Splean Song Lyrics Semantic Similarity

Authors:
Stanislav Krasnov,
Anastasia Yashchenko

Our corpus-based study investigates the value of word frequency in text clustering. We have chosen a corpus of song lyrics in order to show how lexical diversity and word repetition influence the similarity between two given songs and, furthermore, albums.

Hypothesis: lyrics within one album are connected with lyrics within another album and this connection is based on song word frequency.

Data: 203 Splean songs (a popular russian music band) collected manually from the official web-site (http://splean.ru) and annotated with song and album titles.

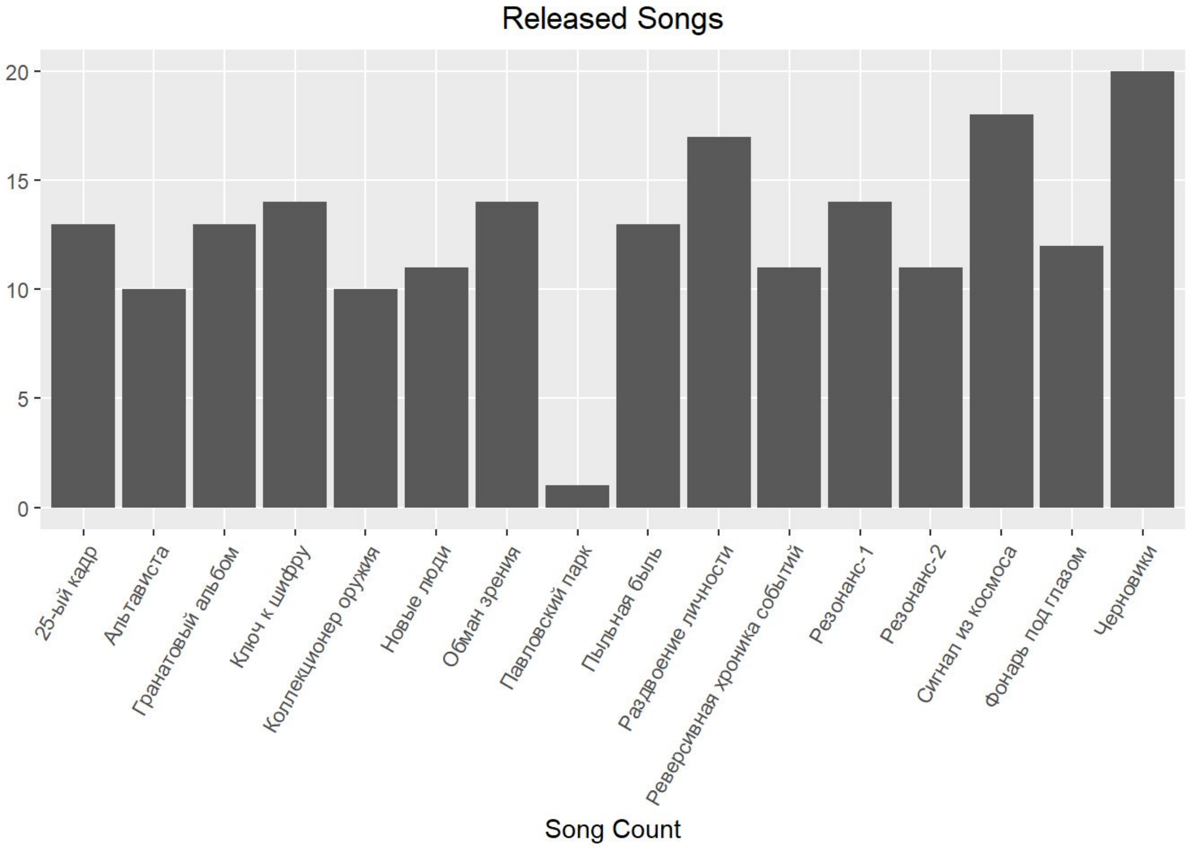
Pre-processing:

Using Python: converting to lower-case, removing punctuation + lemmatization.

