VBM682 Natural Language Processing (Doğal Dil İşleme)

Ödev 1 – Basic Text Processing & Language Models
Teslim Zamanı: 25 Kasım 2021 (Perşembe)

BÖLÜM 1: Yazılı Yanıt Bölümü (40 puan)

Soru 1. (10 puan)
Aşağıdaki her bir karakter dizesi (**string)** kümesi için bir düzenli ifade (**regular expression**) yazın. Yazacağınız düzenli ifadeler sadece kümelerdeki karakter dizeleri ile eşleşmeli.

- **a** ve **b** karakterlerinde oluşun ve ilk ve son harfleri farklı olan karakter dizelerinin kümesi. Yani {ab,ba,aab,abb,baa,bba,…} kümesi.
- Sadece küçük harflerden oluşan kelimeler kümesi.
- **ev** kelimesin çekilmiş hallerinin bir alt kümesi olan {*ev,eve,evi,evde,evden, evler,evlere,evleri,evlerde,evlerden, evim,evime,evimi,evimde,evimden, evlerim,evlerime,evlerimi,evlerimde,evlerimde*n} kümesi. Düzenli ifadeyi sadece kümedeki kelimeleri OR işlemcisi ile birleştirerek yazmayın. Kök kelime ve çekim eklerini kullanarak düzenli ifadeyi yazın.

Soru 2. (10 puan)
kalem ve **akal** kelimeleri arasındaki **Minimum Edit Distance** tablosunu **Backtrace** bilgileri ile birlikte oluşturun. Oluşturacağınız tablo **MinumumEditDistance** ders notlarımızın 16 sayfasındaki gibi olmalıdır.

Soru 3. (20 puan)
Aşağıdaki örnek metinin verildiğini kabul edin. Her satır bir cümleyi tutmaktadır.

<s> Sen dün eve geldin </s>
<s> Sen eve gittin </s>
<s> eve sen gittin </s>
<s> Eve geldin </s>

a) Bu metindeki *unigram* ve *bigram* sayılarını bulunuz. *Unigram* ve *bigramlarınızdaki* kelimeler yukarıdaki metinde gözüken kelimeler (büyük harf ve küçük harf ayrımı yapmadan), özel cümle başı sembolü ve özel cümle sonu sembolü olmalıdır. Bütün *unigramların* ve *bigramların* olasılık değerlerini verin.

b) Metinde bir kere geçen kelimeleri UNK kelimesi ile değiştirin ve metinde gözükmeyen kelimeler için UNK kelimesi olduğunu varsayın. Bu UNK kelimesini metindeki kelime listesine ekleyin ve metinde bir kere geçen kelimeleri kelime listesinden çıkarın. Add-k *smoothing* yöntemini (k değerini 0.5 alarak) düzeltilmiş (*smoothed*) olasılık değerlerini hesaplayarak verin. UNK kelimesinin de geçtiği n-gramlarda dahil olmak üzere bütün *smoothed unigram* ve *bigram* olasılık değerlerini verin.

c) (b) şıkkında elde ettiğiniz *bigram* değerlerini kullanarak aşağıdaki iki cümlenin olasılık değerlerini hesaplayın.

<s> Eve dün sen geldin </s>

<s> Sen sınıfa dün gittin </s>

BÖLÜM 2: Programlama Bölümü (60 puan)

Soru 4. (60 puan)
Programlama ödevi kısmında aşağıdaki işleri yapan bir Python programı yazacaksınız. Bu programlama bölümü için bir Python programı yazmanız tercih ederim. Başka bir programlama dili (Java) kullanabilirsiniz, ama programı o dillerde yazmak daha zor olacağını unutmayın.

