GTU Department of Computer Engineering

Natural Language Processing

Fall 2023 - Homework 2 Report

ENES AYSU 1901042671

Goal of Homework

Developing a statistical language model of Turkish that will use N-grams of Turkish syllables.

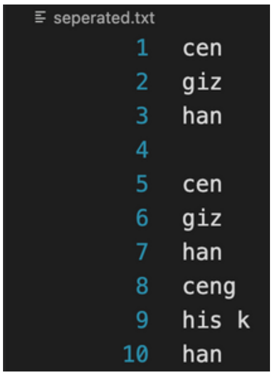
Methods

1) Creating Corpus

The given turkish-wikipedia-dump (wiki_00.txt) text was to big(about 4.5 million lines - 463 mb) to test therefore, used small portion of it which
is about 270_000 lines of data that 95% of the set for ngrams and 5% of the set for test data.

2) Dividing Turkish Words into Syllables

First the text needed a few editing adjustments. Accordingly capital letters were changed to lowercase letters, punctuation marks were removed and lines that disrupted the harmony of the text were removed (lines starting with '<doc')
Afterwards, the 'syllabicate' function of the 'turkishnlp' library was used to seperate the processed text into syllables.



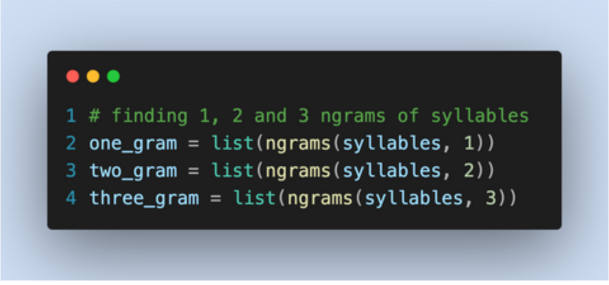
```
cen
giz
han

cen
giz
han
ceng
his k
han
```

a part of seperated.txt

3) Creating N-gram Tables

The 'ngram' function of the 'nltk' library was used to create the n-gram table of the text seperated into syllables.

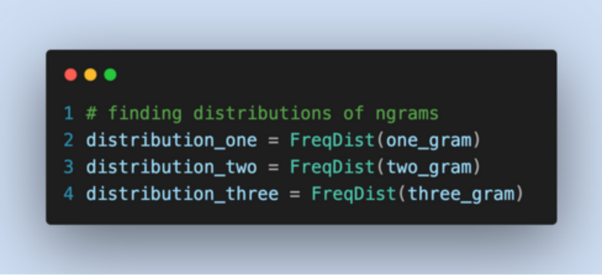


```python
# finding 1, 2 and 3 ngrams of syllables
one_gram = list(ngrams(syllables, 1))
two_gram = list(ngrams(syllables, 2))
three_gram = list(ngrams(syllables, 3))
```

obtaining each n-gram table

4) Counting Syllables

To count syllables, the ‘FreqDist’ class was used.



```python
# finding distributions of ngrams
distribution_one = FreqDist(one_gram)
distribution_two = FreqDist(two_gram)
distribution_three = FreqDist(three_gram)
```

obtaining syllables count for each n-grams

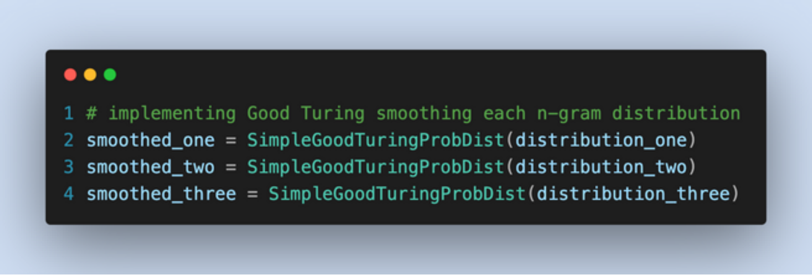
Most common 5 syllables for 1-gram:
[[(('la',), 281446), (('le',), 209010), ((' a',), 154117), (('ya',), 137231), ((' i',), 135980)]

Most common 5 syllables for 2-gram:
[((' o', 'la'), 40654), (('la', 'rı'), 39563), (('le', 'ri'), 38846), (('la', 'rak '), 27671), (('la', 'rı '), 26615)]

Most common 5 syllables for 3-gram:
[((' o', 'la', 'rak '), 25343), (('ta', 'ra', 'fın'), 14143), ((' a', 'ra', 'sın'), 11124), (('yı', 'lın', 'da '), 10775), (('ra', 'fın', 'dan '), 9971)]

5) GT Smoothing

In order to apply Good-Turing Smoothing to the probabilities, ‘SimpleGoodTuringProbDist’ function from ‘nltk’ library was used. That function finds probability for each n-gram.



```python
# implementing Good Turing smoothing each n-gram distribution
smoothed_one = SimpleGoodTuringProbDist(distribution_one)
smoothed_two = SimpleGoodTuringProbDist(distribution_two)
smoothed_three = SimpleGoodTuringProbDist(distribution_three)
```