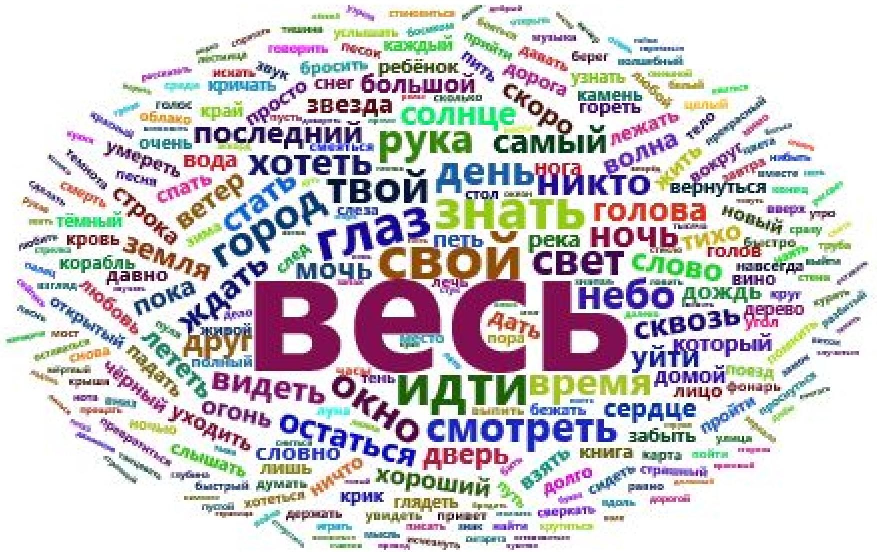
Using R: removing stop-words.

We used Python for lemmatization as it is a usual way for us, since we were not aware of any efficient tool for lemmatization of russian texts in R.

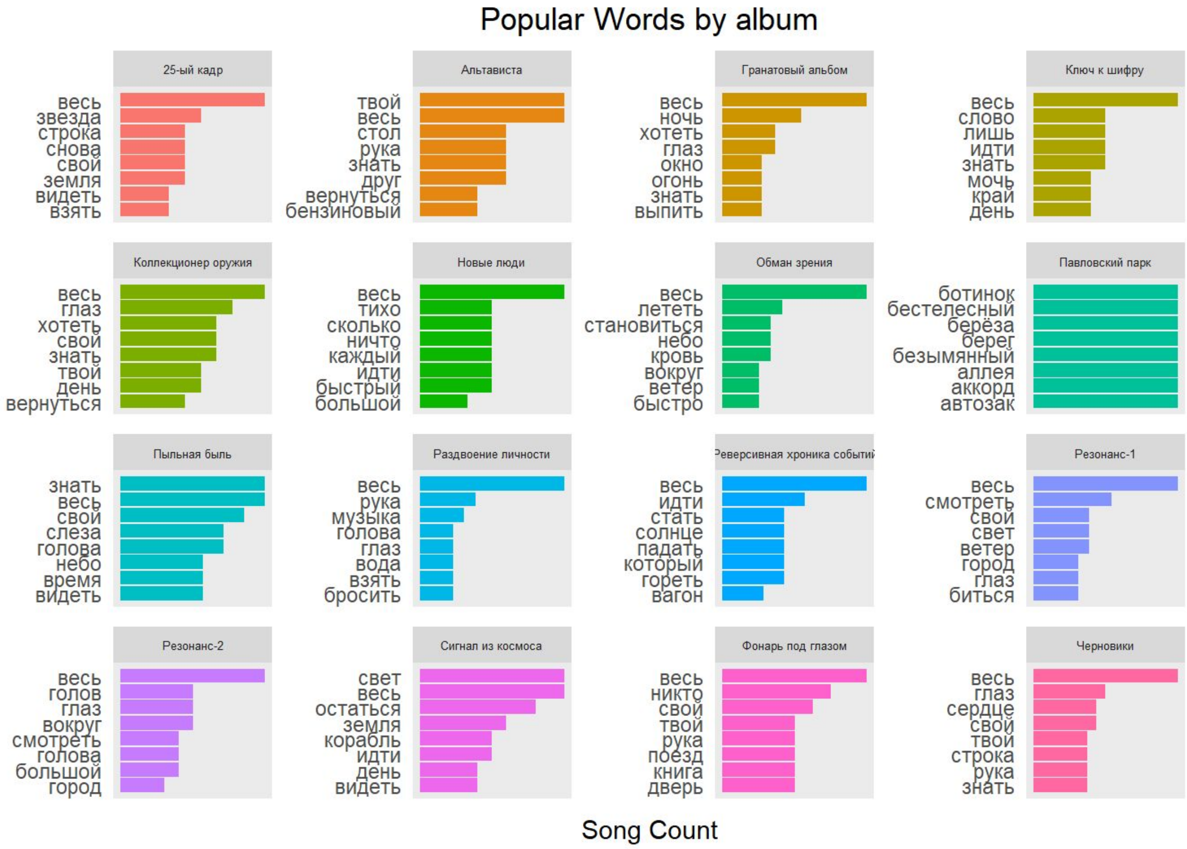
To begin with, we can illustrate our corpus with a histogram of released songs per year.



