GTU Department of Computer Engineering
CSE 484 / 654 - Fall 2022
Homework #3 Report

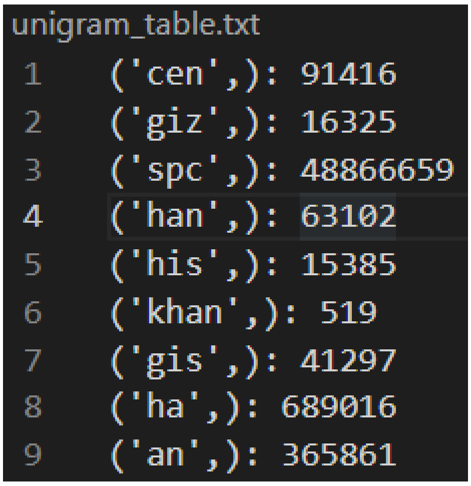
Hasan Mutlu
1801042673

1-2.) Seperating Words into Syllables

The file is seperated into its syllables, spaces and sentence endings in the same way as HW2.

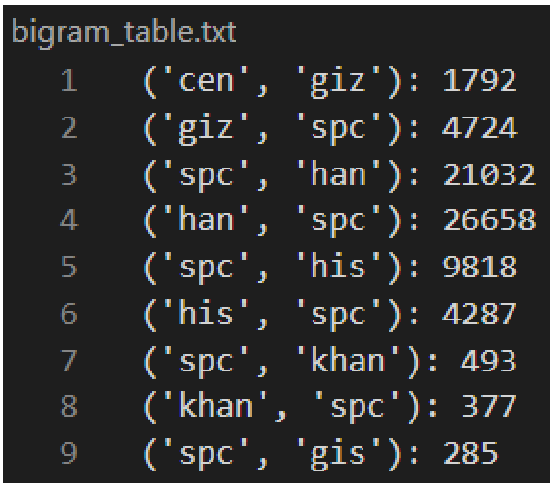
3.) N-Gram Tables

Unlike HW2, N-Grams aren’t manually created. Instead, the “ngrams” method from the “nltk” package is used to create N-Grams. They are saved into files as follows:



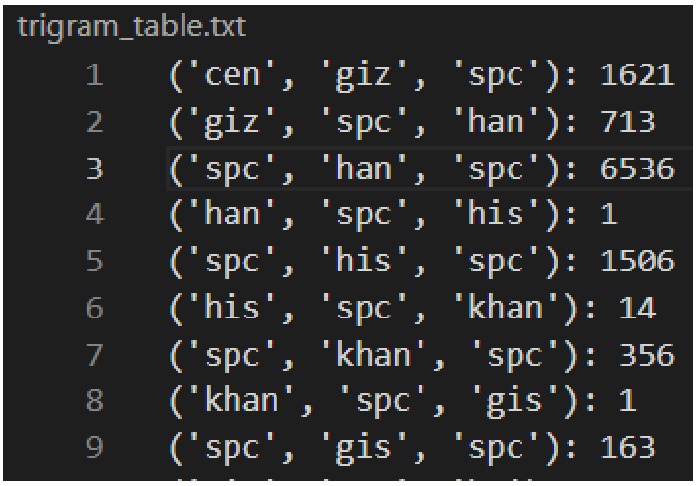
unigram_table.txt

```
('cen',): 91416
('giz',): 16325
('spc',): 48866659
('han',): 63102
('his',): 15385
('khan',): 519
('gis',): 41297
('ha',): 689016
('an',): 365861
```



bigram_table.txt

```
('cen', 'giz'): 1792
('giz', 'spc'): 4724
('spc', 'han'): 21032
('han', 'spc'): 26658
('spc', 'his'): 9818
('his', 'spc'): 4287
('spc', 'khan'): 493
('khan', 'spc'): 377
('spc', 'gis'): 285
```



trigram_table.txt

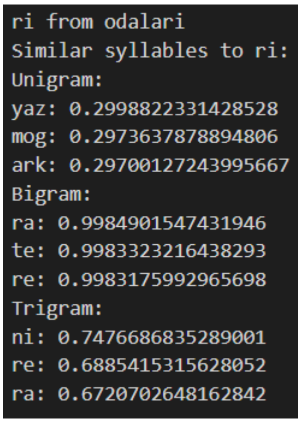
```
('cen', 'giz', 'spc'): 1621
('giz', 'spc', 'han'): 713
('spc', 'han', 'spc'): 6536
('han', 'spc', 'his'): 1
('spc', 'his', 'spc'): 1506
('his', 'spc', 'khan'): 14
('spc', 'khan', 'spc'): 356
('khan', 'spc', 'gis'): 1
('spc', 'gis', 'spc'): 163
```

