Burak Bozdağ
504211552

HMM – NER

1. Bu ödevde "MilliyetNER" adlı veri kümesini kullandım (MilliyetNER). Milliyet gazetesinde 1997-1998 yılları arasındaki makaleleri toplayan bir veri kümesidir.

Eğitim ve sınama veri kümelerinde sırasıyla 419996 ve 49600 kelime bulunmaktadır. Bu kelimeler 7 farklı etikete sahip olabilir:

- B-PERSON
- I-PERSON
- B-LOCATION
- I-LOCATION
- B-ORGANIZATION
- I-ORGANIZATION
- O

Bu etiketleme formatı BIO format olarak da bilinir. B, I ve O harfleri sırasıyla "beginning", "inside" ve "outside" İngilizce tabirlerine karşılık gelir.

2. NER değerlendirme metrikleri temelde 3 farklı skora bağlanır:

- Precision: Modelin isabet oranını belirtir. Belirli bir etiketle işaretlenen verilerin kaçının doğru olduğunun oranını verir.
- Recall: Modelin etiketler arası ayrım yapabilme kabiliyetini belirtir. İlgili etikete sahip olan verilerin kaçının gerçekten o etiketle tahmin edildiğinin oranını verir.
- F1 skoru: Yukarıdaki 2 oranın harmanlanmasıyla elde edilen bir skordur. İkisi arasında bir denge faktörü istediğimizde kullanırız. [F1 = 2 * P * R / (P + R)]

Bu skorlar seçilen bir etiket üzerinde uygulanabilir (entity-level) olmasıyla birlikte model geneli de uygulanabilir (model-level). Model seviyesinde değerlendirme yaparken mantıken "yanlış pozitif" ile "yanlış negatif" değerleri birbirine eşit çıkacaktır.

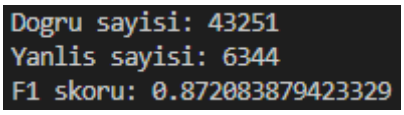
3. HMM denetimli öğrenme modelini eğitirken A ve B matrislerini hesapladım. A (transition) matrisi belli bir etiketten sonra başka bir etiketin gelme olasılığını tutuyor. B (emission) matrisi ise belli bir etiketin belli bir kelime olma olasılıklarını tutar.

```
Train kelime sayisi:  419996
Test kelime sayisi:  49600
Etiketler ['B-PERSON', 'I-PERSON', 'O', 'B-LOCATION', 'B-ORGANIZATION', 'I-ORGANIZATION', 'I-LOCATION']
Farkli kelime sayisi:  59348
B:  7 x 59348
A:  8 x 8
```

Şekil 1. Eğitim sonrası bilgilendirme çıktısı

A matrisinde alışılagelmişin dışında satır ve sütun olarak +1 boyut ekledim. Satırdaki boyut artışı "Start" olarak nitelendirdiğim başlangıç olasılıklarını, sütundaki ise "End" yani bitiş etiketi olasılıklarını gösterir.

Test veri kümesinde sınama işlemi Viterbi algoritması yardımıyla yapılıyor. Her cümleyi bu fonksiyon içinde dinamik programlama mantığıyla kullanarak tahmin çıktısını elimdeki etiketlenmiş test verisiyle karşılaştırdım. Yaklaşık 1 dakikalık sınama işlemleri sonucu doğru ve yanlış sayıları resimdeki gibi oldu:



Şekil 2. Test veri seti sınaması sonrası doğru-yanlış kelime sayıları

F1 skoru olarak gösterilen skor model geneli hesaplandığı için genel isabet oranını gösterir (precision=recall). Modelimiz test verisi üzerinde %87'den fazla başarım elde etmiştir.

Google Colab linki:
https://colab.research.google.com/drive/1dNzGlnEE5sPNLbTczgseWhQGXcIqNS62?usp=sharing