcük verisini (3000 metin instance yani kayıtlardan, son hale getirdiğiniz sözcüklerin de attribute olduğu ve seçtiğiniz yönteme göre her sözcüğün ilgili kayıttaki sayısal temsili değeri (tf-idf, binary vector, vb hangisini kullandıysanız) bulunan veriyi) **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).

	a1	a2	a3	...	...	an	Sınıf
Kayıt1	0	0	2.68	0	0	0	1
Kayıt2	0	1.24	0	0	3.567	0.88	3
..	1.78	0	1.12	4.77	0	0	1
Kayıt3000	0	1.78	0	0	0	0	2

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

- Programınızda eğitim ve test aşaması için **stratified 10-folds cross-validation** yöntemi kullanılacaktır.
- Stratified 10-folds cross-validation sonucunda elde **performans ölçüm değerlerini de aşağıda gösterilen şekilde ayrı bir dosyada teslim etmeniz zorunludur.**

	Sınıf 1 (olumlu)	Sınıf 2 (olumsuz)	Sınıf 3 (nötr)	MACRO Average	Micro Average (weighted average)
Precision					
Recall					
F-Score					
True Positive adedi					
False Positive adedi					
False Negative adedi					

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

- Programınızda**, sınıflandırma algoritması ve k-NN kullanmanız durumunda uzaklık / benzerlik ölçümü yöntemleri kısımlarını tamamen kendiniz kodlamanız gerekmektedir (bu kısımlarda hazır fonksiyon, kütüphane vb. kullanan ödevden 0 alacaktır).** Bunların dışındaki tüm kısımlar için ise, hazır kütüphane / fonksiyon, vb. kullanabilirsiniz, o kısımlarda bir kısıtlama yoktur.
- Ödevinizi **C, C++, C#, Java** programlama dillerinden istediğiniz birisi ile yapabilirsiniz.
- Teslim edilen ve değerlendirmeden geçen ödevler içinde, **en yüksek MACRO Average Fscore değerini elde eden ödeve, + 10 puan ödül verilecektir.** (Yani diğer kısımlar, istenen dosyalar, vb. de eksiksiz ise bu ödev, 110 alacaktır).