NLP – Assignment 1

Group 4

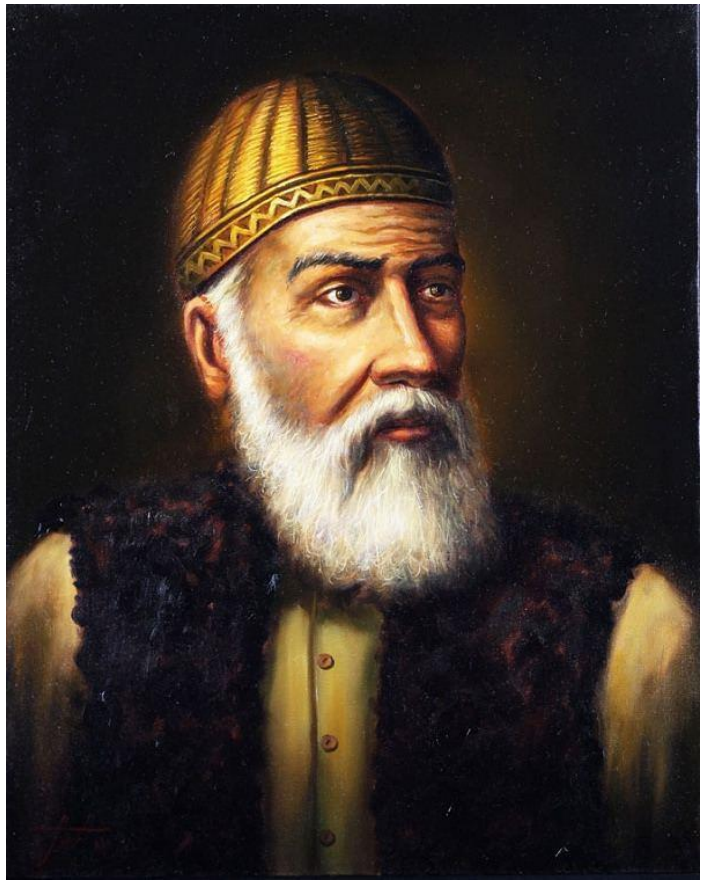
Dataset – Fuzuli’s poems

“Qəd ənarəl-eşqə-lil-üşşaqi minhacəl hüda!

Saliki-rahi-həqiqət eşqə eylər iqtida.

Eşqdir ol nəş’eyi-kamil kim, ondandır müdam

Meydə təşviri-hərarət, neydə tə’siri-səda”...



