**Санкт-Петербургский государственный электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**(СПбГЭТУ «ЛЭТИ»)**

|  |  |  |
| --- | --- | --- |
| **Направление** | 09.04.02 Информационные системы и технологии | |
| **Программа** | Управление IT проектами и продуктами | |
| **Факультет** | КТИ | |
| **Кафедра** | АПУ | |
| *К защите допустить* |  | |
| Зав. кафедрой |  | Шестопалов М.Ю. |

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

МАГИСТРА

Тема: Разработка алгоритма семантической декомпозиции текста на ключевые элементы для решения задачи категоризации

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Студентка |  |  | |  | Терещенко В.Н. |
|  |  | *подпись* | |  |  |
| Руководитель | к.т.н., доцент |  | |  | Кораблев Ю.А. |
|  |  | *подпись* | |  |  |
| Консультанты | к.т.н. |  | |  | Лосева Д.М. |
|  |  | *подпись* | |  |  |
|  | к.т.н. | *По доп разделу* | |  | Заславский М.М. |
|  |  | *подпись* | |  |  |
|  |  | |

Санкт-Петербург

2022**ЗАДАНИЕ**

**на выпускную квалификационную работу**

|  |  |
| --- | --- |
|  | Утверждаю |
|  | Зав. кафедрой АПУ |
|  | \_\_\_\_\_\_\_\_\_\_\_\_ Шестопалов М.Ю. |
|  | «\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_2022 г. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Студентка | Терещенко В.Н. | | | | Группа 6371 |
| Тема работы: Разработка алгоритма семантической декомпозиции текста на ключевые элементы для решения задачи категоризации. | | | | | |
| Место выполнения ВКР: Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), кафедра АПУ | | | | | |
| Исходные данные (технические требования): OC Windows 10 64-bit, Python 3.8.0 и выше 64-bit. | | | | | |
| Содержание ВКР:  Введение, Обзор предметной области, Формулировка требований к решению, Описание процесса разработки, Исследование, Безопасность жизнедеятельности, Заключение. | | | | | |
| Перечень отчетных материалов: пояснительная записка, иллюстративный материал. | | | | | |
| Дополнительные разделы: Коммерциализация результатов исследований в рамках НИР. | | | | | |
|  | | | | | |
| Дата выдачи задания | | | Дата представления ВКР к защите | | |
| «\_01\_» \_\_февраля\_\_\_\_\_2022 г. | | | «\_01\_» \_\_июня\_\_\_\_\_2022 г. | | |
| Студентка | |  | | Терещенко В.Н. | |
| Руководитель к.т.н., доцент | |  | | Кораблев Ю.А. | |

**календарный план выполнения**

**выпускной квалификационной работы**

|  |  |
| --- | --- |
|  | Утверждаю |
|  | Зав. кафедрой АПУ |
|  | \_\_\_\_\_\_\_\_\_\_\_\_ Шестопалов М.Ю. |
|  | «\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_2022 г. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Студентка | Терещенко В.Н. |  | Группа | 6371 |
| Тема работы: Разработка алгоритма семантической декомпозиции текста на ключевые элементы для решения задачи категоризации. | | | | |

|  |  |  |
| --- | --- | --- |
| № п/п | Наименование работ | Срок выполнения |
| 1 | Обзор литературы по теме работы | 20.01 – 21.02 |
| 2 | Разработка алгоритма | 22.02 – 05.04 |
| 3 | Проведение исследования | 06.04 – 09.04 |
| 4 | Внедрение алгоритма в СППР и тестирование | 06.05 – 15.05 |
| 5 | Оформление пояснительной записки | 10.04 – 22.05 |
| 6 | Оформление иллюстративного материала | 23.05 – 26.05 |
| 7 | Предзащита | 26.05 |

|  |  |  |
| --- | --- | --- |
| Студентка |  | Терещенко В.Н. |
| Руководитель к.т.н., доцент |  | Кораблев Ю.А. |

**РЕФЕРАТ**

Пояснительная записка 59 стр., 8 рис., 3 табл., 32 ист., 2 прил.

Ключевые слова: ключевое слово, ключевое словосочетание, ключевая фраза, ключевое предложение, семантическая декомпозиция, классификация, категоризация.

Цель работы – исследование современных методов классификации, создание инструмента, автоматически генерирующего оглавление и указатели терминов для массовых открытых онлайн-курсов платформы Stepik.

Объектом разработки является проектируемое веб-приложение.

В работе описано исследование подходов к решению задачи выделения ключевых слов, описан процесс проектирования веб-приложения для создания оглавления и списка терминов. Приведено практическое сравнение статистического и графового подходов в разработанном веб-приложении. При проектировании приложения были использованы язык программирования Python, веб-фреймворк Flask, СУБД PostgreSQL и REST-интерфейс Stepik API. Полученное приложение исследовано: определены время работы алгоритмов выделения ключевых слов Frequency, TopicRank и TextRank, проверены точность алгоритмов методом сравнения с экспертной оценкой.

**ABSTRACT**

The paper contains a study of approaches to solving the problem of highlighting keywords, describes the design process of a web application to create a table of contents and a list of terms. A practical comparison of statistical and graph approaches in the developed web application is given. When designing the application, the programming language Python, the Flask web framework, the PostgreSQL DBMS, and the Stepik API REST interface were used. The resulting application was investigated: the operating time of the Keyword, Frequency, TopicRank, and TextRank keyword extraction algorithms was determined, the accuracy of the algorithms was verified by comparison with an expert assessment.

**содержание**

[**определения, обозначения и сокращения** 8](#_Toc96536728)

[**введение** 9](#_Toc96536729)

[**1.** **Обзор предметной области** 11](#_Toc96536730)

[**1.1.** **Классификация и категоризация текстов на естественном языке** 11](#_Toc96536731)

[**1.2.** **Анализ основных методов (подходов) обработки текстов на естественном языке** 11](#_Toc96536732)

[**1.3.** **Семантическая декомпозиция как один из инструментов семантического анализа текстов** 12](#_Toc96536733)

[**1.4.** **Сравнительный анализ существующих алгоритмов семантической декомпозиции и выбор базовых алгоритмов для реализации** 12](#_Toc96536734)

[**1.5.** **Выводы** 16](#_Toc96536735)

[**2.** **ПОСТАНОВКА ЗАДАЧИ** 17](#_Toc96536736)

[**3.** **Описание Теоретических основ разрабатываемых алгоритмов** 18](#_Toc96536737)

[**3.1.** **Алгоритм семантической декомпозиции на n-граммы** 18](#_Toc96536738)

[**3.1.1.** **Математические основы алгоритма** 18](#_Toc96536739)

[**3.1.2.** **Графическое описание алгоритма** 18](#_Toc96536740)

[**3.1.3.** **Сценарий применения алгоритма для решения задачи категоризации** 18](#_Toc96536741)

[**3.2.** **Алгоритм семантической декомпозиции на предложения** 18](#_Toc96536742)

[**3.2.1.** **Математические основы алгоритма** 18](#_Toc96536743)

[**3.2.2.** **Графическое описание алгоритма** 18](#_Toc96536744)

[**3.2.3.** **Сценарий применения алгоритма для решения задачи категоризации** 18](#_Toc96536745)

[**3.3.** **Выводы** 18](#_Toc96536746)

[**4.** **ПРОГРАММНАЯ РЕАЛИЗАЦИЯ АЛГОРИТМОВ** 19](#_Toc96536747)

[**4.1.** **Алгоритм семантической декомпозиции на n-граммы** 19](#_Toc96536748)

[**4.1.1.** **Технические подробности программной реализации алгоритма** 19](#_Toc96536749)

[**4.1.2.** **Тестирование алгоритма** 19](#_Toc96536750)

[**4.1.3.** **Сценарий использования** 19](#_Toc96536751)

[**4.1.4.** **Применение алгоритма для решения задачи категоризации** 19](#_Toc96536752)

[**4.2.** **Алгоритм семантической декомпозиции на предложения** 19](#_Toc96536753)

[**4.2.1.** **Технические подробности программной реализации алгоритма** 19](#_Toc96536754)

[**4.2.2.** **Тестирование алгоритма** 19](#_Toc96536755)

[**4.2.3.** **Сценарий использования** 19](#_Toc96536756)

[**4.2.4.** **Применение алгоритма для решения задачи категоризации** 20](#_Toc96536757)

[**4.3.** **Выводы** 20](#_Toc96536758)

[**5.** **исследование** 21](#_Toc96536759)

[**5.1.** **Исследование времени работы** 21](#_Toc96536760)

[**5.1.1.** **Сценарий эксперимента** 21](#_Toc96536761)

[**5.1.2.** **Методика измерения** 21](#_Toc96536762)

[**5.1.3.** **Используемые программные средства** 21](#_Toc96536763)

[**5.1.4.** **Результаты эксперимента** 22](#_Toc96536764)

[**5.1.5.** **Сравнение и оценка результатов** 22](#_Toc96536765)

[**5.2.** **Исследование точности результатов** 22](#_Toc96536766)

[**5.2.1.** **Сценарий эксперимента** 23](#_Toc96536767)

[**5.2.2.** **Входные данные** 23](#_Toc96536768)

[**5.2.3.** **Результат работы алгоритмов** 23](#_Toc96536769)

[**5.2.4.** **Экспертная оценка** 24](#_Toc96536770)

[**5.2.5.** **Сравнение с экспертной оценкой** 24](#_Toc96536771)

[**5.3.** **Выводы** 25](#_Toc96536772)

[**6.** **Коммерциализация результатов исследований в рамках НИР** 26](#_Toc96536773)

[**заключение** 27](#_Toc96536774)

[**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ** 30](#_Toc96536775)

# **определения, обозначения и сокращения**

В настоящей пояснительной записке применяют следующие термины с соответствующими определениями:

Веб-сайт – совокупность взаимосвязанных веб-страниц, используемых одно доменное имя (адрес).

Веб-страница – документ, обычно в формате html, содержащий какой-либо контент (текст, фото, видео и другое).

Документ – коллекция текста (твит, статья и др.).

