**Логика работы программы**

Для выполнения поставленной задачи мы написали программу на языке Python 3, состоящую из нескольких смысловых блоков, выделенных в отдельные файлы: сбор заголовков статей со ссылками на них, обучение модели на корпусе заголовков, выделение кластеров тем на корпусе тестовых заголовков, сбор комментариев к полемичным темам, выделение позиций/мнений по теме.

*Обучение модели*

Автоматическое выделение полемичных тем невозможно без понимания компьютером естественного языка, поэтому мы решили обучить модель Doc2Vec из пакета Gensim на корпусе из 160 тысяч новостных заголовков.

Для качественного обучения модели мы провели предобработку данных: привели заголовки к нижнему регистру, удалили из них знаки пунктуации и стоп-слова и лемматизировали оставшиеся слова. Лемматизацию мы делали с помощью библиотеки pymorphy.

Итоговая модель обучалась по принципу разреженного мешка слов, то есть она не учитывает связи между словами и их порядок в предложении, строит векторы по словам с одной итерацией по корпусу и окном в 2 слова, минимальной частотностью слова 6 и размерностью вектора 80. Проверка качества обученной модели осуществлялась путем передачи ей запросов со словами и оценкой, насколько близкие по смыслу и тематике слова она выдает. Так, например, близкими слову нефть модель считает слова ####, к слову # -- слова ###. Всего в среднем обучение модели занимает около 7 минут и по завершении обучения программа сохраняет модель в файл, в названии которого прописаны основные параметры обучения модели и предобработки данных: noStopLemma\_PV-DBOW\_wrd-vec\_1it\_2win\_6mincount\_alpha25-25\_sz80.

*Выделение кластеров тем на корпусе тестовых заголовков*

Обучив модель, мы использовали её для выделения кластеров тем на корпусе из 2 200 тестовых заголовков, которые мы получили, случайным образом выбрав по n заголовков из каждого файла выкачанных заголовков из тематических разделов сайта Риа-новости. В обучающем корпусе для модели эти заголовки не присутствовали.

Заголовки кластеризовались по методу k-средних, в основе которого лежит минимизация суммарного квадратичного отклонения элементов кластера от его центра, алгоритм кластеризации мы взяли из библиотеки nltk. Кластеризуемые заголовки проходили такую же предобработку, как и заголовки, на которых обучалась модель Doc2Vec -- приведение к нижнему регистру, удаление стоп-слов и пунктуации, лемматизация.

Оптимальным оказалось выделение 35 кластеров с 30-ю попытками случайной кластеризации, что занимает около 10 минут машинного времени. Важно отметить, что, выделяя 35 кластеров мы не ожидаем, что в корпусе из 2 тысяч заголовков мы по найдем 35 тем, вызывающих полемику. В корпусах новостей не бывает так, что все новости можно было бы разделить на группы по узким темам, так сгруппируется только часть новостей, все остальное объединится в группы по более обширным темам. Например, пожары, природные катаклизмы и прочее. Мы глазами выберем узкие темы, потому что это не противоречит названию работы. Главная суть подбора правильного количества кластеров в том, чтобы максимально точно собрать вместе заголовки по узким темам и хорошо отделить от них заголовки по широким и неполемичным темам.На данный момент кажется, что в тестовом датасете должны выделиться следующие темы: выборы во франции, международная борьба с иг, взрыв в метро в питере, (военные) конфликты в украине, днр, лнр, русские хакеры, корейские ракеты. Хотелось бы еще, чтобы выделилось что-то про антироссийские санкции.