А.В. Мельникова, Н.М. Гаврилова

Тюменский государственный университет, г. Тюмень

УДК 004.912

ПРОЕКТИРОВАНИЕ И РАЗРАБОТКА РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ ПО ПОДБОРУ ПЕСЕН ПО СЛОЖНОСТИ ДЛЯ ИЗУЧЕНИЯ АНГЛИЙСКОГО ЯЗЫКА

Аннотация.

Целью данной статьи является разработка рекомендательной системы по подбору песен по сложности. Из подобранного набора текстов были выделены признаки, по которым затем производилась кластеризация. На основе этих данных было разработано приложение для классификации плейлиста пользователя.

Ключевые слова. кластеризация текстов, сложность текстов.

Введение

В настоящее время существует потребность в определении сложности текстов в различных областях, связанных с лингвистикой. Оценка сложности текстов может использоваться в процессе обучения иностранному языку: зная уровень текстов можно их подбирать под конкретный уровень знания ученика, что дает более гибкую систему обучения. Переводчик текстов с иностранного языка может использовать определение сложности текстов, например, для определения времени на перевод, установления цены за перевод и т.д. Определение сложности текстов может применяться в документообороте. Длинные предложения, сложные конструкции внутри предложений затрудняют восприятие текста.

Однако проблема определения сложности состоит в том, что

человеку самостоятельно объективно трудно это сделать по разным причинам: большое количество текстов, большие затраты времени на извлечение признаков вручную и т. д. Поэтому идеей данной работы стала разработка кластеризатора текстов песен по сложности. В работе представлены результаты разработки рекомендательной системы по подбору текстов песен по их сложности.

Набор текстов песен и их признаки

В качестве данных для кластеризации был выбран набор текстов песен, собранных с сайта MetroLyrics. В нем присутствуют полные тексты песен, указаны исполнители и названия песен, а также жанры. В этом наборе оказалось 238202 песни, для каждой из которых был указан один из 10 жанров.

В качестве признаков текстов были выделены следующие признаки:

1. Средняя частота слов в тексте (mfw). Многие слова могут повторяться в песне по нескольку раз. Существуют тексты песен, в которых могут повторяться одни и те же строки по многу раз. Такие тексты обычно хорошо запоминаются.
2. Средняя (meanlw) и медианная (medlw) длины слов в тексте.
3. Общее количество слов в тексте (words). Здесь считается количество уникальных слов (словарь текста). Чем меньше различных слов встречается в тексте, тем проще его запоминать и меньше нагрузка при изучении текста.
4. Число общеупотребимых слов (cw) в тексте. Высчитывается количество слов из словаря исходного текста, которые есть в словаре общеупотребимых слов. Затем вычисляется доля этих слов относительно общего количества слов в словаре текста. Словарь общеупотребимых слов составлялся из слов для новичков и популярных в разговорной речи, его размер составил ~5 тыс. слов [1,2].

5. Доли имен существительных, прилагательных, местоимений, числительных, глаголов, наречий (Noun, Adj, Pron, Num, Verb, Adv) в словаре текста. В английском языке есть много различных частей речи, поэтому для упрощения они были выделены в группы [3].

На рисунках 1 и 2 представлены распределения размера словаря и количества общеупотребимых слов.

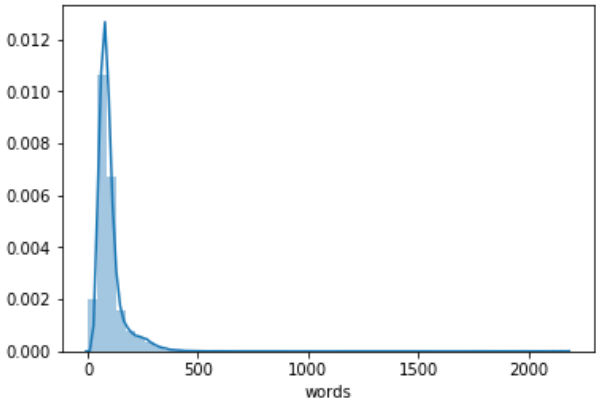

Рис. 1. Распределение размера словаря.

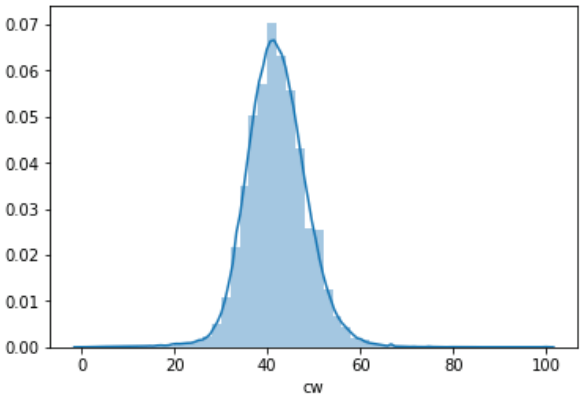

Рис. 2. Распределение количества общеупотребимых слов.