Классификатор – алгоритм, или метод, посредством которого проводится классификация.

Классификация – процедура, в ходе которой объекты (предметы, явления) распределяются по группам, или классам, по каким-либо признакам, удобным для их исследования.

Корпус – набор документов схожей тематики.

Токен – это последовательность символов в документе, имеющая значение для анализа. Обычно это отдельные слова, но не всегда.

Документ – это коллекция текста. (https://nuancesprog.ru/p/5870/)

МООК – массовые открытые онлайн курс (Massive Open Online Course, MOOC).

ОЗУ – оперативное запоминающее устройство.

Реестр – форма систематизации, учета.

СУБД – система управления базами данных.

API – Application Programming Interface (программный интерфейс приложения).

ID – Identifier (идентификатор).

NLP – Natural Language Processing (обработка естественного языка).

REST – Representational State Transfer (передача состояния представления).

SQL – Structured Query Language (язык структурированных запросов).

# **введение**

В настоящее время системы обработки естественного языка используются повсеместно, начиная от переводчика и поисковой системы в браузере, заканчивая умными чат-ботами и голосовыми помощниками. Такие системы внедряются в совершенно различные структуры и упрощают их использование.

Недавно для компании ООО «Газпром переработка» была разработана система поддержки принятия решений для автоматизации работы службы технической поддержки. Данная СППР автоматически обрабатывает запросы, поступающие на 1-ую линию технической поддержки. Обработка главным образом заключается в категоризации запросов и направлении их по различным отделам компании.

Данная тема магистерской диссертации была выбрана с целью исследования влияния различных вариантов разделения текста запроса на результат работы СППР, а именно на результат классификации и, соответственно, категоризации текстовых запросов.

В большинстве случаев наличие среди ключевых слов устойчивых словосочетаний ведет положительно влияет на вероятность правильного категорирования специализированных текстов. **Цель** данной работы – проверить эту гипотезу для текстов обращений, поступающих в техническую поддержку компании ООО Газпром переработка, и предложить вариант доработки проекта с учетов результата исследования.

Для достижения цели были поставлены следующие **задачи**:

* сравнить и проанализировать существующие подходы семантической декомпозиции текста;
* разработать алгоритм и протестировать его (это тоже этап разработки);
* исследовать качество разработанного алгоритма;
* провести внедрение алгоритма в проект (в случае успешного исследования).

**Объектом** исследования являются тексты запросов, поступившие в службу поддержки.

**Предметом** исследования является процесс семантической декомпозиции текста запроса на ключевые элементы.

**Практическая значимость** решения заключается в использовании разработанного алгоритма для улучшения результатов категоризации обращений и, соответственно, для оптимизации работы СППР для службы технической поддержки компании Газпром переработка.

1. **Обзор предметной области**

Разрабатываемый алгоритм является частью системы, предназначенной для категоризации запросов.

* 1. **Классификация и категоризация текстов на естественном языке**

Классификаций или категоризацией называют процесс отнесения входных текстовых документов к одной из групп данных (классу или категории).

Эти две интеллектуальные задачи следует отличать друг от друга. Термин *класс* можно определить, как совокупность предметов или понятий (образов), выделенных и сгруппированных по определенным признакам или правилам. Под *классификацией* понимается отнесение некоторого документа к классу, выполняемое по формальным правилам, исходя из совокупности признаков. Границы классов определены достаточно точно: документ относится к данному классу, если известно, что он обладает необходимым числом признаков, характерных для этого класса.

*Категория* же определяет лишь некоторые общие свойства образов и связи между ними. Задача *категоризации*, т. е. определения отношения данного документа к некоторой категории, гораздо менее определена, чем задача отношения к классу. Границы различных категорий являются нечеткими, расплывчатыми, и обычно сама категория понимается не через формальное определение, а только в сравнении с другими категориями.

Таким образом, задачей систем-классификаторов является установление принадлежности образа к одному из формально определенных классов. Примерами такой задачи является задача классификации растений в ботанике, классификация химических веществ по их свойствам и типам возможных реакций, в которые они вступают, и другие. Формальные признаки могут быть определены посредством правил типа “если-то”, а системы, оперирующие с такими правилами, получили название *экспертных систем*. Традиционной областью применения классификаторов на нейронных сетях является экспериментальная физика высоких энергий, где одной из актуальных задач выступает выделение среди множества зарегистрированных в эксперименте событий с элементарными частицами событий, представляющих интерес для данного эксперимента.

Таким образом, проблема категоризации является более сложной в сравнении с проблемой классификации. В категоризации помимо отнесения документа к какой-либо группе, требуется определить сами эти группы, т. е. сформировать категории.

В случае обучения с учителем формирование категорий происходит методом проб и ошибок на основе примеров с известными ответами, предоставляемыми экспертом. Эксперта обычно называют “супервизором” или учителем. Учитель управляет обучением при помощи изменения параметров связей и, реже, самой топологии сети.

Задачей системы-категоризатора является формирование обобщающих признаков в совокупности примеров. При увеличении числа примеров несущественные, случайные признаки сглаживаются, а часто встречающиеся – усиливаются, при этом происходит постепенное уточнение границ категорий. Хорошо обученная нейросетевая система способна извлекать признаки из новых примеров, ранее неизвестных системе, и принимать на их основе приемлемые решения.

Важно отметить различие в характере неявных “знаний”, запомненных искусственной нейронной сетью, и явных, формальных “знаний”, заложенных в экспертных системах. Некоторые сходства и различия представлены в следующей таблице.

|  |  |  |
| --- | --- | --- |
| Критерий | **Экспертные системы (ЭС)** | **Нейросетевые системы (НС)** |
| Источник знаний | Формализованный опыт эксперта, выраженный в виде логических утверждений (правил и фактов), безусловно принимаемых системой | Совокупный опыт эксперта-учителя, отбирающего примеры для обучения, и индивидуальный опыт обучающейся на этих примерах нейронной сети |
| Характер знаний | Формально-логическое “левополушарное” знание в виде правил | Ассоциативное “правополушарное” знание в виде связей между нейронами сети |
| Развитие знаний | В форме расширения совокупности правил и фактов (базы знаний) | В форме дообучения на дополнительной последовательности примеров, с уточнением границ категорий и формированием новых категорий |
| Роль эксперта | Задает на основе правил полный объем знаний экспертной системы | Отбирает характерные примеры, не формулируя специально обоснование своего выбора |
| Роль искусственной системы | Поиск цепочки фактов и правил для доказательства суждения | Формирование индивидуального опыта в форме категорий, получаемых на основе примеров, и категоризация образов |

Различия в характере экспертных и нейросетевых систем обуславливают и различия в их сферах применения. Экспертные системы применяются в узких предметных областях с хорошо структурированными знаниями, например в классификации неисправностей конкретного типа оборудования, фармакологии, анализе химсостава проб и т. д. Нейронные сети применяются кроме перечисленных областей и в задачах с плохо структурированной информацией, например при распознавании образов, рукописного текста, анализе речи и т. д. (<https://pandia.ru/text/79/574/8277.php#:~:text=Классификация%20и%20категоризация.%20В%20случае%2C,выполняет%20классификацию%20или%20категоризацию%20данных>)

* + 1. **Этапы процесса категоризации текстов**

Решение задач категоризации и классификации состоят из четырех последовательных этапов:

* предобработка и индексация документов;
* уменьшение размерности пространства признаков;
* построение и обучение классификатора с помощью методов машинного обучения;
* оценка качества классификации.

На рисунке 1 представлена общая схема процесса категоризации.



Рисунок 1 – Этапы процесса категоризации текстов (скрин из статьи, можно перерисовать)

Рассмотрим этапы, предшествующие работе классификатора, так как алгоритм семантической декомпозиции будет встроен именно в эту часть системы.

**Предварительная обработка** текста включает нормализацию (перевод всех букв к одному регистру (чаще к нижнему), удаление знаков пунктуации, цифр, пробелов), токенизацию (разбиение текста на более мелкие части – токены), удаление стоп-слов (семантически нейтральных слов, таких как союзы, предлоги, артикли и пр.), морфологический анализ (разметка по частям речи; опционально), лемматизация и стемминг. Это позволяет значительно сократить размерность пространства. В результате в качестве признаков документа выступают все значимые слова, встречающиеся в документе.

Немного о процедурах предварительной обработки.

* **Нормализация и токенизация**. В самом начале обработки текст приводится к единообразному виду (единый регистр слов, отсутствие знаков пунктуации, расшифрованные сокращения, словесное написание чисел и т.д.), что необходимо для применения унифицированных методов обработки текста, и разделяется на более мелкие единицы (на символы, слова, фразы или предложения) для того, чтобы иметь возможность провести последующие процедуры.
* **Лемматизация и стемминг**. При построении классификатора текстов нет смысла различать формы одного и того же слова. Это приводит к неоправданному разрастанию словаря, дроблению статистики, увеличению ресурсоёмкости и снижению качества модели.
  + Лемматизация – приведение каждого слова в документе к его нормальной форме (в русском языке: для существительных – именительный падеж, единственное число; для прилагательных – именительный падеж, единственное число, мужской род; для глаголов, причастий, деепричастий – глагол в инфинитиве). Существуют специальные программы – лемматизаторы, обычно основанные на явном хранении грамматического словаря со всеми формами слов. Недостатком лемматизации является трудоёмкость составления словарей, и, как следствие, их неполнота, особенно по части специальной терминологии и неологизмов, которые как раз и представляют наибольший интерес для тематического моделирования.
  + Стемминг – отбрасывание изменяемых частей слов (главным образом, окончаний). Она не требует хранения словаря всех слов и основана на правилах морфологии языка. Недостатком стемминга является большее число ошибок. Стемминг лучше подходит для английского языка, но хуже для русского.
* **Отбрасывание стоп-слов**. Слова, встречающиеся во многих текстах различной тематики, бесполезны при классификации, и могут быть отброшены. К ним относятся предлоги, союзы, числительные, местоимения, некоторые глаголы, прилагательные и наречия. Число таких слов обычно варьируется в пределах нескольких сотен. Их отбрасывание почти не влияет на длину словаря, но может приводить к заметному сокращению длины некоторых текстов.
* **Отбрасывание редких слов**. Слова, встречающиеся в длинном тексте слишком редко, также можно отбрасывать, полагая, что данное слово не имеет принципиального значения в данном тексте.
* **Выделение ключевых фраз**. При обработке коллекций научных, юридических или других специальных текстов вместо отдельных слов выделяют ключевые фразы – словосочетания, являющиеся устойчивыми оборотами или терминами в данной предметной области. Это отдельная и довольно сложная задача, которая может решаться методами машинного обучения с привлечением экспертов для формирования обучающих выборок и контроля качества автоматического выделения терминов [1].

