Gebze Technical University
Department of Computer Engineering
CSE 654 / 484
Fall 2022

Homework 3
Report

Burak Yıldırım
1901042609

1. Preprocessing

250.000 lines are read from wiki document. All the text is converted to lower case, all the Turkish specific characters are converted into English, all the numbers and all the punctuation marks but dot are deleted.

2. Syllabicating

https://codereview.stackexchange.com/questions/203450/syllabification-function-for-turkish-words is used for syllabication.

3. Calculating N-Grams

2D lists of n-grams are generated by n_grams function which I implemented. 1-Gram, 2-Gram and 3-Gram are generated.

4. Assigning Vectors to N-Grams

I used genism library for Word2Vec algorithm and created models for 1-Gram, 2-Gram and 3-Gram. I saved these models for faster execution time in future.

5. Test Results

```
------MOST SIMILAR SYLLABLE------
------1-GRAM------
1-Gram: la
Most similar syllables: da  0.9998302459716797
                        lar 0.9998270273208618
                        le  0.9998261332511902
                        ' ' 0.999824047088623
                        de  0.9998233318328857

1-Gram: mez
Most similar syllables: yun 0.990431547164917
                        nun 0.9901304244995117
                        u   0.9900603294372559
                        ’in 0.9900489449501038
                        te  0.990011990070343

1-Gram: cik
Most similar syllables: on  0.9803835153579712
                        ruk 0.9803510308265686
                        ad  0.9802456498146057
                        len 0.9801811575889587
                        uy  0.9801328778266907
```

```
------2-GRAM------
2-Gram: larim
Most similar syllables: eskrim  0.8457978367805481
                        lerim   0.8270807266235352
                         esk    0.8237696886062622
                        =evrim  0.795102596282959
                        zirim   0.7924016118049622

2-Gram: tikca
Most similar syllables: arttik  0.9290929436683655
                        ranca   0.84096759557724
                        dikca   0.8292887806892395
                        likca   0.8214527368545532
                        nuca    0.819747269153595

2-Gram: mezler
Most similar syllables: cerler  0.883730411529541
                        lurler  0.8710312843322754
                        rurler  0.8640711307525635
                        surler  0.8634992241859436
                        sirler  0.8593338131904602
```

```
------3-GRAM------
3-Gram: madikca
Most similar syllables: dikca   0.8937404751777649
                        mayip   0.8510788679122925
                        madikla 0.8232616782188416
                        olmadik 0.8116623759269714
                        landikca    0.8077111840248108

3-Gram: larini
Most similar syllables: larinin 0.7992688417434692
                        lari    0.7676640748977661
                        larimiz 0.7525489926338196
                        lari.   0.7347460389137268
                        larina  0.7290388345718384

3-Gram: liyorken
Most similar syllables: debili  0.8961718082427979
                        biliyor 0.889815092086792
                        labili  0.8806048631668091
                        rabili  0.8583105802536011
                        liyorlar    0.8581674098968506
```

```
------SIMILARITY SCORES------
------1-GRAM------
Word 1: la Word 2: le Similarity: 0.9998260736465454
Word 1: ta Word 2: de Similarity: 0.9997570514678955
Word 1: dik Word 2: cik Similarity: 0.9780202507972717

------2-GRAM------
Word 1: lari Word 2: larim Similarity: 0.24288775026798248
Word 1: cakti Word 2: dardi Similarity: 0.06066734343767166
Word 1: daki Word 2: lara Similarity: 0.09891120344400406

------3-GRAM------
Word 1: larini Word 2: lerini Similarity: 0.1894993782043457
Word 1: medigi Word 2: madigi Similarity: 0.48633846640586853
Word 1: dakiler Word 2: diklari Similarity: 0.03665829449892044
```