Doğal Dil İşleme Ödev Raporu

1. GİRİŞ

Bu projede, Reddit platformundan elde edilen gerçek kullanıcı yorumları üzerinde doğal dil işleme teknikleri uygulanarak metin madenciliği yapılmıştır. Amacımız, ham metinlerin uygun ön işlem adımları sonrası TF-IDF ve Word2Vec gibi vektörleştirme yöntemleriyle anlamlı hale getirilmesini sağlamaktır. Metinlerin özellikleri çıkarılmış ve benzerlik analizi gibi görevler için kullanılabilir hale getirilmiştir.

2. VERİ SETİ VE VERİ ÇEKME SÜRECİ

Veri Reddit platformundaki r/soccer subreddit'inden praw kütüphanesi kullanılarak çekilmiştir. Toplamda 200 gönderi taranmış ve bu gönderilere yapılan yorumlar alınarak reddit_soccer_comments.csv dosyası oluşturulmuştur. Çekilen veriler İngilizce olup futbol temalı genel kullanıcı yorumlarını içermektedir.

Özellikler:

Özellik	Değer
Toplam yorum sayısı	~6000+
Dosya formatı	CSV
Kaynak	Reddit API (PRAW)
Dosya adı	reddit_soccer_comments.csv

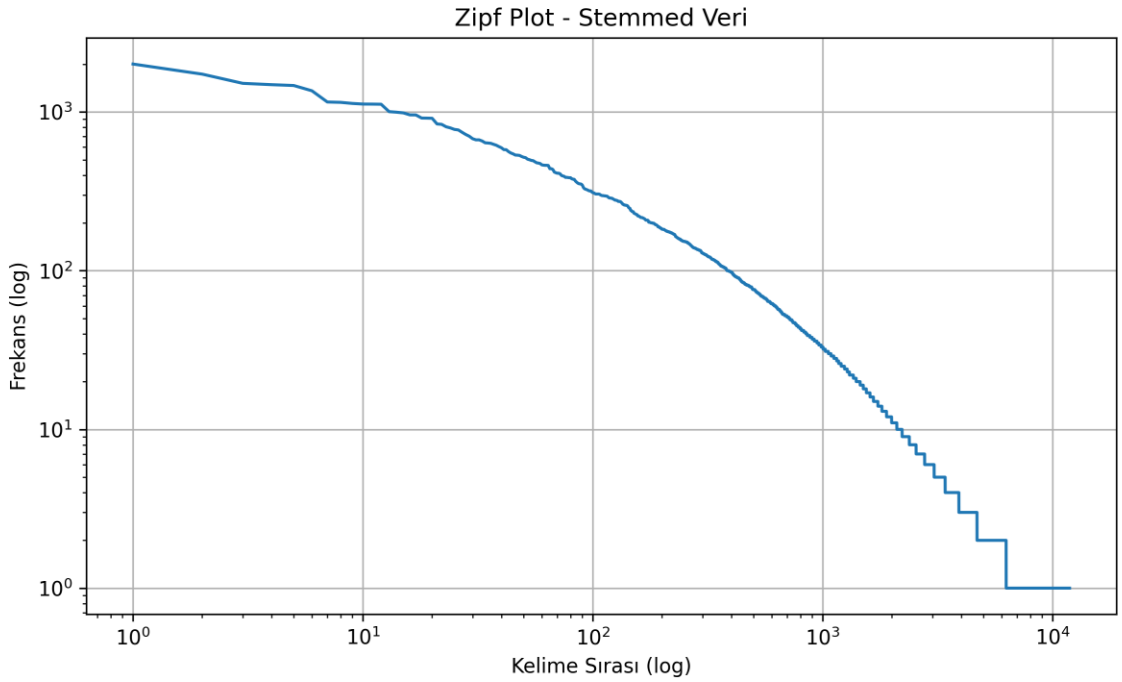
3. ÖN İŞLEME ADIMLARI

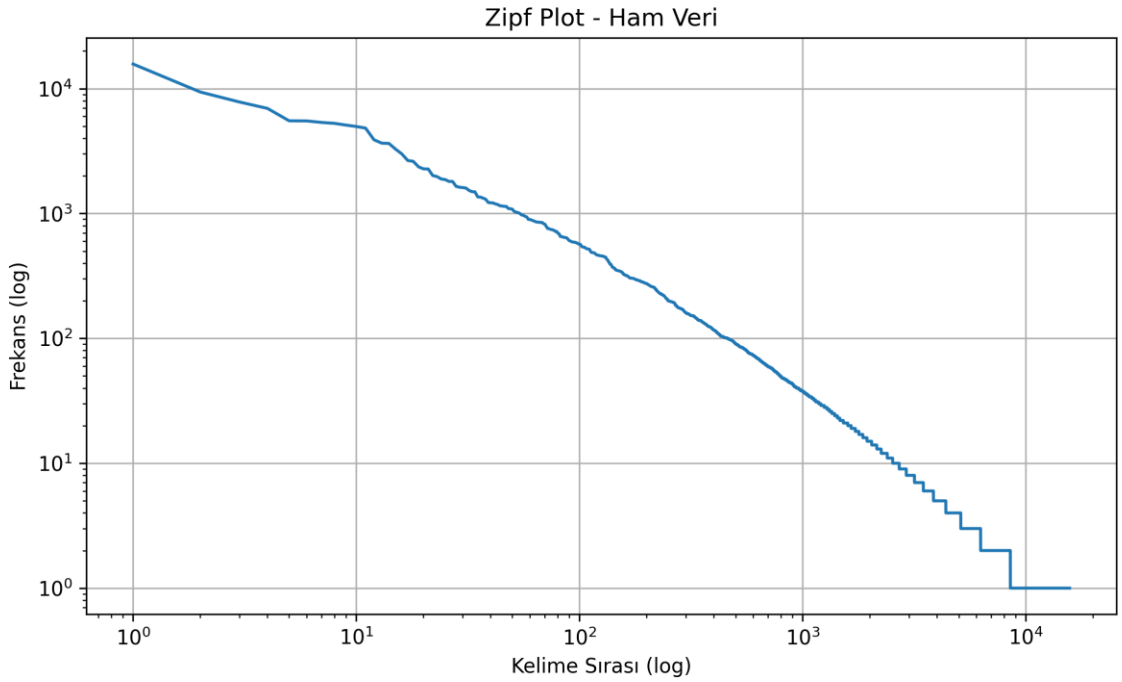
Yorumlar üzerinde aşağıdaki adımlar uygulanmıştır:
- Küçük harfe dönüştürme (lowercasing)
- Noktalama işaretlerini temizleme (regex)
- Stopword (gereksiz kelime) temizliği (nltk.stopwords)
- Tokenization (nltk.word_tokenize)
- Lemmatization (WordNetLemmatizer)
- Stemming (PorterStemmer)