(http://www.machinelearning.ru/wiki/images/0/02/Romanenko2012bach.pdf)

**Индексация** документов представляет собой построение некоторой числовой модели текста, которая переводит текст в удобное для дальнейшей обработки представление.

Распространенные модели индексации:

* модель «мешка слов» (bag-of-words) позволяет представить документ в виде многомерного вектора слов и их весов в документе [2]. Другими словами, каждый документ – это вектор в многомерном пространстве, координаты которого соответствуют номерам слов, а значения координат – значениям весов;
* модель индексации Word2vec [3] представляет каждое слово в виде вектора, который содержит информацию о контекстных (сопутствующих) словах;
* модель индексации, основанная на учете n-грамм [2], то есть последовательностей из соседних токенов.

(https://cyberleninka.ru/article/n/metody-avtomaticheskoy-klassifikatsii-tekstov/viewer)

Этап **извлечения признаков** из текстовых документов необходим для того, чтобы сопоставить каждому документу некий набор характеристик, которые описывают этот документ [1]. (https://cyberleninka.ru/article/n/metody-izvlecheniya-priznakov-iz-tekstovyh-dokumentov/viewer)

Извлечение признаков соответствует выбору элементов признаков и вычислению весов признаков. Основная идея состоит в том, чтобы независимо ранжировать токены в соответствии с определенным индексом оценки, выбирать токены с наивысшими баллами и отфильтровывать оставшиеся. Обычно используемые оценки включают частоту встречаемости токенов, взаимную информацию и др.

Вес признака чаще всего вычисляется классическим методом TF-IDF и расширяющими его методами. Основная идея состоит в том, что важность слова прямо пропорциональна частоте его появления в категории и обратно пропорциональна количеству вхождений во все категории. (https://russianblogs.com/article/52941084692/)

Далее на полученных признаках модель обучается и после успешного прохождения тестирования система категоризации готова к работе.

Исходя из вышеизложенной информации, становится ясно, что разрабатываемый алгоритм семантической декомпозиции на ключевые элементы будет встроен в систему на этапе предобработки.

* + 1. **Ключевые элементы**

Далее можно вставить главу про ключевые элементы (КС и расширение этого понятия с помощью n-грамм).

И потом перейти в главу методов построения n-грамм.

ПЕРЕДЕЛАТЬ РАЗДЕЛ

Основная мысль: ключевые элементы – токены – являются слова, состоящие из букв или цифр. Знаки пунктуации и служебные символы токенами не являются

Под ключевыми элементами в данной работе понимаются токены. Токенами в данной работе могут являться как отдельные слова, так и словосочетания или n-граммы, которые в будущем могут (потенциально) иметь наибольшие показатели весов признаков. То есть, из ключевых элементов во время работы алгоритма выделения ключевых слов, следующего после алгоритма декомпозиции, будут сформированы ключевые слова (термины), словосочетания или фразы.

В общем представлении ключевыми называются важные слова или фразы, дающие высокоуровневое описание содержания текстового документа, позволяющие выявить его тематику, имеющие компактное представление в памяти. Выделенный из текста список ключевых слов (КС) может выступать в качестве метаинформации, представляя текстовый документ при решении задач информационного поиска, классификации, кластеризации, аннотирования и реферирования.

Ключевые словосочетания (фразы) представляют собой сочетание двух или более слов, которые как могут следовать друг за другом в тексте, так и быть разделенными другими языковыми единицами. Не все входящие в состав ключевых фраз слова при отдельном рассмотрении являются ключевыми. Но также вполне очевидно, что выделением отдельных КС затруднительно выразить основной смысл содержимого. Поэтому на практике востребовано выделение именно ключевых фраз. Основное отличие – длина списка на выходе алгоритма. Документу обычно соответствуют единицы ключевых фраз, длина же списка терминов, основой которых также являются КС колеблется от десятков до сотен. Составление перечня ключевых словосочетаний является одной из трудностей в рассматриваемой предметной области.

В результате систематизации данных различных исследователей был выделен перечень существенных свойств и функций ключевых слов в текстах, значимых в контексте моделирования и алгоритмизации процесса их извлечения.

Характеристика КС:

* являются наиболее употребительными (частотными) наименованиями, обозначают признак предмета, состояние или действие;
* представлены значимой лексикой, достаточно обобщены по своей семантике (средней степени абстракции), стилистически нейтральны, не оценочны;
* связаны друг с другом сетью семантических связей, пересечения значений;
* более половины слов ядра тематического компонента состоит из ключевых слов, а минимальный набор КС приближается к инварианту содержания при их логическом упорядочивании;
* набор КС состоит из 5-15 или 8-10 слов, что соответствует объему оперативной памяти человека, в тексте содержится 25-30% ключевых слов;
* набор КС определяет контексты слов, обладающих максимальной предсказуемостью. (https://cyberleninka.ru/article/n/metody-i-algoritmy-izvlecheniya-klyuchevyh-slov/viewer)

Итак, понятие ключевого элемента обозначено. Следующим шагом проведем определение понятия семантической декомпозиции, начиная с более общего понятия – семантического анализа.

* 1. **Семантическая декомпозиция как один из инструментов семантического анализа текстов**

Семантический анализ текста оценивает количество слов или фраз, которые определяют смысл текста (то есть его семантическое ядро, и статистические показатели). <https://cropas.by/seo-slovar/semanticheskij-analiz/>

Так, всю совокупность представленных на сегодняшний день методов анализа текста можно разделить на две группы:

* лингвистический анализ − основан на извлечении смысла текста по его семантической структуре;
  + статистический анализ − основан на извлечении смысла текста по частотному распределению слов в тексте.

Деление на группы условное, так как в реальных задачах и при решении проблем всегда используется сочетание методов для достижения определенного результата. (там есть еще подробное описание -х самых востребованных семантических методов с результатами <http://repo.ssau.ru/bitstream/Informacionnye-tehnologii-i-nanotehnologii/Razrabotka-avtomatizirovannoi-sistemy-semanticheskogo-analiza-tekstovoi-informacii-64154/1/paper%20324_1800-1804.pdf>)

Проблема аналитического анализа семантики и извлечения знаний из текстовых документов является достаточно сложной и требует разработки и применения специализированных интеллектуальных информационных систем (ИИС) [2, 3]. На практике путём введения ряда допущений можно разделить проблему на несколько групп относительно независимых типовых задач.

К типовым задачам исследования семантики одиночных текстовых документов следует отнести:

а) выделение в тексте ключевых слов;

б) выделение в тексте цепочки ключевых слов;

в) выявление контекстных слов (контекста);

г) аннотирование текста.

Для случая обработки набора тестового документа приоритетными являются следующие задачи:

а) классификация текстов по совокупности признаков;

б) каталогизация текстов;

в) поиск заданных информационных фрагментов в потоке ТД;

г) поиск (подбор) текстов, обладающих заданными свойствами. (<https://fundamental-research.ru/ru/article/view?id=41321>)

ТУТ МОЖНО ПРОКОММЕНТИРОВАТЬ ПО ПУНКТАМ, ЧТО ПОДХОДИТ НАМ

Под понятием семантической декомпозиции, в свою очередь, понимают алгоритм, разбивающий значения предложений, фраз или понятий на менее сложные понятия. (<https://translated.turbopages.org/proxy_u/en-ru.ru.65356b33-623e167d-0feb0112-74722d776562/https/en.wikipedia.org/wiki/Semantic_decomposition_(natural_language_processing)>)

В лингвистических экспертизах присутствует огромное множество вариантов компонент, на которые может разбиваться текст. (<https://www.tipse.ru/jour/article/viewFile/220/220>) Однако в рамках данной работы будет вестись исследование семантической декомпозиции на слова и фразы (словосочетания). Поэтому здесь будут рассмотрены три вида алгоритмов, исходя из типа единиц декомпозиции текста:

* алгоритмы семантической декомпозиции текста на слова;
* алгоритмы семантической декомпозиции текста на словосочетания;
* алгоритмы семантической декомпозиции текста на слова и словосочетания.
  1. **Сравнительный анализ существующих алгоритмов семантической декомпозиции и выбор базовых алгоритмов для реализации**

По своей сути алгоритм семантической декомпозиции приближен к алгоритмам токенизации, поэтому для анализа уже существующих решений были подобраны следующие инструменты:

1. **Модуль re**.

Используется для применения регулярных выражений для различных задач поиска. Регулярные выражения – это последовательность символов, которая может использоваться для определения искомого шаблона для поиска текста, изменения или разделения строк. Это самый простой способ токенизации, без каких-либо сопутствующих процедур (лемматизации, распознавания частей речи и так далее). (https://rukovodstvo.net/posts/id\_1304/)

1. **Библиотека Nltk**.

Это самая большая библиотека для NLP на python. Здесь помимо токенизации есть также лемматизация и стемминг, маркировка частей речи, подбор синонимов и антонимов и другое.

В модуле tokenize библиотеки nltk хранится 19 токенизаторов. Ниже приведены наиболее подходящие для поставленной задачи (<https://newtechaudit.ru/obzor-tokenizatorov-vhodyashhih-v-sostav-biblioteki-nltk/>):

* 1. **RegexpTokenizer**

Данный токенизатор разбивает исходный текст на подстроки, используя регулярное выражение, передаваемое ему в качестве параметра.

