Metin Madenciliğine Giriş

Metinden Tema Çıkarma

Mert Arslan

Doğuş Üniversitesi
İstanbul, Türkiye
Mayıs 2021

Dataset json formatında bir kitap.

```
def load_data(file):
    with open (file, "r", encoding="utf-8") as f:
        data = json.load(f)
    return (data)

def write_data(file, data):
    with open (file, "w", encoding="utf-8") as f:
        json.dump(data, f, indent=4)
```

Json tipinde datayı kolayca yüklemek ve yazmak için kullanılan kod bloğu.

```
stopwords = stopwords.words("english")
```

```
print (stopwords)
```

Data set inglizce olduğu için stopwordsü İngilizce seçildi.

```
def lemmatization(texts, allowed_postags=["NOUN", "ADJ", "VERB", "ADV"]):
    nlp = spacy.load("en_core_web_sm", disable=["parser", "ner"])
    texts_out = []
    for text in texts:
        doc = nlp(text)
        new_text = []
        for token in doc:
            if token.pos_ in allowed_postags:
                new_text.append(token.lemma_)
        final = " ".join(new_text)
        texts_out.append(final)
    return (texts_out)


lemmatized_texts = lemmatization(data)
print (lemmatized_texts[0][0:90])
```

```
ıame be bear small town call be bear be very hard work child father mother have small mill
```

Bu kod bloğunda sadece İsim, sıfat, fiil, zarfların kalmasına izin verildi. Parser ve neri daha hızlı çalşıması için kapattıldı. Sadece İsim, sıfat, fiil olan tokenlerin kökleri(lemma) bulundu.

```
def gen_words(texts):
    final = []
    for text in texts:
        new = gensim.utils.simple_preprocess(text, deacc=True)
        final.append(new)
    return (final)

data_words = gen_words(lemmatized_texts)

print (data_words[0][0:20])
```

```
['name', 'be', 'bear', 'small', 'town', 'call', 'be', 'bear', 'be',
```

Stopwordler(my,if,or..) çıkarıldı ve kelimeler tek tek ayrıldı.

```
id2word = corpora.Dictionary(data_words)

corpus = []
for text in data_words:
    new = id2word.doc2bow(text)
    corpus.append(new)

print (corpus[0][0:20])

word = id2word[[0][:1][0]]
print (word)
```

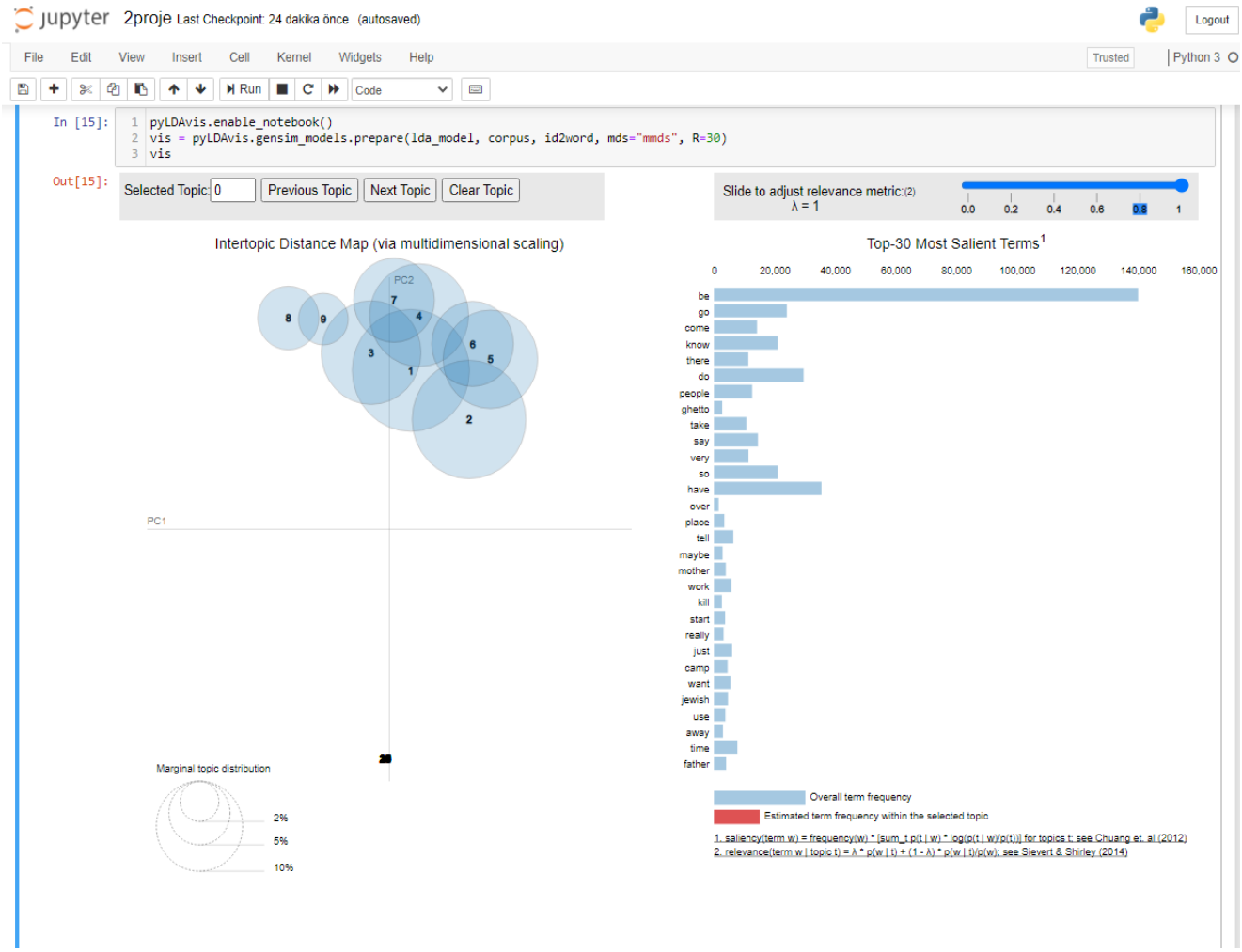
Kelimeleri ve o kelimelerin frekanslarını gösteren bir sözlük oluşturuldu.

```
lda_model = gensim.models.ldamodel.LdaModel(corpus=corpus,
                                            id2word=id2word,
                                            num_topics=30,
                                            random_state=100,
                                            update_every=1,
                                            chunksize=100,
                                            passes=10,
                                            alpha="auto")
```

