Техническое задание

Формулировка задачи: исследовать методами машинного обучения программы политических партий России.

Данные представляют собой текста, размещенные на сайтах партий, либо на веб-страницах, либо в форматах pdf и docx. Данные берутся с сайтов партий.

Всего 30 партий.

Задачи:

1. Провести сбор данных с сайтов партий в текстовый формат (.txt).
2. Разметить партии по попаданию/нет в Госдуму.
3. Провести токенизацию, удалить стоп-слова, выделить основные слова с помощью меры TF-IDF, провести кластеризацию (например, с помощью метода K-means).
4. Посмотреть пересечения множеств слов у различных партий, построить облака слов для различных политических течений.
5. Определить сферы, в которых партии, согласно их программам, собираются проводить политику.
6. Построить классификатор для определения попадания партии в госдуму по программе партии.

Возможно используемые методы классификации: градиентный бустинг, наивный баес, нейронные сети, деревья решений.

Оценка точности. Оценку точности классификации делать с помощью тестового и тренировочного набора, а также кросс-валидации. Для построенных классификаторов строить confusion matrix и считать по ней разные значения, характеризующие качество созданных моделей классификации.

Литература: <https://www.researchgate.net/profile/Dharminder-Yadav/publication/342348549_POLITICAL_SENTIMENT_ANALYSIS_ON_DELHI_USING_MACHINE_LEARNING/links/5ef06b3c92851ce9e7faeab6/POLITICAL-SENTIMENT-ANALYSIS-ON-DELHI-USING-MACHINE-LEARNING.pdf>

Новизна: автоматическая классификации программ политических партий.

Планы на будущее: использование программ партий других стран.