ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
”ВЫСШАЯ ШКОЛА ЭКОНОМИКИ”»

Московский институт электроники и математики

Юткин Дмитрий Игоревич, группа БИВ-144
Вдовкин Василий Алексеевич, группа БИВ-144

Классификация и агрегация новостных статей, используя методы и модели обработки естественного языка

Междисциплинарная курсовая работа
по направлению 09.03.01 Информатика и вычислительная техника
студентов образовательной программы бакалавриата
«Информатика и вычислительная техника»

Студент _____ Д.И. Юткин
Студент _____ В.А. Вдовкин

Руководитель
старший преподователь
Л.Л. Волкова

_____

Москва 2017 г.

Аннотация

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Abstract

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Оглавление

1 Введение

2 Цель и задачи курсовой работы

Целью курсовой работы является разработка веб-сервиса, который

а) в реальном времени получает статьи из новостных источников;

б) классифицирует полученные статьи по общим темам;

в) агрегирует по схожести содержания статьи из различных источников.

Последние два пункта должны быть автоматизированы с помощью алгоритмов анализа данных и машинного обучения.

Для достижения поставленной цели, должны быть выполнены следующие задачи:

а) Изучить подходы автоматической обработки текста и естественного языка;

б) Собрать корпус новостных статей для обучения моделей;

в) Исследовать различные подходы к классификации и агрегации текстовых документов;

г) Разработать back-end и front-end составляющие сервиса;

д) Объединить результаты предыдущих пунктов в единый веб-сервис.

3 Сбор и подготовка данных

3.1 Получение данных из новостных источников

Для получения робастных моделей машинного обучения, требуется достаточно большой корпус новостных статей, содержащий порядка нескольких сотен тысяч документов. Кроме того, статьи должны быть русскоязычными, а также содержать современную лексику и актуальную информацию, соответствующую состоянию дел в мире за последние 10-15 лет. При детальном анализе Интернета, корпус, удовлетворяющий перечисленным требованиям, не был найден и, таким образом, было решено собрать данные для исследования самостоятельно.

В качестве новостных источников были выбраны следующие популярные СМИ: Газета.Ru, Lenta.ru, ТАСС, Новая Газета, ВЕДОМОСТИ и СПОРТ-ЭКСПРЕСС. Последнее было выбран по причине малого количества спортивных новостей от других источников.

При формировании корпуса, все новостные статьи сопровождались различными метаданными, такими как ссылка на статью, дата публикации и название СМИ. В результате было получено более 1,1 млн. новостей, многие из которых имели неправильно проставленные темы. Причин тому может быть несколько, например, ошибки редакторов или технические ограничения веб-сайтов новостных агентств. В итоговую выборку, которая в дальнейшем использовалась для обучения и тестирования алгоритмов вошло 133 тыс. статей. Распределение СМИ и тем на отобранных данных отражено на рисунке 1.

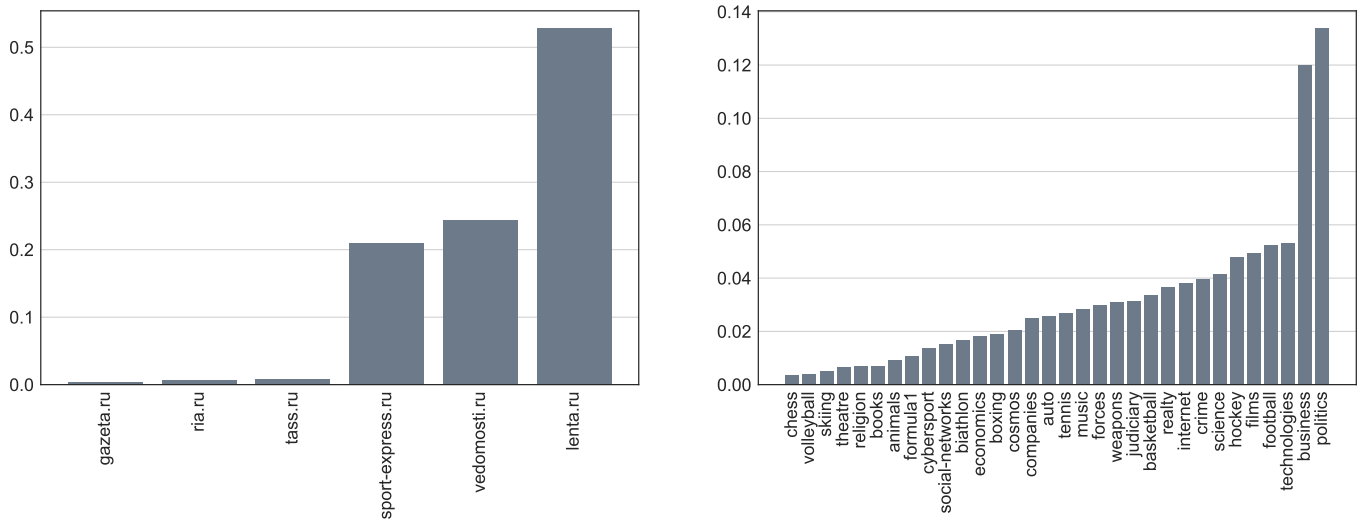

Рисунок 1 — Распределение СМИ и тем в данных

3.2 Препроцессинг новостных статей

Препроцессинг является одной из важнейших стадий анализа данных. В данной работе, обработка данных начинается уже на этапе получения новостей: из них удаляются все лишние HTML, JSON и XTML теги.

Основной конвеер обработки данных состоит из нескольких последовательных этапов:

а) Приведение текста в нижний регистр.

б) Удаление чисел и символов пунктуации. Дефис сохраняется.

в) Удаление стоп-слов. В данный набор входят наиболее часто употребляемые слова русского и английского языков, а также названия новостных агентств («лента», «тасс», «риа» и т.д.), сохранение которых приводит к переобучению моделей.

г) Лемматизация каждого слова с помощью библиотеки MyStem[1].

Исходный код конвеера можно найти в приложении.

[1] https://tech.yandex.ru/mystem/

Список литературы

Приложение