Füzuli

Data description

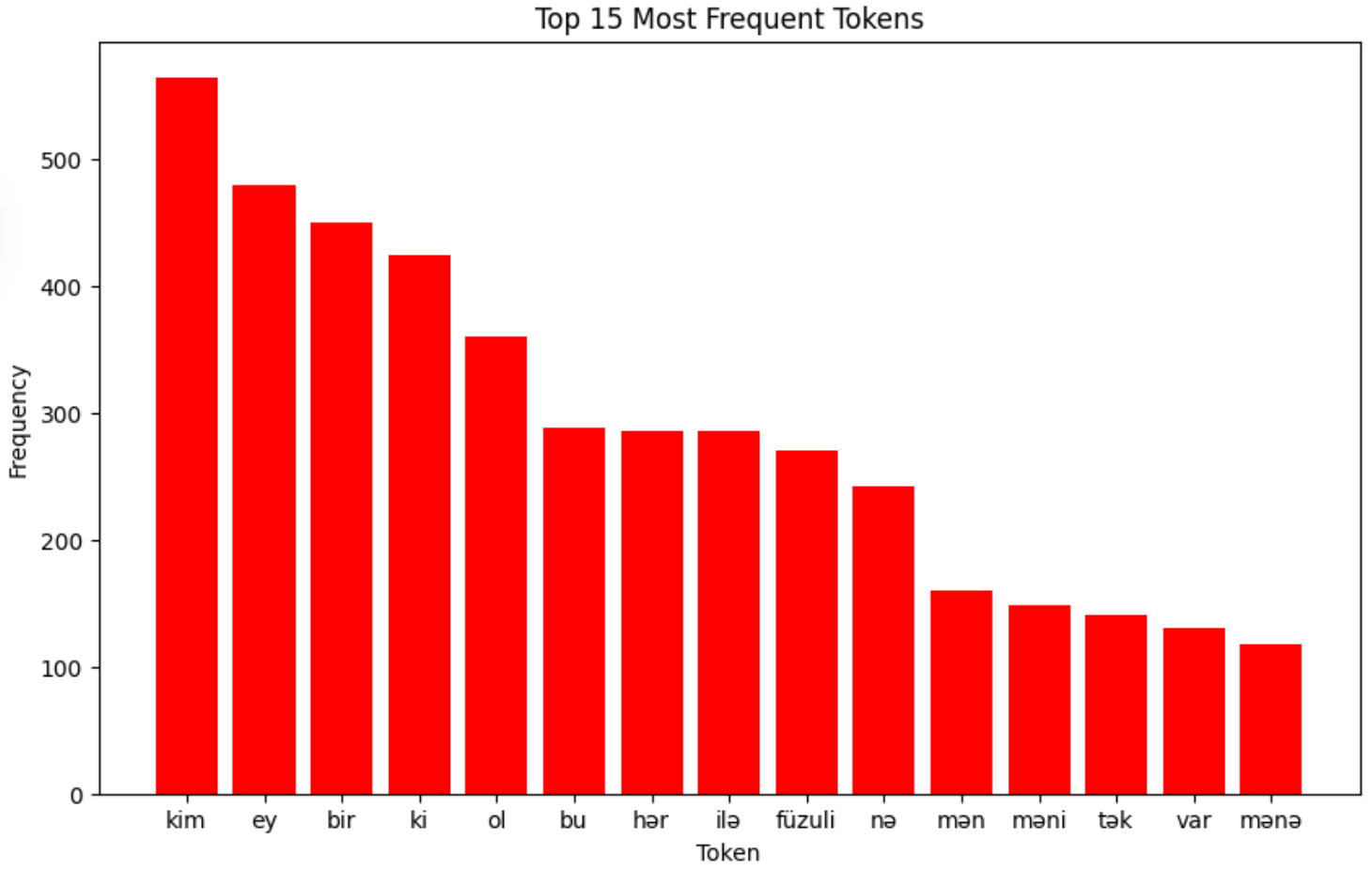
- Motivation : complex word forms, possible usage in linguistic research
- Situation: poetry written in 16 century
- Language: classical Azerbaijani
- Speaker: Demographics: written by Məhəmməd Füzuli
- Collection process: Data was sourced from open presidential library. Conversion from pdf to plain text
- Annotation Process: no annotation
- Distribution: Original text is public domain. The source for the project is © “ŞƏRQ-QƏRB”, 2005.