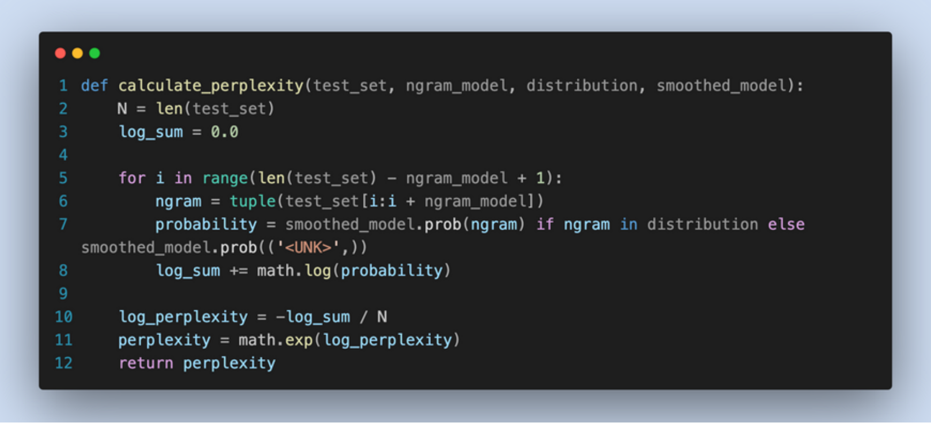
6) Calculating Perplexity with the Markov Assumption

Perplexity formula with the Markov assumption;

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_1 \ldots w_{i-1})}}$$

(Markov Assumption)

Code form of perplexity formula with the Markov assumption;



```python
def calculate_perplexity(test_set, ngram_model, distribution, smoothed_model):
    N = len(test_set)
    log_sum = 0.0

    for i in range(len(test_set) - ngram_model + 1):
        ngram = tuple(test_set[i:i + ngram_model])
        probability = smoothed_model.prob(ngram) if ngram in distribution else smoothed_model.prob(('<UNK>',))
        log_sum += math.log(probability)

    log_perplexity = -log_sum / N
    perplexity = math.exp(log_perplexity)
    return perplexity
```

calculating perplexity for each n-gram

7) 5 Perplexity Values for 5 Example Sentences

7.1) “Cengiz Han, Buhara'da iki gün kaldıktan sonra sonra Semerkant'a ilerledi.”

Perplexity for 1-gram model: 499.387860557625
Perplexity for 2-gram model: 20.21705267748271
Perplexity for 3-gram model: 4.396336816099815

7.2) “Alaaddin Muhammed ölmeden birkaç gün önce oğlu Celaleddin’i veliaht ilan etmişti.”

Perplexity for 1-gram model: 542.3860320085682
Perplexity for 2-gram model: 20.21705267748271
Perplexity for 3-gram model: 4.396336816099815

7.3) “Marx, endüstriyel kapitalistlerin tüccar kapitalistlerden ayrıldığını söyler.”

Perplexity for 1-gram model: 984.1596172303629
Perplexity for 2-gram model: 16.63148275369076
Perplexity for 3-gram model: 3.5581218836823876

7.4) “Beşiktaş Jimnastik Kulübü, 1903 yılında İstanbul'da kurulan spor kulübüdür.”

Perplexity for 1-gram model: 280.99101473065906
Perplexity for 2-gram model: 19.141568340579216
Perplexity for 3-gram model: 4.143499771989139

7.5) “Eski Mısırlılar işlevsel amaçlara hizmet eden bir sanat ürettiler.”

Perplexity for 1-gram model: 308.9509388736363
Perplexity for 2-gram model: 18.456554750211723
Perplexity for 3-gram model: 3.9830726426785503

8) Generating Random Sentences for Each N-Gram Models

Random 1-Gram Sentences (with using 5 best syllables):

Most popular 5 syllables: [('la',), ('le',), (' a',), ('ya',), (' i',)]

Sentence 1: a a ile alayalale aleleyalale
Sentence 2: lalale aleya i aleleyalayaleya
Sentence 3: leya ile a i alelale i a iya a
Sentence 4: la ila i ayaleleya a i ilayaya
Sentence 5: a alale ilaleyalalaya ayalale
Sentence 6: leyale ayala i a iyaleyala i i
Sentence 7: ileya alayalalaleyalelala ile
Sentence 8: leya aya iyaya ilala ayale ala
Sentence 9: yala i a iya i i ilaya a aleya
Sentence 10: la ala alele a i ayale ilelele

Random 2-Gram Sentences (with using 5 best syllables):

Most popular 5 syllables: [(' o', 'la'), ('la', 'rı'), ('le', 'ri'), ('la', 'rak '), ('la', 'rı ')]

Sentence 1: ları grı fonksiden bloch kutmuştur fkfnin siyad aldıbr
Sentence 2: lakonya km konto tor aspir tan erskilar pestitüt
Sentence 3: o önce vpnde yazma platyioles hollynin müjdeki v
Sentence 4: ları václav fay batory ya island tayda kuykulo
Sentence 5: ları pları dbsin vüşür de e yı ayrı künün şart
Sentence 6: lamatnanış veya typeta pek sivri divliktir me
Sentence 7: la evsiz resle catelli jet her pendent in tedir ms
Sentence 8: lahil anest olan mlik babâîlermüzde thyada p sı epl
Sentence 9: okur bir lig yano grainmen ikter herküdardan
Sentence 10: lenucu hâlâ tarlide fromberg kaleşmeksine dns aylım

Random 3-Gram Sentences (with using 5 best syllables):

Most popular 5 syllables: [(' o', 'la', 'rak '), ('ta', 'ra', 'fın'), (' a', 'ra', 'sın'), ('yı', 'lın', 'da '), ('ra', 'fın', 'dan ')]

Sentence 1: tararız hatırları cem yapıldıklara pek rast
Sentence 2: arada plütondurharezmiyi sarmaya atléti
Sentence 3: aracaat nat gibi ret rapova fransa insti
Sentence 4: rafından drúedain cücebenzeti ile frekans
Sentence 5: rafında brika okyanuslar elektrod arada p
Sentence 6: tararak inmişti macalama monteze ika
Sentence 7: olarak rûm şeklinde glauson oxford worlds classics dün
Sentence 8: olarak s standard telephones and david eski en
Sentence 9: arabesk haya ilişki sütun tunçtan bir sorun
Sentence 10: aralık lâkabı koltuk belarussuladapaza