Кластеризация набора текстов песен по сложности

Для кластеризации такого большого набора данных был использован алгоритм KMeans. Это наиболее популярный и быстрый алгоритм кластеризации. Идея алгоритма заключается в том, что на каждой итерации вычисляются центры масс каждого кластера. Затем векторы разбиваются на кластеры снова, исходя из того, к какому из центров кластеров эти

векторы ближе. Для данного метода указывают количество кластеров. Так как в английском языке существует система владения языком, состоящая из 6 уровней, то в данном случае было принято разбивать на такое же небольшое число кластеров.

Данным методом было произведено разбиение на 4,6 и 10 кластеров с помощью библиотеки sklearn для языка Python. В результате было выявлено, что такое разбиение на классы не сбалансировано: если выстроить классы по возрастанию объектов в них, то разница между соседними классами будет более чем в три раза. Поэтому было выбрано разбиение на 7 кластеров, что представлено на рисунке 3.

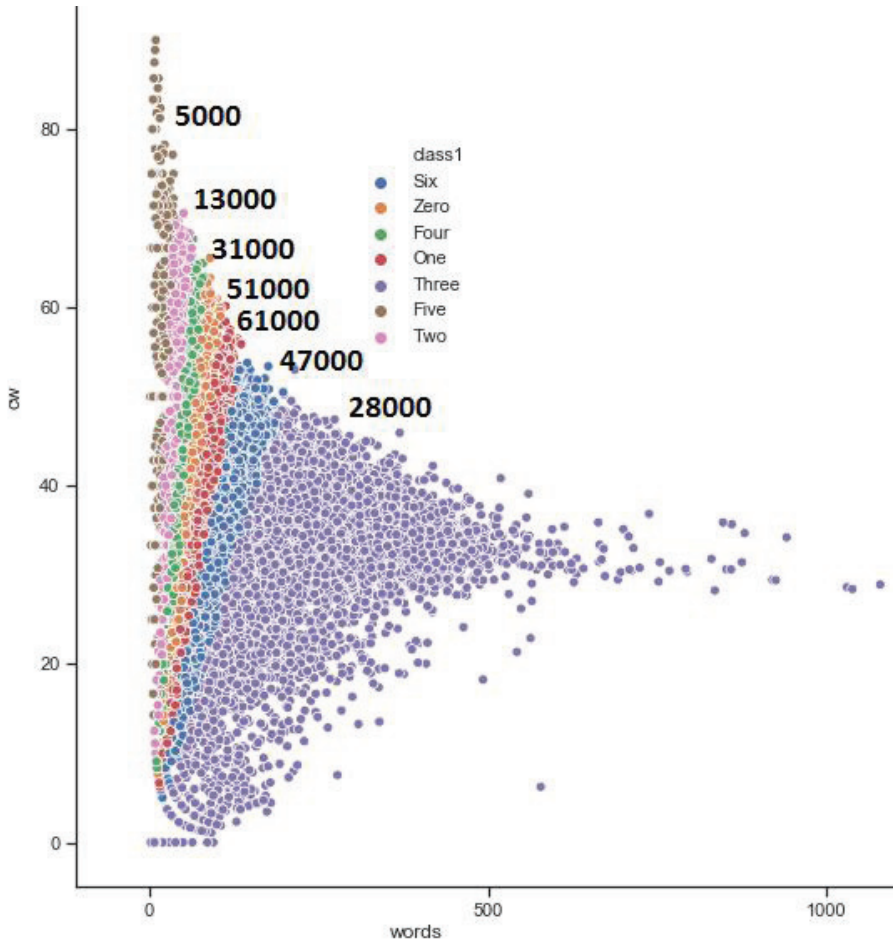

Рис. 3. Разбиение набора на 7 кластеров с обозначением количества текстов в каждом.

Зависимость количества уникальных слов от количества

общеупотребимых слов может говорить о том, что чем больше различных слов в тексте и меньше общеупотребимых, тем сложнее текст.

Классификация текстов песен по сложности

На основе данных, разбитых на кластеры, можно построить классификатор, который позволит определять, к какому из кластеров относится новый текст песни. В таблице 1 представлены результаты работы различных классификаторов на всех признаках. Из нее можно увидеть, что наиболее точным оказался алгоритм k-ближайших соседей.

Таблица 1. Результаты работы классификаторов.

