МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

АНАЛИЗ ТЕКСТОВЫХ ДАННЫХ НА ОСНОВЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ И NLP

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Котумы Андрея Александровича

Научный руководитель:
зав. кафедрой, к.ф.-м.н., доцент _____ М.В. Огнева

подпись, дата

Зав. кафедрой:
к.ф.-м.н., доцент _____ М.В. Огнева

подпись, дата

Саратов 2022

ВВЕДЕНИЕ

Актуальность темы.

Тексты на естественном языке — это одна из наиболее распространенных форм хранения и распознавания информации, которая является также одной из самых понятных для человека. В последние годы накопилось огромное количество текстовой информации. Например, сообщения, комментарии, отзывы на фильмы, музыку, товары и услуги. Поэтому для владельцев компаний очень выгодно проводить анализ отзывов о себе или о товарах для дальнейшего улучшения своей работы, выстраивания линии развития и снижения убытков. Сейчас довольно популярны системы, которые анализируют текст на то, является ли он положительным или отрицательным по тональности, что полезно для анализа комментариев пользователей или подписчиков. Также данная тематика является актуальной для актеров и политиков для более быстрого и полного анализа мнений людей о них. Еще такой вид анализа полезен во время проведения социологических исследований, так как помогает ускорить и упростить обработку данных.

Natural Language Processing (NLP) - подраздел информатики и AI, посвященный тому, как компьютеры анализируют естественные языки.

Одной из областей применения NLP и машинного обучения в настоящее время является анализ кулинарных рецептов. Во-первых, существует огромное количество рецептов на различных сайтах, а также накопленных конкретными людьми и заведениями общественного питания. Следовательно, актуальной является проблема классификации и эффективного поиска по таким данным. Во-вторых, популярным становится направление автоматического анализа вкусов людей и составления рецептов на основе такого анализа. Данное направление ресторанного бизнеса имеет большую перспективу развития в связи с модностью и эффективностью цифровых технологий. Именно от модности блюд или концепта ресторана часто зависит его популярность, а, соответственно, прибыльность. Наконец, опыт анализа рецептов может быть полезен при работе с текстовыми данными другого содержания. Например,

классификация научных и научно-популярных текстов также является актуальной задачей и может решаться подобными методами.

Наиболее очевидно при анализе рецептов анализировать список ингредиентов. В данной работе для анализа будет использоваться еще и описание рецепта (или пошаговое выполнение), а также количество ингредиентов. Анализ будет выполняться с помощью методов машинного обучения (методы кластеризации и классификации) с предварительной обработкой текстовых данных при помощи методов обработки естественных языков.

Цель бакалаврской работы – изучить и реализовать алгоритмы кластеризации и классификации кулинарных рецептов, написанных на естественном языке, и изучить способы обработки такого вида текстов, а также сравнить результаты различных способов классификации и кластеризации.

Поставленная цель определила **следующие задачи**:

1. Изучить методы обработки и анализа текста на естественном языке;
2. Определить основные понятия NLP, необходимые для анализа кулинарных рецептов;
3. Подобрать методы NLP для обработки текста;
4. Изучить способы классификации и кластеризации текстов на естественном языке;
5. Изучить алгоритм LDA и другие алгоритмы для кластеризации набора данных;
6. Изучить алгоритмы классификации набора данных;
7. Реализовать модели кластеризации и классификации кулинарных рецептов;
8. Выполнить сравнительный анализ алгоритмов и сделать выводы о них;
9. Сделать выводы по методам и анализ результатов.

Методологические основы машинного обучения, тематического моделирования и NLP представлены в работах Д. Блея, К. Воронцова, Д.

Ньюмана, М. Дамашека, А. Бугаева, А. Гуренко, М. Нечепоренко, Д. Льюиса, С. Митрофанова, Р. Каруаны.

Структура и объём работы. Бакалаврская работа состоит из введения, 5 разделов, заключения, списка использованных источников и 2 приложений. Общий объем работы – 61 страница, из них 46 страниц – основное содержание, включая 14 рисунков и 2 таблицы, список использованных источников информации – 51 наименование.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Фундаментальные основы тематического моделирования» посвящен анализу фундаментальных основ тематического моделирования.

В данном разделе рассмотрены основные понятия тематического моделирования, описаны такие базовые тематические модели как PLSA и LDA. Здесь приведены основные области применения моделей с латентными переменными. Были изучены такие понятия как интерпретируемость тем и когерентность, а также приведена схема оценки интерпретируемости тем.

