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

В данной работе изучаются алгоритмы классификации и агрегации текста, основанные на методах и моделях обработки языка, и применяются на русскоязычных новостных статьях. В процессе работы удалось собрать и обработать большой корпус новостей, на котором обучены следующие модели: TFIDF, SVM, FastText, word2vec, Kmeans. Классификаторы показали точность 86-88%. Для визуализации работы перечисленных алгоритмов на практике реализован новостной агрегатор в виде web-сервиса, агрегирующий и классифицирующий актуальные новости с сайтов российских СМИ.

Abstract

The main purposes of this work are the natural language processing models and algorithms of the text aggregation and classification and application of these models and algorithms on the russian media news articles. In the course of this work a dataset with the significant number of news was gathered and processed. On this dataset TFIDF, SVM, FastText, word2vec, Kmeans models were trained. The accuracy of the classifiers is between 86 and 88 depending on a model. To show how these algorithms are working on practice the news aggregation web-service was implemented, which aggregates and classifies articles in real-time.

Оглавление

1 Введение

С каждым годом вычислительные мощности современных компьютеров и сервисов, предлагающие облачные вычисления, позволяют обрабатывать всё большие массивы данных. Благодаря этому происходит быстрое развитие алгоритмов анализа данных и машинного обучения. Результат работы этих алгоритмов можно увидеть в нашей повседневной жизни: сервисы прогноза погоды, которые предсказывают направление движения облаков, персонализированная реклама, подстраивающиеся под интересы пользователя, автомобили с автопилотом и т.д. — в основе всех этих разработок лежат алгоритмы интеллектуального анализа данных и машинного обучения.

Один из важнейших вопросов, который стоит перед исследователями и разработчиками подобных систем: где же взять данные для обучения моделей, с помощью которых можно анализировать новую информацию?

Ответ можно найти в Интернете — это текст. Петабайты текста, написанного на естественном языке за последние несколько десятилетий и практически всегда имеющего смысл, доступны любому исследователю. Книги, новостные статьи, блоги, посты в социальных сетях, содержащие пусть и не очень литературный язык — всё это достаточно просто извлечь.

Именно поэтому обработка естественного языка (Natrual Language Processing, NLP) получило такое развитие. Существует большое количество развивающихся проектов, коммерческих и с открытым исходным кодом, которые опираются на NLP. Это, например, IBM Watson — набор продуктов, предлагающий множество инструметов работы с текстом: от извлечения важной информации и классификации, до его генерации. NLP используется в поисковике и переводчике компании Яндекс, использующей для анализа алгоритм MatrixNet. Существуют специально созданные для NLP свободнораспространяемые библиотеки для разработчиков: NLTK, Gensim, Apache OpenNLP, Stanford CoreNLP и многие другие.

Все стандартные и базовые способы обработки естественного языка в исследовательских работах и в примерах использования NLP-инструментов обычно применяются на английском языке, но подразумевается, что все методы применимы и для любого языка, при условии некоторых изменений в нормализации текста. В данной работе решаются две задачи NLP: это классификация и агрегация текста. Для придания работе новизны и оригинальности задача решается для новостных

статей, написанных на русском языке. Востребованность этих задач показывает полностью автоматизированный новостной агрегатор компании Яндекс – сервис Яндекс.Новости, который имеет огромную аудиторию.

Чтобы показать практическую значимость работы, реализован новостной агрегатор в виде web-сервиса, показывающий работу алгоритмов на постоянно обновляющихся актуальных новостях русскоязычных интернет-СМИ. Важно отметить, что точность реализованных решений в процессе данной работы высокая, но несоизмеримая с тем же сервисом Яндекс.Новости, команда которого работает над своими алгоритмами годами, и даже при этом сервис работает не всегда точно. blog/company/novosti-za-voskresene

2 Цель и задачи курсовой работы

Целью курсовой работы является разработка веб-сервиса, который:

а) в реальном времени получает статьи из русскоязычных новостных источников;
б) классифицирует полученные статьи по общим темам;
в) агрегирует по схожести содержания статьи из различных источников.

Агрегация и классификация основана на исследуемых в работе алгоритмы.

Для достижения поставленной цели, должны быть выполнены следующие задачи:

а) Изучить методы и модели автоматической обработки текста и естественного языка;
б) Собрать корпус новостных статей для обучения моделей;
в) Исследовать и реализовать различные подходы к классификации и агрегации текстовых документов;
г) Разработать back-end и front-end инфраструктуру сервиса.

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 нескольких сотен тысяч документов. Обычно существуют готовые коллекции документов, но из-за выбранных ограничений к документам, а именно: новости на литературном русском языке,

содержащие тег – слово, описывающие тему документа, найти такой корпус практически невозможно, поэтому необходимо собрать данные самостоятельно.

