N-gram Analysis of TBMM Corpus

CSE4095 - Introduction to Language Processing

ilker Fener - 150115024
Asem Okby - 150119688
Diala Jubeh - 150119686
Nurhande Akyüz - 150119933

Associate Professor Dr.
Murat Can Ganiz

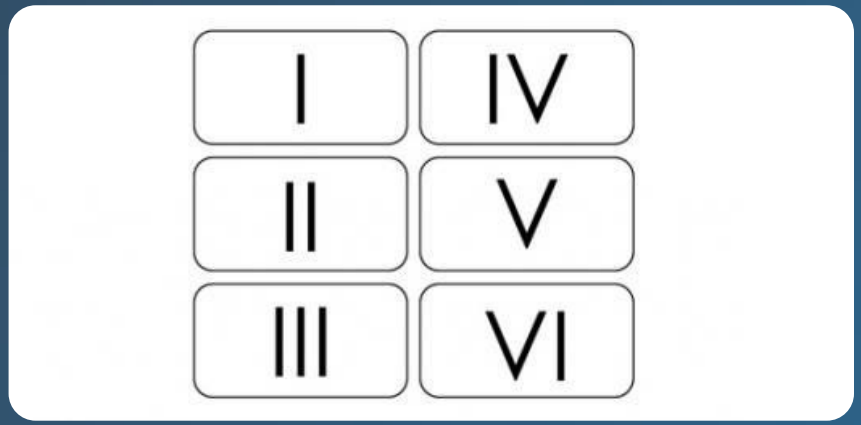
Outline

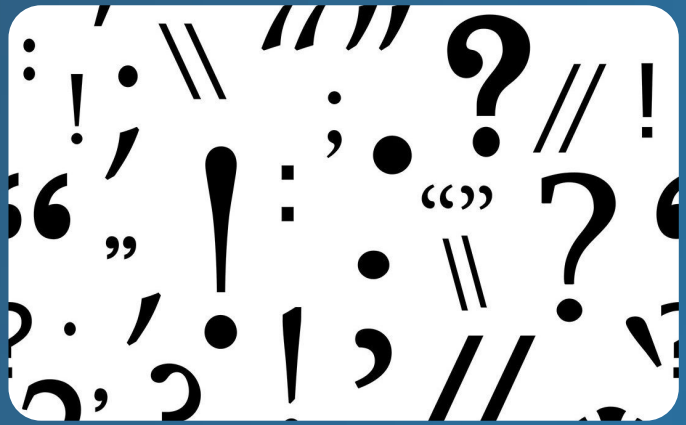
- Introduction
- Preprocessing The Data
- N-Grams Extraction
- Statistical Estimators
- Summary

Introduction

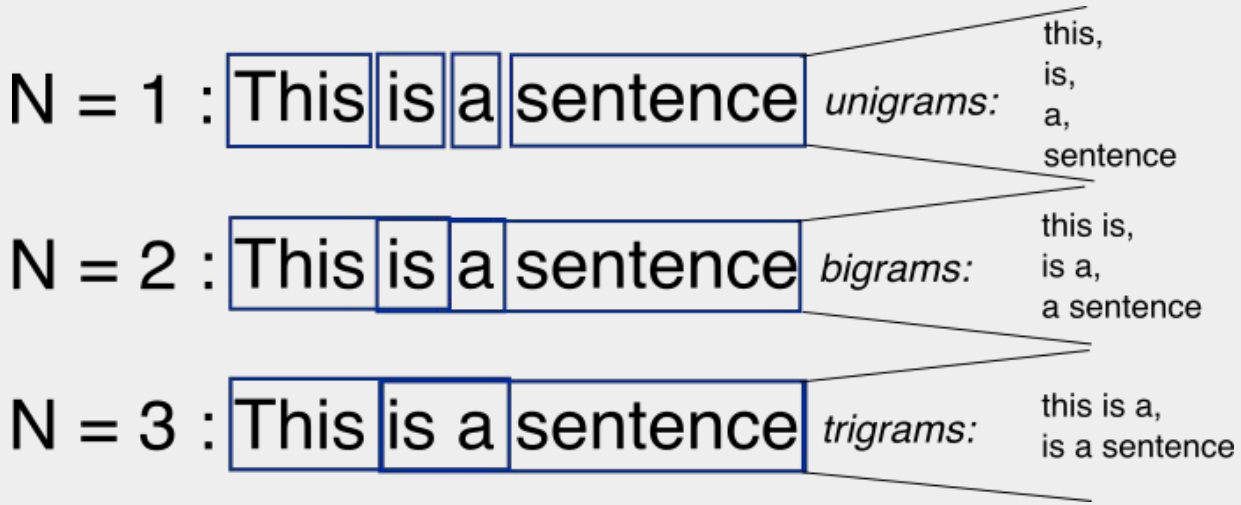
- Aim of the project
- Describing the dataset
- The produced results

