

## 2016-2017 Güz Yarıyılı Makine Öğrenmesi 3. Ödevi

### Konu : Saklı Markov Modeli(HMM) ile yazı düzeltme

Bu ödevde, bazı harfleri hatalı yazılmış bir dokümandaki hataları HMM ile düzelteren bir sistem tasarlanacaktır. Verilen hatalı kelimeler(gözlenen bilgi-observation) için en yüksek olasılıklı harf sırası (gizli durum-hidden state) **Viterbi Algoritması** ile oluşturulacaktır.

#### İşlem Adımları :

1. **Eğitim :** HMM modelinin parametrelerini aşağıdaki gibi hazırlayınız.
  - a.  $I[x]$ : İlk durum olasılıkları. Bir kelimenin  $x$  harfi ile başlama olasılığı
  - b.  $A[x][y]$  : (Transition Probability Matrix) Durum Geçiş Olasılık Matrisi: Eğitim setinde bir kelime  $x$  harfinden sonra  $y$  harfi gelme olasılığı
  - c.  $B[x][o]$ : (Emission Probability Matrix) Çıkış Olasılık Matrisi: Eğitim setinde  $x$  harfi olması gerekirken  $o$  harfinin görülme olasılığı.

$I[x]$  dizisinde indisler harfleri göstermektedir. Örneğin  $i[0]$  adresi bir kelimenin 'a' harfi ile başlama olasılığını göstermektedir.  $A[x][y]$  ve  $B[x][o]$  matrislerinde satırlar ve sütunlar harfleri göstermektedir. Örneğin  $x=0, y=1$  için  $A[0][1]$  indisi bir kelime 'a' harfinden sonra 'b' harfi gelme olasılığını göstermektedir.

2. **Test:** Verilen test dokümanlarına **Viterbi algoritması uygulayarak** en yüksek olasılıklı harf sıralarını elde ediniz.

**Doküman Örnekleri:** docs data dosyasında İngilizce doküman parçaları bulunmaktadır. Dosyadaki ilk 20.000 karakterlik bölüm **test** için, geri kalan bölüm ise **eğitim** için kullanılacaktır. Dokümanların bulunduğu dosyada, ilk sütun doğru yazıyı, 2. sütun yanlış yazılmış halini göstermektedir. Test işlemi için yanlış yazı bölümü kullanılıp doğru yazı oluşturulmaya çalışılacaktır. Dokümandan örnek bir parça:

```
p p
e e
o o
p l
l l
e e

- -
w w
h h
o k

- -
p p
u u
r r
s s
u u
e e

- -
```

Eğitim örneklerini test için kesinlikle **kullanmayınız**.

### **Teslim Edilecekler:**

Aşağıda verilen bütün bilgileri içeren tek bir doküman hazırlayarak e-mail ile gönderiniz.

1. Yaptığınız çalışmayı yöntem, uygulama ve sonuç bölümlerinden oluşan bir raporda anlatınız.
2. Uygulama bölümünde farklı yazı örnekleri için düzeltilmiş çıkış dokümanını vererek HMM ile hatalı harflerin düzeltilme oranını hesaplayınız.
3. Sonuç bölümünde, sistem başarısını yorumlayınız.
4. Yazdığınız programların **kaynak kodlarını** açıklamalı olarak hazırlayarak veriniz.

### **Teslim Edilecekler:**

#### **1. Rapor:**

1. Yaptığınız çalışmayı giriş, yöntem, uygulama ve sonuç bölümlerinden oluşan bir **makale** yapısında anlatınız.
2. **Yöntem** bölümünde işlem adımlarını kısaca anlatınız.
3. **Uygulama** bölümünde :
  - a. I dizisini, A ve B matrislerini veriniz.
  - b. Yazıda başlangıçtaki
    - i. doğru kelime sayısını, yanlış kelime sayısını,
    - ii. doğru harf sayısını, yanlış harf sayısını,
    - iii. Viterbi uygulanması sonrasında doğru kelime sayısını, yanlış kelime sayısını,
    - iv. Viterbi uygulanması sonrasında doğru harf sayısını, yanlış harf sayısını hesaplayınız.
    - v. Yukarıda hesapladığınız değerleri kullanarak başarı oranlarını hesaplayınız
  - c. Farklı kelime örnekleri için doğru ve yanlış düzeltilmiş kelimelere örnekler veriniz.
4. **Sonuç** bölümünde sistemin başarısını değerlendiriniz. Başarısızlık sebeplerini kısaca yorumlayınız.

**2. Kaynak Kod :** Yazdığınız programların **kaynak kodlarını** açıklamalı olarak hazırlayarak veriniz. **Viterbi için hazır kod kullanmayınız.** Kendiniz yazınız.

**Teslim Tarihi:** Ödevlerinizi **27 Aralık 2016** tarihine kadar e-maile gönderiniz. Teslim edilecekleri tek bir **pdf** dosya içinde birleştirip dosyanın ismini okul numaranız ve ödev numarası olarak veriniz (09501002hw1 gibi). Geç teslimlerde hergün 10 puan değerindedir.