Библиотека NLTK содержит также несколько подклассов RegexpTokenizer, использующих заранее определенные регулярные выражения: WhitespaceTokenizer, BlanklineTokenizer, WordPunctTokenizer.

* + 1. **WhitespaceTokenizer**

При помощи данного токенизатора исходный текст разделяется на группы символов, ограниченных пробелами (обычно, это отдельные слова, либо слова со следующими за ними знаками препинания).

Вместо WhitespaceTokenizer можно использовать строковый метод split().

* + 1. **BlanklineTokenizer**

Токенизатор разбивает текст, рассматривая в качестве разделителя любую последовательность пустых строк (содержащих только символы пробела или табуляции). Для этого используется регулярное выражение ‘\s\*\n\s\*\n\s\*’.

* + 1. **WordPunctTokenizer**

Токенизатор преобразует текст в последовательность буквенных и неалфавитных символов с помощью регулярного выражения ‘\w+|[^\w\s]+’. Отличается от WhitespaceTokenizer тем, что отделяет знаки препинания и иные символы пунктуации от слова, за или перед которым они идут, и возвращает их в качестве отдельных токенов.

* 1. Следующая группа токенизаторов, входящих в состав библиотеки NLTK, носит название «простых» (simple) токенизаторов. Она включает в себя 3 токенизатора: SpaceTokenizer, TabTokenizer и LineTokenizer.
     1. **SpaceTokenizer**

Разбивает строку на токены, используя в качестве разделителя пробел. Результат работы этого метода полностью совпадает с результатом метода s.split(‘ ‘).

* + 1. **TabTokenizer**

Аналогичен SpaceTokenizer, только для разделения строки использует символ табуляции, что совпадает с применением метода s.split(‘\t’).

* + 1. **LineTokenizer**

Аналогичен SpaceTokenizer, за исключением того, что для разделения строки использует символ новой строки (при необходимости может отбрасывать пустые строки). Его применение похоже на применение метода s.split(‘\n’).

* 1. **MWETokenizer (Multi-Word Expression Tokenizer)**

Токенизатор берет строку, которая уже была разделена на токены, и повторно токенизирует ее, объединяя выражения из нескольких слов в отдельные токены, используя словарь MWE. Это может быть полезно в том случае, когда в тексте встречаются устойчивые выражения или именованные сущности, состоящие из нескольких слов.

Из недостатков: токенизатору необходим предварительно заданный список выражений из нескольких слов для определения, существует ли какое-либо из них в тексте. Кроме того, он чувствителен к регистру слов, а также не определяет различные словоформы и написания одного и того же слова.

* 1. **ToktokTokenizer**

Данный токенизатор позиционируется как простой универсальный токенизатор. На вход ему подается по одному предложению в строке. Работа основана на последовательном применении к исходному тексту списка регулярных выражений. С их помощью происходят, например, замена неразрывных пробелов обычными, замена знака многоточия единичной точкой, расстановка пробелов после открывающих и перед закрывающими знаками препинания. В итоговой строке удаляются начальные и конечные пробелы, а сама строка преобразуется в кодировку Unicode.

Согласно документации, он был протестирован и дает достаточно хорошие результаты для английского, персидского, русского, чешского, французского, немецкого, вьетнамского, таджикского и некоторых других языков. Входные данные должны быть представлены в кодировке UTF-8.

* 1. **TweetTokenizer**

Это токенизатор с поддержкой особенностей коротких сообщений Twitter, разработанный для гибкой и простой адаптации к новым предметным областям и задачам. Основная логика его работы заключается в следующем:

* определяется список регулярных выражений, для поиска в исходном тексте гиперссылок, смайликов, телефонных номеров, имен пользователей Twitter, хештегов и другого;
* регулярные выражения помещаются по порядку в скомпилированный объект регулярного выражения, называемый word\_re;
* выполняется токенизация с помощью метода word\_re.findall(s), где s – строка, исходного текста.
  1. **TreebankWordTokenizer**

Токенизатор Treebank использует регулярные выражения для разделения текста на слова так, как это реализовано в Penn Treebank. Для этого последовательно выполняются следующие шаги:

1. разделяются стандартные сокращения, например, «don’t» **->** «do n’t» или «they’ll» **->** «they ‘ll» для английского языка;
2. большинство знаков препинания, разделяются на отдельные токены;
3. разделяются запятые и одинарные кавычки, если за ними следует пробел;
4. отделяются точки, которые проставлены в конце строк.
   1. **NLTKWordTokenize**r

Это улучшенный вариант TreebankWordTokenizer, включающий несколько адаптированных списков сокращений.

Стоит отметить, что, по сравнению с TreebankWordTokenizer, NLTKWordTokenizer является «деструктивным» токенизатором. То есть, применяемые регулярные выражения приводят исходный текст в состояние, не поддающееся восстановлению.

* 1. **StanfordSegmenter**

Некоторые языки требуют более обширной предварительной обработки токенов, которая обычно называется сегментацией. Например, StanfordSegmenter предназначен для «токенизации» или «сегментации» слов китайского или арабского текста (также можно использовать для английского, французского и испанского языков).

Исходя из описания токенизаторов библиотеки nltk, в данной работе могут быть использованы для сравнения следующие токенизаторы: WhitespaceTokenizer, WordPunctTokenizer, SpaceTokenizer, ToktokTokenizer, **TweetTokenizer,** NLTKWordTokenizer**.**

1. **Библиотека SpaCy.**

Одна из самых популярных библиотек NLP наряду с NLTK. Основное различие между двумя библиотеками заключается в том, что NLTK содержит широкий спектр алгоритмов для решения одной проблемы, тогда как spaCy содержит только один, но лучший алгоритм для решения проблемы.

При создании текстового документа SpaCy автоматически разбивает документ на токены, распознает части речи токенов и их зависимости в предложении. Так же с помощью spaCy можно обнаруживать именованные сущности (имена, названия компаний, мест, зданий, учреждений, валют и другие). (<https://rukovodstvo.net/posts/id_1131/>) Поиск n-грамм с помощью texstacy

1. **Библиотека StanfordCoreNLP.**

Здесь тоже присутствует токенизация по словам и предложениям, лемматизация, тегирование и распознавание именных сущностей. (<https://rukovodstvo.net/posts/id_1092/>) Написана для java

1. **Библиотека Pattern.**

Это многоцелевая библиотека, способная решать следующие задачи:

* Обработка естественного языка: выполнение таких задач, как токенизация, стемминг, теги POS, анализ настроений и т. Д.
* Data Mining: он содержит API-интерфейсы для сбора данных с таких сайтов, как Twitter, Facebook, Wikipedia и т. Д.
* Машинное обучение: содержит модели машинного обучения, такие как SVM, KNN и персептрон, которые можно использовать для задач классификации, регрессии и кластеризации.

Касаемо решаемой задачи, в данной библиотеке нет отдельной функции для токенизации, но есть функция, которая включает токенизацию, маркировку частей речи, ролей и лемматизацию. Также есть возможность поиска n-грамм, преобразования чисел в текстовое представление и некоторые другие функции. (<https://rukovodstvo.net/posts/id_1099/>) https://stackabuse.com/python-for-nlp-introduction-to-the-pattern-library/

1. **Библиотека Gensim.**

Gensim – это библиотека, которая используется для тематического моделирования без учителя и обработки естественного языка (NLP). Она предназначена для извлечения семантических тем из документов, где под темой понимается повторяющаяся группа слов, часто встречающихся вместе.

Gensim может работать с большими текстовыми коллекциями. Этим она отличается от других программных библиотек машинного обучения, ориентированных на обработку в памяти. GenSim также предоставляет эффективные многоядерные реализации различных алгоритмов для увеличения скорости обработки. В нее добавлены более удобные средства для обработки текста, чем у конкурентов, таких как Scikit-learn, R и т. д.

В библиотеке реализована токенизация на слова (униграммы) и словосочетания (биграммы и триграммы), а также другие процедуры обработки токенов. (https://pythonru.com/biblioteki/gensim)

1. **Библиотека Scikit-learn.**

Scikit-learn – это библиотека машинного обучения с открытым исходным кодом, которая поддерживает контролируемое и неконтролируемое обучение. Здесь также представлены различные инструменты для подгонки модели, предварительной обработки данных, выбора модели, оценки модели и многих других утилит.

Касаемо токенизации, в Scikit-learn имеется возможность токенизации как на униграммы, так и на n-граммы. (https://scikit-learn.ru/6-2-feature-extraction/)

1. **Библиотека Keras.**

Класс Tokenizer Keras используется для токенизации текста на отдельные слова, но не работает для n-грамм. Конечной целью использования класса является векторизация текстового корпуса. (https://machinelearningknowledge.ai/keras-tokenizer-tutorial-with-examples-for-fit\_on\_texts-texts\_to\_sequences-texts\_to\_matrix-sequences\_to\_matrix/)

1. **Rutokenizer.**

Пакет содержит набор классов, которые умеют выделять предложения из текста (сегментация) и разбивать предложения на слова (токенизация) с учетом многословных единиц текста типа "из-за" или "какой-то". Однако здесь не поддерживается распознавание n-грамм. (<https://github.com/Koziev/rutokenizer>)

1. **TextBlob**

Здесь предложено множество встроенных методов для обычных задач обработки естественного языка, среди которых: извлечение именных фраз, разметка части речи, анализ настроений, классификация текста, токенизация (в том числе n-граммы), частота слов и фраз, словофлексия, исправление орфографии. (https://biconsult.ru/products/textblob-v-python-rukovodstvo)

В качестве резюмирования вышеперечисленных характеристик инструментов были составлены таблицы (см. табл. 1-5). В них наглядно продемонстрированы их сходства и различия.

