Collocations Extraction of Judiciary Dataset

CSE4095 Natural Language Processing Project

150117010 Süleyman Ahmet SÖNMEZ
150117036 Muhammed Enes AKTÜRK
150117064 Yunus Emre ERTUNÇ
150117004 Güneş YÜZAK
150117048 Fatih BAŞ

Outline

- Preprocessing the Data
- Collocations Extraction
- Summary

Preprocessing the Data

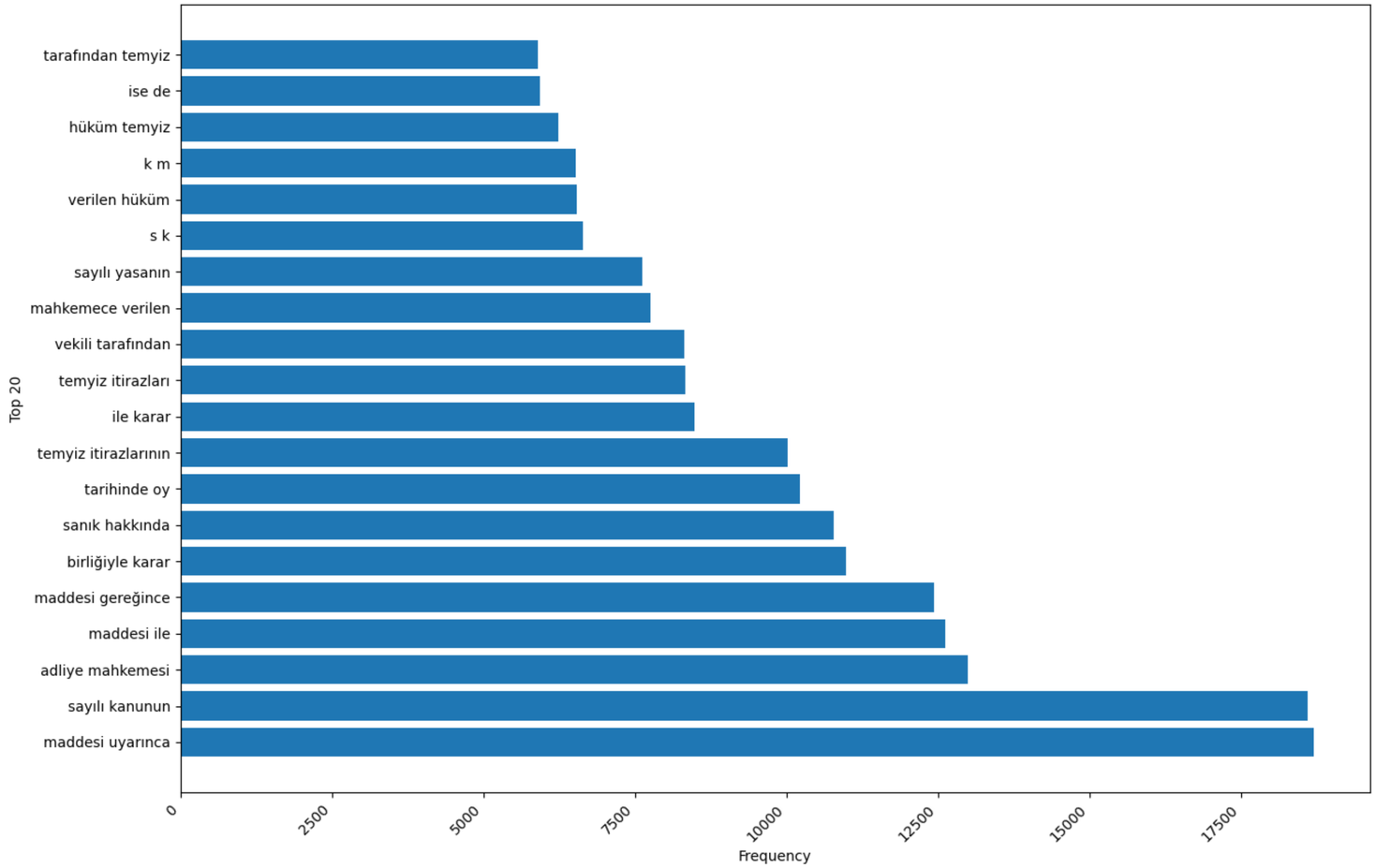
We analyzed the dataset. As a result, we got two different tokenized data outputs. One is reduced to the root of the words, the other is the raw format.