Preprocessing The Data





N-Grams Extraction



Most Frequent 10 Unigrams

with stopwords

n-gram	frequency
('ve',),	214589
('bir',),	122458
('bu',),	120480
('sayın',),	75078
('başkan',),	51861
('da',),	46926
('de',),	45148
('türkiye',),	37773
('milletvekili',),	36544
('için',),	35309

Donem - 20

without stopwords

n-gram	frequency
('sayın',),	75078
('başkan',),	51861
('türkiye',),	37773
('milletvekili',),	36544
('büyük',),	25584
('ilişkin',),	25143
('genel',),	23211
('kabul',),	22784
('soru',),	22544
('yazılı',),	21019

Most Frequent 10 Bigrams

with stopwords

n-gram	frequency
('sayın', 'başkan'),	27219
('sıralarından', 'alkışlar'),	18196
('büyük', 'millet'),	18005
('türkiye', 'büyük'),	17929
('değerli', 'milletvekilleri'),	17827
('soru', 'önergesi'),	17528
('başkan', 'sayın'),	15317
('millet', 'meclisi'),	13174
('yazılı', 'soru'),	11618
('başkanlığa', 'geliş'),	11238

Donem - 21

without stopwords

n-gram	frequency
('sayın', 'başkan'),	27219
('sıralarından', 'alkışlar'),	18196
('büyük', 'millet'),	18005
('türkiye', 'büyük'),	17929
('değerli', 'milletvekilleri'),	17827
('soru', 'önergesi'),	17528
('başkan', 'sayın'),	15317
('millet', 'meclisi'),	13174
('yazılı', 'soru'),	11618
('başkanlığa', 'geliş'),	11238

Most Frequent 10 Trigrams

with stopwords

Donem - 22

n-gram	frequency
('mil', 'let', 've'),	42194
('ve', 'ki', 'li'),	31622
('let', 've', 'ki'),	29844
('ar', 'ka', 'daş'),	22557
('tür', 'ki', 'ye'),	22071
('de', 'ğer', 'li'),	20510
('de', 'vam', 'la'),	20130
('başkanlığa', 'geliş', 'tarihi'),	19192
('yazılı', 'soru', 'önergesi'),	19016
('türkiye', 'büyük', 'millet'),	17819

without stopwords

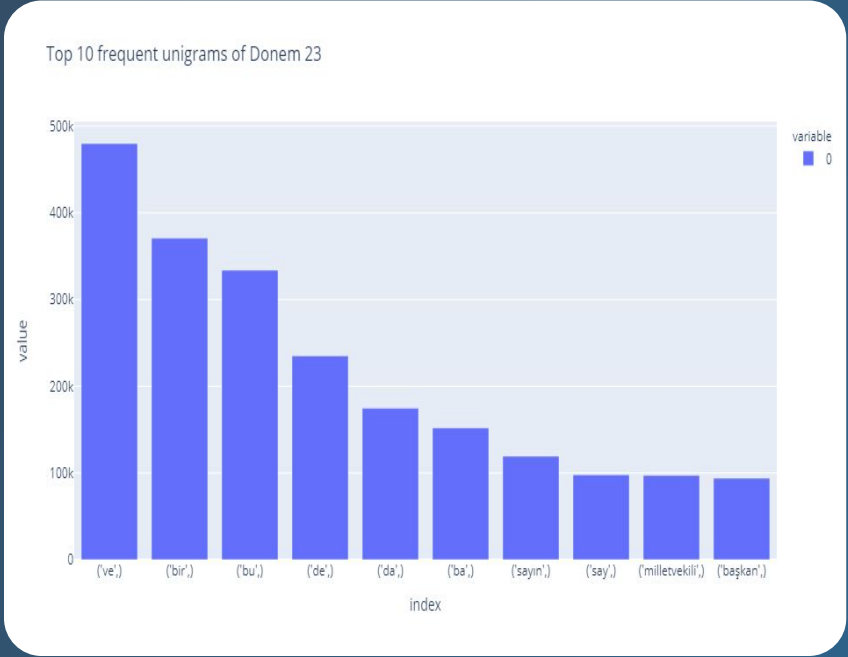
n-gram	frequency
('mil', 'let', 'li'),	29549
('ar', 'ka', 'daş'),	22557
('başkanlığa', 'geliş', 'tarihi'),	19192
('yazılı', 'soru', 'önergesi'),	19020
('türkiye', 'büyük', 'millet'),	17821
('önergesi', 'başkanlığa', 'geliş'),	16772
('li', 'ar', 'ka'),	16502
('soru', 'önergesi', 'başkanlığa'),	16349
('sunuyorum', 'kabul', 'edenler'),	14735
('oylarınıza', 'sunuyorum', 'kabul'),	14678