Whitespace Tokenization

```
r"\p{L}+(?:[-’']\p{L}+)*"
```

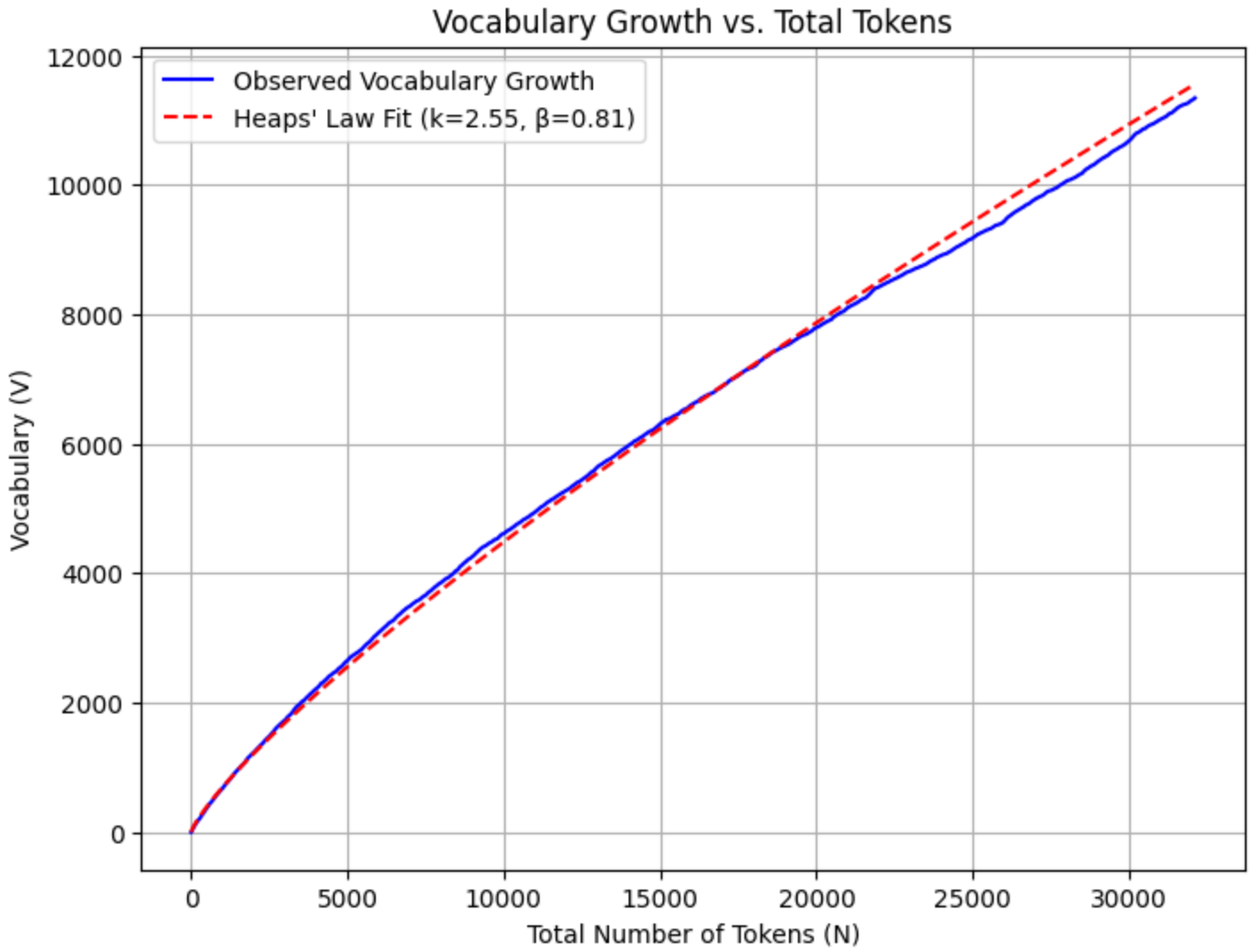
Tokenizer regex

- 32,000 Tokens
- 11,000 Unique tokens



Heaps' Law

- k = 2.55
- β = 0.81



BPE Tokenization