Второй раздел «Способы подготовки данных» посвящен основным подходам к подготовке данных перед использованием методов машинного обучения.

В данном разделе были изучены классические подходы к преобразованию документов в вектор: Bag of Words, Bag of Ngrams и TF-IDF. Еще здесь были рассмотрены основные способы работы с текстовыми данными (NLP): удаление «стоп-слов», токенизация и стемминг.

Третий раздел «Алгоритмы классификации» посвящен теоретическим основам различных методов классификации текстов.

В данном разделе были изучены описания, основные формулы, а также преимущества и недостатки популярных алгоритмов классификации:

- Наивный метод Байеса — алгоритм, основанный на теореме Байеса с допущением о независимости признаков.

- Метод k-ближайших соседей — метрический метод классификации, для работы которого достаточно определить функцию расстояния между документами и иметь некоторый набор классифицированных заранее документов.
- Деревья решений — алгоритм, принципом которого является построение разрешающего дерева на обучающей выборке документов.
- Метод опорных векторов — алгоритм, основой которого является построение оптимальной разделяющей гиперплоскости в пространстве векторов.
- Случайные леса — ансамблевый метод машинного обучения, представляющий собой ансамбль деревьев решений.
- Логистическая регрессия — метод построения линейного классификатора, позволяющий оценивать апостериорные вероятности принадлежности объектов классам.

Четвертый раздел «Алгоритмы тематического моделирования» посвящен теоретическим основам алгоритмов вероятностного тематического моделирования.

В данном разделе изучено описание задачи тематического моделирования, а также даны определения моделей скрытого размещения Дирихле (LDA) и иерархического скрытого размещения Дирихле (hLDA).

Пятый раздел «Программная реализация моделей машинного обучения» посвящен описанию программной реализации методов NLP, алгоритмов классификации и алгоритма LDA.

В качестве исходных данных для практической части работы был выбран датасет из 500 тысяч кулинарных рецептов с такими данными как название, описание, список ингредиентов, шаги приготовления и теги. Целью практической части является определение темы или метки каждого рецепта.

Прежде чем применить методы машинного обучения к текстовым данным, необходимо провести предварительную обработку. В первую очередь, при помощи регулярных выражений (формального языка поиска и

осуществления манипуляций с подстроками в тексте, основанного на использовании метасимволов) из текста удаляются знаки препинания, пробелы и символы переноса строки. Затем используются стемминг и токенизация, а также очистка строк датасета от стоп-слов. С целью того, чтобы увеличить значимость рядом идущих слов в список слов каждого рецепта были добавлены биграммы. В результате применения стемминга, токенизации и добавления биграмм каждый рецепт теперь представляет собой список слов, за исключением стоп-слов. Далее необходимо преобразовать список слов в формат мешка слов. А также для алгоритма LDA необходимо создать словарь всех слов, при этом имеет смысл провести фильтрацию, удалив из рассмотрения слова, встречающиеся в большом количестве документов и слова, встречающиеся в малом количестве документов.

Для классификации данных не создаются словарь и корпус, а используется матрица признаков. Для превращения списков слов в разреженную матрицу признаков используется объект CountVectorizer (преобразует коллекцию текстовых документов в матрицу количества токенов) библиотеки scikit-learn. В результате применения данного объекта и фильтрации слишком редких слов было выделено 8295 уникальных признаков и соответственно матрица имеет высокую размерность. В исходных данных поисковые теги представляют собой список меток, из которого для однозначной классификации следует оставить только одну метку. Для этого были выбраны 40 самых популярных тегов, при этом из них были удалены такие теги как «low», «sodium», «carb», «calorie», «fat». Они наиболее часто встречаются и не передают особой информации о том, к чему относятся блюда, а только описываются соблюденные ограничения при приготовлении. Затем для каждого рецепта выбиралась первая совпавшая метка из обратного списка топ-40 тегов. И если ни один тег не совпадал, то присваивался тег «healthy». Обратный список популярных тегов используется с целью балансировки количества рецептов в классах, так как в случае прямого списка около 65% блюд будут отнесены лишь к 2 тегам – «dinner» и «dessert».

