СТАТИСТИЧЕСКИЙ АНАЛИЗ ТЕКСТА И ЗАДАЧА ОПРЕДЕЛЕНИЯ АВТОРСТВА

Е. С. Жданович

Белорусский государственный университет, г. Минск;
jewgenija27@gmail.com;
науч. рук. – Е. М. Радыно, канд. физ.-мат. наук, доц.

В последнее время наблюдается тенденция поиска и определения структур, характерных для текстов, принадлежащих различным авторам. Для решения данной задачи применялись статистические, формально-количественные методы, позволяющие выявить характерные для авторов черты. Предыдущие исследования ставили перед собой одну цель: установление авторства неизвестных текстов, полагаясь на имеющуюся выборку авторов. Но ранее не исследовалась динамика точности классификации текстов при изменении длины классифицируемого текста. Для анализа данной динамики производился отбор признаков, частотный анализ текстов различной длины и обучение построенных моделей.

Ключевые слова: машинное обучение, статистический анализ текста, определение авторства, анализ данных, обработка естественного языка.

ПОСТАНОВКА ЗАДАЧИ И ЕЕ ДИФФЕРЕНЦИАЦИЯ

Имеется два множества: множество текстов (фрагменты произведений) и авторов (классов), которым они принадлежат. Необходимо установить авторство текстов, которым не присуждены метки о принадлежности конкретному автору. Также целью является изучение зависимости точности классификации текстов от длины классифицируемого текста.

Основными этапами в решении задач являются:

1. Подготовка текста к анализу.
2. Поиск и выделение статистических признаков, позволяющих охарактеризовать текст.
3. Получение векторного представления текста.
4. Визуальное представление полученных признаков.
5. Построение и обучение моделей, осуществляющих классификацию текстов.
6. Сравнение моделей, построенных на разных корпусах текстов, и определение пороговых значений длин текстов, при которых точность классификации изменяется незначительно, либо, наоборот, претерпевает спад или подъем точности классификации.

ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА И ВЕКТОРИЗАЦИЯ ТЕКСТА

Для обучения моделей необходима векторизация текста, т.е. каждому тексту ставится в соответствие вектор признаков, имеющий фиксированную длину. Сначала текст подвергается предварительной обработке, основными этапами которого являются:

- Токенизация – разбиение текста на отдельные единицы-токены;
- Лемматизация – приведение слов к начальной форме;
- Удаление «стоп-слов» – слов, которые встречаются в большом объеме текстов и не несут особой смысловой нагрузки;
- Представление текста в виде «мешка слов» (bag-of-words): единицы «мешка слов» – слова, каждое из которых имеет атрибут (частоту встреч данного слова в тексте);
- TF-IDF (TF – частота слова, IDF – обратная частота документа) метод векторизации – оценка важности слова в констексте документа, который является частью корпуса документов;
- N-граммы – наборы из n идущих подряд токенов. В качестве токенов могут выступать как слова, так и буквы.

Статистические признаки текста

В качестве признаков, позволяющих выявить особенности текста, были использованы следующие:

- Отношение количества прописных к количеству строчных букв;
- Распределение различных знаков препинания по тексту;
- Распределение длин предложений;
- Распределение длин слов;
- Водность текста – разность между единицей и отношением количества слов после очистки текста от «стоп-слов» к количеству слов в исходном тексте;
- Разнообразие речи – отношение количества уникальных слов к общему количеству слов в тексте;
- Распределение наиболее часто встречающихся частей речи;
- Частоты буквенных биграмм.

КЛАССИФИКАЦИЯ И СРАВНЕНИЕ МОДЕЛЕЙ

Генерировались выборки из текстовых фрагментов различной длины L. Ввиду того, что многие алгоритмы ожидают на входе центрированные признаки с центром в нуле и с одинаковым распределением, признаки подвергались соответствующей обработке.

Визуализация и методы понижения размерности

Для изображения полученных признаков в двух- либо трехмерном пространстве использовались методы понижения размерности: метод главных компонент и t-SNE (t-distributed stochastic neighbor embedding). Исходя из оригинальных меток классов, точкам присваивались цвета, каждая точка обозначает конкретный текст, а точнее вектор, который был получен после отображения вектора признаков в пространство меньшей размерности (рис. 1).

