

Разработка системы категоризации текстов

Студент 325 группы Протасов Александр Васильевич
Научный руководитель: Волкова Ирина Анатольевна

Цели работы



По имеющемуся набору веб-страниц требуется разработать методы:


- сбора информации с веб-страниц,
- предварительной обработки текстов,
- формирования датасетов, на которых будет произведено сравнение результатов классификаторов и выбор оптимального из них.

Постановка задачи



- Изучить существующие подходы категоризации текстов различных авторов, их преимущества и недостатки, методы обработки текстовой информации на веб-страницах, методы классификации в машинном обучении.
- Разработать и реализовать алгоритм, позволяющий получать текстовые данные с веб-страниц.
- Подготовить исходные данные для решения поставленной задачи.
- Сравнить полученные результаты с результатами существующих решений задачи.
- Выбрать наиболее оптимальный метод для системы категоризации текстов.

Используемые методы машинного обучения для классификации текстов

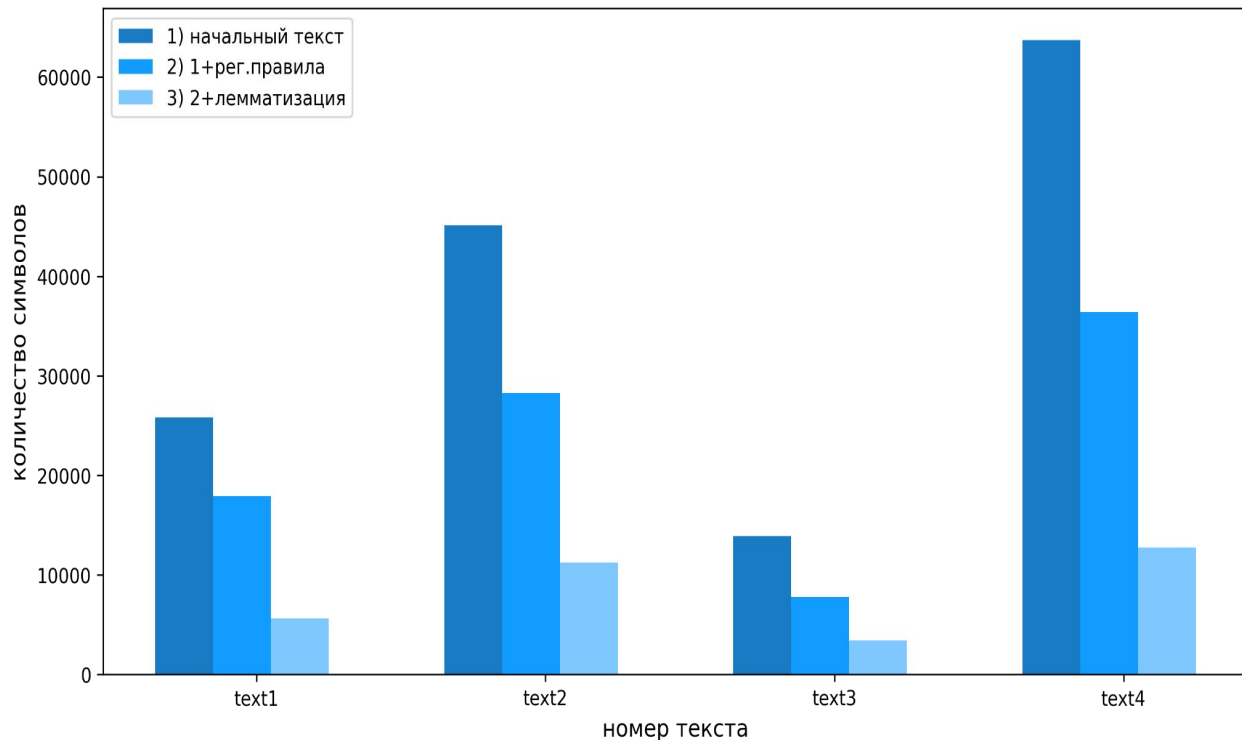
- 
- Logistic Regression (LR)
 - Decision Tree (DT)
 - Multinomial Naive Bayes (MNB)

Соответствующие модули взяты из библиотеки sklearn.