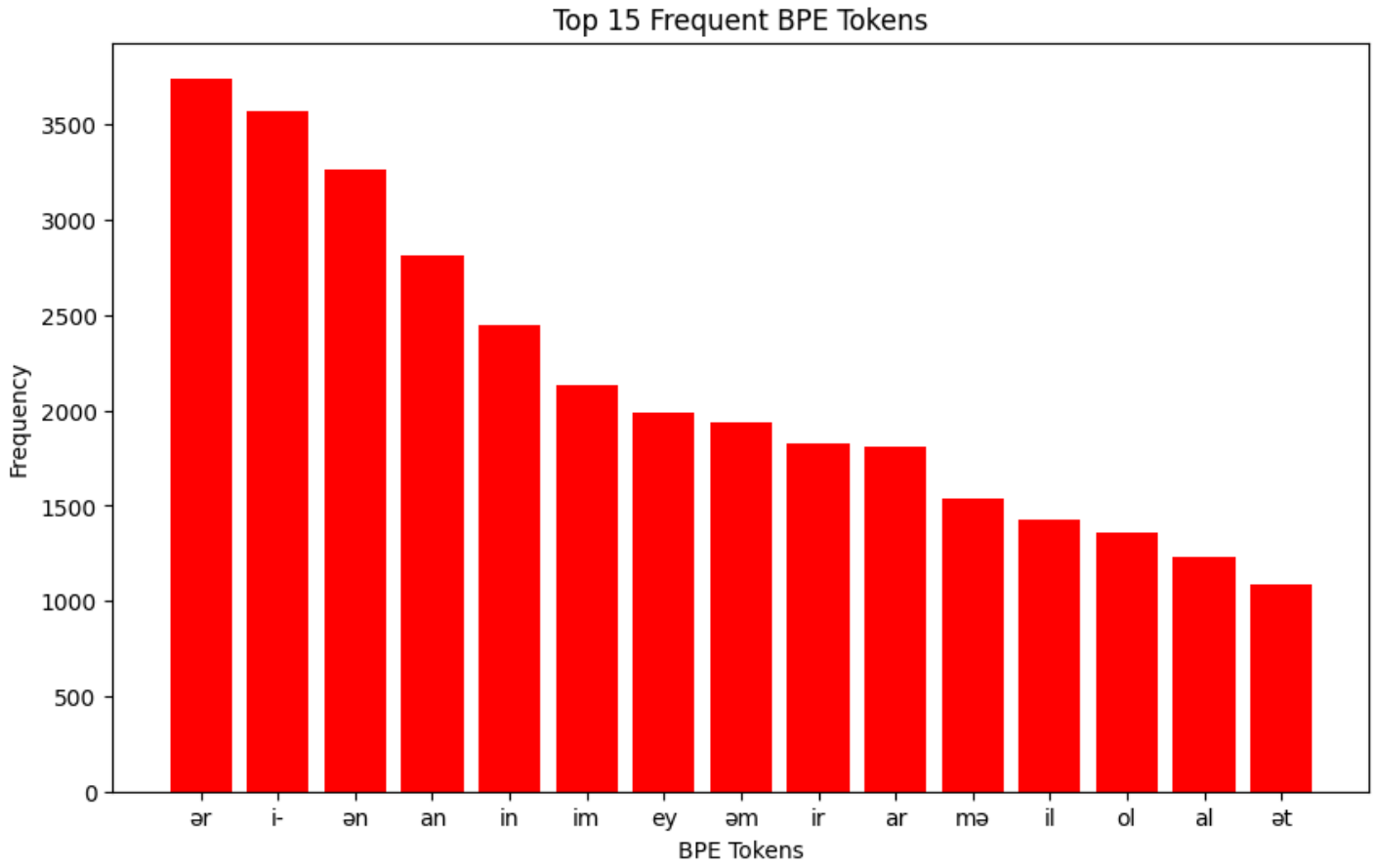
- 1000 merges

```
çox   lə z z ət li   ş or ba dır ! 11 tokens

t e z - t e z   me n yu mu z a   q ola q   olur . 18 tokens

bu y ur un ,   siz   de   da dın a   bax ın . 13 tokens
```