Таблица 1 – Характеристика инструментов токенизации библиотеки NLTK

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Инструменты | Критерии | | | |
| Поиск униграмм | Поиск n-грамм (где n>1) | Поддержка автоматической обработки | Поддержка обработки русского языка |
| **RegexpTokenizer** | Есть | Нет | Есть | Есть |
| WhitespaceTokenizer | Есть | Нет | Есть | Есть |
| BlanklineTokenizer | Нет | Нет | Есть | Есть |
| WordPunctTokenizer | Нет | Нет | Есть | Есть |
| SpaceTokenizer | Есть | Нет | Есть | Есть |
| TabTokenizer | Нет | Нет | Есть | Есть |
| LineTokenizer | Нет | Нет | Есть | Есть |
| MWETokenizer | Нет | Есть | Нет | Есть |
| ToktokTokenizer | Есть | Нет | Есть | Есть |
| **TweetTokenizer** | Есть | Нет | Есть | Есть |
| **TreebankWordTokenizer** | Есть | Нет | Есть | Есть |
| NLTKWordTokenizer | Есть | Нет | Есть | Есть |
| StanfordSegmenter | Есть | Нет | Есть | Нет |

В первой таблице (см. табл. 1) представлены токенизаторы библиотеки NLTK и их характеристики по критериям возможности поиска n-грамм, применения инструмента к русскоязычному тексту и наличие автоматизации работы инструмента.

Из всех токенизаторов из таблицы 1 только один не работает с русским языком – StanfordSegmenter.

Инструменты BlanklineTokenizer, WordPunctTokenizer, TabTokenizer и LineTokenizer предназначены для токенизации на более крупные единицы, такие как предложения, абзацы и так далее.

MWETokenizer не поддерживает автоматизированную работу: **для работы** MWETokenizer требуется предварительное составление словаря устойчивые выражения или именованные сущности для их дальнейшего поиска в тексте.

При рассмотрении **TreebankWordTokenizer** и NLTKWordTokenizer необходимо отметить, что они очень похожи, но второй представляет собой улучшенный вариант первого. Поэтому было решено из двух инструментов выбрать один **–** NLTKWordTokenizer.

По итогу осталось 7 токенизаторов, применимых для разделения на униграммы (**RegexpTokenizer,** WhitespaceTokenizer, SpaceTokenizer, ToktokTokenizer, TweetTokenizer, NLTKWordTokenizer) и ни одного для n-грамм.

Таблица 2 – Обобщенная характеристика инструментов токенизации

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Инструменты | Критерии | | | |
| Поиск униграмм | Поиск n-грамм (где n>1) | Поддержка обработки русского языка | Возможность реализации дополнительных процедур (кроме токенизации) |
| Re | Есть | Нет | Есть | Нет |
| Nltk | Есть | Есть | Есть | Есть |
| SpaCy | Есть | Есть | Есть | Есть |
| StanfordCoreNLP | Есть | Нет | Есть | Есть |
| Pattern | Есть | Есть | Есть | Есть |
| Gensim | Есть | Есть | Есть | Есть |
| Scikit-learn | Есть | Есть | Есть | Есть |
| Keras | Есть | Нет | Есть | Есть |
| Rutokenizer | Есть | Нет | Есть | Нет |
| TextBlob | Есть | Есть | Есть | Есть |

В таблице 2 рассмотрены характеристики инструментов токенизации по следующим критериям: поиск 1- и n-грамм, поддержка обработки русскоязычных текстов и наличие возможностей реализации дополнительных процедур обработки токенов.

Модули библиотеки NLTK объединены в данной таблице, так как по многим свойствам, важных для данной таблицы, они схожи. Однако в NLTK есть модули (ngrams, bigrams, trigrams), с помощью которых можно распознавать n-граммы. Есть также модуль everygrams, однако в нем нельзя регулировать значение n, n всегда равно количеству слов в предложении.

По итогам сравнения, все перечисленные в таблице инструменты поддерживают работу с русским языком. Однако два из них (Re и Rutokenizer) проигрывают на фоне остальных из-за невозможности реализации дополнительных процедур обработки токенов.

Также было принято решение исключить StanfordCoreNLP из дальнейшего исследования, потому что библиотека написана на java, что затрудняет ее использование.

Оставшиеся инструменты можно разделить на подходящие для поиска униграмм (Nltk, SpaCy, StanfordCoreNLP, Pattern, Gensim, Scikit-learn, Keras, TextBlob) и n-грамм (Nltk, SpaCy, Pattern, Gensim, Scikit-learn, TextBlob).

Далее будут отдельно рассмотрены кандидаты двух этих групп.

Первая группа будет сравниваться по времени работы и по чистоте токенов (см. табл. 3). Время работы будет измерено с помощью библиотеки time, числом в таблице будет среднее между 5-ю измерениями (см. Приложение А). Под чистотой токенов понимаются списки токенов, состоящих из слов и не содержащих знаков препинания, которые некоторые алгоритмы определяют как токены.

В таблице 3 будут сравниваться все инструменты из таблицы 2, так как кандидаты из таблицы 1 также не имеют возможности реализации дополнительных процедур для токенов, как и Re и Rutokenizer.

Таблица 3 – Итоговая таблица сравнения для токенизации по словам

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Инструменты | Критерии | | | | |
| Время работы, с | Чистота токенов | | | |
| 1 | 2 | 3 | 4 |
| **RegexpTokenizer** | 1,8399 | + | + | - | - |
| WhitespaceTokenizer | 0,0002 | + | - | + | + |
| SpaceTokenizer | 0,0002 | + | - | + | + |
| ToktokTokenizer | 0,0004 | + | + | + | + |
| TweetTokenizer | 0,0088 | + | + | - | + |
| NLTKWordTokenizer | 0,0090 | + | + | + | + |
| Re | 0,0010 | + | + | - | - |
| SpaCy | 5,5951 | - | + | + | + |
| Pattern | --- | --- | --- | --- | --- |
| Gensim | 0,1307 | - | + | - | - |
| Scikit-learn | 0,0012 | - | + | - | - |
| Keras | 0,0038 | - | + | - | - |
| Rutokenizer | 0,0199 | + | + | - | + |
| TextBlob | 0,1300 | + | + | + | + |

Исследование времени проводилось на предложении «Не ветер, а какой-то ураган!».

Исследование по критерию «Частота токенов» проводилась по трем предложениям: 1 – «Я-то из-за угла вышел.» (слова, которые пишутся через дефис должны быть одним токеном, но постфиксы могут писаться отдельно, как незначительная часть слова), 2 – «Не ветер, а какой-то ураган!» (проверка распознавания знаков препинания), 3 – «Я - к.т.н., живу в США.» (проверка на распознавание аббревиатур), 4 – «В Нью-Йорке (США) хорошие маффины стоят $3.88.» (проверка на распознавание цифр, чисел с плавающей точкой).

Расшифровка «Частоты токенов»:

* по итогу первого эксперимента «+» получают инструменты, которые распознают слова, пишущиеся через дефис, как одно слово, при этом постфиксы могут отделяться от основной части слова. Остальные инструменты получают «-»;
* по итогу второго эксперимента «+» получают инструменты, распознающие знаки препинания как отдельные токены или исключающие их из списка токенов сразу. Остальные инструменты получают «-»;
* по итогу третьего эксперимента «+» получают инструменты, распознающие обе аббревиатуры («к.т.н.» и «США») как токены. Остальные инструменты получают «-»;
* по итогу четвертого эксперимента «+» получают инструменты, распознающие число с плавающей точкой как один токен. Остальные инструменты получают «-».

По количеству плюсов в графе «Чистота токенов»:

* 1 «+» у Gensim, Scikit-learn, Keras;
* 2 «+» у **RegexpTokenizer, Re**;
* 3 «+» у WhitespaceTokenizer, SpaceTokenizer, TweetTokenizer, Rutokenizer;
* 4 «+» у ToktokTokenizer, NLTKWordTokenizer и TextBlob.

Токенизатор Pattern не получилось проверить, из-за его несовместимости с новыми версиями Python (проверка осуществлялась с версией 3.10).

SpaCy и **RegexpTokenizer** исключается из сравнения по причине сравнительно большого времени работы, превышающее среднее время работы других инструментов более, чем в 50 и 10 раз соответственно.

WhitespaceTokenizer и SpaceTokenizer исключаются из сравнения по причине того, что в результатах их работы к токенам «приклеиваются» знаки препинания. В будущем это помешает правильному распределению признаков и может привести к неверно выполненной задаче основного алгоритма.

Инструменты Gensim, Scikit-learn, Keras исключаются из сравнения из-за того, что у них из четырех возможных всего один плюс.

Оставшиеся инструменты были ранжированы по времени работы (см. табл. 4):

Таблица 4 – Оставшиеся инструменты токенизации на униграммы

|  |  |  |
| --- | --- | --- |
| Инструменты | Время работы, с | Количество «+» |
| ToktokTokenizer | 0,0004 | 4 |
| TweetTokenizer | 0,0088 | 3 |
| NLTKWordTokenizer | 0,0090 | 4 |
| Re | 0,0010 | 2 |
| Rutokenizer | 0,0199 | 3 |
| TextBlob | 0,1300 | 4 |

Из оставшихся инструментов у TextBlob самое большое время работы, а у Re самое маленькое количество «+», поэтом у они исключаются из сравнения. У TweetTokenizer и NLTKWordTokenizer время работы отличается всего на 0,0002 с, но при этом у NLTKWordTokenizer больше плюсов, поэтому TweetTokenizer выбывает из сравнения.

Все три оставшихся инструмента (ToktokTokenizer, NLTKWordTokenizer, Rutokenizer) являются достойными, однако здесь, учитывая время работы и количество «+», можно однозначно распределить между ними призовые места:

* + - 1. ToktokTokenizer
      2. NLTKWordTokenizer
      3. Rutokenizer

Далее (см. табл. 6) инструменты токенизации на 2 и 3-граммы сравниваются по наличию возможностей дополнительной обработки токенов и по времени токенизации.

Однако в библиотеке NLTK содержатся несколько подходящих модулей. Их свойства исследуются в таблице ниже (см. табл. 5). Полная таблица с проведением проверки времени предложена в Приложении Б.