Collocations Extraction

- Frequency Method
- Mean and Variance Method
- Hypothesis Testing: T-test

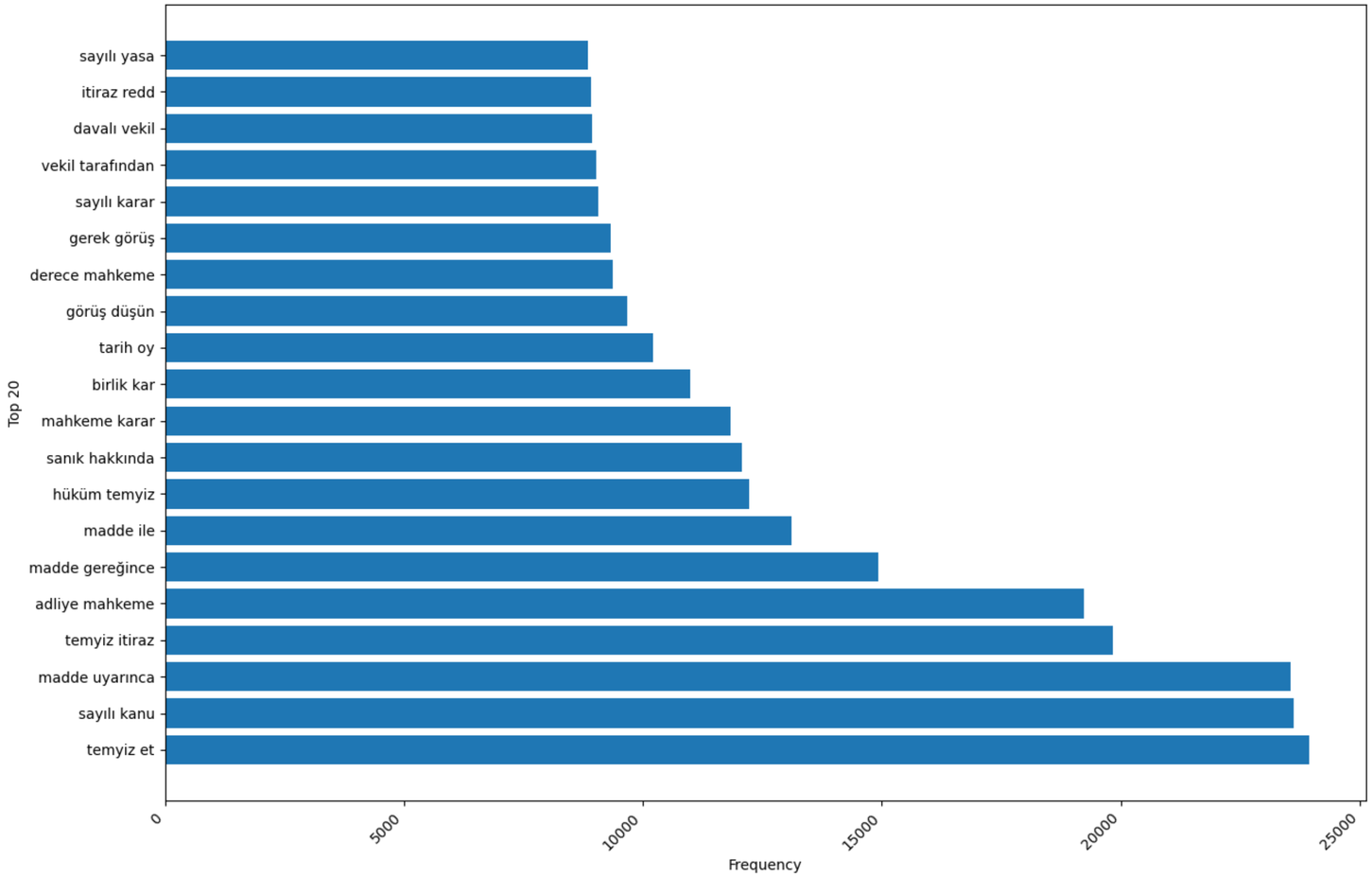
Frequency Method

with raw data



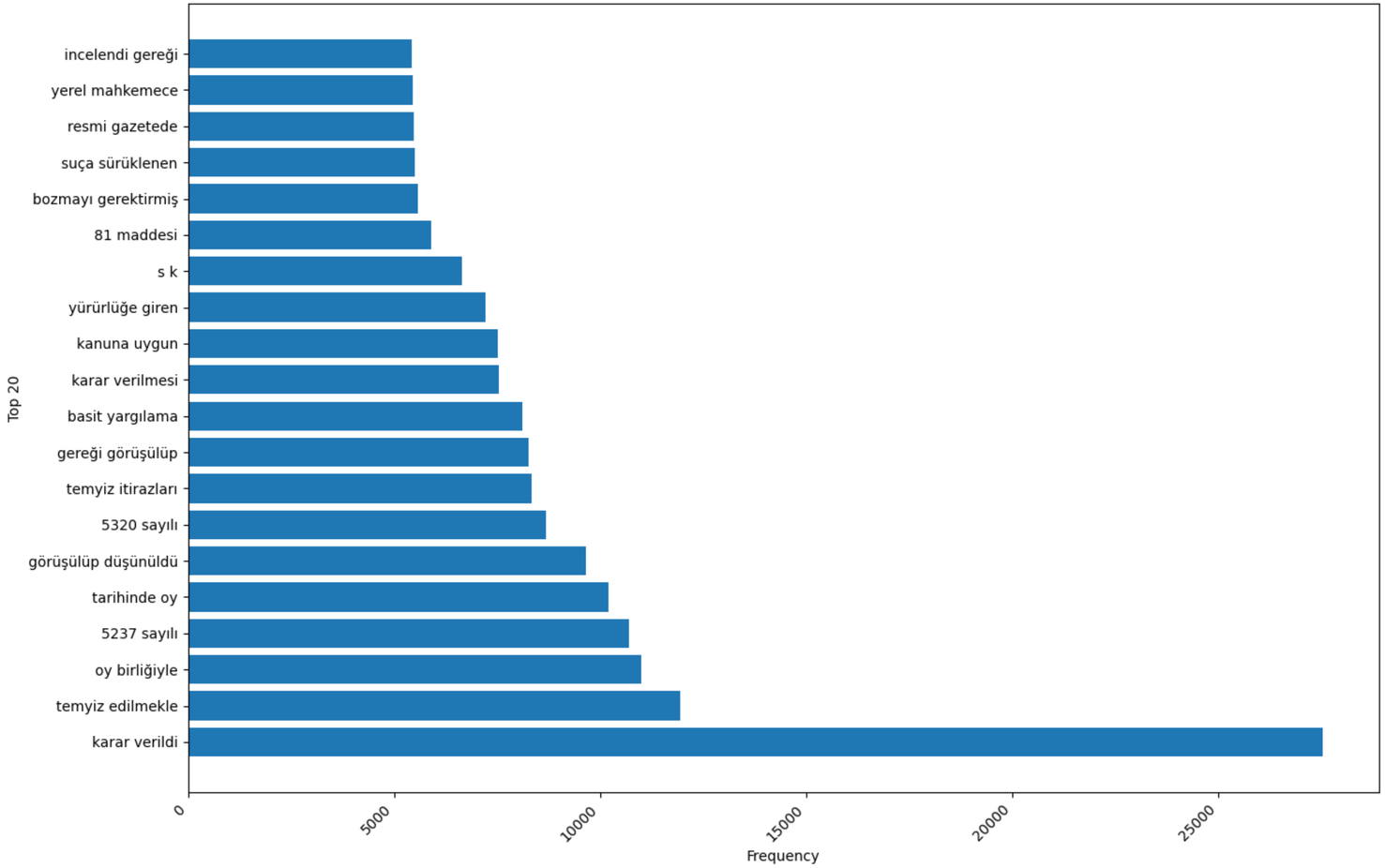
Frequency Method