Result from a test input



Sentence Segmentation

```
sample_text = "Bakıdan əlimə çatan qadın jurnallarının birində Kələkoş sözü diqqətimi cəlb etdi. " \
"Bu, elə reseptini axtardığım Kələkoş idi! Amma bu dəfı adın yanında resept də var idi! " \
"Beləliklə sirr dolu Kələkoşu bişirmək fürsətini əldə etmiş oldum. Çox ləzzətli şorbadır! " \
"Tez-tez menyumuza qolaq olur. Buyurun, siz de dadına baxın."
```

```
1) ba k ı dan   əl imə   ç at an   q ad ın   j ur nal ların ın   bir ində   k əl ək o ş   söz ü   di q qət imi   c əl b   etdi . 35 tokens

2) bu ,   e lə   r e s e p t ini   ax tar dı ğ ım   k əl ək o ş   idi ! 23 tokens

3) am ma   bu   dəf ı   ad ın   yan ında   r e s e p t   də   var   idi ! 19 tokens

4) b el əl ik lə   sir r   d ol u   k əl ək o şu   b iş ir mək   f ür s ət ini   əl də   etmiş   oldu m . 30 tokens

5) çox   lə z z ət li   ş or ba dır ! 11 tokens

6) t e z - t e z   me n yu mu z a   q ola q   olur . 18 tokens

7) bu y ur un ,   siz   de   da dın a   bax ın . 13 tokens

Total number of tokens: 149
```

Spelling checker

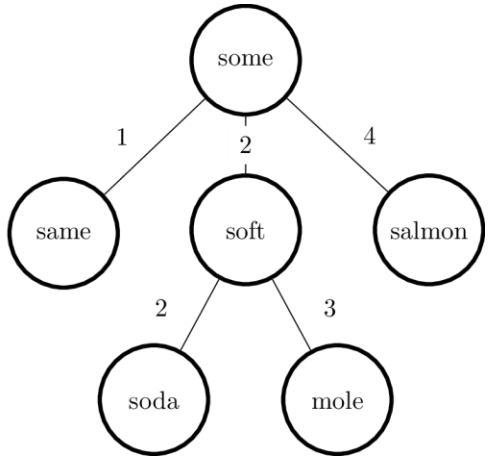
- Levenstein Distance

$$D[i,j] = \begin{cases} 0 & \text{if } i = 0 \text{ and } j = 0 \\ j & \text{if } i = 0 \\ i & \text{if } j = 0 \\ \min \begin{cases} D[i-1,j] + 1 & \text{(deletion)} \\ D[i,j-1] + 1 & \text{(insertion)} \\ D[i-1,j-1] + \text{cost}(s_i, t_j) & \text{(substitution)} \end{cases} & \text{otherwise} \end{cases}$$

where

$$\text{cost}(s_i, t_j) = \begin{cases} 0 & \text{if } s_i = t_j \\ 1 & \text{if } s_i \neq t_j \end{cases}$$