As we use word frequency to cluster our songs, let us see which words are common for our corpus:



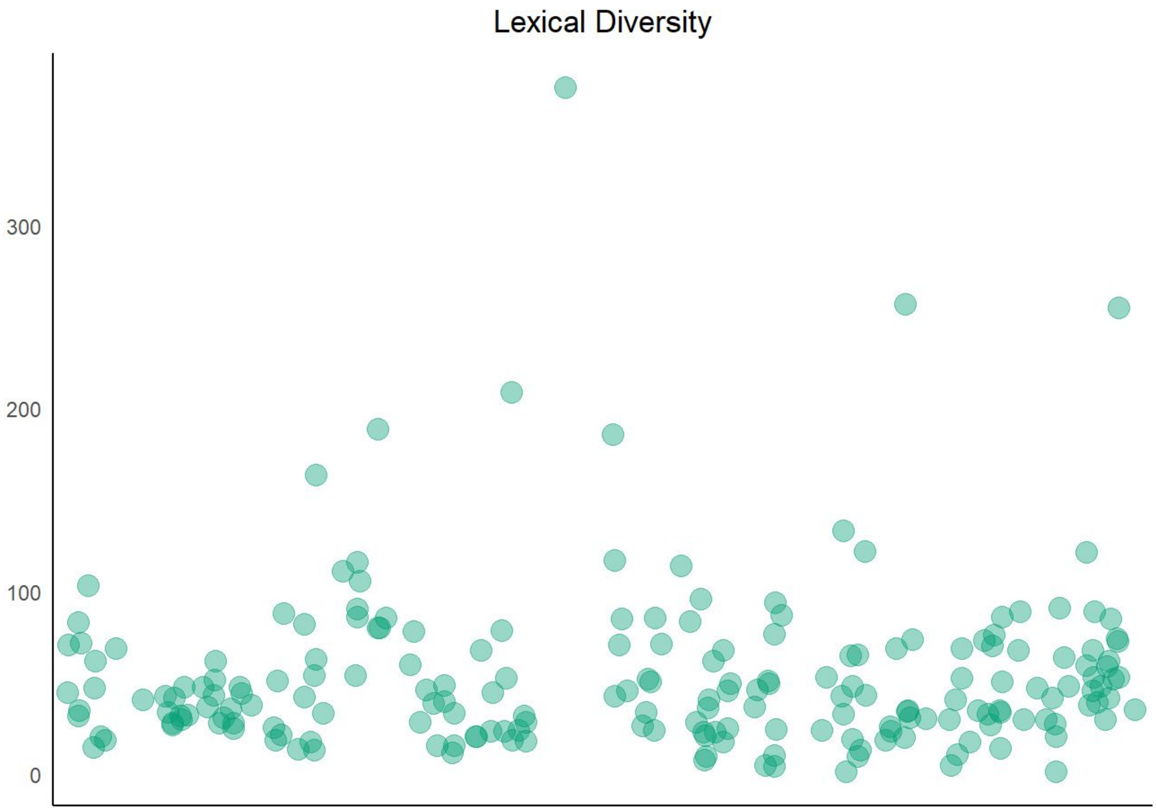
Our data have the information on the album of a given song, so we use it to show the most frequent words for each album:



As we can see, in the album ‘Павловский парк’ the set of popular words is rather different from other albums (due to the fact ‘Павловский парк’ consists of only one song).

