Assignment 3

TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. In comparison with the most frequent word, IDF part of this measure decrease the score for common words, so TF-IDF mostly shows the rare “valuable” words.

More about tf-idf: https://en.wikipedia.org/wiki/Tf%E2%80%93idf

Working process:

1. Make a list of English-speaking (songs lyrics should be in English) artists from genius.com (choose whoever you want). The number of artists should be at least 20.

2. Collect at least 200 lyrics from the website. Try not to use selenium.

3. Clean text corpus. Make all the words lowercase, reduce all the punctuation, so each lyrics text should be presented as string. Normalize texts (e.g. with pymorhy2).

4. Implement TF-IDF. Do not use ready-to-use methods from different libraries.

5. Apply your function for your dataset.

6. Present the results in a user-friendly form.

7. Analyze the results. How you can interpreter them? What are the most popular words for each of the artists? What are the words with the biggest score for each artist? Each songs?