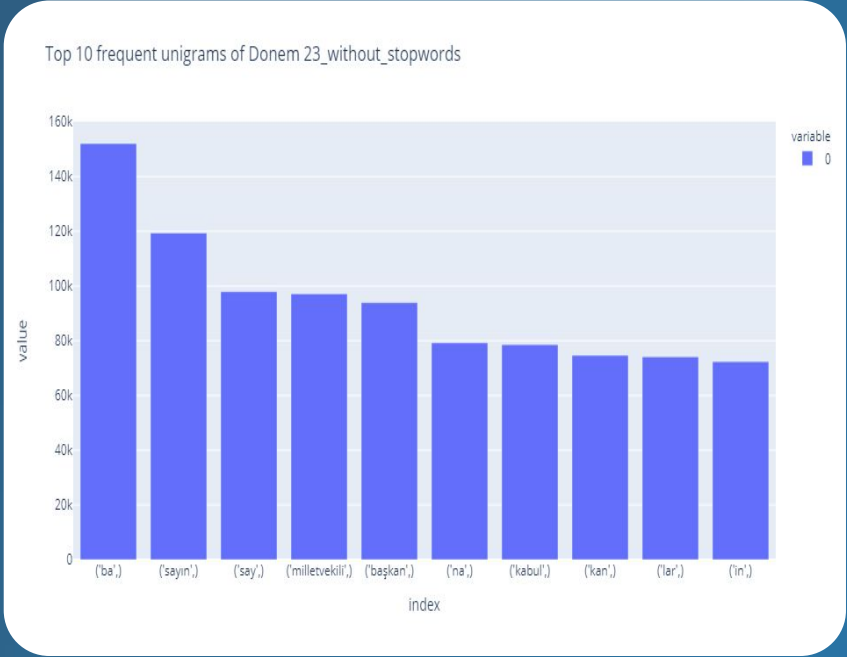
Most Frequent 10 Unigrams

with stopwords

without stopwords

Donem - 23



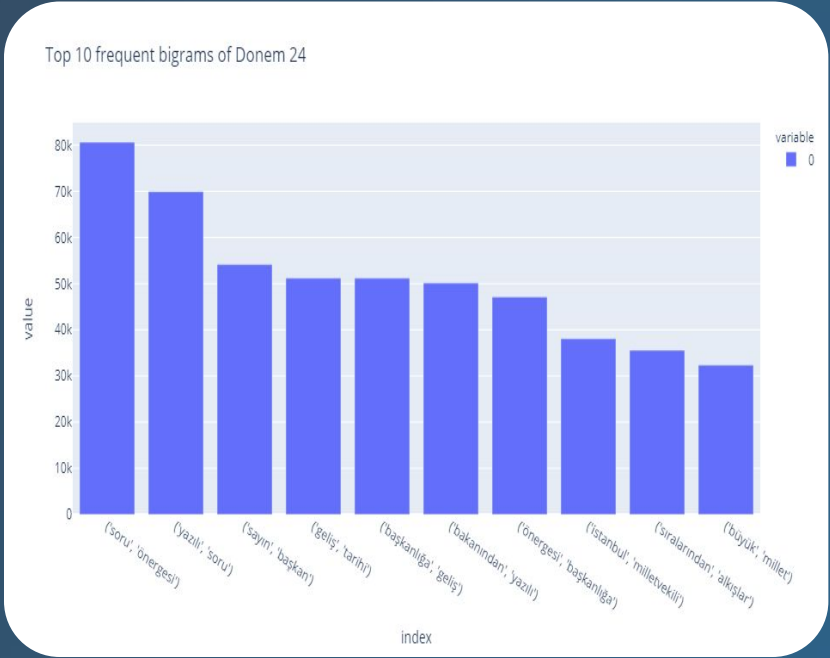


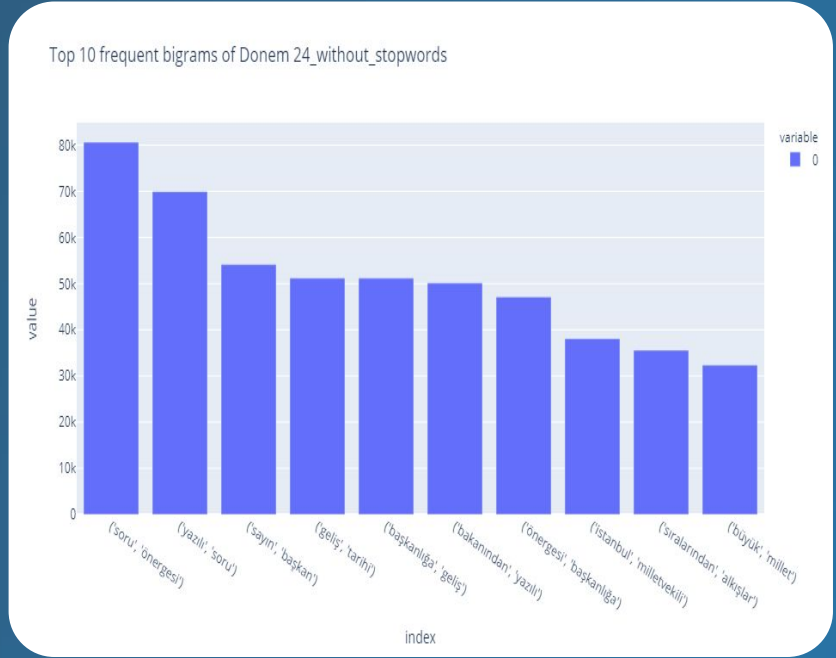
Most Frequent 10 Bigrams

with stopwords

without stopwords

Donem - 24



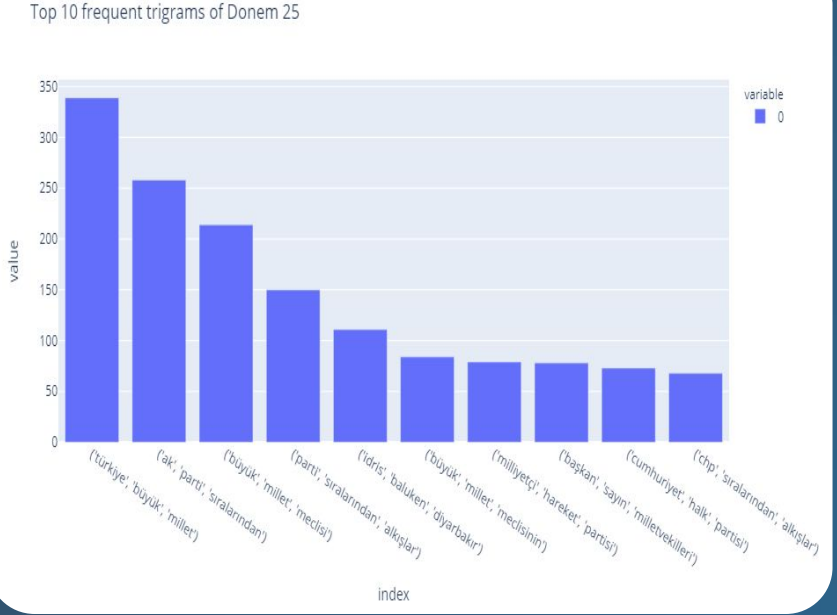


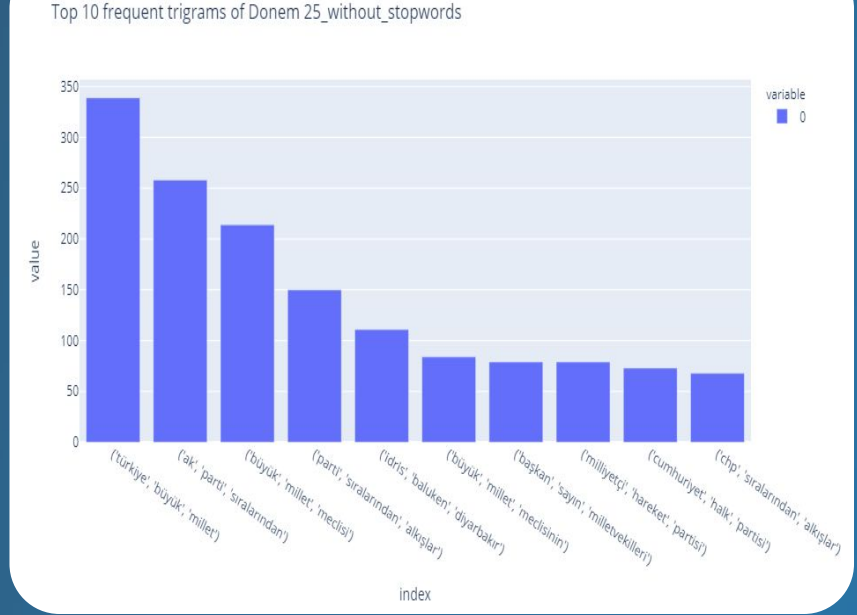
Most Frequent 10 Trigrams

with stopwords

without stopwords

Donem - 25



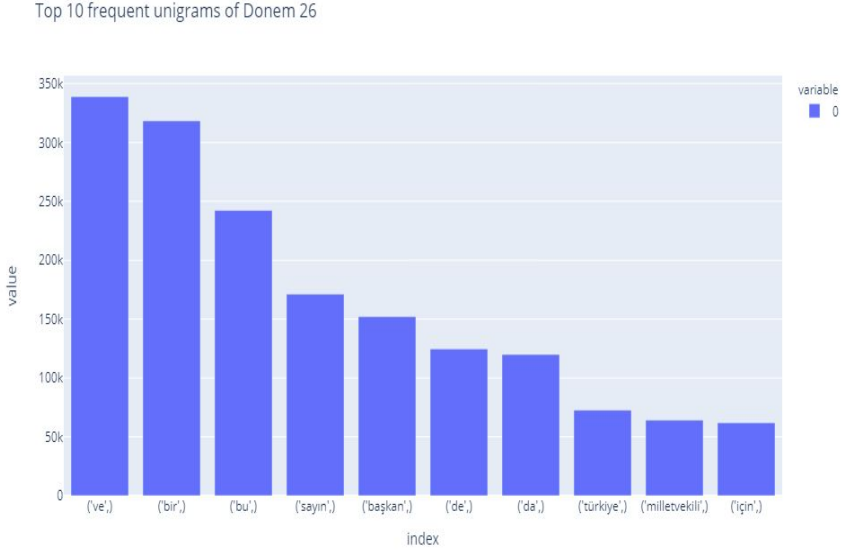


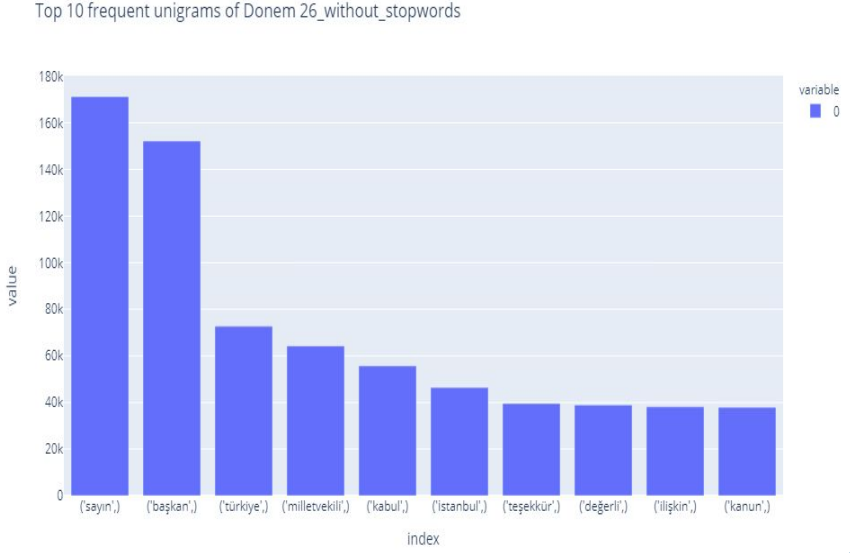
Most Frequent 10 Unigrams

with stopwords