Для оценки модели LDA и оптимальности количества тем в данной работе используется когерентность. Принято считать, что чем больше значение когерентности, тем более оптимальным было выбрано число тем для модели. Для нашего датасета с рецептами оптимальным количеством тем (с ограничением, что количество тем не более 50) оказалось 40 тем. Когерентность для данного количества тем является максимальной при этом ограничении. Как оказалось, тремя наиболее релевантными темами после обучения алгоритма LDA на данном наборе данных оказались темы, ключевые слова которых описывают рецепты ванильной выпечки, итальянской пасты и шоколадных десертов.

По результатам обучения модели LDA можно сделать следующие выводы:

- Данный метод является хорошо интерпретируемым.
- Используя оценку когерентности, можно найти оптимальное количество тем для датасета.
- Темы, определяемые алгоритмом LDA, пересекаются лишь в малой мере между собой.
- Данный алгоритм трудно автоматизировать для определения темы, а не набора ключевых слов.

При обучении моделей классификации (перечисленных в разделе 3) измерялись их точность и время обучения, а также была построена матрица спутанности. По результатам анализа матрицы спутанности можно сделать выводы, что основная часть предсказаний модели являются верными тегами, а остальные предсказания относительно равномерно распределены между другими тегами. Важным фактором является время обучения модели, так как в реальных условиях она должна дообучаться при добавлении определенного количества новых рецептов. Ниже представлена таблица со значениями точности и временем обучения и предсказания использованных в работе классификаторов.

Таблица 1 - Результаты классификаторов

Классификатор	Значение точности	Время обучения
MultinomialNB	0.5198	3 секунды
LogisticRegression	**0.7681**	**6 минут (362 секунды)**
KNeighborsClassifier	0.5642	82 минуты (4911 секунд)
SVC	-	Более 13 часов
DecisionTreeClassifier	0.7150	46 минут (2781 секунда)
RandomForestClassifier	0.6887	45 минут (2668 секунд)

Данные вычисления проводились на виртуальной машине со следующими характеристиками: 12.6 гигабайт ОЗУ и 2 ядра.

Для сравнения с моделями классификации время обучения модели LDA составило 1950 секунды с учетом подсчета когерентности или 453 секунды без подсчета когерентности. По результатам обучения различных видов классификации можно сделать следующие выводы:

- Самым быстрым и точным классификатором на разреженных матрицах больших размеров оказалась логистическая регрессия.
- Немного меньшую точность показали деревья решений и случайные леса — их время обучения приблизительно равно, а точность отдельного дерева решений оказалась даже выше.
- Плохую точность (около 50%) показали наивный байесовский классификатор и классификатор k-ближайших соседей. Причем первая модель оказалась быстрее более чем в 1500 раз, имея при этом точность всего на 5% ниже.
- Классификатор опорных векторов проигрывает по времени более чем в 4 раза самому медленному из остальных алгоритмов. Этот недостаток делает его плохо применимым для реальных задач, потому что в них часто необходимо дообучать модель. При этом допустимым временем дообучения модели можно считать максимум 12 часов во время проведения технических работ на сайте, например.

ЗАКЛЮЧЕНИЕ

Целью данной работы было изучение и реализация алгоритмов тематического моделирования и классификации кулинарных рецептов, написанных на естественном языке, а также изучение способов обработки текстов такого вида. Для ее достижения были изучены и применены алгоритмы NLP, тематического моделирования и классификации данных.

В главе 1 были изучены фундаментальные основы тематического моделирования. В ней приводились выводы ряда исследований и публикаций по данной теме.

В главе 2 были рассмотрены способы подготовки данных. В ее пунктах и подпунктах были подробно разобраны методы NLP для очистки документов и приведения слов к начальной форме (удаление стоп-слов, стемминг и токенизация). Также в данной главе были изучены способы векторного представления документа, такие как Bag of Words, Bag of Ngrams и TF-IDF, и причины необходимости представления документа в виде вектора его слов.

В главе 3 были подробно разобраны алгоритмы классификации, их преимущества и недостатки. В данной работе были изучены и применены следующие модели машинного обучения:

- наивный метод Байеса,
- метод k-ближайших соседей,
- деревья решений,
- метод опорных векторов,
- случайные леса,
- логистическая регрессия.

В главе 4 были подробно рассмотрены алгоритмы тематического моделирования, такие как:

- скрытое размещение Дирихле,

- иерархическое скрытое размещение Дирихле (основанное на вложенном процессе китайского ресторана).

