Анализ служебных документов для оценки качества работы персонала

Е.С. Рыков[1], О.А. Невзорова[1]

[1]*Казанский (Приволжский) Федеральный Университет, Институт вычислительной математики и информационных технологий, Кремлевская 35, Казань, Россия, 420111*

Аннотация

В статье рассмотрена проблема анализа отчетных документов по критериям схожести и разработана система аналитики, в которой последовательно решаются задачи обработки большого массива текстов служебных документов. Система протестирована в экспериментальном режиме и полученные результаты согласуются с мнением экспертов об оценке близости выбранных групп отчетов.

Ключевые слова

Схожесть отчетов, сравнение строк, отчетные документы

1. Введение

Целью данной работы является разработка общей архитектуры и базового модуля, реализующего контроль уникальности служебных документов-отчётов сотрудников за определенный период. Новизна построенных решений заключается в применимости реализованного подхода к реальным неочищенным большим данным. В литературе имеется большое число статей, посвященных описанию различных методов кластеризации для выявления сходства в текстах, а также методам очистке текстовых данных [1]. Для решения поставленной задачи был разработан конвейер преобразований исходных текстов (рис. 1). Для реализации проекта выбран язык Python, включающий в себя набор мощных научных и вычислительных библиотек.

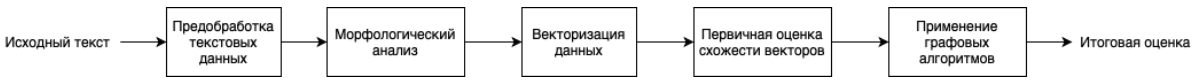

Рисунок 1: Конвейер преобразования текстовых данных

Для предварительной очистки данных был разработан модуль, который устраняет типовые ошибки. Следующий этап обработки данных – этап морфологического анализа, на котором использовался морфологический анализатор pymorphy2. Далее следуют этапы векторизации текстов отчётов и сравнения полученных векторов, а затем применение графовых алгоритмов для оценки групп похожих документов. Для формального сравнения строк в качестве метрики была выбрана косинусная мера. Ввиду того, что косинусная мера сравнивает всего два вектора, был разработан алгоритм на основе графов, позволяющий сравнивать множество строк. Суть алгоритма заключается в попарном сравнении всех векторов и построении матрицы смежности A, в которой i и j – номера векторов. Вычисление значения для ячейки i j представлено в формуле (1).

$$A_{ij} = \begin{cases} 1, sim(i,j) \geq threshold \\ 0, sim(i,j) < threshold \end{cases} \quad (1)$$

Результирующий параметр score рассчитывается по формуле (2), где k(G) – число компонент связности графа G, а n – максимальное значение компонент связности графа G:

$$score = 1 - \frac{k(G)}{n} \quad (2)$$

Эксперименты проводились на группе отчетов одного сотрудника, поданных за три недели (143 отчета). Первым экспериментом были проведены замеры результирующего параметра

score для разных пороговых параметров threshold. Результаты исследования представлены в Таблице 1, оптимальным параметром порога выбран 0.4.

Таблица 1

Зависимость параметра Score от Threshold

Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Score	0.95	0.91	0.77	0.63	0.5	0.45	0.41	0.39	0.36

В качестве второго эксперимента массив отчётов одного сотрудника был поделён на три группы. В первую группу вошли отчёты, начинающиеся со слова *совещание*, во вторую группу – идентичные отчёты, а в третью группу – все оставшиеся. Значения параметра Score для группы каждого типа приведены в Таблице 2. Показано, что среднее и медианное значение score для трёх разных групп выше, чем значение оценки score всех отчётов целиком.

Таблица 2

Параметр Score для разных групп отчётов

Группа отчётов	Score
Все отчёты	0.63
Первая группа	0.74
Вторая группа	0.97
Третья группа	0.34

Проведенное исследование показало, что если при группировке выделять отчеты по некоторым ключевым словам, например, *совещание, обсуждение* и др., то анализ схожести внутри группы можно проводить более точно с учетом семантических параметров, например, именований тем совещаний внутри соответствующей группы. Отдельный сравнительный эксперимент был проведен для группы отчётов, начинающихся со слова *совещание*. Результаты с учётом и без учёта семантического фактора приведены в Таблице 3.

Таблица 3

Параметр Score с учётом и без учёта семантического фактора

Группа отчётов	Score
Без учёта семантики	0.74
С учётом семантики	0.51

Модуль обрабатывает входные данные: удаляет пунктуацию и стопслова, приводит всё в нижний регистр. Третьим экспериментом был замерен итоговый параметр схожести отчётов без предобработки строк: был получен score равный 0.70. При проведении предобработки текстовых данных параметр был равен 0.63.

2. Заключение

В статье представлены основные результаты по анализу служебных документов-отчетов по критериям схожести. Дальнейшие исследования будут направлены на включение в анализ различных семантических элементов для выделения семантических составляющих отчетов. Кроме того, предлагается расширить возможности системы по предобработке текстов, включив модуль исправления ошибок и другой функционал обработки текстов.

3. Литература

[1] Неелова, Н.В. Исследование лексического метода вычисления схожести строк с учетом предварительной обработки / Н.В. Неелова // Известия ТулГУ. Технические науки. – 2009. – № 2(2). – С. 202-212.

[2] Кормен, Т.М. Алгоритмы: построение и анализ / Т.М. Кормен, Ч. Лейзерсон, Р. Ривест, К. Штайн. – М: И. Д. Вильямс, 2013. – 1328 с.

[3] Бенгфорт, Б. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка / Б. Бенгфорт, Р. Билбро, Т. Охеда. – СПб: Питер, 2019. – 368 с.