Московский государственный университет имени М. В. Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

КУРСОВАЯ РАБОТА СТУДЕНТА 517 ГРУППЫ

«Ранжирование текстов литературных произведений»

Содержание

Аннотация

1 Введение

Автоматическая обработка текстов становится все более востребована в связи с постоянно растущим объемом информации в Интернете и потребностью в ней ориентироваться. Ранжирование текстов — важная задача автоматической обработки текстов, исследованиями в которой активно занимаются все поисковые системы. Наиболее популярные поисковые системы используют методы машинного обучения для решения данной задачи.

В работе исследованы основные подходы к ранжированию текстов на примере задачи ранжирования литературных произведений. Для этого с сайта [2] выбраны тексты стихотворений различных русских поэтов и, используя методы машинного обучения, построен алгоритм способный ранжировать тексты одного автора в порядке возраста, в котором он написал произведения.

Для решении данной задачи ранжирования был применен попарный подход. Был построен бинарный классификатор, принимающий на вход пары стихотворений одного и того же автора, и определяющий, какое стихотворение было написано раньше. Качество классификации измерялось по метрике Accuracy.

В работе приведены результаты решения задачи ранжирования литературных произведений с использованием как нейросетевых методов машинного обучения [3,4], получивших особую популярность в последнее время, так и традиционных методов автоматической обработки текстов.

В работе показано, что возможно предположить возраст автора, в котором было написано стихотворение, если известна информация о других произведениях данного автора.

1.1 Постановка задачи ранжирования

Ранжирование решает следующую задачу. Имеется множество объектов, для конечного подмножества которых известен их правильный порядок. Это подмножество называется обучающей выборкой. Порядок остальных объектов не известен. Требуется построить алгоритм, способный упорядочивать произвольные объекты из исходного множества.

Т. е. ранжирование – это класс задач машинного обучения с учителем заключающихся в подборе ранжирующей модели по обучающей выборке. Ранжирующая модель должна наилучшим образом обобщить способ упорядочивания объектов в обучающей выборке на новые данные.

Задачу ранжирования можно формализовать следующим образом [1]: $X-$ множество объектов; $X^\ell = \{x_1, ..., x_\ell\}-$ обучающая выборка; $i \prec j-$ правильный порядок на парах $(i, j) \in \{1, ..., \ell\}^2$; необходимо построить ранжирующую функцию $a : X \to \mathbb{R}$ такую, что $i \prec j \Rightarrow a(x_i) < a(x_j)$.

В задаче ранжирования текстов объекты — это текстовые документы.

1.2 Методы ранжировния

Выделяют три основных подхода в задаче ранжирования:

1. Поточечный подход (pointwise approach). В поточечном подходе предполагается, что каждому объекту поставлена в соответствие численная оценка. Задача обучения ранжированию сводится к построению регрессии: для каждого отдельного объекта необходимо предсказать его оценку.

   В рамках этого подхода могут применяться многие алгоритмы машинного обучения для задач регрессии. Когда оценки могут принимать лишь несколько значений, также могут использоваться алгоритмы классификации.

2. Попарный подход (pairwise approach). В данном подходе обучение ранжированию сводится к построению бинарного классификатора, которому на вход поступают два объекта, соответствующих одному и тому же запросу, и требуется определить, какой из них должен стоять выше в упорядоченном списке.

3. Списочный подход (listwise approach). Списочный подход заключается в построении модели, на вход которой поступают сразу все объекты, а на выходе получается их перестановка.

2 Ранжирование текстов

В данном разделе будут описаны основные алгоритмы ранжирования текстов.

2.1 Ранжирование текстов литературных произведений

Для исследования задачи ранжирования текстов были выбраны тексты литературных произведений, так как это естественным образом сформировавшийся набор данных, упорядоченный во времени. Не удалось найти статей, где бы решалась задача упорядочивания текстов литературных произведений методами машинного обучения.

Возникает вопрос, существует ли закономерность между тем, о чем, и как пишет автор, и его возрастом. А главное, смогут ли алгоритмы машинного обучения найти такую закономерность.

В данной работе исследованы три различных задачи ранжирования :

- По имеющимся литературным произведениям с известным порядком, различных авторов , необходимо упорядочить тексты нового автора.
- По части текстов с известным порядком одного автора, необходимо восстановить порядок остальных произведений данного автора.
- По имеющимся литературным произведениям с известным порядком во времени различных авторов, а так же по части текстов одного автора с известным порядком , необходимо восстановить порядок остальных произведений данного автора.

3 Вычислительные эксперименты

3.1 Исходные данные

Для решения задачи ранжирования текстов литературных произведений использовались тексты стихотворений русских поэтов. Необходимые данные были выкачены с сайта [1], используя язык Python и библиотеку scrapy [2]. Из полученной выборки использовались лишь тексты содержащие не более 1014 символов. Полученные данные содержат 8871 стихотворений 112-ти русских поэтов.

