Дорожкин Денис, 425 группа

Отчет

20 исходных статей про политику и спорт

Politics

https://www.vedomosti.ru/politics/news/2017/11/16/741978-rossiisko-turetsko-iranskogo-sammita
http://www.rosbalt.ru/world/2017/11/16/1661309.html
https://iz.ru/671649/2017-11-16/iavlinskii-rasskazal-o-tceliakh-svoego-uchastiia-v-prezidentskikh-vyborakh
https://regnum.ru/news/2345987.html
https://www.vedomosti.ru/business/news/2017/11/16/741985-turetskih-tomatov
https://www.kp.ru/online/news/2933182/
https://rg.ru/2017/11/16/reg-ufo/medinskij-dopustil-razryv-muzejnyh-otnoshenij-s-gollandiej.html
https://rg.ru/2017/11/16/v-rossii-vvedut-pozhiznennoe-zakliuchenie-za-verbovku-terroristov.html
https://riafan.ru/997482-putin-nazval-zadachu-po-snizheniyu-smertnosti-ot-tuberkuleza-na-95-ambicioznoi
http://www.tatar-inform.ru/news/2017/11/16/583731/

Sports

https://rsport.ria.ru/football/20171116/1128740440.html
https://www.liveresult.ru/news/%D0%A5%D0%BE%D0%BA%D0%BA%D0%B5%D0%B9/c_55007/
https://www.kaliningrad.kp.ru/online/news/2933298/
https://russian.rt.com/sport/news/450147-peru-chm-shapka-ushanka
https://www.gazeta.ru/sport/news/2017/11/16/n_10824014.shtml
https://matchtv.ru/football/matchtvnews_NI799651_Nejmar_pytalsa_sorvat_transfer_iz_Barselony_v_PSZh
https://www.championat.com/football/news-2955577-visla-snova-poprobuet-priobresti-vratarja-vasjutina-prinadlezhaschego-zenitu.html
https://russian.rt.com/sport/news/450057-korintians-chempion-brazilii
https://www.niann.ru/?id=516800
https://www.liveresult.ru/news/%D0%A5%D0%BE%D0%BA%D0%BA%D0%B5%D0%B9/c_55020/

Предобработка

1) Все статьи были заранее преобразованы морфологическим анализатором.
2) К каждой из 2х категорий статей в файл была дописана соответствующая классификация ‘politics’, ‘sports’.
3) Был сформирован список файлов всех статей all.txt

Программа

classificator.py на вход принимает:
1) Имя файла с названиями всех файлов со статьями (уже обработанными морфологическим анализатором)
2) Имя файла со статьей, обработанной морфологическим анализатором, которую необходимо классифицировать.

Далее программа вычисляет вероятности принадлежности статьи к определенным классам (вес, имя класса).
В конце программа выводит наиболее подходящий класс статьи.

Пример работы

Стоит заметить, что если необходимо классифицировать уже классифицированную ранее статью (политика или спорт), то классификатор классифицирует корректно.

Некоторая статья о политике уже классифицированная
https://rg.ru/2017/11/16/reg-ufo/medinskij-dopustil-razryv-muzejnyh-otnoshenij-s-gollandiej.html

[(0.84763795731821179, 'politics'),
(0.75288664254383619, 'sports')]
Класс текста: politics

Произвольная новая статья о политике (раннее не классифицированная)
https://riafan.ru/997610-vks-rf-za-nedelyu-unichtozhili-svyshe-1250-obektov-boevikov-v-sirii

[(0.73309287356560393, 'politics'),
(0.7045627710966814 2, 'sports')]
Класс текста: politics

Некоторая статья о спорте уже классифицированная
https://russian.rt.com/sport/news/450147-peru-chm-shapka-ushanka\

[(0.74928024121732761, 'sports'),
(0.7225100704717532, 'politics')]
Класс текста: sports

Произвольная новая статья о спорте (раннее не классифицированная)
https://rg.ru/2017/11/17/video-hokkeist-kamenev-slomal-ruku-v-pervom-zhe-matche-za-kolorado.html

[(0.82522348125275335, 'sports'),
(0.79496507162675312, ‘politics')]
Класс текста: sports