ГЛАВА 2

Решение задачи

2.1 Источники и структура данных

Исходными данными в задаче выступают множество статей в формате pdf, источником данных выступил Общероссийский портал Math-Net.Ru содержащий архивы статей математических изданий. Были рассмотрены математические журналы, приведенные в таблице 2.1:

Таблица 2.1 – Журналы — источники статей

Название журнала	**Код в Math-Net**	**Число статей**
Сибирский математический журнал	smj	998
Алгебра и логика	al	928
Математический сборник	sm	951
Дифференциальные уравнения	de	993

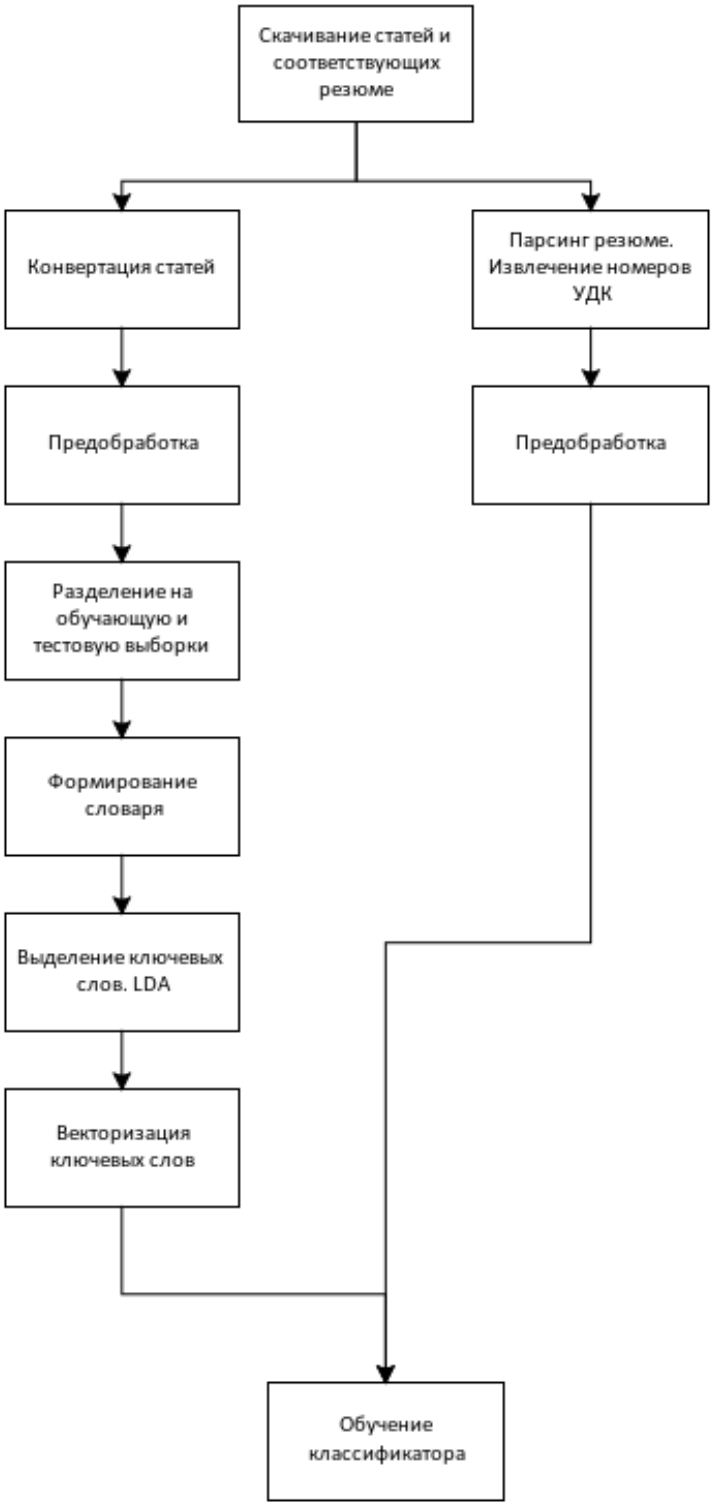
Каждой статье соответствует номер по универсальной десятичной классификации (УДК), определяющий основное направление, освещенное в работе.

На выходе необходимо получить классификатор, обученный на входящих данных, который будет способен с достаточной точностью определить номер УДК, соответствующий новой статье.

2.2 Структура модели

На рисунке 2.1 приведена структурная схема модели:

Рисунок 2.1 – Структурная схема модели



2.3 Алгоритм обучения

Обучение модели будем проводить в следующем порядке:

1. Сбор данных.
2. Подготовка данных.
3. Формирование обучающей и тестовой выборки.
4. Формирование словаря и обучение LDA-модели.
5. Сопоставление тексту списка ключевых слов.
6. Векторизация списка ключевых слов, получение векторного представления текста.
7. Обучение классификатора.