- BK-tree



Example

- Candidates sorted according to Levenstein distance in ascending order

- Search word: lə'li-nabın
- Results: lə'li-nabın; lə'li-nabin; lə'li-nab

- Search word: ölmaz
- Results: ölməz; olmaz; almaz; olmaq; açmaz; qılmaz; urmaz; olman; olma; alma; qomaz; namaz; bulmaz; salmaz; ölmək; qımaz; qalmaz; almaq; nəmaz

Confusion matrix

- Based on keyboard layout
- Probabilities depend on Euclidean distance

```
keyboard_az = {

    'q': (0, 0), 'ü': (1, 0), 'e': (2, 0), 'r': (3, 0),
    't': (4, 0), 'y': (5, 0), 'u': (6, 0), 'i': (7, 0),
    'o': (8, 0), 'p': (9, 0), 'ö': (10, 0), 'ğ': (11, 0),
    '-': (12, 0),


    'a': (0.5, 1), 's': (1.5, 1), 'd': (2.5, 1), 'f': (3.5, 1),
    'g': (4.5, 1), 'h': (5.5, 1), 'j': (6.5, 1), 'k': (7.5, 1),
    'l': (8.5, 1), 'ı': (9.5, 1), 'ə': (10.5, 1),
    "'": (11.5, 1),



    'z': (1, 2), 'x': (2, 2), 'c': (3, 2), 'v': (4, 2),
    'b': (5, 2), 'n': (6, 2), 'm': (7, 2), 'ç': (8, 2),
    'ş': (9, 2)
}
```

Weighted Levenstein Distance

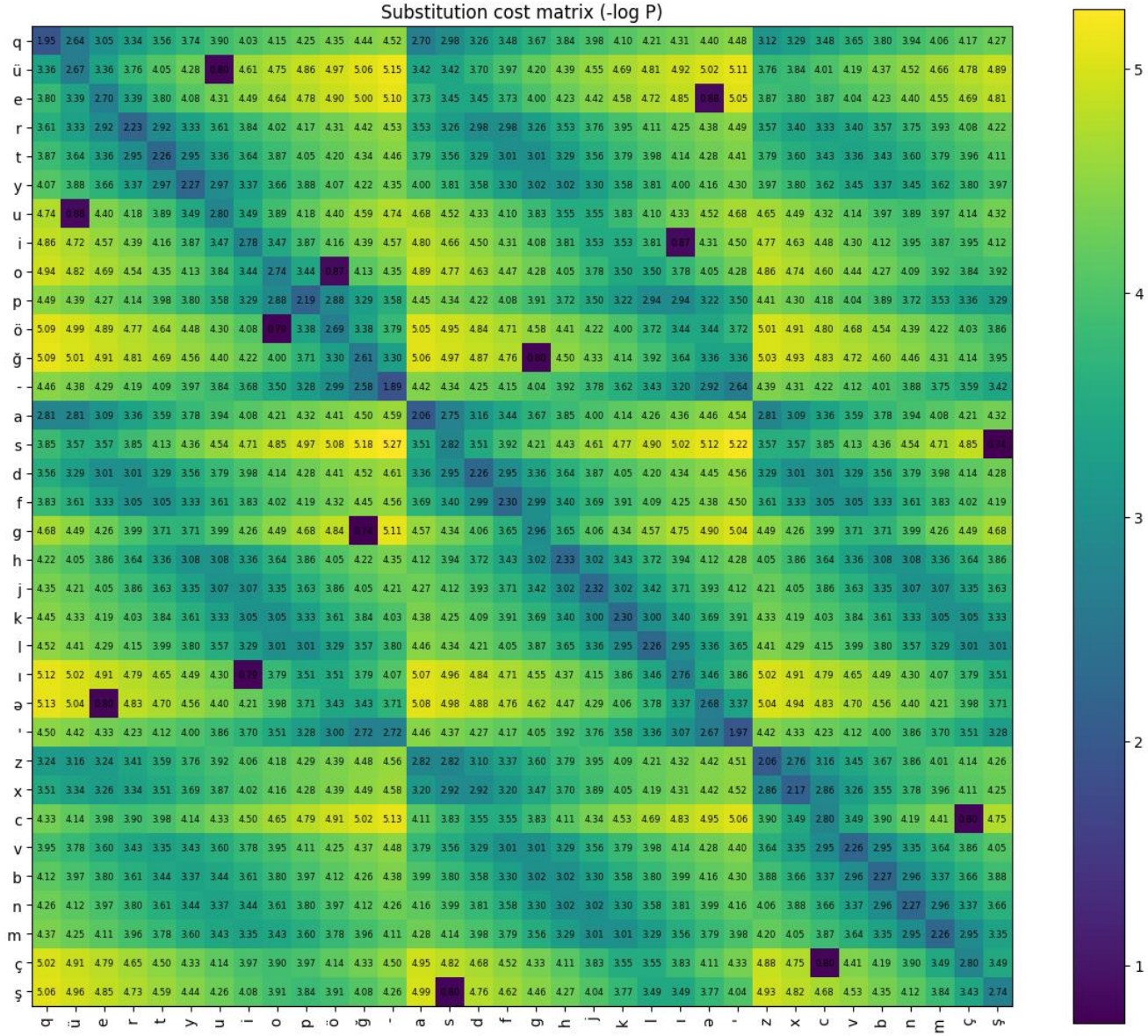
$$
D[i,j] = \begin{cases}
0 & \text{if } i = 0 \text{ and } j = 0 \\
j & \text{if } i = 0 \\
i & \text{if } j = 0 \\
\min \begin{cases}
D[i-1,j] + 1 & \text{(deletion)} \\
D[i,j-1] + 1 & \text{(insertion)} \\
D[i-1,j-1] + w(s_i, t_j) & \text{(substitution)}
\end{cases} & \text{otherwise}
\end{cases}
$$