without stopwords

Donem - 26



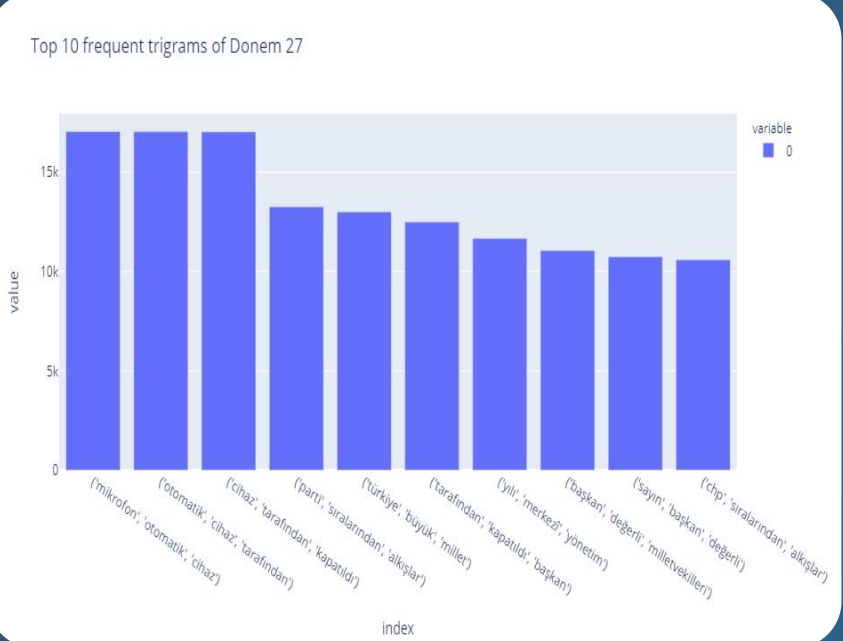


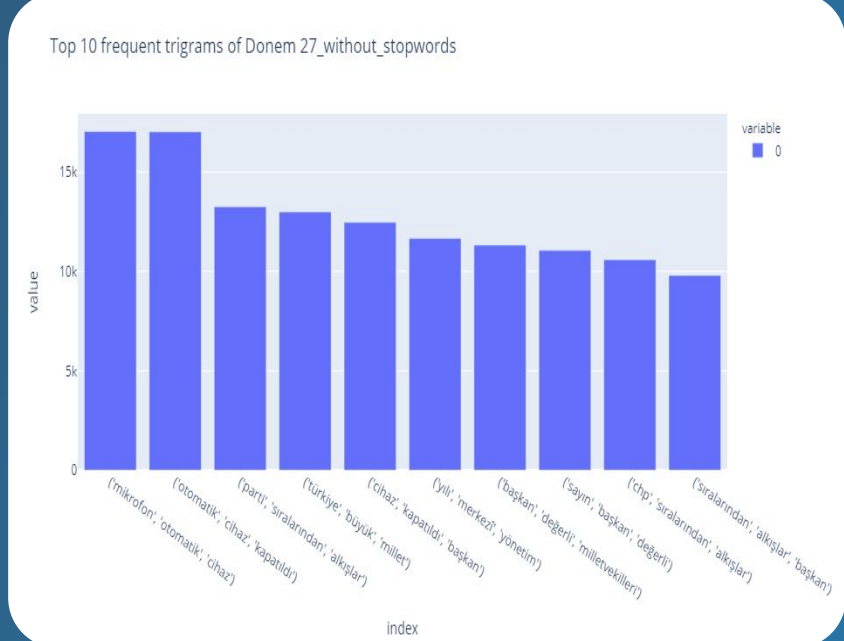
Most Frequent 10 Trigrams

with stopwords

without stopwords

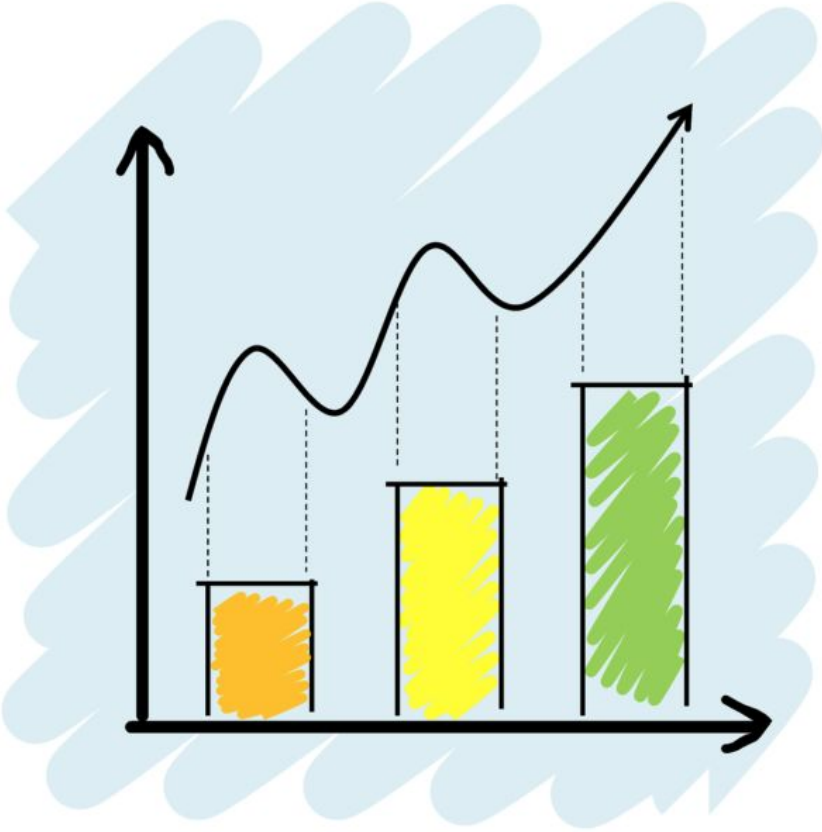
Donem - 27





Statistical Estimators

- Maximum Likelihood Estimator
- LaPlace’s Law
- Lidstone’s Law
- Jeffreys-Perks Law



All Corpus (All Donems) - Maximum Likelihood Estimator

with stopwords

n-gram	frequency	p_est
('sayın', 'başkan'),	227619.0	0.0018796933494997692
('soru', 'önergesi'),	160663.0	0.0013267660986590812