Таблица 5 – Модули NLTK для токенизации на 2 и 3-граммы

|  |  |  |
| --- | --- | --- |
| Инструменты | n | Среднее время работы инструмента, с |
| Bigram | 2 | 0.00041 |
| Trigram | 3 | 0.00052 |
| Ngrams | 2 | 0.00060 |
| 3 | 0.00055 |

Для проверки инструментов NLTK внутри них использовался токенизатор для униграмм ToktokTokenizer, являющийся лучшим по итогам предыдущего исследования.

Из таблицы 5 видно, что для n = 2 более оптимальным будет использование инструмента Bigram, а для n = 3 – Trigram. При этом результат декомпозиции текста сравниваемых инструментов одинаков. Исходя из данного исследования, в таблицу 6 для модуля NLTK введены данные Bigram и Trigram.

Таблица 6 – Итоговая таблица сравнения для токенизации на 2 и 3-граммы

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Инструменты | Критерии | | | | | | |
| Время работы, с | | Дополнительные шаги внутри токенизатора | Лемматизация | Распознавание частей речи | Распознавание именованных сущностей | Распознавание зависимостей слов |
| n = 2 | n = 3 |
| Nltk | 0.00041 | 0.00052 | - | + | + | + | + |
| SpaCy | 0.00060 | 0.00043 | + | + | + | + | + |
| Pattern | 2.33229 | 2.46947 | + | + | + | - | - |
| Gensim | --- | --- | + | + | - | - | - |
| Scikit-learn | 0.00101 | 0.00177 | + | + | - | - | - |
| TextBlob | 0.00966 | 0.00895 | + | + | + | + | + |

Инструмент Gensim имеет инструмент распознавания n-грамм, но в его документации под n-граммами понимаются значимые словосочетания, такие как «искусственный интеллект», «машинное обучение» и так далее. Для русского языка данный инструмент пока не обучен до уровня, чтобы его можно было применять в прикладной разработке, в отличие от исследовательской. Данный инструмент выбывает из сравнения.

Необходимый шаг для работы алгоритма поиска ключевых слов, следующего за алгоритмом декомпозиции, – лемматизация, есть у всех кандидатов. Остальные распознаватели (частей речи, именных сущностей, зависимостей слов) понадобятся для оптимизации предобработки текста, поэтому в дальнейшее сравнение были взяты те инструменты, в которых присутствуют данные распознаватели. На данном шаге выбыли Scikit-learn и Pattern.

При рассмотрении колонки со временем, инструмент Pattern снова выделяется, и это вторая причина его исключения из сравнения. Напротив Pattern можно увидеть сильно отличающиеся от остальных в худшую сторону значения времени.

При сравнении времени работы оставшихся инструментов получается следующее ранжирование при n = 2: 1 место – NLTK, 2 место – SpaCy, 3 место – TextBlob; при n = 3: 1 место – SpaCy, 2 место – NLTK, 3 место – TextBlob.

Что касается дополнительных шагов внутри токенизатора, то в NLTK они отсутствуют (их нужно подключать и настраивать отдельно), в SpaCy они подключены сразу (сюда входит фильтрация стоп-слов, лишних знаков препинания), в TextBlob есть фильтрация второстепенных знаков препинания.

Таким образом, по итогам анализа таблицы 6 в SpaCy самая хорошая обработка (в сравнении с аналогами) в совокупности с самым малым временем работы.

Далее (см. табл. 8) инструменты токенизации на 1-2-3-граммы сравниваются по наличию возможностей дополнительной обработки токенов и по времени токенизации.

В библиотеке NLTK содержатся несколько подходящих модулей. Их свойства исследуются в таблице ниже (см. табл. 7). Полная таблица с проведением проверки времени предложена в Приложении Г.

Таблица 7 – Модули NLTK для токенизации на 1-2-3-граммы

|  |  |
| --- | --- |
| Инструменты | Среднее время работы инструмента, с |
| Unigram + Bigram + Trigram | 0.00080 |
| Ngrams | 0.00093 |
| Everygrams | 0.00079 |

Для проверки инструментов NLTK внутри них использовался токенизатор для униграмм ToktokTokenizer, являющийся лучшим по итогам предыдущего исследования.

Из таблицы 7 видно, что более оптимальным будет использование инструмента Everygrams. При этом результат декомпозиции текста сравниваемых инструментов одинаков. Исходя из данного исследования, в таблицу 8 для модуля NLTK введены данные Everygrams.

Таблица 8 – Итоговая таблица сравнения для токенизации на 1-2-3-граммы

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Инструменты | Критерии | | | | | |
| Время работы, с | Дополнительные шаги внутри токенизатора | Лемматизация | Распознавание частей речи | Распознавание именованных сущностей | Распознавание зависимостей слов |
| Nltk | 0.00079 | - | + | + | + | + |
| SpaCy | 0.00081 | + | + | + | + | + |
| Pattern | 2.45547 | + | + | + | - | - |
| Gensim | --- | + | + | - | - | - |
| Scikit-learn | 0.00178 | + | + | - | - | - |
| TextBlob | 0.01079 | + | + | + | + | + |

В данном случае инструмент Gensim выбывает из сравнения по тем же причинам, что и в предыдущий раз (пока не готов к применению в прикладной разработке).

Наличие и отсутствие различных распознавателей и лемматизатора в сравниваемых инструментах также описано внутри анализа таблицы 6, поэтому можно сразу сказать, что из сравнения исключаются инструменты Scikit-learn и Pattern.

При сравнении времени работы оставшихся инструментов получается следующее ранжирование при n = 2: 1 место – NLTK, 2 место – SpaCy, 3 место – TextBlob. При рассмотрении дополнительных шагов внутри токенизатора, влияющих на качество результата, видно, что в NLTK они отсутствуют, однако при практически таком же времени работы SpaCy фильтрует стоп-слова и лишние знаки препинания. TextBlob при значительно большем времени работы, чем SpaCy, имеет меньше дополнительной фильтрации (здесь она направлена только на знаки пунктуации).

Таким образом, по итогам анализа таблицы 8 в SpaCy самая хорошая обработка текста (в сравнении с аналогами) в совокупности с самым малым временем работы.

В результате обзора и сравнения инструментов декомпозиции текста были выбраны два из них в качестве базовых для дальнейшей реализации разрабатываемых алгоритмов. В качестве первого инструмента (для токенизации на униграммы) был выбран ToktokTokenizer из библиотеки NLTK, отлично прошедший тестирование качества результатов декомпозиции и имеющий лучшее время работы среди аналогов, также хорошо прошедших тестирование качества. В качестве второго (для токенизации на 2-, 3-граммы и для смешанной токенизации – на 1-2-3-граммы) – SpaCy, из-за улучшенной декомпозиции за счет дополнительных процедур обработки текста, а также из-за отличного времени работы в сравнении с аналогами.

* 1. **Выводы**

В результате обзора предметной области были определены основные термины, необходимые для понимания рассматриваемой темы. При рассмотрении существующих аналогов для решения задачи декомпозиции были определены: реализуемые процедуры обработки текста, наличие токенизации на слова и словосочетания, наличие поддержки обработки русского языка и автоматизированной обработки, временные затраты на применение и качество результатов относительно поставленных требований. Два инструмента (ToktokTokenizer для униграмм и SpaCy для n-грамм) были выбраны в качестве базовых для реализации алгоритмов декомпозиции. Таким образом, было проведено сравнение и анализ существующих инструментов декомпозиции текста.

1. **ПОСТАНОВКА ЗАДАЧИ**

Исходя из обзора предметной области и анализа проблемы, описанной во введении, принято решение о создании веб-приложения.

Программное решение должно обладать следующими свойствами:

* извлечение терминов из содержимого разделов русскоязычных курсов;
* наличие для каждого найденного термина определения;
* наличие интерфейса для взаимодействия пользователя с приложением;
* наличие интерфейса для взаимодействия сервера приложения с базой данных;
* наличие интерфейса для взаимодействия сервера приложения с платформой Stepik.

Уже не раз отмечалось, что наличие ключевых словосочетаний (фраз) помогает более точно определить тематику текста и, соответственно, улучшить его классификацию/категоризацию. Поэтому в данной работе будут реализованы алгоритмы семантической декомпозиции для выделения как ключевых слов, так и ключевых словосочетаний, чтобы в последствии сравнить результаты и сделать выводы. (думаю сделать один для ключевых слов, второй для словосочетаний и третий гибридный)

Исходя из обзора предметной области и анализа проблемы, описанной во введении, созданы критерии к разрабатываемому алгоритму.

Алгоритм должен решать следующие проблемы:

* основные:

1. декомпозиция текста на ключевые элементы;
2. минимально возможное время работы, не превышающее 10 секунд;
3. токенами являются слова, состоящие из букв или цифр. Знаки пунктуации и служебные символы токенами не являются.

* второстепенные:

1. декомпозиция может осуществляться по униграммам и/или по n-граммам
2. алгоритм может содержать вспомогательные процедуры обработки токенов: лемматизацию, удаление стоп слов,

<https://proglib.io/p/fun-nlp>

1. **Описание Теоретических основ разрабатываемых алгоритмов**

Если рассмотреть общую структуру алгоритма категоризации (см. рис. Такой-то), можно понять, что алгоритм семантической декомпозиции будет находиться в самом его начале, среди процедур предобработки текста.

Рисунок такой-то

Общая схема построения трех алгоритмов представлена на рис. Таком-то

Рисунок такой-то

(текст поступает на вход, он отделяется от пунктуации и пробелов, делится на отдельные слова, слова приводятся к начальной форме, удаляются стоп-слова, удаляются короткие слова, удаляются цифры, удаляются имена(?), удаляются слова, встречающиеся малое количество раз, на выходе получаем список токенов (отдельных слов) в начальной форме, претендующих на звание ключевых слов)

Первым этапом работы алгоритма семантической декомпозиции является

* 1. **Алгоритм семантической декомпозиции на ключевые слова**

Перед проектированием веб-приложения необходимо понять, из чего оно должно состоять.

