BİL3102 METİN VE WEB MADENCİLİĞİNE GİRİŞ ARA SINAV

Bu sınav, iki bölümden oluşmaktadır. İlk bölümde, 75 puanlık bir proje verilmiştir. İkinci bölümde ise toplam 25 puanlık 2 adet soru sorulmuştur.

Proje ve soruların yanıtları için son teslim tarihi: 10 Mayıs 2021 Pazartesi, saat 20:00

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

Sınavın 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 sınav **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.**

1. Bölüm

"araSnv-veriler.rar" adlı sıkıştırılmış dosya içerisinde, "metinler" klasörü ve bu klasör **altında farklı uzunluklarda 135 adet kayıt** bulunmaktadır ve her bir kayıt (yazı metni) ayrı bir txt dosyasındadır *(1.txt, 2.txt, vb.)*. Bu kayıtlar, Internet'teki çeşitli sosyal ağ, blog, forum sitelerinden alınmıştır. **Duygu analizinde kategorik ayırım şu şekilde yapılacaktır:**

1- **Olumlu**
2- **Olumsuz**
3- **Belirsiz**

Yapılacaklar ve İstenenler:

- **1. Aşama:** Bu veri kümesini kullanarak, derslerde size aktarılan bir veya birkaç yöntemi kullanarak 135 farklı kaydın her birisini 3 kategoriden birisine atayacaktır. Bu atamaları, ya program ekranında liste şeklinde görüntülenecek, ya da bilgisayarda bir dosya yaratıp oraya liste olarak yazdırılacaktır.
  - *Örnek:*
    21.txt Olumsuz
    22.txt Belirsiz
    ...
    117.txt Olumlu

- **2. Aşama:** Programın ekranından herhangi bir sözcük girildiğinde, bunun hangi olasılık değeri ile hangi duygu kategorisine atandığı hesaplanıp ekrana yazılacaktır.
  - *Örnek: ekrandan "kalite" sözcüğü girilir.*
    Olumlu
    p=0.0000342

  - *Örnek 2: ekrandan "tebeşir" sözcüğü girilir.*
    Belirsiz
    p=0.00000157
    vb...

- Programınızda, hazır kütüphane, fonksiyon, vb kullanabilirsiniz, bu konuda bir kısıtlama yoktur.
- Kodlama kısmında, sadece, **aşağıdaki programlama dillerinden birini kullanabilirsiniz, aşağıdakiler harici bir dil kullanılamaz:**

  C, C++, C#, .Net, Java, Python.

1. Bölümle ilgili olarak teslim edilecekler listesi:

1-Programın tüm kaynak kodları, bağlantılı kütüphane, dizinler, vb.
2-Kullanılan algoritmalar, vb. ile ilgili kısa bilgiler / notlar (istenirse kaynak kod içine de açıklamalar olarak eklenebilir).

---

2. Bölüm

Take-home şeklindeki ara sınavın ikinci bölümünde, aşağıda 3 adet soru verilmiştir. Bu soruların yanıtları ve çözümleri, net, ayrıntılı ve düzgün bir biçimde, bir Word belgesi içerisine yazılacaktır (Eğer, öğrenciler çözümleri el yazısıyla kâğıda çözerse, scanner ile taranıp ya da akıllı telefonla fotoğrafı çekilip, vb. Word içinde resim olarak kaydedilip de iletebilir).

1. Aşağıda verilmiş olan tabloda, 3 farklı sınıf (A, B, C) için her bir kaydın gerçek ait olduğu sınıf ve algoritmanın tahmin ettiği sınıf bilgileri bulunmaktadır. Bu tablodan giderek aşağıdaki şıkları yanıtlayınız.
Her şıkta çözümünüzü hesaplayarak açıklayınız ve işlemlerinizi gösteriniz.

		Algoritmanın Tahmin Ettiği Sınıf		
		A	**B**	**C**
Gerçek Sınıf	**A**	8	4	6
	B	3	9	1
	C	0	2	5

1.1. (1 puan) A sınıfının **Recall** (Duyarlılık) değeri nedir**?**
1.2. (1 puan) B sınıfının **Precision** (Kesinlik) değeri nedir**?**
1.3. (1 puan) C sınıfının **F-Score** (F1 Score) değeri nedir**?**
1.4. (1 puan) Üç sınıfın **Precision** için **micro average** (mikro ortalama) değeri nedir?

2. (21 puan) N-gram ile dil modelleme analizi yapıyorsunuz. Aşağıdaki tablolarda, bazı yazışmalardan elde edilen sözcüklerin unigram ve bigram dağılım tabloları (adet / count değerleri şeklinde) verilmektedir. Yazışmalara ait **V** yani Vocabulary (sözcük dağarcığı) değeri **= 165**.

Unigram tablosu

kar	ve	zarar	büyük	sorun	çok	yok
12	28	20	30	16	32	12

Bigram tablosu

	kar	ve	zarar	büyük	sorun	çok	yok
kar	0	2	2	1	0	1	2
ve	1	0	3	2	0	1	0
zarar	0	0	0	2	1	2	1
büyük	1	1	2	0	3	0	0
sorun	0	1	0	1	1	1	3
çok	0	1	2	2	1	2	1
yok	0	1	0	0	0	2	0

Maximum Likelihood Estimate yöntemi ile aşağıdaki **cümlenin bigram tahmini olasılığını** bulunuz. **Çözümünüzde gerekli tüm hesaplamaları ve işlemleri göstermek zorundasınız.**

çok büyük zarar yok