Example

- Candidates sorted according to Weighted Levenstein distance in ascending order

- Search word: lə'li-nabın
- Results: lə'li-nabın; lə'li-nabin; lə'li-nab

- Search word: ölmaz
- Results: olmaz; olma; ölməz; almaz


Substitution cost matrix (-log P)
q 1.95 2.64 3.05 3.34 3.56 3.74 3.90 4.03 4.15 4.25 4.35 4.44 4.52 2.70 2.98 3.26 3.48 3.67 3.84 3.98 4.10 4.21 4.31 4.40 4.48 3.12 3.29 3.48 3.65 3.80 3.94 4.06 4.17 4.27
ü 3.36 2.67 3.36 3.76 4.05 4.28 0.80 4.61 4.75 4.86 4.97 5.06 5.15 3.42 3.42 3.70 3.97 4.20 4.39 4.55 4.69 4.81 4.92 5.02 5.11 3.76 3.84 4.01 4.19 4.37 4.52 4.66 4.78 4.89
e 3.80 3.39 2.70 3.39 3.80 4.08 4.31 4.49 4.64 4.78 4.90 5.00 5.10 3.73 3.45 3.45 3.73 4.00 4.23 4.42 4.58 4.72 4.85 0.88 5.05 3.87 3.80 3.87 4.04 4.23 4.40 4.55 4.69 4.81
r 3.61 3.33 2.92 2.23 2.92 3.33 3.61 3.84 4.02 4.17 4.31 4.42 4.53 3.53 3.26 2.98 2.98 3.26 3.53 3.76 3.95 4.11 4.25 4.38 4.49 3.57 3.40 3.33 3.40 3.57 3.75 3.93 4.08 4.22
t 3.87 3.64 3.36 2.95 2.26 2.95 3.36 3.64 3.87 4.05 4.20 4.34 4.46 3.79 3.56 3.29 3.01 3.01 3.29 3.56 3.79 3.98 4.14 4.28 4.41 3.79 3.60 3.43 3.36 3.43 3.60 3.79 3.96 4.11
y 4.07 3.88 3.66 3.37 2.97 2.27 2.97 3.37 3.66 3.88 4.07 4.22 4.35 4.00 3.81 3.58 3.30 3.02 3.02 3.30 3.58 3.81 4.00 4.16 4.30 3.97 3.80 3.62 3.45 3.37 3.45 3.62 3.80 3.97
u 4.74 0.88 4.40 4.18 3.89 3.49 2.80 3.49 3.89 4.18 4.40 4.59 4.74 4.68 4.52 4.33 4.10 3.83 3.55 3.55 3.83 4.10 4.33 4.52 4.68 4.65 4.49 4.32 4.14 3.97 3.89 3.97 4.14 4.32
i 4.86 4.72 4.57 4.39 4.16 3.87 3.47 2.78 3.47 3.87 4.16 4.39 4.57 4.80 4.66 4.50 4.31 4.08 3.81 3.53 3.53 3.81 0.87 4.31 4.50 4.77 4.63 4.48 4.30 4.12 3.95 3.87 3.95 4.12
o 4.94 4.82 4.69 4.54 4.35 4.13 3.84 3.44 2.74 3.44 0.87 4.13 4.35 4.89 4.77 4.63 4.47 4.28 4.05 3.78 3.50 3.50 3.78 4.05 4.28 4.86 4.74 4.60 4.44 4.27 4.09 3.92 3.84 3.92
p 4.49 4.39 4.27 4.14 3.98 3.80 3.58 3.29 2.88 2.19 2.88 3.29 3.58 4.45 4.34 4.22 4.08 3.91 3.72 3.50 3.22 2.94 2.94 3.22 3.50 4.41 4.30 4.18 4.04 3.89 3.72 3.53 3.36 3.29
ö 5.09 4.99 4.89 4.77 4.64 4.48 4.30 4.08 0.79 3.38 2.69 3.38 3.79 5.05 4.95 4.84 4.71 4.58 4.41 4.22 4.00 3.72 3.44 3.44 3.72 5.01 4.91 4.80 4.68 4.54 4.39 4.22 4.03 3.86
ğ 5.09 5.01 4.91 4.81 4.69 4.56 4.40 4.22 4.00 3.71 3.30 2.61 3.30 5.06 4.97 4.87 4.76 0.80 4.50 4.33 4.14 3.92 3.64 3.36 3.36 5.03 4.93 4.83 4.72 4.60 4.46 4.31 4.14 3.95
- 4.46 4.38 4.29 4.19 4.09 3.97 3.84 3.68 3.50 3.28 2.99 2.58 1.89 4.42 4.34 4.25 4.15 4.04 3.92 3.78 3.62 3.43 3.20 2.92 2.64 4.39 4.31 4.22 4.12 4.01 3.88 3.75 3.59 3.42
a 2.81 2.81 3.09 3.36 3.59 3.78 3.94 4.08 4.21 4.32 4.41 4.50 4.59 2.06 2.75 3.16 3.44 3.67 3.85 4.00 4.14 4.26 4.36 4.46 4.54 2.81 3.09 3.36 3.59 3.78 3.94 4.08 4.21 4.32
s 3.85 3.57 3.57 3.85 4.13 4.36 4.54 4.71 4.85 4.97 5.08 5.18 5.27 3.51 2.82 3.51 3.92 4.21 4.43 4.61 4.77 4.90 5.02 5.12 5.22 3.57 3.57 3.85 4.13 4.36 4.54 4.71 4.85 0.74
d 3.56 3.29 3.01 3.01 3.29 3.56 3.79 3.98 4.14 4.28 4.41 4.52 4.61 3.36 2.95 2.26 2.95 3.36 3.64 3.87 4.05 4.20 4.34 4.45 4.56 3.29 3.01 3.01 3.29 3.56 3.79 3.98 4.14 4.28