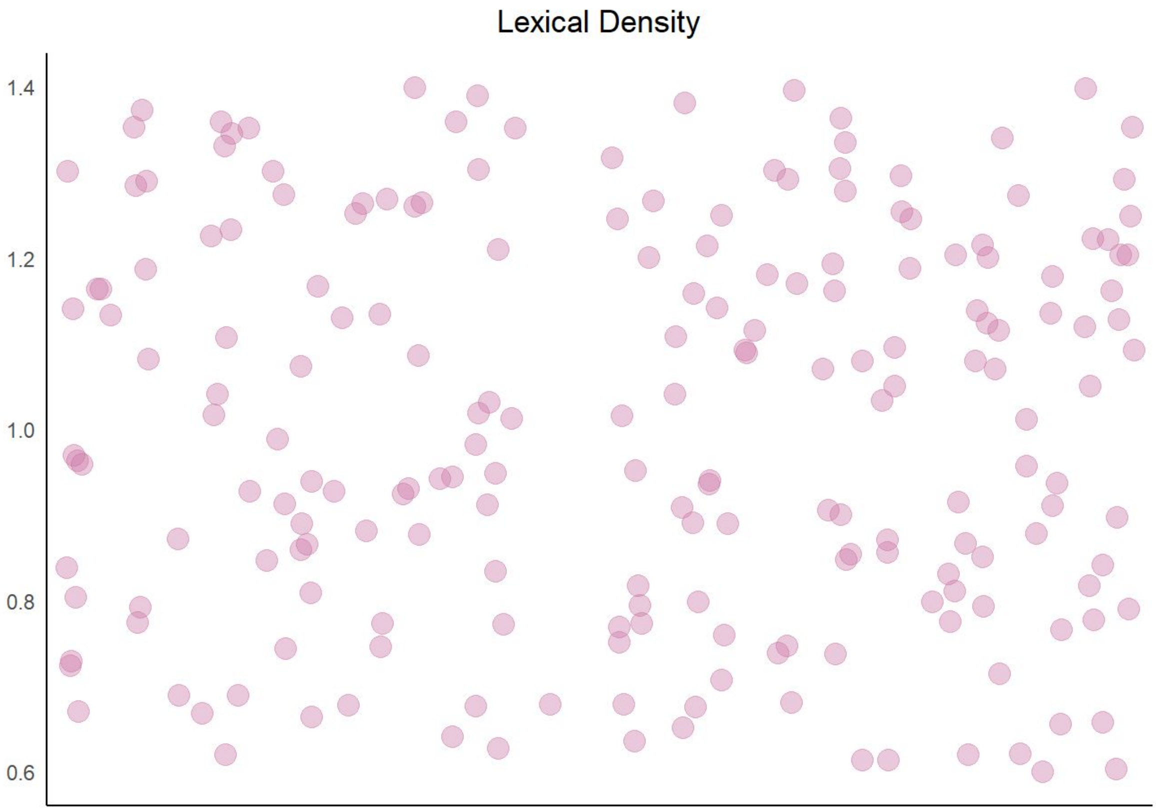
Since we are dealing with lyrics analysis, we should have a closer look at such parameters as lexical diversity and lexical density.

Lexical diversity is the number of unique words in a given song.



So, there is one song with a lot of unique words and several songs with above the average LD.

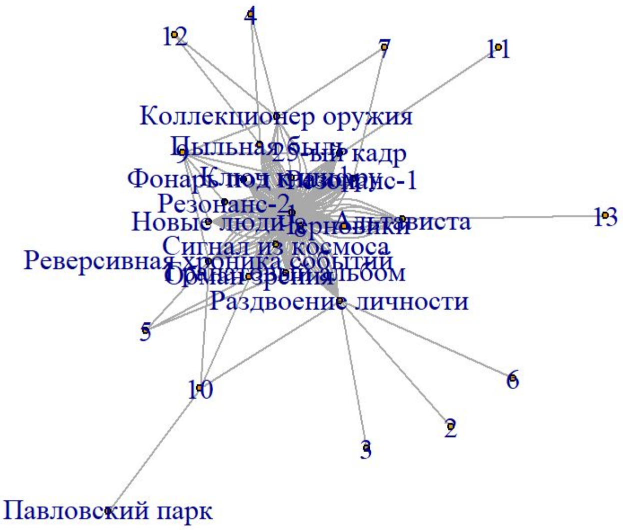
Lexical density is the measure of word repetition within one song (the number of unique words divided by total word number).



As we can see, a lot of songs have lexical density more than 0.8, what tells us that it is usual for author of these lyrics to repeat words in song.

Lyrics clustering is based on term frequency throughout the whole corpus. After creating document term matrix (documents multiplied by all unique words), we used k-means algorithm to cluster texts (k=13).

The final visualization shows that there is a cluster with a lot of songs in it (the number of edges shows how many songs in this album are connected with the cluster). An album with unpopular (among other albums) frequent words has only one connection to its cluster.


4
12
7
11
Коллекционер оружия
Пыльная быль
9
25-ый кадр
Фонарь под
Ключ к шифру
Резонанс-1
Резонанс-2
1
Новые люди
Черновики
Альтависта
13
Сигнал из космоса
Реверсивная хроника событий
Семейный альбом
Обман зрения
Раздвоение личности
5
10
6
2
3
Павловский парк