Programınız aşağıdaki işleri yapmalıdır:

- Programınız klavyeden bir metin kütüğün (.txt kütüğü) ismini okuyacaktır. Bu kütük bir İngilizce metin saklayan bir kütük olacaktır.
- Programınız okuduğu kütükteki metini cümlelere ayırmalıdır. Programınız metindeki cümle sayısını bularak bastırmalıdır. Cümle sonlarını nokta, soru ve ünlem işaretini belirtiğini kabul edebilirsiniz.
- Programınız okunan metinin içindeki kelimeleri bularak toplam kelime sayısını ve tekil kelime sayısını bulmalıdır. Bir kelime sadece harflerden oluşmalıdır. Kelime sayılarını bulmadan önce bütün kelimeleri küçük harfe çevirin ve kelime sayılarını kelimeleri küçük harfe çevirdikten sonra bulunuz. Bulunan toplam kelime sayısını ve tekil (*unique*) kelime sayısını yazdırınız.
- Programınız metindeki *unigram* ve *bigram* sayılarını ve onların olasılık değerlerini bulmalıdır. Sayıları en fazla olan ilk 10 *unigramı* olasılık değerleri ile birlikte bastırın. Sayıları en fazla olan ilk 10 *bigramı da* olasılık değerleri ile birlikte bastırın.
- Metinde en az geçen 3 kelimeyi UNK kelimesi ile değiştirin ve metinde gözükmeyen kelimeler için UNK kelimesi olduğunu varsayın. Bu UNK kelimesini metindeki kelime listesine ekleyin ve metinde en az geçen 3 kelimeyi kelime listesinden çıkarın. *Bigram* olasılık değerlerini *Add-k smoothing* (k değerini 0.5 alarak) düzeltin. En fazla olasılık değerine sahip ilk 10 *bigaramı* tekrar yeni *smoothed* olasılık değerleri ile bastırın.
- Programız klavyeden bir cümle okumalı ve o cümlenin olasılık değerini *smoothed bigram* olasılık değerlerini kullanarak bulmalıdır. Programınız bulunan olasılık değerini bastırmalıdır. Eğer girilen cümle metinde hiç görülmeyen bir kelime içeriyorsa, programınız o kelimin UNK kelimesi olduğunu varsayması gerekir.

Programınızı en aşağı verilen üç örnek kütükle (`hw01_tiny.txt`, `hw01_AMemorableFancy.txt`,`hw01_FireFairies.txt`) deneyiniz. Programınızın her bir kütük için çalışmasını ve ürettiği sonuçları ödev ile birlikte teslim edin. Ödevle birlikte teslim etmek için, her bir örnek kütük için üretilen sonuçları tutan bir pdf kütüğü (hw01_tiny_Sonuc.pdf) veya png kütüğü yaratın. Bir sonuç kütüğün yaklaşık içeriği aşağıdaki gibi olmalıdır.

Kütükteki Cümle Sayısı: .....
Toplam Kelime Sayısı: .....
Unique Kelime Sayısı (Vocubalary Size): .....

En Sık Geçen İlk 10 Unigram:
unigram1 sayısı olasılığı
....

En Sık Geçen İlk 10 Bigram:
bigram1 sayısı olasılığı
....

UNK değiştirme ve Smooting işleminden sonra En Sık Geçen İlk 10 Bigram Değerleri:
bigram1 olasılığı
....

Örnek Cümle 1 olasılığı

Örnek Cümle 2 olasılığı

....

Teslim Edilecekler:

- Ödevinizi EVDEKAL sistemi üzerinden teslim edeceksiniz. Toplam 5 kütük yüklemelisiniz
- Bölüm 1 (Yazılı Yanıt Bölümü) deki ilk üç sorunun yanıtını içeren bir pdf kütüğü (**`hw01_bolum1.pdf`**) yükleyeceksiniz.
- Bölüm 2 (Programlama Ödevi) için yazdığınız programla ilgili 4 kütük yükleyeceksiniz.
    - Programınızın kodunu içeren *source* kütüğü (adı **`hw01_bolum2.txt`** olsun). Bu sadece sizin Python kodunuzu içeren bir kütük olmalı.
    - `hw01_tiny.txt` kütüğünün sonuçlarını ve iki örnek cümle olasılığını içeren **`hw01_tiny_Sonuclari.pdf`**.
    - `hw01_AMemorableFancy.txt` kütüğünün sonuçlarını ve iki örnek cümle olasılığını içeren **`hw01_AMemorableFancy_Sonuclari.pdf`**.
    - `hw01_FireFairies.txt` kütüğünün sonuçlarını ve iki örnek cümle olasılığını içeren **`hw01_FireFairies_Sonuclari.pdf`**.

- **ÖDEVLERİNİZİ KENDİNİZ YAPIN.**