f 3.83 3.61 3.33 3.05 3.05 3.33 3.61 3.83 4.02 4.19 4.32 4.45 4.56 3.69 3.40 2.99 2.30 2.99 3.40 3.69 3.91 4.09 4.25 4.38 4.50 3.61 3.33 3.05 3.05 3.33 3.61 3.83 4.02 4.19
g 4.68 4.49 4.26 3.99 3.71 3.71 3.99 4.26 4.49 4.68 4.84 0.74 5.11 4.57 4.34 4.06 3.65 2.96 3.65 4.06 4.34 4.57 4.75 4.90 5.04 4.49 4.26 3.99 3.71 3.71 3.99 4.26 4.49 4.68
h 4.22 4.05 3.86 3.64 3.36 3.08 3.08 3.36 3.64 3.86 4.05 4.22 4.35 4.12 3.94 3.72 3.43 3.02 2.33 3.02 3.43 3.72 3.94 4.12 4.28 4.05 3.86 3.64 3.36 3.08 3.08 3.36 3.64 3.86
j 4.35 4.21 4.05 3.86 3.63 3.35 3.07 3.07 3.35 3.63 3.86 4.05 4.21 4.27 4.12 3.93 3.71 3.42 3.02 2.32 3.02 3.42 3.71 3.93 4.12 4.21 4.05 3.86 3.63 3.35 3.07 3.07 3.35 3.63
k 4.45 4.33 4.19 4.03 3.84 3.61 3.33 3.05 3.05 3.33 3.61 3.84 4.03 4.38 4.25 4.09 3.91 3.69 3.40 3.00 2.30 3.00 3.40 3.69 3.91 4.33 4.19 4.03 3.84 3.61 3.33 3.05 3.05 3.33
l 4.52 4.41 4.29 4.15 3.99 3.80 3.57 3.29 3.01 3.01 3.29 3.57 3.80 4.46 4.34 4.21 4.05 3.87 3.65 3.36 2.95 2.26 2.95 3.36 3.65 4.41 4.29 4.15 3.99 3.80 3.57 3.29 3.01 3.01
ı 5.12 5.02 4.91 4.79 4.65 4.49 4.30 0.79 3.79 3.51 3.51 3.79 4.07 5.07 4.96 4.84 4.71 4.55 4.37 4.15 3.86 3.46 2.76 3.46 3.86 5.02 4.91 4.79 4.65 4.49 4.30 4.07 3.79 3.51
ə 5.13 5.04 0.80 4.83 4.70 4.56 4.40 4.21 3.98 3.71 3.43 3.43 3.71 5.08 4.98 4.88 4.76 4.62 4.47 4.29 4.06 3.78 3.37 2.68 3.37 5.04 4.94 4.83 4.70 4.56 4.40 4.21 3.98 3.71
' 4.50 4.42 4.33 4.23 4.12 4.00 3.86 3.70 3.51 3.28 3.00 2.72 2.72 4.46 4.37 4.27 4.17 4.05 3.92 3.76 3.58 3.36 3.07 2.67 1.97 4.42 4.33 4.23 4.12 4.00 3.86 3.70 3.51 3.28
z 3.24 3.16 3.24 3.41 3.59 3.76 3.92 4.06 4.18 4.29 4.39 4.48 4.56 2.82 2.82 3.10 3.37 3.60 3.79 3.95 4.09 4.21 4.32 4.42 4.51 2.06 2.76 3.16 3.45 3.67 3.86 4.01 4.14 4.26
x 3.51 3.34 3.26 3.34 3.51 3.69 3.87 4.02 4.16 4.28 4.39 4.49 4.58 3.20 2.92 2.92 3.20 3.47 3.70 3.89 4.05 4.19 4.31 4.42 4.52 2.86 2.17 2.86 3.26 3.55 3.78 3.96 4.11 4.25
c 4.33 4.14 3.98 3.90 3.98 4.14 4.33 4.50 4.65 4.79 4.91 5.02 5.13 4.11 3.83 3.55 3.55 3.83 4.11 4.34 4.53 4.69 4.83 4.95 5.06 3.90 3.49 2.80 3.49 3.90 4.19 4.41 0.80 4.75
v 3.95 3.78 3.60 3.43 3.35 3.43 3.60 3.78 3.95 4.11 4.25 4.37 4.48 3.79 3.56 3.29 3.01 3.01 3.29 3.56 3.79 3.98 4.14 4.28 4.40 3.64 3.35 2.95 2.26 2.95 3.35 3.64 3.86 4.05
b 4.12 3.97 3.80 3.61 3.44 3.37 3.44 3.61 3.80 3.97 4.12 4.26 4.38 3.99 3.80 3.58 3.30 3.02 3.02 3.30 3.58 3.80 3.99 4.16 4.30 3.88 3.66 3.37 2.96 2.27 2.96 3.37 3.66 3.88
n 4.26 4.12 3.97 3.80 3.61 3.44 3.37 3.44 3.61 3.80 3.97 4.12 4.26 4.16 3.99 3.81 3.58 3.30 3.02 3.02 3.30 3.58 3.81 3.99 4.16 4.06 3.88 3.66 3.37 2.96 2.27 2.96 3.37 3.66
m 4.37 4.25 4.11 3.96 3.78 3.60 3.43 3.35 3.43 3.60 3.78 3.96 4.11 4.28 4.14 3.98 3.79 3.56 3.29 3.01 3.01 3.29 3.56 3.79 3.98 4.20 4.05 3.87 3.64 3.35 2.95 2.26 2.95 3.35
ç 5.02 4.91 4.79 4.65 4.50 4.33 4.14 3.97 3.90 3.97 4.14 4.33 4.50 4.95 4.82 4.68 4.52 4.33 4.11 3.83 3.55 3.55 3.83 4.11 4.33 4.88 4.75 0.80 4.41 4.19 3.90 3.49 2.80 3.49
ş 5.06 4.96 4.85 4.73 4.59 4.44 4.26 4.08 3.91 3.84 3.91 4.08 4.26 4.99 0.80 4.76 4.62 4.46 4.27 4.04 3.77 3.49 3.49 3.77 4.04 4.93 4.82 4.68 4.53 4.35 4.12 3.84 3.43 2.74
q ü e r t y u i o p ö ğ - a s d f g h j k l ı ə ' z x c v b n m ç ş
5
4
3
2
1