Домашнее задание 5
Машинное обучение, БИБ20

Кластеризация: о вкусах не спорят

Наверняка у каждого из нас есть любимый жанр музыки, песня или исполнитель.

Также наверняка мы хоть раз в жизни спорили с другими людьми о том, что именно наш любимый жанр или исполнитель – самый лучший и слушабельный.

Но так ли велика разница между различными жанрами музыки, если мы будем основываться только на фрагментах текста? В этом задании вам предлагается провести небольшое исследование на подобную тему.

Выберите несколько жанров музыки и самостоятельно составьте датасет, включающий в себя следующую информацию: отрывок некоторой песни (например, припев) и жанр этой песни. Вместо жанра можно использовать также определенных исполнителей.

Возможный датасет (пример №1):

text	genre
death blood satan...	rock
money girls money...	rap
you are the best...	pop
death again death...	rock

Возможный датасет (пример №2):

text	artist
death blood satan...	SlipKnot
money girls money...	Lil Pump
you are the best...	Katy Perry
death again death...	SlipKnot

В данном примере жанр (или исполнитель) являются **кластером**, а наша задача – по исходным данным провести **кластеризацию**, то есть отнести песню к определенному жанру (исполнителю) по отрывку из её текста.

Требования к датасету: минимум 150 объектов (песен) и от 4 до 20 кластеров (жанров или исполнителей).

Как набирать данные? Попробуйте найти тематические сайты с текстами песен, напишите небольшой веб-парсер, выкачивающий тексты для

тех исполнителей, которых вы укажете. Также можно делать это руками без скриптов – смотрите сами, какой вариант Вам кажется быстрее.

Что использовать? Кластеризацию необходимо проводить методом k-means с использованием любых вспомогательных средств (TF-IDF, MiniBatch, нейронные сети, ...).

На что ещё обратить внимание? Для хорошего результата все тексты должны быть приведены к нижнему регистру и очищены от всех знаков препинания ("слово1 слово2 слово3 ..."). Перед самой кластеризацией тексты необходимо прогнать через Encoder (OneHot, Label, ...) – вспоминаем Д/З №2.

После разделения на кластеры сверьте то, насколько предсказание алгоритма совпало с реальным жанром (исполнителем) песни.

Согласны ли Вы с тем, что разделять песни по жанрам только на основе текста – это плохая идея?