**Министерство науки и высшего образования Российской Федерации**
Федеральное государственное автономное образовательное учреждение высшего образования

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Инженерная школа информационных технологий и робототехники
Отделение информационных технологий
Направление подготовки 09.04.04 Программная инженерия

# Отчёт по лабораторной работе №3

## TF-IDF

по дисциплине Представление знаний в системах искусственного интеллекта

Выполнил студент гр. 8ПМ4Л _____ _____ Сокуров Р.Е.
Подпись Дата Фамилия И.О.

Проверил лаборант ОИТ _____ _____ Сапегин А.А.
Подпись Дата Фамилия И.О.

Томск 2024 г.

**Задание**

1. Составить список песен на русском языке одного исполнителя (можно использовать ресурс genius.com). Всего список должен включать не менее 20 исполнителей. Количество песен для каждого исполнителя одинаково. Для каждого исполнителя должно быть не менее 10 песен.
2. Собрать список из текстов песен с ресурса автоматически (без использования selenium).
3. Очистить текстовый корпус. Перевести все тексты в нижний регистр, отфильтровать пунктуацию и т.д. Каждый текст должен быть представлен как строка. Нормализовать текст (например, с помощью pymorphy2).
4. Реализовать TF-IDF. Не использовать готовые методы из библиотек.
5. Применить созданную функцию на датасете.
6. Представить результаты в доступной для пользователя форме.
7. Проанализировать результаты. Как их можно интерпретировать? Какие слова являются самыми «важными» (имеют максимальное значение метрики) для каждого исполнителя? Какое слово самое важное для каждой песни? В отчете необходимо представить информацию, отвечающую на данные вопросы. Если данных много, необходимо вынести их в приложения в конце отчета.

**Ход работы**

Код работы и подробно расписанные этапы представлены в приложении А. Обработаем полученные данные.

Видно, что самые популярные слова по метрике TF-IDF по всем песням это «la», «na», «ah» «ooh» и тому подобные популярные в англоязычной музыке междометия. Связано это с тем, что эти междометия не зависят от текста песни, её темы, настроения и т.д., и используется для использования голоса в качестве музыкального инструмента, не передавая прямую смысловую нагрузку.

Если же смотреть по конкретным песням, то в песне Eminem – Mockingbird самыми популярными словами является «daddy» и «mama», ведь в произведении автор обращается к своей дочери.

И это характерно для всех песен: популярными словами в них являются непосредственно основная идея произведения. Так, в песне Abba – Happy new year популярными словами являются «happy», «year», «new», а в песне Abba – Waterloo самым популярным словом является «waterloo».

**Вывод**

В ходе лабораторной работы была реализован веб-скрейпер для считывания 10 самых популярных песен каждого 10 артистов. Текст этих песен далее был проанализирован с помощью TF-IDF и были найдены самые популярные слова по всем песням и для каждой песни отдельно.

Приложение А

Код файла «Code.ipynb»