Слайд 1

Тема моей дипломной работы – тексты поп и рэп исполнителей: количественный анализ и автоматическое определение авторства.

Слайд 2

Основная цель данного исследования – провести количественный анализ текстов рэп исполнителей, которая делится на 3 части: сравнение текстов рэп исполнителей, сравнение поп и рэп стиля и автоматическое определение авторства.

Основным инструментом исследования является подсчет расстояний между текстами на основе частотных списков лемм, скипграмм, ключевых слов и символьных n-грамм.

Стоит отметить, что все сборы и вычисления, сведения в таблицы автоматизированы с помощью программы, которую мы написали.

Слайд 3

Материал – автоматически собранный корпус из двух подкорпусов текстов поп и рэп исполнителей. 40 исполнителей, более 2к песен, более 450к словоформ. Вся информация о корпусе на слайде.

Слайд 4

Символьные n-граммы последовательности символов заданной длины. В нашем случае от 2 до 5. Выбраны такие длины, тк они охватывают сразу несколько уровней языка. Более того, последовательности выше 6 по статистике менее действенны

- Биграммы включают в себя предлоги, часть местоимений и аффиксы
- 3-4 граммы включают в себя уровень основ и корней слов, более того, согласно [Piperski 2019], n-граммы длины 4 являются лучшим показателем в автоматическом определении авторства
- N-граммы длины 5 могут включать в себя больше пунктуации, что даёт рассматривать тексты не только на уровне слов.

Скипграммы – пары слов внутри одного предложения, расстояния между которыми не более двух. Показывают вариации комбинаций слов и конструкций, коллокации

ARF – частотность с учетом расстояния между словами в тексте – поскольку в рэп стиле часто одно слово повторяется внутри припева.

Ключевые слова – особенности текстов со стороны лексики – тематика, характерные слова

Слайд 5

Расстояние по ключевым словам рассчитывается как среднее арифметическое рангов, которые находятся по формуле, представленной на слайде

Слайд 6

Расстояние по остальным параметрам находится как расстояние между векторами из начала координат и частотностями в качестве координат конца. В нашей работе использовались 3 расстояния – косинусное (от 0 до 1), манхэттенское (от 0), евклидово (от 0). Хотя при нормализации показывают одинаковые результаты, вторые расстояния имеют большие числа и тем самым сравнивать их проще. Основной упор всё же делался на косинусное. Формулы приведены на слайде.

Слайд 7

Все результаты представлены в виде heatmaps. Разберем результаты

Слайд 8

Более-менее одинаковые результаты, однако достаточно сильно отличающаяся группу 2rbina2rista, что связано с особым стилем и тематикой их текстов. На облаке слов можно увидеть наличие высокочастотной лексики, связанной со смертью, насилием.

Также удалось выделить групу исполнителей Noize MC, Каста, Кровосток, Баста, Макс Корж, Лизер, Скриптонит.

Слайд 9

Ключевые слова не дали никаких особых результатов по расстояниям.

Слайд 10

N-граммы. Диаграмма показывает наличие нескольких отличающихся исполнителей (высокие линии, есть пики графика). Соответственно:

1. 2-граммы дали короткие строки и более длинные тексты у 2rbina2rista, Хлеб, Big Russian Boss
2. 3-4 граммы дали особенности этих исполнителей по большому количеству звукоподражаний
3. 5-граммы -аналогично

Слайд 11

О чем говорилось в начале о ARF – высокая частотность пар одинаковых слов – группа Хлеб

2rbina2rista и Big Russian Boss отличаются большим количеством своих коллокаций

Слайд 12

Был результат о наличие двух групп исполнителей, которые отличаются группами – новая и старая школа. Рецензент указал на недостаток работы – отсутствие метода кластеризации, поэтому для визуализации и проверки данного результата были составлены дендрограммы (графы кластеров). На данном графе видно наличие двух групп, однако кластеризация нуждается в более детальном анализе

Слайд 13

Сравнение корпусов поп и рэп стиля.

Слайд 14

Леммы. Близкое расстояние, тк большие корпуса, большая часть списков – высокочастотные слова русского языка. Даже за вычетом 100 самых частотных списки похожи

Слайд 15

Данный параметр, как и ожидалось, дал результат – главное отличие – большое кол-во нецензурной лексики в рэп стиле при почти полном её отсутствии в поп. При высоком значении параметра N расстояние уменьшается, но наполнение списков показывает разную тематику отличающих корпуса слов

Слайд 16

н-граммы – аналогично частотным спискам лемм, тем более их еще больше. Соответственно, расстояния маленькие. Большие расстояния при 5-граммах – появление слов и словоформ, а по ключевым словам есть различия, а значит, расстояние чуть больше

Слайд 17

Высокие позиции – топы частотных списков биграмм русского, расстояния меньше. Однако композиции из слов, которые отличают, увеличивают расстояние

Слайд 18

Получено два алгоритма определения авторства, которые сочетают в себе сразу несколько параметров, которые необходимо подбирать для каждого типа корпуса.

Слайд 19

Первый – основан на минимальной сумме дельт. Второй – длина минимального вектора

Слайд 20

На базе имеющегося корпуса второй алгоритм оказался лучше. При тестировании на корпусе стихов показал не такой хороший результат, однако при подборе должных параметров, этот алгоритм имеет право на жизнь.