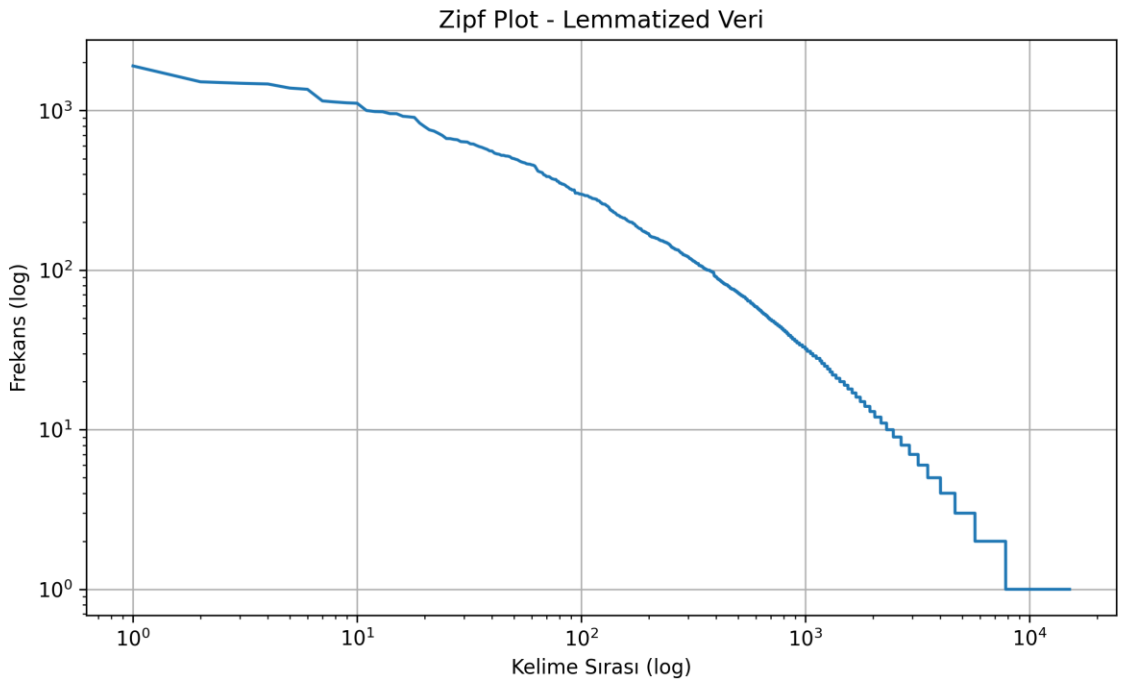
Lemmatize edilmiş yorumlar reddit_soccer_clean_comments.csv dosyasında; stemmed yorumlar ise doğrudan TF-IDF işlemi sırasında elde edilmiştir.

4. ZIPF YASASI ANALİZİ

Ham, lemmatized ve stemmed veriler üzerinde Zipf yasası log-log grafikleri çizilmiştir.







Yorumlar:
- Lemmatization en dengeli ve anlamlı dağılımı vermiştir.
- Temizleme sonrası kelime sayısı azalmış ama bilgi yoğunluğu artmıştır.
- Zipf eğrisi verinin istatistiksel anlamda uygunluğunu göstermektedir.

5. VEKTÖRLEŞTİRME – TF-IDF

Her iki veri setine TF-IDF vektörizasyonu uygulanmıştır. Elde edilen DataFrame'ler:

Lemmatized: tfidf_lemmatized.csv
Stemmed: tfidf_stemmed.csv

6. VEKTÖRLEŞTİRME – WORD2VEC

Toplam 16 farklı model eğitilmiştir (lemma & stem × CBOW & SkipGram × 2 window × 2 boyut). Model açıklamaları generate_model_descriptions.py ile otomatik oluşturulmuştur.

Kelime	Benzerlik Skoru
4	0.9919
month	0.9918
minute	0.9917
10	0.9916
two	0.9915

7. SONUÇ VE DEĞERLENDİRME

Bu projede veri çekiminden model eğitimine kadar tüm süreçler otomatikleştirilmiş ve hem istatistiksel (TF-IDF) hem de anlamsal (Word2Vec) metin temsilleri başarıyla elde edilmiştir. Lemmatized veri ile eğitilen modeller, stemmed veriye göre daha tutarlı sonuçlar vermiştir. Zipf analizleri, verinin doğal dil yapısına uygun olduğunu kanıtlamıştır. Bu yapı, gelecekte metin sınıflandırma ve öneri sistemlerinde temel oluşturabilecek güçlü bir altyapıdır.