VBM682 Natural Language Processing (Doğal Dil İşleme)

Ödev 2 – Part of Speech Tagging
Teslim Zamanı: 29 Aralık 2023

Bu ödevde küçültülmüş BrownCorpus'u kullanarak HMM PartofSpeechTagger yaratacaksınız. Bu programlama ödevi için bir Python programı yazmanızı tercih ederim. Başka bir programlama dilde (Java) kullanabilirsiniz.

DataSet

DataSet Train (*brown_train_hw_noemptyline.txt*) ve Test (*brown_test_hw_noemptyline.txt*) olmak üzere iki kütük içermektedir. Kütüklerinin her bir satırı bir cümle tutuyor (boş satırlarda var ise; boş satırları göz önüne almayacaksınız). Örnek bir cümle aşağıdaki gibidir.

```
Attorneys/nns for/in the/at mayor/nn said/vbd that/cs an/at
amicable/jj property/nn settlement/nn has/hvz been/ben
agreed/vbn upon/rb ./.
```

Her kelime / karakteri ile işaretlenmiş olan Part-Of-Speech'inden ayrılmıştır.

Programınız aşağıdaki işleri yapmalıdır:

1. POS-Tagging için First-Order-HMM'in Yaratılması:

- Programınız Train kütüğünü okuyarak First-Order-HMM'u yaratacaktır. Bu demektir HMM için gerekli olan bütün olasılık verileri Train kümesinden toplanacaktır. HMM için gerekli olan bilgilere ek olarak başka bilgilerde (kelime frekansları, toplam kelime sayısı, …) toplamanız gerekebilir.
- Bütün kelimeleri önce küçük harfe çevirin. Böylece modelinizdeki bütün kelimeler sadece küçük harfler (ve diğer karakterlerden) oluşacaktır.
- HMM modeli yaratırken Train kümesinde frekansı en az olan 10 kelimeyi UNK (unknown) kelimesi olarak varsayın. Böylece bu kelimeler yerine UNK kelimesi kullanılmış gibi varsayacaksınız. Test kümesinin POS'lerini yaratılan HMM ile bulurken, eğer oradaki kelime Train kümesinde yok ise o kelime için UNK kelimesinin bilgilerini kullanacaksınız.

2. Train Kütüğünden Toplanan HMM Verilerin Kütüklere Yazılması:

- **Vocabulary.txt:** Train kümesinde geçen bütün kelimeleri, frekansları ve olası taglarını (olası tagların frekansı ile birlikte) ile birlikte bu kütüğe yazın. Eğer bir kelimenin k tane olası tagı var ise, bu kütüğün bir satırı aşağıdaki gibi olmalıdır.

  kelime kelimeFrekansı tag1:tag1frekansı … tagk:tagkfrekansı

  Bu kütüğün ilk satırında Train kümesindeki toplam kelime sayısını ve tekil kelime sayısını (Vocabulary Size) yazdırın.

  Bu kütüğün satırları kelimelerin alfabetik sırasına göre sıralanmış olmalıdır.

- **PosTags.txt**: Train kümesinde bulunan POStagleri frekansları ile birlikte PosTags.txt kütüğüne yazdırın. Bu kütüğün her bir satırı aşağıdaki gibi olmalıdır.

  tag tag_frekansı

  Bu kütüğün satırları **tag**'e göre alfabetik olarak sıralanmış olmalıdır.

- **TransitionProbs.txt:** HMM için yarattığınız TagTransitionProbability değerlerini bu kütüğe yazacaksınız. Bu kütüğün her bir satırı aşağıdaki verileri tutmalıdır.

  taga tagb P(tagb|taga)

  Bu kütüğün satırları önce **taga**'ya sonrada **tagb**'ye göre alfabetik olarak sıralanmış olmalıdır.

- **InitialProbs.txt:** POSTaglerin cümle başında gözükme olasılıklarını bu kütüğe yazın. Bu kütüğün her bir satırı aşağıdaki gibi olmalıdır.

  tag P(tag|<s>)

  Bu kütüğün satırları **tag**'e göre alfabetik olarak sıralanmalıdır.

- **EmissionProbs.txt:** POSTaglerin EmissionProbability değerlerini bu kütüğe yazın. Bu kütüğün her bir satırı aşağıdaki gibi olmalıdır.

  tag kelime P(kelime|tag)

  Bu kütüğün satırları önce **tag**'e sonrada kelimeye göre alfabetik olarak sıralanmalıdır.

3. Test Kümesinin HMM kullanılarak POSTaglerin bulunması ve elde edilen sonuçların kütüğe yazılması.

- Test kütüğün içindeki kelimelerin POSTaglerini yaratmış olduğunuz HMM'u kullanarak bulacaksınız ve başarı oranlarını hesaplayarak sonuç kütüğüne yazacaksınız. Kelimeleri ilk önce küçük harfe çevirmelisiniz.
- Test kütüğünün içindeki her cümlenin en olası POSTaglarını sırasını yaratmış olduğunuz HMM ve Viterbi algoritması yardımıyla bulacaksınız. Böylece cümledeki her kelimenin POSTagini bulmuş olacaksınız. Test kümesinde doğru PosTaglerde olduğu için modeliniz başarı değerini bulabilirsiniz.
- **Sonuc.txt:** Bu kütüğe Test kümesi için elde ettiğiniz başarı sonuçlarını yazacaksınız. Bu kütük sırasıyla aşağıdaki bilgileri içermelidir.
  - Test kütüğündeki toplam kelime sayısı
  - Test kütüğünde POSTagleri doğru bulunan kelime sayısı
  - Test kütüğünün cümleleri, bulunmuş POSTagleri ile birlikte Test kütüğüne benzer şekilde **Sonuc.txt** kütüğüne yazılmalıdır.

Teslim Edilecekler:

- Ödevinizi LİSANSÜSTÜ sistemi üzerinden teslim edeceksiniz. Toplam 7 kütük yüklemelisiniz

- Programınızın kodunu içeren ***source*** kütüğü.
- HMM Kütükleri ve sonuç kütüğü (6 kütük):
  - Vocabulary.txt
  - PosTags.txt
  - TransitionProbs.txt
  - InitialProbs.txt
  - EmissionProbs.txt
  - Sonuc.txt

- **Ödevlerinizi kendiniz yapın.**