* + 1. **Графическое описание алгоритма**

Схема построения алгоритма представлена на рис. Таком-то

Рисунок такой-то

(текст поступает на вход, он отделяется от пунктуации и пробелов, делится на отдельные слова, слова приводятся к начальной форме, удаляются стоп-слова, удаляются короткие слова, удаляются цифры, удаляются имена(?), удаляются слова, встречающиеся малое количество раз, на выходе получаем список токенов (отдельных слов) в начальной форме, претендующих на звание ключевых слов)

* + 1. **Математические основы алгоритма**

дпдшмлм

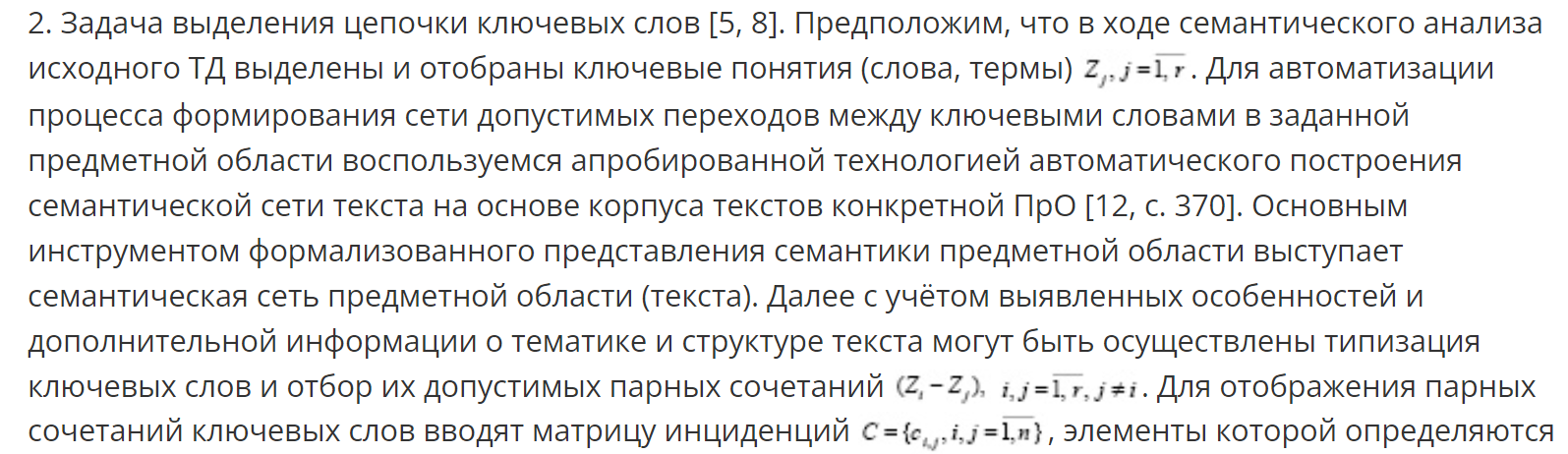
* + 1. **Сценарий применения алгоритма для решения задачи категоризации**

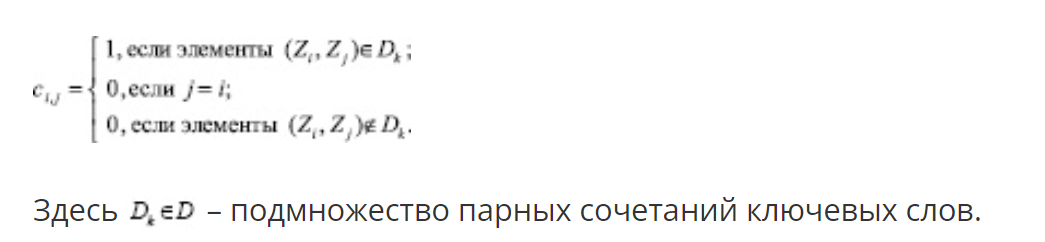
Текст поступает на вход в систему, далее он проходит через алгоритм выделения ключевых слов (внутри которого предобработка + алгоритм семантической декомпозиции + основной этап алгоритма выделения КС), далее ключевые слова с весами отправляются на вход следующему алгоритму – алгоритму категоризации.

* 1. **Алгоритм семантической декомпозиции на ключевые словосочетания**
     1. **Математические основы алгоритма**

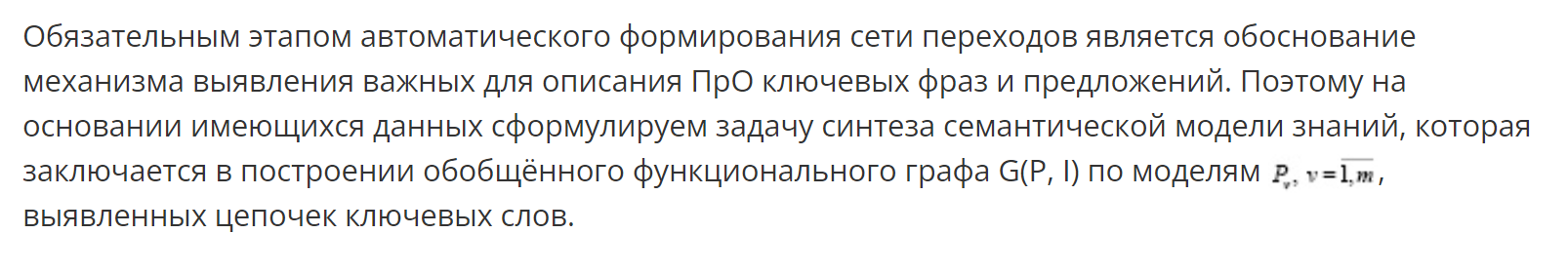
Дпдшмлм

Задача выделения цепочки ключевых слов [5, 8]. Предположим, что в ходе семантического анализа исходного ТД выделены и отобраны ключевые понятия (слова, термы) . Для автоматизации процесса формирования сети допустимых переходов между ключевыми словами в заданной предметной области воспользуемся апробированной технологией автоматического построения семантической сети текста на основе корпуса текстов конкретной ПрО [12, с. 370]. Основным инструментом формализованного представления семантики предметной области выступает семантическая сеть предметной области (текста). Далее с учётом выявленных особенностей и дополнительной информации о тематике и структуре текста могут быть осуществлены типизация ключевых слов и отбор их допустимых парных сочетаний  . Для отображения парных сочетаний ключевых слов вводят матрицу инциденций , элементы которой определяются по правилу:

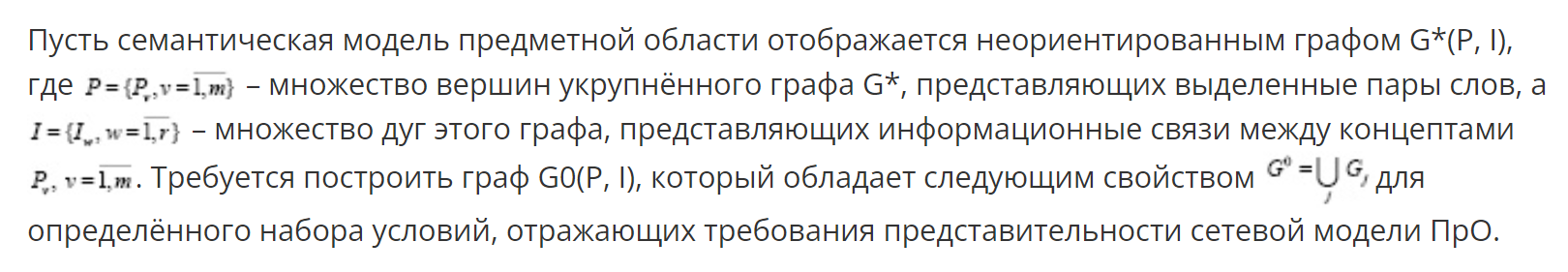
по правилу:



Здесь  – подмножество парных сочетаний ключевых слов.



Обязательным этапом автоматического формирования сети переходов является обоснование механизма выявления важных для описания ПрО ключевых фраз и предложений. Поэтому на основании имеющихся данных сформулируем задачу синтеза семантической модели знаний, которая заключается в построении обобщённого функционального графа G(P, I) по моделям , выявленных цепочек ключевых слов.



Пусть семантическая модель предметной области отображается неориентированным графом G\*(P, I), где  – множество вершин укрупнённого графа G\*, представляющих выделенные пары слов, а  – множество дуг этого графа, представляющих информационные связи между концептами . Требуется построить граф G0(P, I), который обладает следующим свойством для определённого набора условий, отражающих требования представительности сетевой модели ПрО.

Для решения задачи 2 можно воспользоваться методом, изложенным в авторской статье [8]. После выбора ключевых слов и значимых предложений текста на основе этих предложений строится ассоциативная сеть, которая является основой для идентификации сети переходов между ключевыми словами для этой группы предложений. Сеть, построенная на всех предложениях корпуса текстов, описывающего предметную область с рангом выше порогового, является сетью переходов между ключевыми словами для всей ПрО.

* + 1. **Графическое описание алгоритма**

дпдшмлм

* + 1. **Сценарий применения алгоритма для решения задачи категоризации**

Идпмдлид

* 1. **Алгоритм семантической декомпозиции на ключевые слова и словосочетания**
     1. **Математические основы алгоритма**

дпдшмлм

* + 1. **Графическое описание алгоритма**

дпдшмлм

* + 1. **Сценарий применения алгоритма для решения задачи категоризации**

В результате предварительной обработки будет составлен словарь, содержащий как отдельные слова, так и ключевые фразы. Элементы словаря далее будут называться терминами.