На Рис. 1 и Рис. 2 показано распределение следующих признаков: год рождения авторов и возраст авторов, в котором было написано произведение. Видно, что данные содержат стихи поэтов начиная с 16 века.

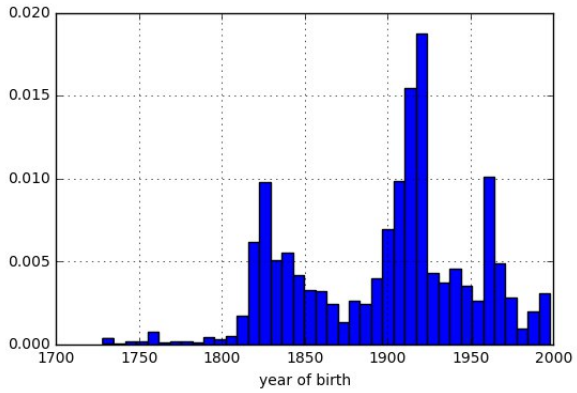

Рис. 1: Год рождения авторов

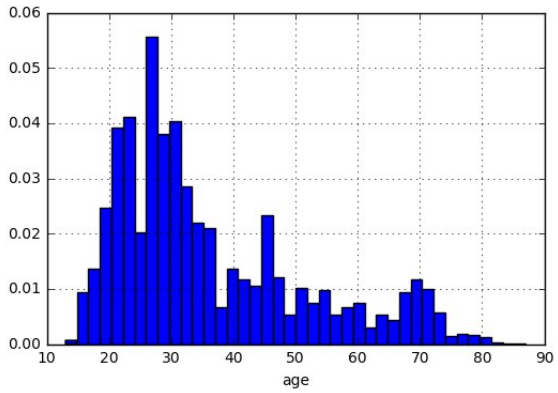

Рис. 2: Возраст написания произведения

3.2 Подходы к решению

Решая задачу ранжирования текстов, использовался попарный подход, в нем 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 подходе доля объектов класса 1 равна доле объектов класса 0.

Разбиение на обучение и контроль происходило следующими способами:

1. Выбиралось несколько 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 выбирались 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.

3. Для каждого автора в отдельности случайно выбиралось 20% его произведений и из них формировалась тестовая выборка. Из остальных 80% произведений данного автора формировалась обучающая выборка. В таком способе в обучающую и тестовую выборку попадают тесты одного поэта. Итоговое качество усреднялось по всем авторам.

3.3 Традиционные методы

Для получения простого базового качества 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 с помощью логистической регрессии. Качество оценивалось по метрике Accuracy. Тем самым мы оцениваем долю верно упорядоченных пар стихотворений внутри автора.

Другой способ решения данной задачи основан на эвристических признаках. В ходе работы удалось придумать 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

Используя полученную признаковую матрицу решалась задача бинарной классификации с помощью алгоритмов логистической регрессии и случайного леса.

Ниже в таблице приведено качество базового решения задачи, усредненное по трем запускам (это необходимо, так как выборка на обучение и контроль разбивалась случайным образом) для первого и второго способа разбиения на обучение и контроль.

	TFIDF+Ngram 4 букв	TFIDF+мешок слов	Эвристич. признаки
Способ №1	0.55	0.54	0.51
Способ №2	0.59	0.57	0.53

В способе №3 изменен принцип 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. В данном способе использовались тексты авторов, написавших более 100 произведений.

По мимо базового решения с использованием мешка слов и TFIDF, и решения, основанного на эвристических признаках, применялся метод LSA, для снижения размерности.

Ниже в таблице приведены качество работы полученных методов.

	Среднее Accuracy по всем авторам
TFIDF + мешок слов + log regression	0.68
lsa + log regression	0.68
lsa + log regressio + эвристич. признаки	0.69

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

3.4 Нейронные сети

Нейронные сети на данный момент активно применяются для задачи автоматической обработки текстов. Для данной задачи использовалась сверточная нейронная сеть 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.

Известно, что использование нейронной сети с архитектурой, описанной в статье, дает лучшее качество по сравнению с традиционными методами, если порядок числа объектов в выборке не менее 10^6. В способе 1 и 2 в обучающей выборке более 10^6 объектов. В способе 3 для каждого автора обучение происходит отдельно и количество объектов может быть от 10^2 до 10^4.

Качество по метрике Accuracy приведено в таблице ниже. Из таблицы видно, что данная нейронная сеть не показала результат лучший, чем традиционные методы ни для одного из способов.

	Accuracy
Способ №1	0.56
Способ №2	0.58
Способ №3	0.50

3.5 Выводы

4 Заключение