Эксперименты: данные и их обработка



1. Набор веб-страниц, 23 категории. Сделано несколько датасетов с разным количеством веб-страниц и количеством категорий.



Эксперименты: методы машинного обучения



test_size = 0.33 от общего количества документов в датасете, остальное – обучающая выборка.
Все разбиения автоматические.
Метрика f1_score(micro).

Метод / Обучающий датасет	DT	MNB	LR
dataset_150_10	0.34	0.56	0.63
dataset_200_15	0.31	0.52	0.61
dataset_75_15	0.32	0.62	0.67
dataset_350_4	0.61	0.77	0.82
dataset_200_23	0.27	0.50	0.58

Эксперименты: бинарная классификация



Выделение групп миноритарных
и мажоритарных классов.

Метод / Обучающий датасет	DT	MNB	LR
dataset_max_23	0.87	0.90	0.91


true	0	1412	11
	1	128	57
		0	1
		predicted	

Анализ результатов. Часть 1

- Таблица с наиболее значимыми словами для нескольких категорий. Такие слова выделены с помощью метода LR, показавший наилучший результат.


sysadm	dev	multimedia	databases	security
аналогично разделяться пространство вывести соединение флаг указывать опция сокет сетевой	среда компилировать указатель кроссплатформенный статический синтаксис программирование редактор компилятор отладчик	медиацентр изображение звук формат эффект графический рисование аудио медиа видео	бд распределенный хранилище небольшой восстановление база хранить запрос транзакция репликация	атаковать устранить затрагивать выявить выпустить безопасность обнаружить атака исследователь злоумышленник


Анализ результатов. Часть 2

- 
- Методы MNB и DT работают быстрее. Для DT очень много признаков, и на разреженных данных работает плохо.
 - Полученные результаты аналогичны результатам работ [1-3].

Метод / Обучающий датасет	DT	MNB	LR
полученные результаты	0.34	0.56	0.63
работа [1]	0.76	0.91	0.92
работа [2]		0.56	
работа [3]			0.6

Основные результаты

- 
1. Изучены методы машинного обучения для решения задачи категоризации текстов;
 2. Разработаны и реализованы методы сбора текстовых данных с веб-страниц и предварительной обработки текстов;
 3. Реализованы алгоритмы, использующие методы Logistic Regression, Decision Tree, Multinomial Naive Bayes;
 4. Проведен сравнительный анализ методов машинного обучения. По полученным результатам выполнено сравнение с результатами других аналогичных решений. И выбран наиболее подходящий метод, на основе которого будет разработана система категоризации веб-страниц.



Дальнейшее развитие

- Рассмотреть различные методы отбора признаков классификации и экспериментально проверить их на эффективность.
- Реализовать систему категоризации веб-страниц с графическим интерфейсом, которая включает классификацию с веб-страницами на английском языке.
- Рассмотреть модели для получения эмбеддингов и включить их в свой алгоритм классификации веб-страниц.



- [1] Patrick Dave P. Woogue, Gabriel Andrew A. Pineda. Automatic Web Page Categorization Using Machine Learning and Educational-Based Corpus // IJCTE, Vol. 9, No. 6, 2017, URL:
<http://www.ijcte.org/vol9/1180-IT026.pdf>
- [2] Tobias Eriksson. // Project. – 2013. URL:
<https://www.diva-portal.org/smash/get/diva2:700316/FULLTEXT01.pdf>
- [3] Результаты работы логистической регрессии на датасете 20newsgroups. URL:
https://scikit-learn.org/stable/auto_examples/linear_model/plot_sparse_logistic_regression_20newsgroups.html