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

При анализе сайтов СМИ стало понятно, что они имеют схожую структуру: для отображения ссылок на статьи, используются страницы со списком новостей («Лента новостей»), по которым можно итерироваться, изменяя параметры запросов, например, дату последней новости на странице или количество показанных статей. Для извлечения данных из источников реализован набор парсеров, которые опираются на описанную структуру.

Стоит отметить, что во многих случах для получения чистого текста приходится ждать ответа от сервера и обрабатывать HTML содержание страниц, и это сильно замедляет работу, поэтому парсеры одновременно обрабатывают множество ссылок на статьи в параллельных процессах, что значительно ускоряет работу. Впоследствии эти парсеры удалось также внедрить в web-сервер для получения недавних новостей.

При формировании корпуса, все новостные статьи сопровождались различными метаданными: название СМИ, ссылка на статью, дата публикации, тег и заголовок. В результате было получено более 1,1 млн. новостей от 1999 до 2017 года, многие из которых имели неправильно проставленные темы или не имели тем вовсе. Причин тому может быть несколько, например, ошибки редакторов или технические ограничения веб-сайтов новостных агентств. Например, большинство новостей на сайте « Новой Газеты» помечены тегом «политика», хотя речь в них идет часто совсем о другом. После детального анализа тегов стало ясно, что относить новости к рубрикам редакторы стали только после определённого времени, а все уже имеющие статьи на сайте отнесли в «политику». Некорректные данные пришлось удалить.

В итоговую выборку, которая в дальнейшем использовалась для обучения и тестирования алгоритмов, вошло 133 тыс. статей, помеченных 32 различными тегами. Распределение СМИ и тем на отобранных данных отражено на рисунке 1.

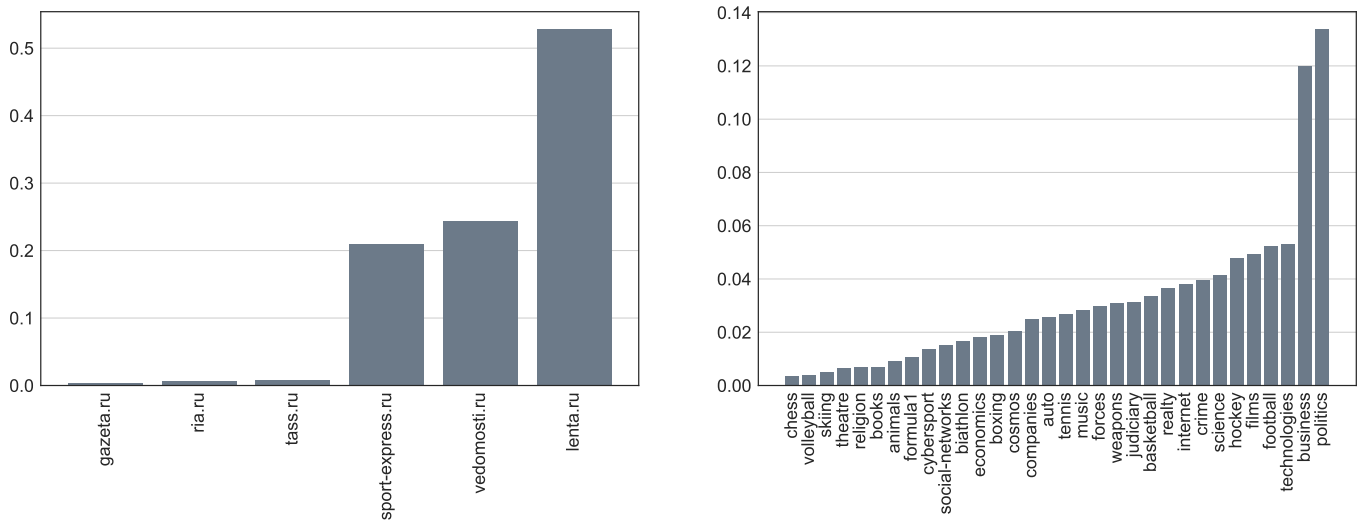

Рисунок 1 — Распределение СМИ и тем в данных

3.2 Нормализация новостных статей

Нормализация является одной из важнейших стадий обработки естественного языка. Вкратце, нормализация приводит текст в более информативный для моделей вид. В зависимости от языка процесс нормализации может отличаться. Для русского языка, например, достаточно удалить пунктуацию, стоп-слова и лемматизировать. Стоп-слова – это слова, которые примерно одинаково распределены по всему корпусу языка, чаще всего ими являются местоимения, предлоги и союзы. Лемматизация – приведение слова в начальную форму. Часто вместо лемматизации используется стемминг – алгоритм, который убирает части слова, влияющие на его форму, например, окончание. Данные алгоритмы не только опираются на словари, но и на определённые правила, зависящие от языка, так как в корпусе могут встречаться слова в разных формах, которых нет в словаре, например, неологизмы, но образованы они по правилам языка.

Процесс обработки текста в данной работе 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].

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

Список литературы

Приложение