Отчет по научной работе

Рысьмятова Анастасия

19.03.2017

1 Сентябрь-Октябрь

1. Выбрала данные.

2. Выкачала данные с сайта, используя язык Python и библиотеку scrapy.

2 Ноябрь

Пыталась обучать линейную регрессию на данных, использовала в качестве ответа возраст писателя. Текст преобразовывала с помощью TF IDF используя слова и ngram. Достичь хороших результатов не удалось.

3 Декабрь

Для формирования обучающей и тестовой выборки был опробован следующий метод:

Каждый объект представляет собой тексты двух произведений одного автора. Ответ для данного объекта 1, если текст первого произведения был написан раньше, чем текст второго, и 0, если иначе. При таком способе количество объектов возрастает до 2 млн.

Разбиение на обучение и контроль происходило следующими двумя способами:

1. Выбиралось 12 случайных авторов, и из их стихотворений формировалась тестовая выборка вышеописанным способом. Из оставшихся произведений формировалась обучающая выборка. Т. е. при таком способе в обучающую и тестовую выборку попадают произведения разных авторов.

2. Случайно выбиралось 800 произведений и из них формировалась тестовая выборка вышеописанным способом (Путем сопоставления каждого произведения одного автора с каждым). Из оставшихся произведений формировалась обучающая выборка.

При таком способе в обучающую и тестовую выборку могут попадать произведения одних и тех же авторов.

Традиционные методы

Для получения простого базового решения данной задачи тексты были переведены в матрицу признаков путем использования TFIDF и буквенных Ngram. Перед формированием матрицы признаков текст был предобработан: символы приведены к нижнему регистру, удалены знаки препинания, проведена лемматизация. Далее, используя полученную разряженную матрицу, решалась задача бинарной классификации. Качество оценивалось по метрике Accuracy. Тем самым мы оцениваем долю верно упорядоченных пар стихотворений внутри автора алгоритмом.

Ниже в таблице приведено качество базового решения задачи.

	TF IDF + Ngram из 4 букв	TF IDF + мешок слов
Способ №1	0.55	0.54
Способ №2	0.62	0.6

Нейронные сети

Следующий метод решения данной задачи использует сверточные нейронный сети с посимвольным подходом. Архитектура данной сети описана в статье [3]. Из каждого стихотворения было выбрано первые 507 символов, затем каждые два укороченных стихотворения одного автора объединялись в 1 текст длиной 1014 символов. В данном подходе использовался алфавит из символа перевода строки , цифр, русских букв и знаков препинания. Качество по метрике Accuracy пока что не удалось достичь лучше, чем в прошлом методе.

4 Январь

Метод формирования обучающей и тестовой выборки был прежним. Добавила следующие признаки:

- количество слов в тексте
- количество слов в заголовке
- количество точек, запятых, восклицательных и вопросительных знаков
- среднее количество слов в строке
- среднее количество точек, запятых, восклицательных и вопросительных знаков в строке

- количество существительных, прилагательных и глаголов в тексте

- среднее количество существительных, прилагательных и глаголов в строке

Использовала различные алгоритмы классификации из sklearn, но не удалось превзойти результаты базового решения.

5 Начало февраля

Изменила способ разбиения на обучающую и тестовую выборку. Теперь в обучение и в тестирование попадают тексты лишь одного из авторов. Т. е. для каждого автора задача решается отдельно, независимо от остальных. Объект представляет собой тексты двух произведений одного автора. Ответ для данного объекта 1, если текст первого произведения был написан раньше, чем текст второго, и 0, если иначе.

Выкинула из выборки тексты всех авторов у которых меньше 10 произведений.

Для получения простого базового решения данной задачи тексты были переведены в матрицу признаков путем использования TFIDF и буквенных Ngram. Перед формированием матрицы признаков текст был предобработан: символы приведены к нижнему регистру, удалены знаки препинания, проведена лемматизация. Далее, используя полученную разряженную матрицу, решалась задача бинарной классификации.

Пыталась использовать LSA, для снижения размерности и использовать различные алгоритмы классификации.

Ниже в таблице приведено качество базового решения задачи.

	Среднее Accuracy по всем авторам
log regression	0.69
lsa + log regression + RF	0.73

Пробовала обучать на тем же способом на придуманных признаках, но пока что не удалось достичь качества лучше, чем в TFIDF.

6 Конец февраля

Сформулировала и подала тезисы на конференцию Ломоносов. Экспериментировала с параметрами в задаче. Для каждого из авторов, имеющих в выборке более 100 стихотворений в таблице приведен результат работы полученного алгоритма.

Размер тестовой выборки	Accuracy	Автор
(1892, 50)	0.70	Александр Блок
(16002, 50)	0.63	Александр Пушкин
(420, 50)	0.45	Алексей Апухтин
(21170, 50)	0.701	Анна Ахматова
(462, 50)	0.755	Аполлон Майков
(2756, 50)	0.745	Афанасий Фет
(506, 50)	0.616	Булат Окуджава
(1722, 50)	0.687	Валерий Брюсов
(2550, 50)	0.605	Владимир Высоцкий
(420, 50)	0.652	Владислав Ходасевич
(1056, 50)	0.646	Илья Эренбург
(3906, 50)	0.754	Иосиф Бродский
(3540, 50)	0.722	Константин Бальмонт
(15006, 50)	0.73	Марина Цветаева
(420, 50)	0.633	Михаил Лермонтов
(2162, 50)	0.677	Николай Гумилев
(420, 50)	0.7190	Осип Мандельштам
(2652, 50)	0.719	Федор Сологуб
(2970, 50)	0.673	Федор Тютчев
(2162, 50)	0.544	Эдуард Асадов

7 Март

Искала статьи связанные с ранжированием текстов и с обработкой текстов литературных произведений.

Разобралась в статье [1]. В данной работе показано, как с помощью модели word2vec вычислять близость между текстами. Реализация метода описанного в статье выложена авторами на github [2].

8 Что буду делать дальше

- Искать новые статьи
- Использовать информацию полученную от остальных авторов
- Придумывать новые признаки
- Применять метод из статьи [1] к задаче

Список литературы

[1] From Word Embeddings To Document Distances http://jmlr.org/proceedings/papers/v37/kusnerb15.pdf

[2] Word Mover's Distance (WMD) from Matthew J Kusner https://github.com/mkusner/wmd