4.) Creating Syllable Vectors

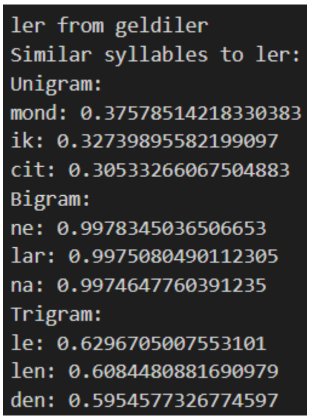
In order to create syllable vectors, the “gensim.models.Word2Vec” method is used from the gensim package. Each n-gram dictionary is converted into required form and the word2Vec algorithm is applied. These models are also saved into files so they can be obtained easily once they are created.

5.) Word Similarity Tests

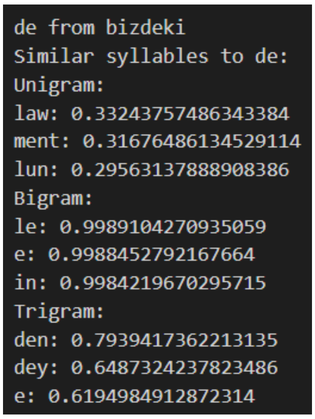
Word similarity tests are run for 5 example sequences. One syllable is given and most similar 3 syllables from each n-gram vector models are displayed. For example, when most similar syllables to “ler”, plurality suffix, are searched, “lar” is expected to be found. Here are the outputs:



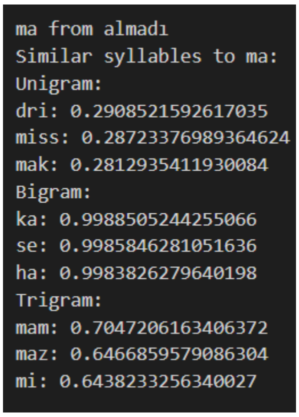
```
ri from odalari
Similar syllables to ri:
Unigram:
yaz: 0.2998822331428528
mog: 0.2973637878894806
ark: 0.2970012724399566
Bigram:
ra: 0.9984901547431946
te: 0.9983323216438293
re: 0.9983175992965698
Trigram:
ni: 0.7476686835289001
re: 0.6885415315628052
ra: 0.6720702648162842
```



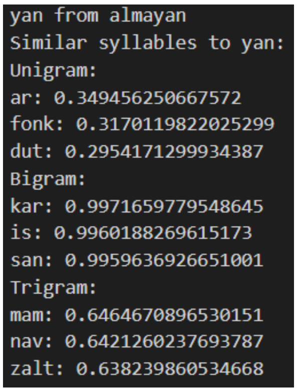
```
ler from geldiler
Similar syllables to ler:
Unigram:
mond: 0.37578514218330383
ik: 0.32739895582199097
cit: 0.30533266067504883
Bigram:
ne: 0.9978345036506653
lar: 0.9975080490112305
na: 0.9974647760391235
Trigram:
le: 0.6296705007553101
len: 0.6084480881690979
den: 0.5954577326774597
```



```
de from bizdeki
Similar syllables to de:
Unigram:
law: 0.33243757486343384
ment: 0.31676486134529114
lun: 0.29563137888908386
Bigram:
le: 0.9989104270935059
e: 0.9988452792167664
in: 0.9984219670295715
Trigram:
den: 0.7939417362213135
dey: 0.6487324237823486
e: 0.6194984912872314
```



```
ma from almadı
Similar syllables to ma:
Unigram:
dri: 0.2908521592617035
miss: 0.28723376989364624
mak: 0.2812935411930084
Bigram:
ka: 0.9988505244255066
se: 0.9985846281051636
ha: 0.9983826279640198
Trigram:
mam: 0.7047206163406372
maz: 0.6466859579086304
mi: 0.6438233256340027
```

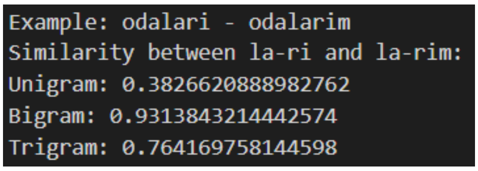


```
yan from almayan
Similar syllables to yan:
Unigram:
ar: 0.349456250667572
fonk: 0.3170119822025299
dut: 0.2954171299934387
Bigram:
kar: 0.9971659779548645
is: 0.9960188269615173
san: 0.9959636926651001
Trigram:
mam: 0.6464670896530151
nav: 0.6421260237693787
zalt: 0.638239860534668
```

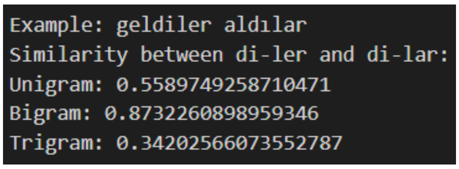
It is observed that most consistent outputs are given from the bigram vectors.

