2024-2025 GÜZ DÖNEMİ

İSTANBUL BEYKENT ÜNİVERSİTESİ
MÜHENDİSLİK MİMARLIK FAKÜLTESİ

Öğrenci Adı ve Soyadı	
Öğrenci Numarası	
Ders Adı	Doğal Dil İşlemeye Giriş
Sınav Tarihi ve saati	6.12.2024 11.00
Sınav Süresi	50 dakika
Sınav Salonu	E1- 31, E1-41 , E1-51
Öğretim Elemanı Adı	Dr.Öğr.Üyesi Zeynep Altan

Soru Numarası	1	2	3	4	5	6	Toplam
Alınan Not							

1)Aşağıdaki düzgün ifadelerin her birinin hangi dizgilerle eşleşebileceğini { } içerisinde yazın. Her bir dizgiyi nasıl elde ettiğinizi açık olarak yazın. Sadece cevap değerlendirilmez (20 puan)

i)[Aa] (n (in)? | yem?) {Anin, An, anın, an, Aye, aye, Ayem, ayem}

'A' ya da 'a' dan sonra 'n' gelir. Sonrasında 'in' imi olur ya da olmaz. 'A' ya da 'a' dan sonra 'ye' imi gelir. Sonrasında 'm' durur ya da olmaz.

ii) [Aa] kıl ** ı { Akıl**ı, akıl**ı }

* anlamı '\' den sonra ne varsa o sembolün gelmesidir. Burada * sembolü 2 kere gelecektir

iii) A(ra)+ba { Araba, Ararabe, Arara.....raba }

'ra' imi bir ya da daha fazla sayıda tekrar edilir. Öncesinde 'A', sonrasında 'ba' vardır

iv) \d+(\.\d+)?(\w+) {15.49kod15, 25kod25}

\d herhangi bir decimaldir. '+' bir veya daha fazla sayıda decimaldir. ( )? ifadesi ya vardır ya da olmaz. ( ) anlamı . imi bir veya daha fazla decimalin izlemesidir. \w yanındaki imi ... da bir veya daha fazla alfanumeriktir.

2) *Sarı* | *sarı* düzgün ifadesi aynı dizgi ya da dizgilerle eşleşecek şekilde verilenden farklı iki şekilde daha ifade edilebilir. Bu düzgün ifadeler nelerdir? Hangi dizgiler eşleşir? Cevabınızı mutlaka açıklayarak vermelisiniz (10 puan).

[Ss]arı
(S|s)arı
her ikisi de S ya da s anlamına gelir

Düzgün ifade, Sarı veya sari dizgileri ile eşleşir

3) i) Satır başında (dizginin başında) alfanümerik karakterlerden oluşan bir sıralanışı herhangi bir tamsayı izlemektedir ve satır (dizgi) sonlanmaktadır. Bu düzgün ifadeyi 2 farklı şekilde nasıl yazarsınız? Yaptığınız işlemleri açıklamanız gerekir. Bu düzgün ifade ile eşleşen 2 farklı dizgi yazın (10 puan) .

^[a-zA-Z0-9_]+ [0-9]+$ ya da
^\w+\d+$ 1. dizgi 15kod17 ve -All45 eşleşen dizgilerdir

ii) i) deki düzgün ifadeyi olumsuz olarak yazdığınızda eşleşen dizgileriniz neler olacaktır? Cevap için düzgün ifadenin de yazılması zorunludur. Yapılan işlemler açıklamalıdır (10 puan)

^[^\w+\d+]$ ya da ^[^(a-zA-Z0-9)*(0-9 ...
Klavyedeki özel karakterlerin bir dizilişini (örneğin !+) , klavyedeki rakamların bir dizilişi dışındaki karakterler, örneğin, za işler => !+za

4) Çalışmaya ait derlem örneği "01" "12" "123" "234" olarak verilmektedir. Bu derleme byte pair encoding (BPE) belirteç(token) öğrenici algoritmasını 2 kere uyguladığınızda sonucunuz ne olur? Bu algoritma ne amaçla kullanılmaktadır? (Belirteç sonu imini hesaplamalarınıza eklemeyiniz) (15 puan).

sözlük = {0, 1, 2, 3, 4}

01 0 1 defa
12 1 3 defa
123 2 3 defa
234 3 2 defa
4 1 defa
12 2 defa
23 2 defa

sözlük {0, 1, 2, 3, 4, 12, 23, ...}
12 birleştirilir ve sözlüğe eklenir
23 birleştirilir ve sözlüğe eklenir
Algoritmanın 2 kere uygulanması istenmiş

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Sözlüğün kısıtlı elemandan oluşmasını önler. Sözlüğe orijinalinden türetilmiş yeni elemanlar (öğeler) eklenir.

5) Çok terimli Naive Bayes algoritması ile ne ifade edilmektedir? Açıklayınız. (10 puan) .

Maksimum likelihood estimation → sınıflandırıcı
Çok terimli bir Naive Bayes algoritmasıdır

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Özgeli Markov varsayımıdır. Sonraki sözcüğün gelme olasılığıdır n-gram modeller olarak genelleştirilir. Metin sınıflandırıcıdır
Bir c_j sınıflandırıcısında w_i sözcüğünün olma olasılığı hesaplanır. Sonuca sayma ile ulaşılır. Tüm dokümanlar arda arda eklenerek mega doküman elde edilir.
w_i sözcüğünün c_j sınıfında (konusunda) kaç kere geçtiği pay ifadesidir.
Payda ise, sözlükteki (V) yani tam sözcük tiplerinin birleşimindeki sözcüklerin c_j sınıfında (konusunda)