2.3.1 Сбор данных

Для организации сбора данных было решено воспользоваться API портала MathNet.ru. Информация об API этого портала не опубликована, поэтому опытным путем была получена следующая ключи:

— `jrnid=` — код журнала.
— `paperid=` — номер статьи.
— `wshow=paper` — информация о статье.
— `what=full` — Выгрузить полный текст статьи.
— `option_lang=rus` — язык страницы — русский.

Отметим, что ключи `wshow` и `what` одновременно не используются.

Для реализации на языке Python был использован модуль `requests` [16], позволяющий отправлять запросы HTTP/1.1. Процедура загрузки статьи приведена в листинге 2.1.

Листинг 2.1 – Процедура получения статьи

```
import requests
def get_paper(paper_id, journal_code):
url_full_text = 'http://www.mathnet.ru/php/getFT.phtml' +
 f'?jrnid={journal_code}&paperid={paper_id}&what=fullt&option_lang=rus'
url_summary = 'http://www.mathnet.ru/php/getFT.phtml' +
 f'?wshow=paper&jrnid={journal_code}&paperid={paper_id}&option_lang=rus'
```

```
filename_pdf = f'{journal_code} – {paper_id}.pdf'
filename_summary = f'{journal_code} – {paper_id}–resume.html'
response_text = requests.get(url_full_text)
response_summary = requests.get(url_summary)
f = open(filename_pdf, 'w+b')
f.write(response_text.content)
f.close()
f = open(filename_summary, 'w+b')
f.write(response_summary.content)
f.close()
```

В нем `paper_id` — номер статьи, `journal_code` — код журнала, приведенный в таблице 2.1.

Собранные статьи представлены в виде pdf-файлов, поэтому необходимо выделить из них текстовый слой. Его будем сохранять в формате простого текста `txt`.

Конвертация производилась средствами библиотеки aspose.words, предназначенной для обработки текстовых данных и преобразования их из одного формата в другой. Далее, из сопутствующих статьям файлов с краткой информацией, посредством использования библиотеки beautifulsoup были извлечены номера УДК.

2.3.2 Подготовка данных

При помощи регулярных выражений из текстов были удалены математические символы, латинские и греческие буквы, а также пунктуация. Был сформирован датафрейм, содержащий тексты статей и их номера УДК, и произведено сохранение на диске резервной копии в формате `csv`.

Из наиболее часто встречаемых во всех статьях, а оттого малоинформативных, слов был сформирован список стоп-слов. К полученному списку были добавлены стоп-слова из библиотеки nltk. На следующем этапе все стоп-слова были удалены из текстов, а также произведен стемминг с помощью SnowballStemmer из библиотеки nltk.

Был выбран стемминг, по той причине, что из-за математической специа-

лизации текстов лемматизация производилась недостаточно корректно, однако недостатком этого способа предобработки текста является невозможность восстановления исходного слова после подвержения его процессу стемминга.

2.3.3 Формирование обучающей и тестовой выборки

При помощи train test split из библиотеки sklearn было произведено разделение подготовленных данных на тестовую и обучающую выборки, где доля тестовой составила 0.14 от всего объема данных.

2.3.4 Формирование словаря и обучение LDA-модели

Далее были сформированы биграммы и триграммы слов средствами библиотеки gensim. Полученные данные были добавлены в словарь к односложным токенам. К сформированным таким образом для каждого текста словарям была применена модель LDA также из библиотеки gensim. Для каждого текста было выделено по 4 наиболее преобладающих темы. Данное число тем было выбрано по причине того, что дальнейшие темы были малоинформативны за счет небольшого количества терминов, кластеризованных в них.

Таким образом произведено выделение ключевых слов и словосочетаний для каждого текста.

2.3.5 Сопоставление тексту списка ключевых слов

Из полученного путём применения модели LDA списка кортежей с ключевыми слова соответственно с их частотами, был выделен список, содержащий лишь ключевые слова для каждого текста.

Этот список был объединен во фрейм данных с соответствующими текстам номерами УДК, который в дальнейшем был сохранены в качестве резервной копии на диске в формате `csv`. Пример выделения ключевых слов представлен на рисунке На рисунке 2.2 приведен пример выделения ключевых слов из статьи с номером УДК: 512, то есть её темой является «Алгебра».

Рисунок 2.2 – Пример выделения ключевых слов

"['групп', 'произведен', 'секц', 'сибирск', 'ран', 'гов', 'множ', 'положительн', 'последн', 'люб', 'нор', 'мальн', 'общност', 'силовск', 'следовательн', 'полож', 'ул', 'соответствен', 'котор', 'подгрупп', 'счита', 'фонд', 'рассужден', 'замен', 'номер', 'постро', 'содерж', 'редколлег', 'групп', 'нормальн', 'порядк', 'изоморфн', 'доказательств', 'конечн', 'теор', 'цеп', 'нов', 'одн', 'сил', 'секц', 'либ', 'определ', 'групп', 'нормальн', 'циклическ', 'явля', 'секц', 'произведен', 'существ', 'вопрос', 'чисел', 'аабутурлакин', 'стар', 'положительн', 'хухр', 'иском']"

2.3.6 Векторизация списка ключевых слов, получение векторного представления текста

Для произведения классификации текстовые данные необходимо привести к числовому виду, то есть векторизовать. Наиболее предпочтительным способом векторизации в данной задаче является векторизация посредством преобразования набора ключевых слов в матрицу функций TF-IDF. Данное значение вычисляется для каждого ключевого слова в каждом тексте. Его основная идея функции TF-IDF состоит в том, чтобы больший вес получают слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах. [9] Итоговый вес термина в документе относительно всей коллекции документов вычисляется по формуле:

$V_{t,d} = TF \cdot IDF$

TF - оценка важности слова t в пределах одного документа d.