В главе 5 было приведено описание программной реализации изученных алгоритмов классификации и алгоритма LDA. Для их написания был выбран язык программирования Python, так как он очень удобен для работы с большим объемом данных, а также к нему уже написаны библиотеки для работы с моделями машинного обучения, тематического моделирования и обработки текстов. Для анализа работы алгоритмов использовался датасет с 500 тысячами кулинарных рецептов. В этой главе были приведены методы очистки текста от стоп-слов, стемминга и токенизации, а также обоснование и особенности их практического применения. Также здесь были разобраны применение алгоритма LDA, способ его оптимизации, описание подбора оптимального количества тем, а также был проведен подробный анализ полученных тем и их интерпретация. Еще в данной главе была рассмотрена практическая реализация алгоритмов классификации. После чего был проведен сравнительный анализ их точности и времени выполнения.

По итогам выполнения практической части данной работы были сделаны следующие выводы:

- Методы NLP хорошо помогают с подготовкой текстовых данных для дальнейшего использования алгоритмов классификации и тематического моделирования.
- Модели тематического моделирования удобны для интерпретации результатов и нахождения оптимального количества тем для набора документов.
- Алгоритм LDA нужно усовершенствовать для автоматизации определения тем текстов на реальных проектах.
- Для применения алгоритмов тематического моделирования необходим дополнительный анализ результатов человеком.

- Среди классификаторов оптимальном выбором для применения на большом объеме текстовых документов оказался классификатор логистической регрессии, показав самый точный результат и обучившись всего за 6 минут на 500000 документов.
- Хорошие результаты показали деревья решений и случайные леса, показав высокую точность и среднее время обучения.
- Классификатор опорных векторов не применим для реальных задач с частым обновлением набора документов, так как имеет крайне высокое время обучения и не годится для частого дообучения.
- Оценку когерентности в методах тематического моделирования труднее интерпретировать, чем оценку точности в алгоритме классификации.
- Алгоритмы тематического моделирования дают хорошо интерпретируемые результаты, но при этом алгоритмы классификации проще автоматизировать и применить в реальной задаче.
- Время обучения модели LDA лучше, чем у большинства моделей классификации.

Изучение алгоритмов NLP перспективно в настоящее время, так как количество текстовой информации на просторах сети интернет с годами растет экспоненциально. Практическая значимость исследования заключается в том, что была изучена эффективность применения алгоритмов классификации и тематического моделирования на большом объеме данных и были получены выводы об их применимости в реальных задачах. В заключении отметим, что использование методов машинного обучения с каждым годом становится всё более эффективным и распространенным.

Основные источники информации:

1. Воронцов, К.В. Вероятностное тематическое моделирование [Электронный ресурс] / К.В. Воронцов // MachineLearning.ru. – Режим доступа: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf (Дата обращения: 02.02.2022)

2. David M. Blei. «Probabilistic topic models». В: Commun. ACM 55.4 (2012), с. 77–84.
3. David Newman и др. «Automatic Evaluation of Topic Coherence». В: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. HLT ’10. Los Angeles, California: Association for Computational Linguistics, 2010, с. 100–108. ISBN: 1-932432-65-5.
4. Damashek, M. Gauging similarity with n-grams: Language-independent categorization of text / Marc Damashek // Science, New Series. – 1995.
5. Бугаев А.А., Гуренко А.В., Нечепоренко М.А. Методы обработки естественного языка // ИННОВАЦИОННЫЕ НАУЧНЫЕ ИССЛЕДОВАНИЯ: ТЕОРИЯ, МЕТОДОЛОГИЯ, ПРАКТИКА. Материалы Международной (заочной) научно-практической конференции. под общей редакцией А.И. Вострецова. 2019. С. 9–18.
6. Lewis, David D, 1998, Naive (Bayes) at forty: The independence assumption in information retrieval. In Proe, of the European Conference on Machine Learning (ECML) p, 4–15.
7. Mitrofanov S.A., Semenkin E.S. Tree Retraining in the Decision Tree Learning Algorithm // IOP Conference Series: Materials Science and Engineering. Krasnoyarsk Science and Technology City Hall., Krasnoyarsk, Russian Federation, 2021. С. 12082.
8. Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In Machine learning, proceedings of the twenty-third international conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25–29, 2006, pp. 161–168.