with stemmed data



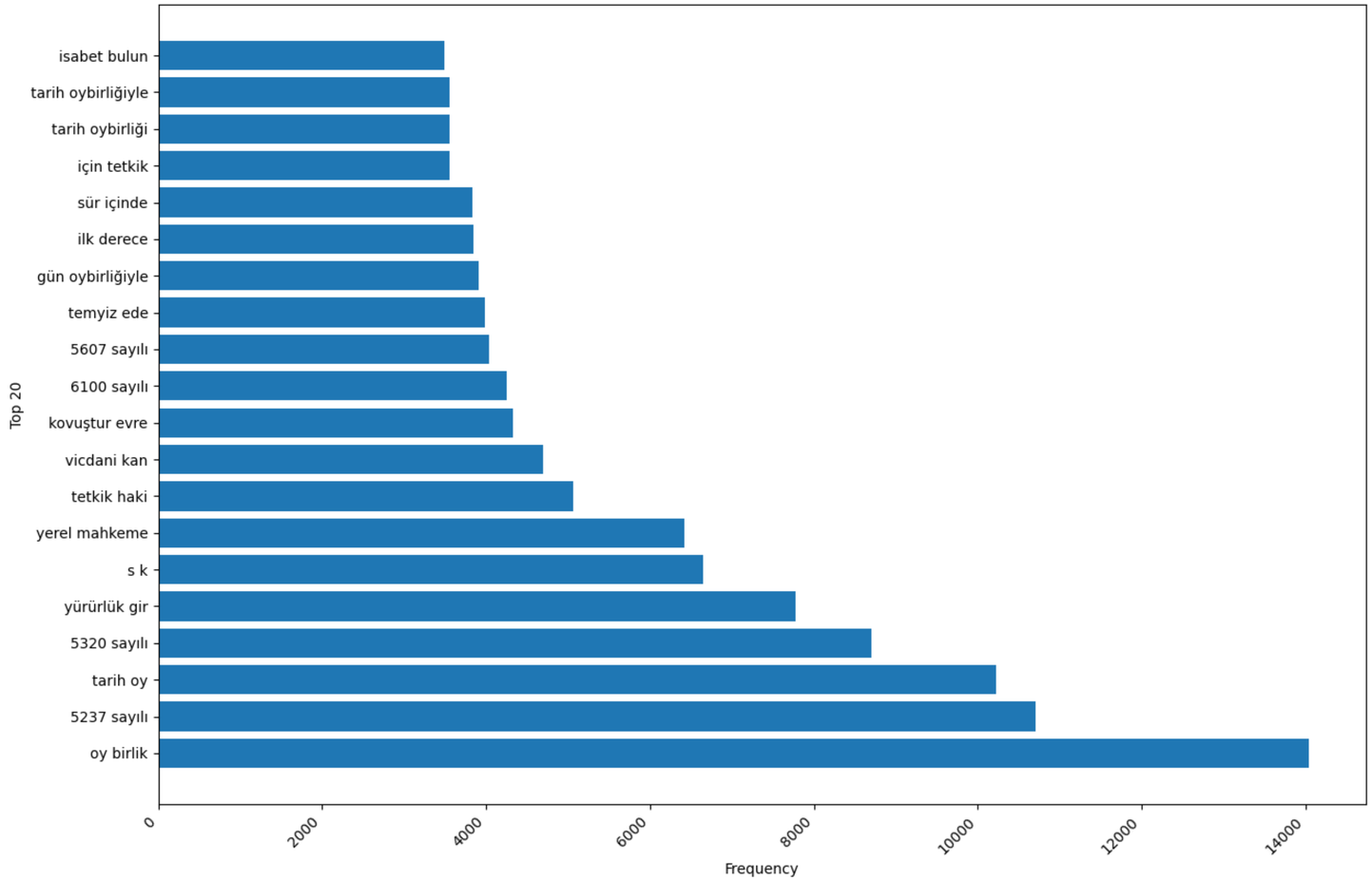
Mean and Variance Method

with raw data