('sıralarından', 'alkışlar'),	160423.0	0.0013247841621604587
('büyük', 'millet'),	130856.0	0.0010806178435989165
('yazılı', 'soru'),	130669.0	0.0010790735847437398
('türkiye', 'büyük'),	130376.0	0.0010766539706016716
('başkan', 'sayın'),	127859.0	0.0010558684115723686
('değerli', 'milletvekilleri'),	127724.0	0.0010547535722918936
('kabul', 'edenler'),	109905.0	0.0009076030453379205
('geliş', 'tarihi'),	103917.0	0.0008581537296972903

without stopwords

n-gram	frequency	p_est
('sayın', 'başkan'),	227620.0	0.0022995079040962355
('soru', 'önergesi'),	160667.0	0.001623122029819128
('sıralarından', 'alkışlar'),	160427.0	0.0016206974542239117
('türkiye', 'büyük'),	131101.0	0.0013244345212851268
('büyük', 'millet'),	130883.0	0.0013222321984528056
('yazılı', 'soru'),	130721.0	0.0013205956099260347
('başkan', 'sayın'),	128863.0	0.0013018253538597364
('değerli', 'milletvekilleri'),	127728.0	0.0012903591317740268
('kabul', 'edenler'),	109905.0	0.0011103040866342886
('geliş', 'tarihi'),	103917.0	0.001049810925533646

MLE = Count / N

All Corpus (All Donems) - Laplace Estimator

with stopwords

n-gram	frequency	p_est
('sayın', 'başkan'),	227619.0	4.5648125912069773e-07
('soru', 'önergesi'),	160663.0	3.2220413415063605e-07
('sıralarından', 'alkışlar'),	160423.0	3.217228253808049e-07
('büyük', 'millet'),	130856.0	2.6242759039081427e-07
('yazılı', 'soru'),	130669.0	2.620525706409875e-07
('türkiye', 'büyük'),	130376.0	2.6146497285115196e-07
('başkan', 'sayın'),	127859.0	2.5641724712754774e-07
('değerli', 'milletvekilleri'),	127724.0	2.561465109445177e-07
('kabul', 'edenler'),	109905.0	2.2041134023776209e-07
('geliş', 'tarihi'),	103917.0	2.0840268643047479e-07