$TF = \frac{n_{t,d}}{n_d}$,

где $n_{t,d}$ - количество употреблений слова t в документе d, n_d - общее число слов в документе d

IDF - инверсия частоты, с которой слово встречается в документах коллекции.

$IDF = log(\frac{|D|}{D_t})$,

где $|D|$ – общее количество документов в коллекции, D_t – количество всех документов, в которых встречается слово t

2.3.7 Обучение классификатора

Был обучен классификатор случайный лес: RandomForestClassifier из библиотеки sklearn, количество деревьев ограничено 100 штуками, так как после этого значения качество на тестовой выборке выходит на асимптоту.

2.4 Описание результатов и их анализ

На выходе была получена классификация тестовой выборки в размере 510 статей. Точность классификации была проверена при помощи матрицы ошибок(confusion matrix) и оценки точности(accuracy score). В таблице 2.2 представлен отчет по произведенной классификации.

Из 510 статей в верные классы были отнесены 324 штуки, что составляет 63%. 8 статей, обладающих устаревшими номерами УДК, были классифицированы в те классы, которые соответствуют их номерам в обновленной классификации УДК. Наиболее популярной тематикой является «Анализ», соответствующая коду УДК 517.

Классификация производится со значением точности(accuracy) = 0.63, большая часть выборки классифицируется в верные классы, однако, у модели ещё остается большой потенциал к усовершенствованию качества производимой классификации.

Значения presicion/recall для тестовой выборки представлены на рисунке 2.3

Рисунок 2.3 – Значения параметров для тестовой выборки

	precision	recall	f1-score	support
510	0.58	0.13	0.22	52
512	0.60	0.59	0.59	129
513.7	0.00	0.00	0.00	2
513.74	0.00	0.00	0.00	1
513.8	0.00	0.00	0.00	1
513.83	0.00	0.00	0.00	1
513.88	0.00	0.00	0.00	1
513.881	0.00	0.00	0.00	2
514	0.00	0.00	0.00	18
515.1	0.00	0.00	0.00	6
517	0.65	0.92	0.76	261
518	0.00	0.00	0.00	1
518.321.1	0.00	0.00	0.00	1
518.322.2	0.00	0.00	0.00	1
519.1	0.00	0.00	0.00	4
519.2	0.00	0.00	0.00	9
519.3	0.00	0.00	0.00	1
519.44	0.00	0.00	0.00	1
519.45	0.00	0.00	0.00	1
519.48	0.00	0.00	0.00	2
519.54	0.00	0.00	0.00	1
519.542	0.00	0.00	0.00	2
519.6	0.00	0.00	0.00	4
519.7	0.00	0.00	0.00	4
537.84	0.00	0.00	0.00	1
62-505	0.00	0.00	0.00	1
917.941.9	0.00	0.00	0.00	1
917.941.92	0.00	0.00	0.00	1
accuracy			0.63	510
macro avg	0.07	0.06	0.06	510
weighted avg	0.54	0.63	0.56	510

0.6333333333333333

Таблица 2.2 – Отчет по классификации тестовой выборки

Predicted True	**510**	**512**	**515.1**	**517**	**All**
510	7	17	0	28	52
512	4	76	0	49	129
513.7	0	0	0	2	2
513.74	0	0	0	1	1
513.8	0	0	0	1	1
513.83	0	0	0	1	1
513.88	0	0	0	1	1
513.881	0	0	0	2	2
514	0	4	0	14	18
515.1	0	0	0	6	6
517	1	17	2	241	261
518	0	0	0	1	1
518.321.1	0	0	0	1	1
518.322.2	0	0	0	1	1
519.1	0	3	0	1	4
519.2	0	3	0	6	9
519.3	0	0	0	1	1
519.44	0	1	0	0	1
519.45	0	0	0	1	1
519.48	0	1	0	1	2
519.54	0	1	0	0	1
519.542	0	2	0	0	2
519.6	0	0	0	4	4
519.7	0	2	0	2	4
537.84	0	0	0	1	1
62-505	0	0	0	1	1
917.941.9	0	0	0	1	1
917.941.92	0	0	0	1	1
All	12	127	2	369	510

Пример выделения ключевых слов из статьи Д. С. Аниконова, Д. С. Коноваловой, «Краевая задача для уравнения переноса с чисто комптоновским рассеянием» [14] приведен на рисунке 2.4

Для данной статьи путем классификации был определён УДК: 517, что соответствует теме «Анализ», истинный УДК данной статьи: 517.958, что также соответствует данной теме.

Рисунок 2.4 – Пример