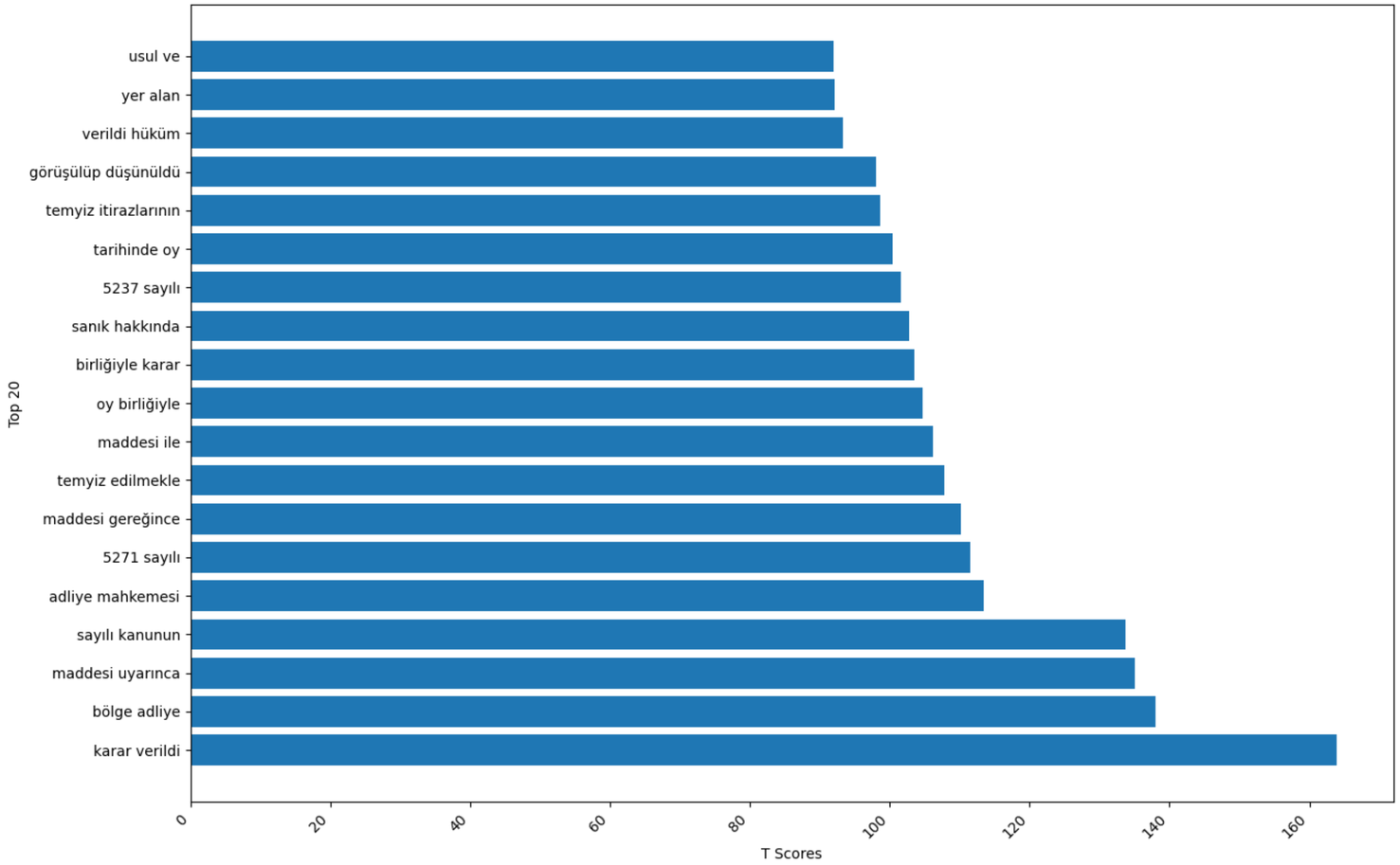
Mean and Variance Method

with stemmed data



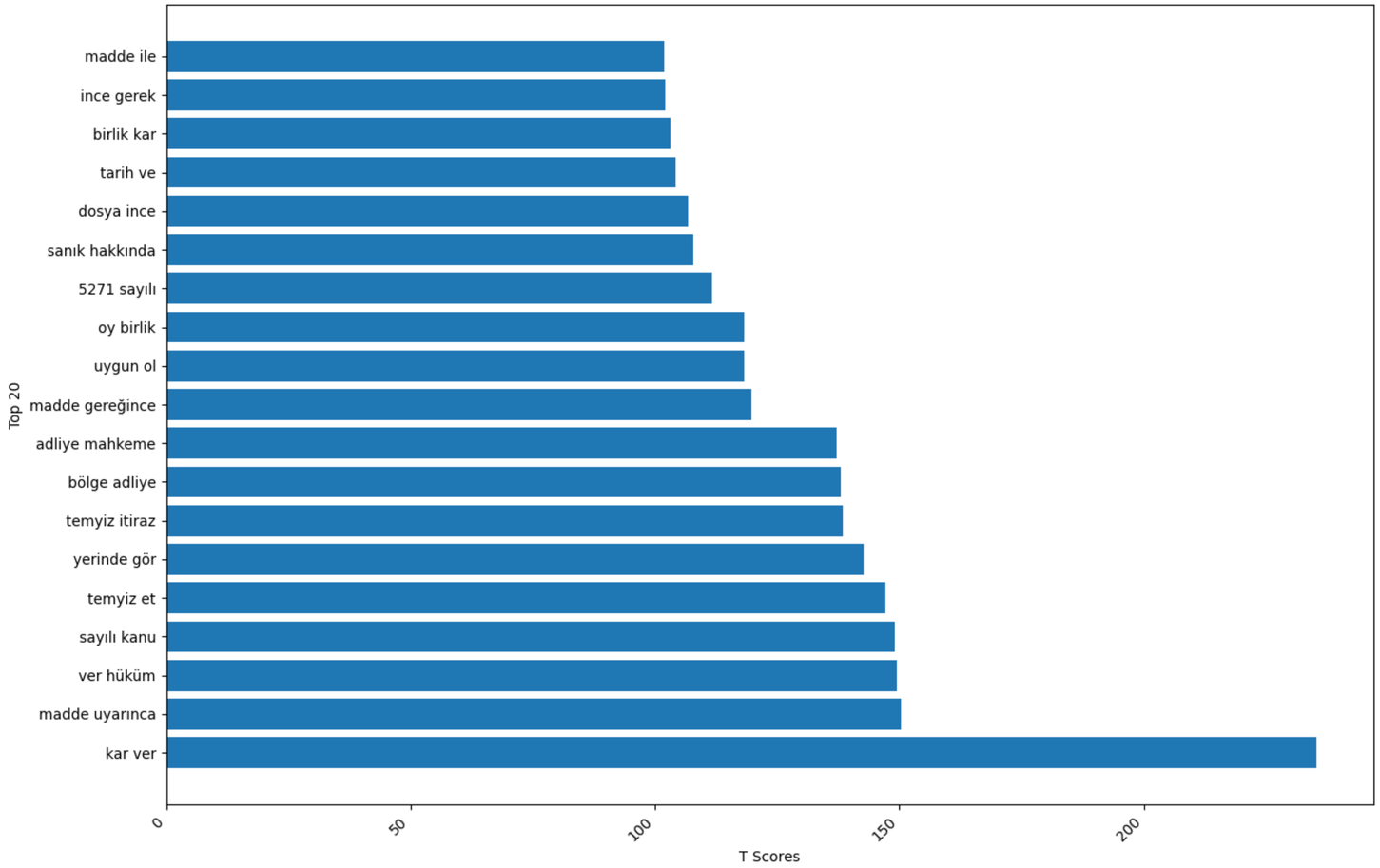
Hypothesis Testing: T-test

with raw data



Hypothesis Testing: T-test

with stemmed data



Summary

- Comparison of methods