without stopwords

n-gram	frequency	p_est
('sayın', 'başkan'),	227620.0	4.567725345926252e-07
('soru', 'önergesi'),	160667.0	3.22416339388404e-07
('sıralarından', 'alkışlar'),	160427.0	3.2193472561681777e-07
('türkiye', 'büyük'),	131101.0	2.630855361770766e-07
('büyük', 'millet'),	130883.0	2.626480703345524e-07
('yazılı', 'soru'),	130721.0	2.623229810387317e-07
('başkan', 'sayın'),	128863.0	2.5859448775703497e-07
('değerli', 'milletvekilleri'),	127728.0	2.563168559622417e-07
('kabul', 'edenler'),	109905.0	2.2055101324981905e-07
('geliş', 'tarihi'),	103917.0	2.0853474964874253e-07

LAP = (c + 1) / (N + B)

B = V ^ n

All Corpus (All Donems) - Lidstone Estimator

with stopwords

n-gram	frequency	p_est
('sayın', 'başkan')	227619.0	1.8245897277297277e-06
('soru', 'önergesi')	160663.0	1.2878722497050368e-06
('sıralarından', 'alkışlar')	160423.0	1.2859484162214666e-06
('büyük', 'millet')	130856.0	1.048940147018467e-06
('yazılı', 'soru')	130669.0	1.0474411600958519e-06
('türkiye', 'büyük')	130376.0	1.0450924800513266e-06
('başkan', 'sayın')	127859.0	1.0249162763923843e-06
('değerli', 'milletvekilleri')	127724.0	1.023834120057876e-06
('kabul', 'edenler')	109905.0	8.809974998756374e-07
('geliş', 'tarihi')	103917.0	8.32997854460561e-07

without stopwords

n-gram	frequency	p_est
('sayın', 'başkan')	227620.0	1.8259959795414172e-06
('soru', 'önergesi')	160667.0	1.288891267556273e-06
('sıralarından', 'alkışlar')	160427.0	1.2869659597899827e-06
('türkiye', 'büyük')	131101.0	1.0517093949806937e-06
('büyük', 'millet')	130883.0	1.0499605737596466e-06
('yazılı', 'soru')	130721.0	1.0486609910174007e-06
('başkan', 'sayın')	128863.0	1.0337559000600365e-06
('değerli', 'milletvekilleri')	127728.0	1.024650798748622e-06
('kabul', 'edenler')	109905.0	8.816726307544884e-07
('geliş', 'tarihi')	103917.0	8.336362019855454e-07

$$LDSTN = (c + \lambda=0.25) / (N + B*\lambda=0.25)$$

$$B = V \wedge n$$

All Corpus (All Donems) - Jeffrey Perks Estimator

with stopwords

n-gram	frequency	p_est
('sayın', 'başkan'),	227619.0	9.127388562334111e-07
('soru', 'önergesi'),	160663.0	6.442498082477848e-07
('sıralarından', 'alkışlar'),	160423.0	6.432874244208455e-07
('büyük', 'millet'),	130856.0	5.247257468745313e-07
('yazılı', 'soru'),	130669.0	5.239758894760411e-07
('türkiye', 'büyük'),	130376.0	5.228009792206526e-07
('başkan', 'sayın'),	127859.0	5.127079788356263e-07
('değerli', 'milletvekilleri'),	127724.0	5.12166637932973e-07
('kabul', 'edenler'),	109905.0	4.407136487153393e-07
('geliş', 'tarihi'),	103917.0	4.167021722332028e-07

without stopwords

n-gram	frequency	p_est
('sayın', 'başkan'),	227620.0	9.133616337847206e-07
('soru', 'önergesi'),	160667.0	6.447026093700111e-07
('sıralarından', 'alkışlar'),	160427.0	6.437395731227999e-07
('türkiye', 'büyük'),	131101.0	5.260645690156535e-07
('büyük', 'millet'),	130883.0	5.251898110911033e-07
('yazılı', 'soru'),	130721.0	5.245397616242358e-07
('başkan', 'sayın'),	128863.0	5.170842560104092e-07
('değerli', 'milletvekilleri'),	127728.0	5.125298970913063e-07
('kabul', 'edenler'),	109905.0	4.4101241778278 59e-07
('geliş', 'tarihi'),	103917.0	4.169846634148669e-07

LDSTN = (c + λ=0.5) / (N + B*λ=0.5)

B = V ^ n

Summary

- The results found
- The statistical estimators
- What’s next?

References

- Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing (1st ed.). The MIT Press.

Thank you
for listening ! :)