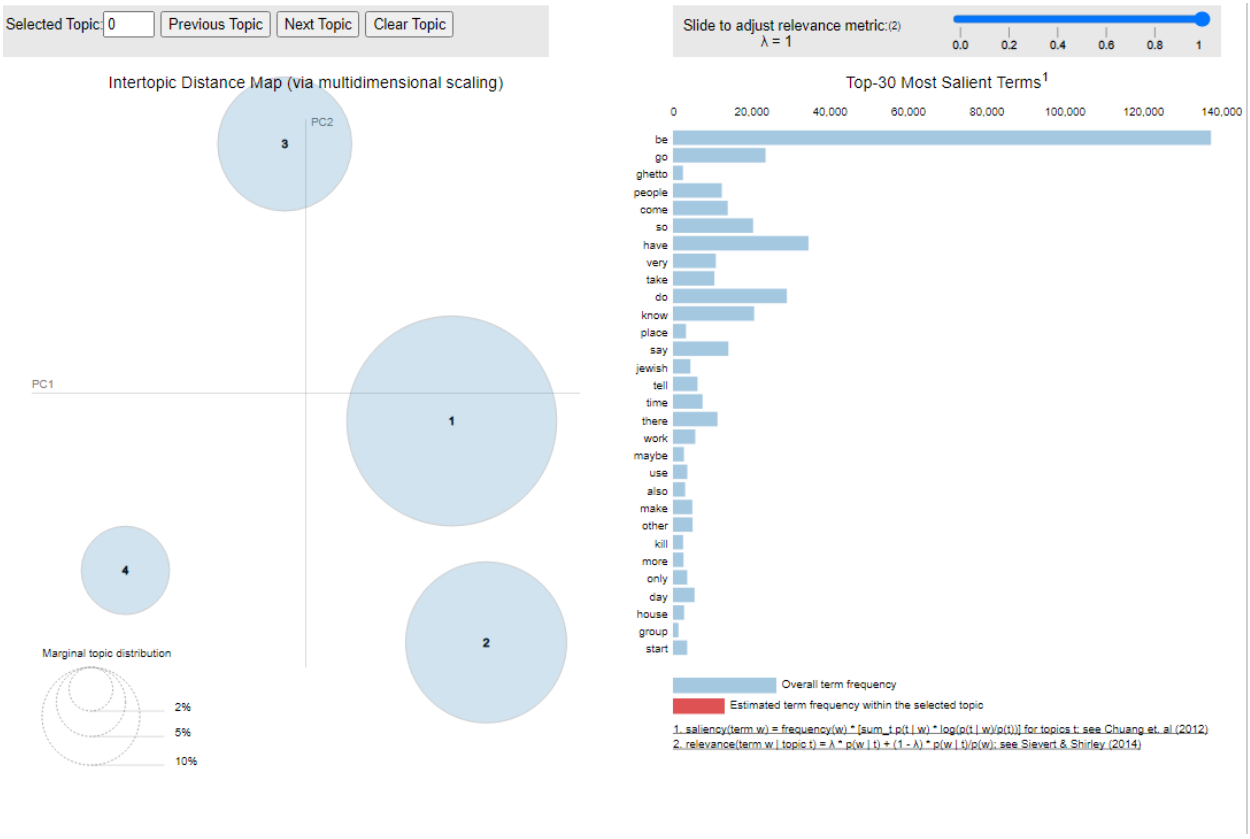
LDA topic modeli oluşturuldu.

```
pyLDAvis.enable_notebook()
vis = pyLDAvis.gensim_models.prepare(lda_model, corpus, id2word, mds="mmds", R=30)
vis
```

LDA topic modeli gösterildi.



LDA topic modelinde clusterlar üst üste bindi preproscesing kısmında sorun olablir. Tema olabilecek kelimeler know, go, take gibi kelimeler gösterdi bunlar filtrelenerek daha iyi bir sonuç alınabilir.



Number of topics kısmını 4 e düşürdüğümde clusterlar daha iyi bir şekilde ayırıldı. Tema olabilecek kelimelerde anlamsız olanlar filtrelenmeli.