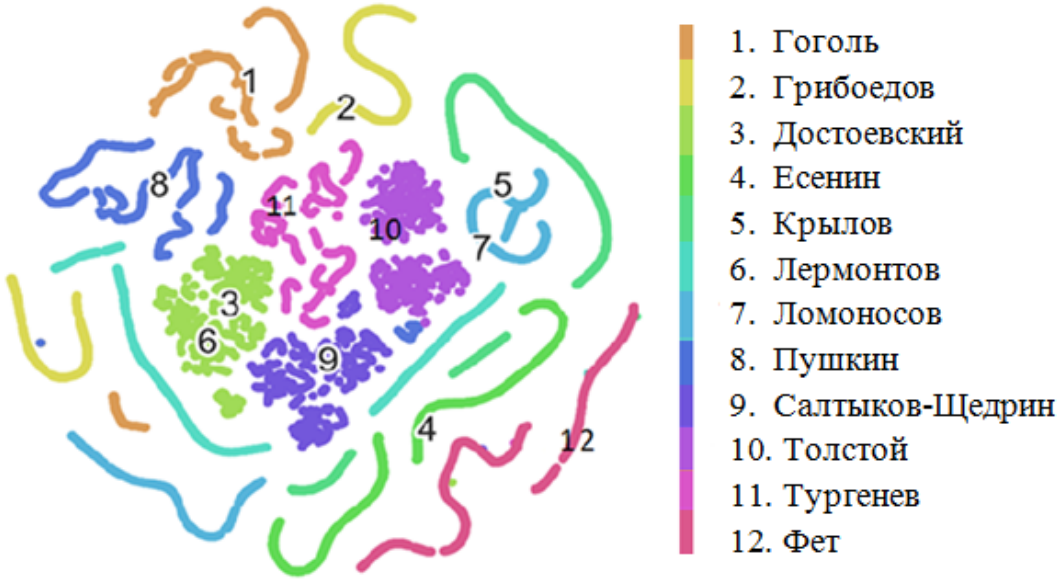

Рис 1. Алгоритм t-SNE для текстов длины 20000 символов тренировочного набора

Матрица ошибок и изучение динамики классификации

Для классификации использовалась следующие алгоритмы: логистическая регрессия, метод k-ближайших соседей, метод опорных векторов, стохастический градиентный спуск. Точность оценивалась отношением верно классифицированных объектов ко всем объектам, так как выборка предполагалась сбалансированной (табл.).

Таблица

Алгоритмы классификации текстов различной длины

	L = 20000	L = 10000	L = 5000	L = 1000	L = 500	L = 200
Logistic Regression	0.813	0.751	0.725	0.536	0.467	0.317
KNN	0.774	0.718	0.665	0.332	0.236	0.133
SVM	0.776	0.743	0.702	0.503	0.475	0.280
Tfidf + SGDClassifier	0.709	0.721	0.733	0.606	0.534	0.332
SGDClassifier	0.753	0.721	0.664	0.492	0.395	0.255

Матрица ошибок дает подробное представление о классификации объектов. Главная диагональ матрицы показывает количество верно

предсказанных значений для каждого из классов. Неверно предсказанные элементы будут располагаться вне главной диагонали (рис. 2).

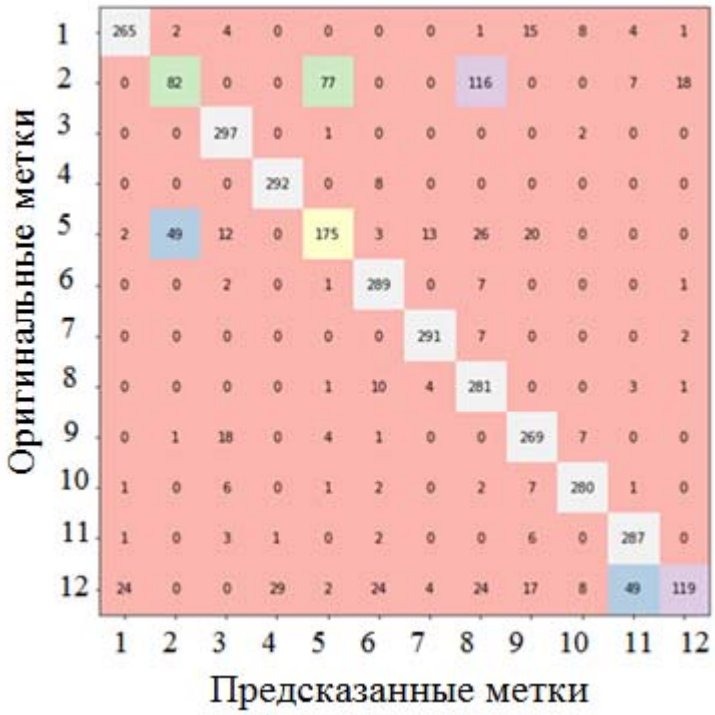

Рис 2. Матрица ошибок на текстах длиной 20000 символов:
1 – Гоголь, 2 – Грибоедов, 3 – Достоевский, 4 – Есенин, 5 – Крылов,
6 – Лермонтов, 7 – Ломоносов, 8 – Пушкин, 9 – Салтыков-Щедрин,
10 – Толстой, 11 – Тургенев, 12 – Фет.

Вывод: качество классификации значительно снижается при работе с текстами длины менее 5000 символов. При увеличении длины текстов с 5000 до 20000 символов качество повышается незначительно. Векторизация исходных текстов – достаточно трудоемкая задача. Поэтому при отсутствии достаточных ресурсов, можно не использовать тексты наибольшей длины, не ощущая при этом особых потерь качества.

Библиографические ссылки

1. *Морозов Н. А*. Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного известного автора. Стилеметрический этюд. // Известия отд. русского языка и словестности Имп.Акад.наук, Т.XX, кн.4, 1915.
2. *Марков А. А*. Об одном применении статистического метода. // Известия Имп.Акад.наук, серия VI, Т.X, N4, 1916.
3. *Марков А. А*. Пример статистического исследования над текстом «Евгения Онегина» иллюстрирующий связь испытаний в цепь. // Известия Имп.Акад.наук, серия VI, Т.X, N3, 1913, с.153.
4. *Хмелев Д. В*. Распознавание автора текста с использованием цепей А. А. Маркова // Вести. МГУ. Сер. 9. Филология. 2000. № 2.