6.) Syllable Analogy Tests

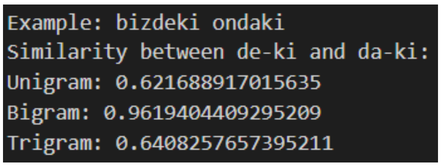
In the word analogy tests, the combining vectors for the similar sequences with the same or similar morphological structures are created. Pearson correlation coefficient test is used to determine their similarity. In order to calculate Pearson correlation coefficient, “pearsonr” method from “scipy” package is used. Here are the outputs:



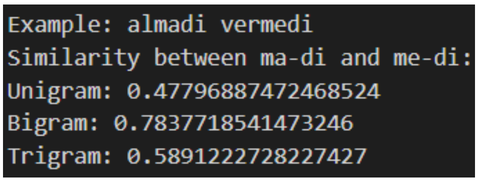
```
Example: odalari - odalarim
Similarity between la-ri and la-rim:
Unigram: 0.3826620888982762
Bigram: 0.9313843214442574
Trigram: 0.764169758144598
```



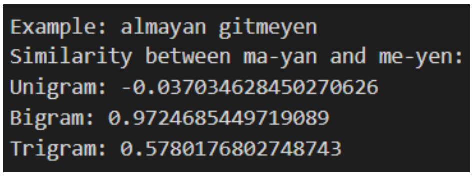
```
Example: geldiler aldılar
Similarity between di-ler and di-lar:
Unigram: 0.5589749258710471
Bigram: 0.8732260898959346
Trigram: 0.34202566073552787
```



```
Example: bizdeki ondaki
Similarity between de-ki and da-ki:
Unigram: 0.621688917015635
Bigram: 0.9619404409295209
Trigram: 0.6408257657395211
```



```
Example: almadi vermedi
Similarity between ma-di and me-di:
Unigram: 0.47796887472468524
Bigram: 0.7837718541473246
Trigram: 0.5891222728227427
```

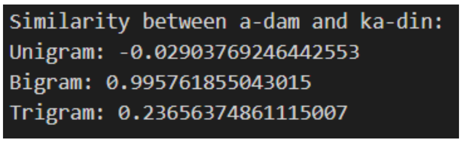


```
Example: almayan gitmeyen
Similarity between ma-yan and me-yen:
Unigram: -0.037034628450270626
Bigram: 0.9724685449719089
Trigram: 0.5780176802748743
```

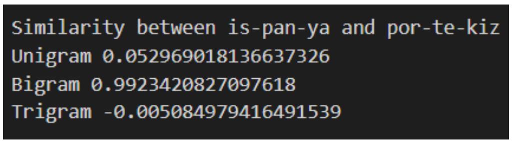
It is observed that the bigram vectors gave better results.

Bonus Word Analogy Test

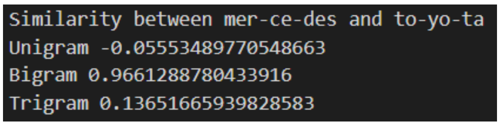
Out of curiosity, I tested correlation between the syllables of related words that have same number of syllables. The bigram model gave good results so I decided to include it in the homework submission as well. Here are some examples:



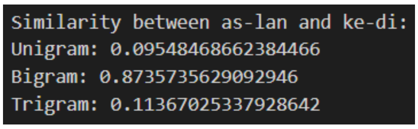
```
Similarity between a-dam and ka-din:
Unigram: -0.02903769246442553
Bigram: 0.995761855043015
Trigram: 0.23656374861115007
```



```
Similarity between is-pan-ya and por-te-kiz
Unigram 0.052969018136637326
Bigram 0.9923420827097618
Trigram -0.005084979416491539
```



```
Similarity between mer-ce-des and to-yo-ta
Unigram -0.05553489770548663
Bigram 0.9661288780433916
Trigram 0.13651665939828583
```



```
Similarity between as-lan and ke-di:
Unigram: 0.09548468662384466
Bigram: 0.8735735629092946
Trigram: 0.11367025337928642
```