**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 общим темам. Например, пожары, природные катаклизмы и прочее. Мы вручную выберем узкие темы, потому что это не противоречит теме 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.

В результате мы отобрали 10 кластеров, которые довольно однозначно выделяются кластеризатором и могут вызвать полемику: выборы во франции, международная борьба с иг, взрыв в метро в Петербурге, кибератаки русских хакеров, запуск ракет КНДР, международные отношения между США и Россией, ядерная программа Ирана, военные столкновения в днр и лнр, Русско-Украинские отношения, антироссийские санкции.

*Выделение мнений по выбранным темам*

Собрав для каждой из десяти выделенных тем корпус комментариев к статьям мы решили, что оптимальным способом выделить них мнения также будет кластеризация. Мы предположили, что люди, придерживающиеся похожего мнения, будут пользоваться схожей лексикой, и их комментарии будут выделяться в один кластер.

Оценив количество комментариев и новостных заголовков по каждой из тем, мы пришли к выводу, что действительно полемичными стоит считать только те темы, в которых было не меньше 100 комментариев и на каждую статью в среднем приходилось не менее 10 комментариев. Однако, при первых попытках кластеризации мы не исключали из анализа ни одну из тем и оценив результаты кластеризации, а также содержание комментариев, убедились, что наше предположение верно. Таким образом, из десяти отобранных тем действительно полемичными были признаны пять: запуск ракет КНДР, международные отношения между США и Россией, военные столкновения в днр и лнр, Русско-Украинские отношения, антироссийские санкции.

Для кластеризации мы снова воспользовались методом k-средних в реализации библиотеки sklearn. Перед кластеризацией комментарии проходили ту же предобработку, что и новостные заголовки при выделении тем. Опытным путем, исходя из общих представлений о том, каким может оказаться распределение мнений по выбранным темам, мы установили, что оптимальнее всего будет выделять пять кластеров. В качестве параметров кластеризации также использовались 8 попыток выбора центра кластера алгоритмом и максимум 100 повторений.

В конечном счете мы можем выделить несколько основных точек зрения, которых придерживаются комментаторы. Например, по теме запуска корейских ракет мнения разделились на: 1) угрозы Кореи – реакция на попытки вмешательства во внутренние дела страны другими странами, в частности США – сочувствие Корее; 2) Китаю и России необходимо изменить политику взаимодействия с Кореей, препятствуя вмешательству в него со стороны США; 3) Россия не должна вмешиваться, для нас они не угроза.

Кластер мнений об отношениях России и США оказался сильно смещенным в сторону обсуждения русско-болгарских отношений, так как 75% комментариев были оставлены под статьей с заголовком Президент Болгарии опасается, что США наладят диалог с Россией, а так как пользователей из России комментирует намного больше, комментарии во всех кластерах выражают примерно одну и ту же мысль – «Болгары поздно опомнились, что Россия из этого конфликта выйдет победительницей, мы их уже не простим».

По теме военных столкновений в днр и лнр мнения были примерно следующими: 1) действия Украины в этом вопросе нелогичны и не обоснованы; 2) Россия и Украина – братские народы с одним прошлым и совершающие одинаковые ошибки; 3) вопрос, возможно, придется решать силой и с привлечением военной мощи России.

По Русско-Украинским отношениям и Крымскому вопросу в частности выделились следующие позиции: 1) Крым – исконно русская территория; 2) Крымчане должны сами выбирать, в состав какой страны им входить; 3) Крым принадлежит Украине, так как это более древняя страна, равно как и украинская нация – более древняя нация.

О санкциях комментаторы говорили, что: 1) санкции выгодны России; 2) санкции не оказывают на Россию никакого влияния; 3) нужно предпринимать меры против санкций.