Doğal Dil İşleme
ÖDEV 1

İbrahim Okan Akveç

İçerik

- Ödev İçeriği
- Stanza
- Google Colab

Ödev İçeriği

1. Cümlelere ayırma (Sentence Segmentation)
2. Birimlere ayırma (Tokenization)
3. Kök indirgeme (Stemming)
4. Sözcük indirgeme (Lemmatization)
5. Etkisiz sözcükleri (stopwords) çıkarma
6. Sözcük türü (Part-of-speech) etiketleme
7. Noktalama işaretlerini kaldırma (Remove punctuations)

Stanza Kütüphanesi

- Stanza, Stanford NLP Group tarafından Python için geliştirilmiş doğal dil analiz paketidir.
  - Metin içeren bir diziyi cümle ve kelime listelerine dönüştürmek (Sentence Segmentation, Tokenization)
  - Kelimelerin temel biçimlerini, konuşma bölümlerini ve morfolojik özelliklerini oluşturmak (Lemmatization)
  - Sözdizimsel bir yapı bağımlılığı ayrıştırması vermek (Part-of-speech)
  - Adlandırılmış varlıkları tanımak (Recognize named entities)
- gibi görevler için bir pipeline kullanılabilecek araçları içerir.
- 70'ten fazla dil ile çalışacak şekilde tasarlanmıştır.
(Universal Dependencies formalism)

Stanza & Java

- Stanza client&server mimarisi ile Standford NLP Group tarafından Java ile geliştirilmiş olan CoreNLP kütüphanesine erişilmesine olanak sağlar.
- CoreNLP şuanda sadece 8 dil ile çalışmaktadır(Arapça, Çince, İngilizce, Fransızca, Almanca, Macarca, İtalyanca ve İspanyolca)

Stanza İstatistikleri

- Stanza vs NLTK vs spaCy

Stanza Kullanımı

pip install stanza

```
import stanza

stanza.download('en') # download English model

nlp = stanza.Pipeline('en') # initialize English neural pipeline

doc = nlp("Barack Obama was born in Hawaii.") # run annotation over a sentence
```

Google Colab Link

- Ödev Linki: https://colab.research.google.com/drive/1ZwlGDD4gGToTJ4VY-ybDHhZSjG8ulO7m#scrollTo=jgwRydxBj7tn