

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 преподаватель
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	4
2	Цель и задачи курсовой работы	5
3	Сбор и подготовка данных	5
3.1	Получение данных из новостных источников	5
3.2	Нормализация новостных статей	6
4	Классификация новостных статей	7
4.1	Линейная модель на TF-IDF признаках	8
4.2	Градиентный бустинг деревьев на word2vec	9
5	Агрегация новостных статей	9
5.1	Кластеризация с помощью алгоритмов машинного обучения	9
5.2	Объединение в связные компоненты графа	9
6	Разработка веб-приложения	9
7	Заключение	9
	Список литературы	10
	Приложение	11

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.

Одна из важнейших проблем, возникающая перед исследователями и разработчиками подобных систем, заключается в поиске данных для обучения моделей, которые в будущем будут использоваться для анализа новой информации.

Решением данной проблемы являются тексты из сети Интернет. Петабайты информации, записанной на естественном языке за последние несколько десятилетий, доступны любому исследователю. Книги, новостные статьи, блоги, посты в социальных сетях — всё это является источником легкодоступных данных.

Именно поэтому обработка естественного языка (Natural Language Processing, NLP) получила большое развитие. За последние несколько лет появилось множество специализированных библиотек и сервисов для анализа естественного языка. К таким сервисам, например, относится IBM Watson — набор платных продуктов, предлагающий различные инструменты работы с текстом: от извлечения важной информации и классификации, до его генерации. Яркими примерами свободнораспространяемых библиотек являются: NLTK, Gensim, MyStem, rumorphy и многие другие.

В данной работе решаются две задачи:

- а) классификация тем новостных статей;
- б) агрегация новостных статей по смысловой близости.

Для придания работе новизны и оригинальности задача решается для русскоязычных новостей.

Практическая значимость работы доказывается на примере реализации в виде web-сервиса новостного агрегатора, который в реальном времени, с помощью

алгоритмов и подходов предложенных в данной работе, обрабатывает публикации русскоязычных интернет-СМИ.

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 несколько сотен тысяч документов. Обычно существуют готовые коллекции текстов, но из-за выбранных ограничений к документам, а именно: новости на русском языке вместе с тег — словом, описывающим тему документа, найти такой корпус не удалось, поэтому было принято решение собрать данные самостоятельно.

В качестве новостных источников были выбраны следующие популярные СМИ: «Газета.Ru», «Lenta.ru», «ТАСС», «Новая Газета», «ВЕДОМОСТИ», «РИА Новости» 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 алгоритмов, которые опираются на описанную структуру.

Стоит отметить, что во многих случаях для получения чистого текста приходится ждать ответа от сервера и обрабатывать HTML содержание страниц, что сильно замедляет работу, поэтому алгоритмы получения новостей одновременно обрабатывают в параллельных процессах множество веб-страниц, что значительно ускоряет работу. Данные алгоритмы также используются в основном web-сервисе для получения недавних новостей.

При формировании корпуса, все новостные статьи сопровождалась различными метаданными: название СМИ, ссылка на статью, дата публикации, тег и заголовок. В результате было получено более 1,1 млн. новостей с 1999 по 2017 год, 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», что зачастую не совпадает с истинным содержанием статьи. После детального анализа тегов стало ясно, что относить новости к рубрикам редакторы стали только после определённого времени, а все уже имеющиеся 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.

3.2 Нормализация новостных статей

Нормализация является одной из важнейших стадий обработки естественного языка. Не формально, нормализация приводит текст в более информативный для моделей вид. В зависимости от языка процесс нормализации может отличаться. Например, для русского языка зачастую удаляют пунктуацию и стоп-слова, а также производят лемматизацию. 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

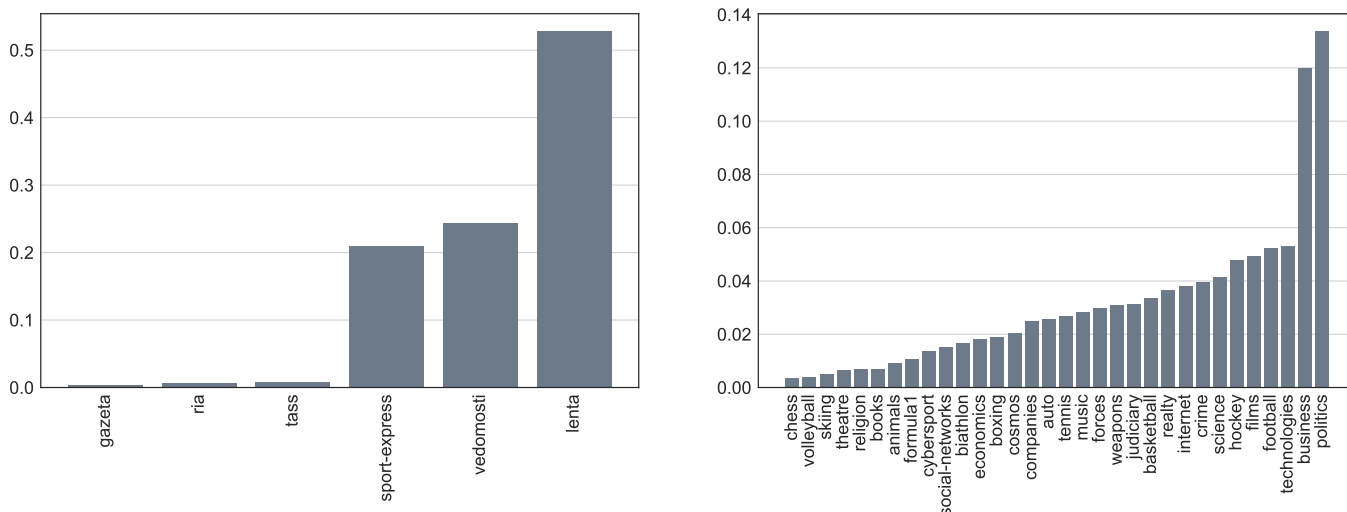


Рисунок 1 — Распределение СМИ и тем в данных

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 MyStem¹.

