**Предобработка**

Тексты были скачаны с данных ссылок и предобработаны с помощью инструментов из библиотек nltk и Mystem. Код Находится в ноутбуке

**Обзор примененных подходов**

1. **TF-IDF** (ноутбук TF-IDF)
2. **Количественный и качественный анализ частей речи с помощью Mystem** (ноутбуки Part\_of\_Speech counter и Verb tense similarity)
3. **Тематическое моделирование с помощью LdaMulticore** (ноутбук Topic modelling)
4. **Сравнение векторов слов с помощью word2vec** (ноутбук word2vec )
5. **Визуализация результатов с помощью WordCloud** (ноутбук TF-IDF)

**Ответы на каждый из поставленных вопросов**

***Задача — с помощью анализа текстов рассказать, в чем особенности каждого текста и какие тексты схожи и в чем.***

Особенности каждого из текстов были отображены с помощью TF-IDF, топ-10 важных слов были выведены в соответствующем ноутбуке

В каждом из ноутбуков было произведено сравнение каждого из текстов друг с другом с точки зрения пересекающихся слов (выведенных тем или иным способом) и с точки зрения соответствующих метрик

***Какое из посланий в чем-то схоже с посланием президента?***

|  |  |  |
| --- | --- | --- |
| ***Регион*** | ***Метод выявивший наибольшее сходство*** | ***Количество голосов*** |
| Рязань | word2vec по всему тексту;  тематическое моделир-е(наравне с Тулой);  удельное количество глаголов непрошедшего, прошедшего и настоящего времени | 2,5 |
| Тула | Персечение топ-100 td-idf слов;  тематическое моделирование(наравне с Рязанью) | 1,5 |
| Тюмень | Пересечение частотных существительных, прилагательных и глаголов | 1 |
| Ямал | word2vec по абзацам | 1 |

***Какие идеи по анализу текстов можно взять из этого исследования?***

|  |  |
| --- | --- |
| ***Идеи Яндекса*** | ***Реализация в данном задании*** |
| Пятьсот глаголов, существительных и прилагательных, которые звучат в рэпе чаще всего | Топ-10 тех же частей речи (ноутбук Part\_of\_Speech counter) |
| Частотность использования того или иного слова в рэпе и в других жанрах | Анализ важных слов с помощью tf\_idf(ноутбук TF-IDF) |

***Что можно еще дополнить? Возможно какие-то общие словесные обороты, выражения, фразы, знаки препинания, длина предложений?***

Дополнительной идеей как раз было рассмотрение того кто чаще говорит о каком времени. Это было рассмотрено с помощью анализа времени глаголов.

Из еще нереализованных идей есть 1) Более подробное рассмотрение свойств частей речи с помощью Mystem. В идеале можно сделать вектор к каждому слову, каждое измерение которого будет соответствовать тому или иному грамматическому параметру рассматриваемой части речи. 2) Извлечение именованных сущностей (например, библиотекой «Наташа») и анализ пересечения полученных значений

***Расскажите, как бы вы визуализировали полученные данные (не для науки, а чтобы это было интересно СМИ и читателям)?***

Инструмент WordСloud, использованный в статье Яндекса, ссылка на которую давалась в задании, показался мне самым подходящим.