**Сбор данных**

Выделение полемичных тем проводилось по заголовкам статей на сайте Риа-новости из следующих разделов: политика, общество, экономика, происшествия, в мире, а также по рубрике Пхенчхан-2018 раздела спорт. Разделы наука, культура и религия не присутствуют в нашем корпусе из-за отсутствия в них достаточного количества статей. В корпус вошли статьи за период 1 января 2017 года – 10 февраля 2018 года, из раздела пхенчхан-2018 были взяты статьи за период 12 июля 2017 года – 28 февраля 2018 года.

Обработка сайта осуществлялась с помощью библиотеки Beautiful Soup для Python 3.6, и состояла из следующих этапов:

* Переход на страницу раздела
* Сбор заголовков статей с первой страницы
* Автоматический поиск кнопки «загрузить еще» и соответствующей ей ссылки на страницу с более ранними статьями
* Переход на предыдущую страницу и выкачивание заголовков содержащихся в ней статей
* Рекурсивное повторение до тех пор, пока не будут собраны заголовки всех статей за выбранный период

Для раздела пхенчхан-2018 использовалась та же последовательность действий, но несколько отличалось устройство самих функций, потому что весь раздел спорт находится на дочернем сайте Р-спорт, у которого исходные коды html-страниц частично отличаются от кодов страниц на главном сайте.

В итоге мы получили 6 текстовых файлов с заголовками статей и ссылками на сами статьи.

Для выделения мнений мы собирали комментарии к статьям, заголовки которых наш алгоритм выделил в одну тему. Комментирование на сайте РИа-новости устроено так, что на каждый комментарий можно ответить, то есть таким образом сразу выделяются связанные обсуждения по теме.

Чтобы собрать комментарии и ответы на них, мы с помощью библиотеки Beautiful Soup и ссылки на новостную статью автоматически находили на странице раздел с комментариями. Затем мы собирали необходимую метаинформацию: id статьи, дату последнего комментария на странице, дату и id комментария, к которому есть ответы, и с ее помощью собирали комментарии и ответы, если они занимают больше одной отображаемой страницы. На выходе для каждой выделенной темы мы получали список с заголовком статьи, ссылкой на нее, датой первого комментария, всеми комментариями и ответами на них и списком всех пользователей, участвовавших в обсуждениях по теме статьи.

Всего мы собрали корпус из 167 тысяч новостных заголовков и 6 800 комментариев к статьям в выделенных темах.