4 Классификация новостных статей

За последние два десятка лет, в результате активных исследований в области машинного обучения, было изобретено множество успешных алгоритмов классификации. Например, такие модели как support vector machines (SVM) [1], градиентный бустинг деревьев и нейронные сети [2], были успешно применены к задачам классификации текстов. В данной работе для классификации новостных статей

¹<https://tech.yandex.ru/mystem/>

использовались две из перечисленных выше модели — это SVM и градиентный бустинг деревьев.

Анализ текста является важной частью машинного обучения, однако сырые данные, а именно последовательности символов переменной длины, не могут быть переданы на вход алгоритму в явном виде т.к. большинство моделей ожидают численный вектор признаков фиксированной длины.

Векторизация — метод трансформации коллекции текстовых документов в числовые вектора признаков. Существуют различные подходы к векторизации текста, в данной работе были применены два самых популярных: TF-IDF и word2vec.

4.1 Линейная модель на TF-IDF признаках

Одним из самых простых подходов к решению задачи классификации текстовых документов является обучение линейного классификатора на TF-IDF признаках, посчитанных на корпусе документов.

TF-IDF — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. [3] TF-IDF — это произведение двух статистик: TF (term frequency) и IDF (inverse document frequency). На сегодняшний день, TF-IDF один из самых популярных способов взвешивания слов, входящих в корпус документов. Например, 83% рекомендательных систем цифровых библиотек используют TF-IDF [4].

Существует множество способов подсчёта TF-IDF, в данной работе использовался следующий:

$$\text{tf}(t, d) = \frac{n_t}{\sum_k n_k},$$

где n_t есть число вхождений слова t в документ, а $\sum_k n_k$ — общее число слов в данном документе.

$$\text{idf}(t, D) = \log \frac{|D|}{|\{d_i \in D \mid t \in d_i\}|},$$

где $|D|$ — число документов в корпусе, $|\{d_i \in D \mid t \in d_i\}|$ — число документов из коллекции D , в которых встречается t (когда $n_t \neq 0$).

Таким образом, мера TF-IDF является произведением двух сомножителей:

$$\text{TF-IDF}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D).$$

Признаковым описанием одного объекта $d \in D$ будет вектор

$$(\text{TF-IDF}(t, d, D))_{t \in V},$$

где V — словарь всех слов, встречающихся в коллекции D .

Для преобразования новостных статей в числовые признаки, использовался класс `TfidfVectorizer` из библиотеки машинного обучения `Scikit-learn` [5]. В качестве параметров векторизации использовались следующие значения: `min_df=3` — учитываются слова, встретившиеся суммарно во всех документах минимум 3 раза и `ngram_range=(1, 2)` — учитываются как отдельные слова, так и би-граммы.

После векторизации новостных статей, была получена разрежённая матрица размера 133 529 строк на 567 890 столбцов. По причине большого количества признаков (\gg обучающих примеров), в качестве классификатора было решено использовать линейный SVM. Не смотря на то, что данный алгоритм был впервые описан более пятидесяти лет назад, сегодня он по прежнему показывает одни из самых высоких результатов в задачах классификации текста. Как было показано в [6], особенно высокое качество удаётся получить при использовании би-грамм.

4.2 Градиентный бустинг деревьев на word2vec

5 Агрегация новостных статей

5.1 Кластеризация с помощью алгоритмов машинного обучения

5.2 Объединение в связные компоненты графа

6 Разработка веб-приложения

Веб-приложение состоит из двух основных частей: `back-end` и `front-end`. `Back-end` — это сам веб-сервер, который осуществляет обработку запросов пользователей, получение и обработку данных. `Front-end` — это пользовательский интерфейс, визуализирующий полученные данные от `back-end` в понятный вид. С помощью этого интерфейса пользователь способен не только получать, но и передавать данные на `back-end`.

7 Заключение

Список литературы

1. *Weston Jason, Watkins Chris*. Support Vector Machines for Multi-class Pattern Recognition // Proc. European Symposium on Artificial Neural Networks. — 1999. — Pp. 219–224.
2. Very Deep Convolutional Networks for Natural Language Processing / Alexis Conneau, Holger Schwenk, Loïc Barrault, Yann LeCun // *CoRR*. — 2016. — Vol. abs/1606.01781. <http://arxiv.org/abs/1606.01781>.
3. *Jones Karen Sparck*. A statistical interpretation of term specificity and its application in retrieval // *Journal of Documentation*. — 1972. — Vol. 28, no. 1. — Pp. 11–21. <http://dx.doi.org/10.1108/eb026526>.
4. Research-paper recommender systems: a literature survey / Joeran Beel, Bela Gipp, Stefan Langer, Corinna Breitingner // *International Journal on Digital Libraries*. — 2016. — Vol. 17, no. 4. — Pp. 305–338.
5. Scikit-learn: Machine Learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort et al. // *Journal of Machine Learning Research*. — 2011. — Vol. 12. — Pp. 2825–2830.
6. *Wang Sida I., Manning Christopher D*. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification // Proceedings of the ACL. — 2012. — Pp. 90–94.

Приложение