Классификатор	Precision (точность)	Recall (полнота)	F-метрика
Байесовский классификатор (NB)	0,91	0,85	0,87
Линейный дискриминатный анализ (LDA)	0,68	0,58	0,60
K-ближайших соседей (KNN)	1	1	1

Разработка приложения

Приложение для классификации и подбора песен по сложности разработано на языке C# на платформе WPF с использованием различных библиотек. Некоторый функционал реализован на языке Python.

Функционал приложения включает в себя:

1. Выбор плейлиста, для песен которого будет определяться сложность, и они станут основой рекомендаций.
2. Определение сложности текстов загруженных песен.
3. Рекомендации новых n заданных песен по жанрам пользовательского плейлиста. Новые песни выбираются с учетом жанрового состава плейлиста пользователя

На рисунке 4 представлено главное окно программы.

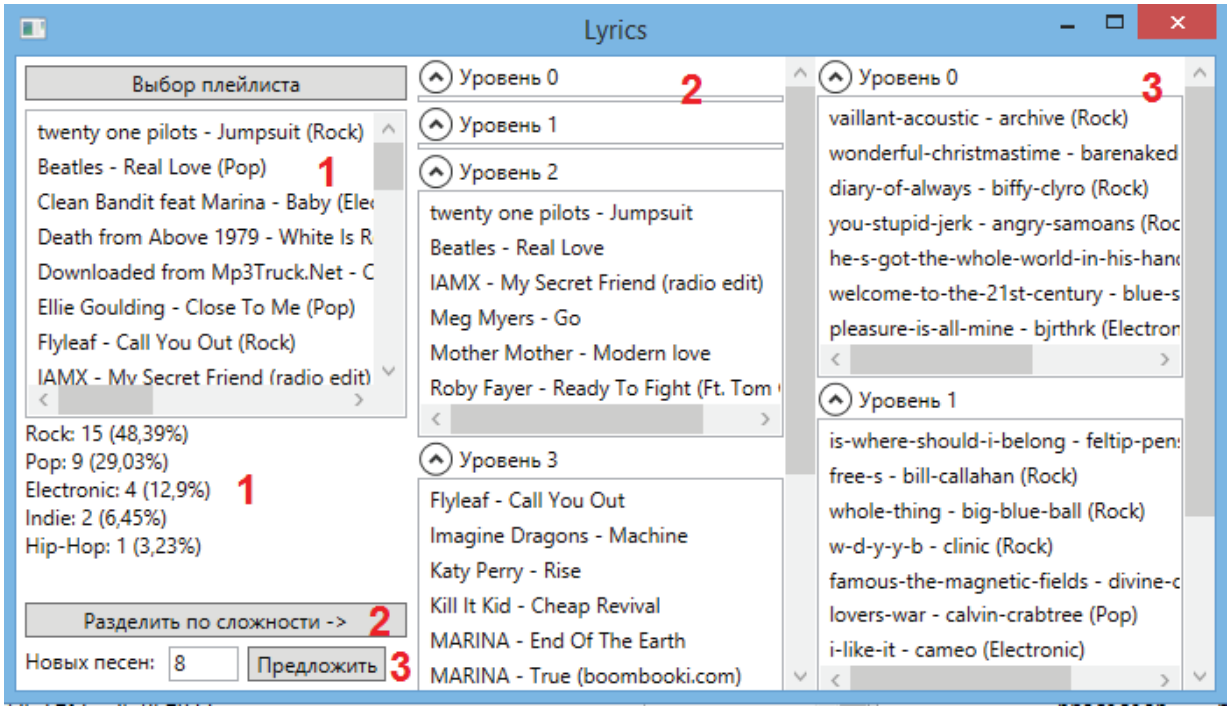

Рис. 4. Главное окно программы.

Здесь область 1 – загрузка плейлиста, его отображение и жанровый состав, область 2 – отображение классифицированного по сложности плейлиста, область 3 – поле для ввода количества новых песен и отображение нового плейлиста.

Заключение

В данной работе был сформирован набор текстов песен. Данный набор состоял из ~238000 текстов песен, относящихся к 10 жанрам, и из него были выделены 5 групп признаков. Данный набор был кластеризован на 7 классов с помощью метода KMeans. В качестве классификатора для новых текстов был выбран алгоритм k-ближайших соседей, так как он показал один из наилучших результатов. В итоге было разработано приложение на языках C# и Python, которое позволяет классифицировать плейлист пользователя по сложности и рекомендовать новые песни.

СПИСОК ЛИТЕРАТУРЫ

1. 1000 популярных слов на английском языке. URL:https://puzzle-english.com/directory/1000-popular-words (дата последнего обращения 9.04.2020).

2. 5000 часто используемых английских слов (список). URL:https://studynow.ru/dicta/allwords (дата последнего обращения 9.04.2020).

3. Части речи в английском языке. URL:https://engblog.ru/parts-of-speech (дата последнего обращения 9.04.2020).