* 1. **Выводы**

адплпщжиди

1. **ПРОГРАММНАЯ РЕАЛИЗАЦИЯ АЛГОРИТМОВ**
   1. **Алгоритм семантической декомпозиции на n-граммы**
      1. **Технические подробности программной реализации алгоритма**

Лалалалал

NLTK (Natural Language Toolkit) – ведущая платформа для создания NLP-программ на Python. У нее есть легкие в использовании интерфейсы для многих [языковых корпусов](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D1%80%D0%BF%D1%83%D1%81_%D1%82%D0%B5%D0%BA%D1%81%D1%82%D0%BE%D0%B2), а также библиотеки для обработки текстов для классификации, токенизации, [стемминга](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%B5%D0%BC%D0%BC%D0%B8%D0%BD%D0%B3), [разметки](https://ru.wikipedia.org/wiki/%D0%A7%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%B5%D1%87%D0%BD%D0%B0%D1%8F_%D1%80%D0%B0%D0%B7%D0%BC%D0%B5%D1%82%D0%BA%D0%B0), фильтрации и [семантических рассуждений](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BC%D0%B0%D0%BD%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9_%D0%BC%D0%B5%D1%85%D0%B0%D0%BD%D0%B8%D0%B7%D0%BC_%D1%80%D0%B0%D1%81%D1%81%D1%83%D0%B6%D0%B4%D0%B5%D0%BD%D0%B8%D0%B9). Ну и еще это бесплатный опенсорсный проект, который развивается с помощью коммьюнити. (<https://habr.com/ru/company/Voximplant/blog/446738/>)

NLTK-это сокращение от **Natural Language ToolKit** . Это библиотека, написанная на Python для символьной и статистической обработки естественного языка. (https://pythobyte.com/tokenization-in-python-using-nltk-96642092/)

* + 1. **Тестирование алгоритма**

лалалалал

* + 1. **Сценарий использования**

лалалалал

* + 1. **Применение алгоритма для решения задачи категоризации**

лалалалал

* 1. **Алгоритм семантической декомпозиции на предложения**
     1. **Технические подробности программной реализации алгоритма**

Лалалалал

Для реализации веб-приложения был выбран язык Python версии 3.7. Решение использовать Python для реализации решения было принято потому, что данный язык программирования является универсальным языком, применяемым в работе с системами, обрабатывающими HTTP-запросы, с хранилищами баз данных, с шаблонами веб-страниц, с анализом данных и со многим другим [12]. Ниже дается описание основных библиотек и фреймворков, задействованных при проектировании веб-приложения

* + 1. **Тестирование алгоритма**

лалалалал

* + 1. **Сценарий использования**

Лалалала

После запуска веб-приложения работа с ним может проводиться по следующему сценарию:

1. На начальной странице пользователь вводит ID курса.
2. Веб-приложение загружает список разделов курса в порядке, в котором они находятся в МООК.
3. Пользователь выбирает интересующий раздел, кликая по нему.
4. На экране появляются три группы по 10 терминов из выбранного раздела в порядке от слова с самой большой вероятностью того, что оно является термином в данном разделе, к слову с наименьшим значением вероятности. Во главе каждой группы терминов выводится название метода, которым было проведено их выделение.
5. Пользователь кликает на термин.
6. Под выбранным термином появляется его определение.
   * 1. **Применение алгоритма для решения задачи категоризации**

лалалалал

* 1. **Выводы**

Результатом разработки является готовое веб-приложение для генерации оглавления и списка терминов для МООК, в котором были реализованы алгоритмы выделения ключевых слов Frequency, TextRank и TopicRank. С помощью данных алгоритмов приложение проводит анализ русскоязычных курсов Stepik, в результате которого из содержимого разделов МООК извлекаются термины. К терминам подбираются определения из Википедии. Приложение интегрировано со Stepik API и с PostgreSQL и имеет пользовательский интерфейс.

1. **исследование**
   1. **Исследование времени работы**

Было проведено исследование времени работы алгоритмов Frequency, TextRank и Topic Rank.

* + 1. **Сценарий эксперимента**

Измеряемой величиной является время, затраченное на работу каждого из алгоритмов. Единица измерения – секунды.

Измерение проводилось путем фиксирования времени начала выполнения алгоритма и окончания его выполнения, затем найденная разность данных значений приравнивалась ко искомому значению затраченного времени.

Процедура проводилась 100 раз для каждого эксперимента.

* + 1. **Методика измерения**

Измерение происходит машине со следующими характеристиками:

* операционная система – Windows 10 Pro;
* системные характеристики:
  + процессор – Intel Core i3-7100U CPU 2.4GHz (2 ядра);
  + ОЗУ – 8Гб;
  + тип системы – х64.

Эксперимент проводился для текстового содержимого раздела «Функции. Словари. Интерпретатор. Файлы. Модули» курса «Программирование на Python» (ID 67).

* + 1. **Используемые программные средства**

Эксперимент проводился с использованием средств Python, а именно с помощью библиотеки time и ее метода monotonic() [32].

* + 1. **Результаты эксперимента**

Было рассчитано минимальное, максимальное и среднее время выполнения каждого алгоритма. Полученные данные, приведенные в секундах с точностью до тысячных, представлены в табл. 2.

Таблица 2 – Обработанные данные (для алгоритмов)

|  |  |  |  |
| --- | --- | --- | --- |
| Данные | Алгоритмы | | |
| Frequency | TopicRank | TextRank |
| Минимальное время работы | 8,032 | 5,892 | 7,504 |
| Среднее время работы | 8,761 | 6,556 | 8,165 |
| Максимальное время работы | 13,100 | 11,014 | 12,563 |

Таблица с необработанными данными представлена в прил. Б.

* + 1. **Сравнение и оценка результатов**

Как видно из расчетов, разность между максимальным и средним временем работы гораздо больше, чем разность между средним и минимальным временем, что говорит о том, что итерации, показавшие сравнимое с максимальным время, достигалось малое количество раз. Данную гипотезу можно проверить, ознакомившись с таблицей в приложении Б. Подобное поведение может быть связано с внешними факторами, такими как скорость интернет-соединения, время ответа сервера Stepik.

Сравнивая среднее время работы алгоритмов, видно, что лучшее время работы показывает TopicRank, алгоритм TextRank оказывается на втором месте, а Frequency на третьем. Все три алгоритма реализованы с применением различных готовых решений, имеющихся в библиотеках обработки текстов, что объясняет разную скорость работы алгоритмов. Работа методов таких библиотек не отличается быстродействием, что объясняет невысокую скорость выполнения алгоритмов.

* 1. **Исследование точности результатов**

В рассмотрении задач выделения ключевых слов большую роль играет точность их выделения.

* + 1. **Сценарий эксперимента**

В ходе эксперимента запускается веб-приложение, результатом чего становится получение трех списков с терминами. Далее слова в каждом из списков сравниваются со словами эталонного списка, состоящего из такого же количества ключевых слов, таким образом выделяется точность алгоритма в процентах.

Под эталонным списком понимается список отобранных вручную ключевых слов из того же документа.

В списках терминов и в эталонном списке содержится по 10 слов.

* + 1. **Входные данные**

Эксперимент проводится на базе текстового содержимого раздела «Введение» курса «Как писать научные статьи» (ID 10524).

* + 1. **Результат работы алгоритмов**

Результатом работы алгоритмов стали следующие ключевые слова:

1. Frequency:
   * научный;
   * статья;
   * работа;
   * курс;
   * комментарий;
   * исследование;
   * результат;
   * источник;
   * текст;
   * пример;
2. TopicRank:
   * научный;
   * работа;
   * курс;
   * комментарий;
   * исследование;
   * результат;
   * источник;
   * текст;
   * список;
   * пример;
3. TextRank:
   * метод;
   * явно;
   * вывод;
   * введение;
   * аннотация;
   * слово;
   * state;
   * анализ;
   * название;
   * совет.
     1. **Экспертная оценка**

Эталонный список содержит следующие ключевые слова:

* научный;
* статья;
* исследование;
* обоснованность;
* анализ;
* аннотация;
* курс;
* автор;
* публикация;
* источник.

Словосочетание «научная статья» было разделено ввиду того, что алгоритмы ведут поиск юниграмм – N-грамм, состоящих из одного слова.

* + 1. **Сравнение с экспертной оценкой**

В ходе сравнения было вычислено отношение числа совпавших слов к общему количеству слов в списке. Результаты сравнения представлены в табл. 3.

Таблица 3 – Сравнение с экспертной оценкой

|  |  |  |  |
| --- | --- | --- | --- |
| Данные | Алгоритмы | | |
| Frequency | TopicRank | TextRank |
| Точность, % | 50 | 50 | 20 |

Согласно таблице 3, алгоритм TextRank не применим к задаче выделения ключевых слов, так как его точность составляет всего 20%. Это можно объяснить тем, что данный алгоритм основан на алгоритме PageRank, для построения графа которого используются связи между элементами. В данной же реализации алгоритма на вход подается список уже отделенных друг от друга слов, что мешает работе алгоритма.

В связи с этим же фактом, алгоритм TopicRank, показавший средние результаты (50%), может улучшить свой результат при иной реализации этапа предобработки.

Алгоритм Frequency, показавший также средний результат (50%), тоже можно улучшить с помощью использования N-грамм. Однако, даже в этом случае, он с малой вероятностью покажет стопроцентный результат, так как не предназначен для специализированных текстов.

* 1. **Выводы**

Таким образом, было проведено исследование свойств программного продукта, таких как время работы алгоритмов, выполняющих выделение терминов из текста МООК, и их точность. В ходе исследования было выявлено, что алгоритм TopicRank быстрее всех справляется с задачей (6,556 с.) и имеет среднюю точность (50%); алгоритм TextRank на втором месте по времени (8,165 с.), но на последнем по точности (20%); Frequency на третьем месте по времени (8,761 с.) и имеет такую же точность, как и TopicRank – 50%.

1. **Коммерциализация результатов исследований в рамках НИР**

Здесь о том, что разработанный алгоритм до настоящего времени не внедрялся в программные продукты организации, но в будущем может быть использован в СППР для службы технической поддержки ООО ГП.

Пример: «Разработанное решение до настоящего момента не внедрялось в организации, но было опубликовано на Docker Hub. Docker Hub – это общедоступный реестр, поддерживаемый Docker, хранящий большую коллекцию Docker-образов».

# **заключение**

В результате работы над выпускной квалификационной работой были изучены походы к задаче выделения ключевых слов, выполнено проектирование программного продукта, представляющее собой веб-приложение, автоматически генерирующее оглавление и указатели терминов для массовых открытых онлайн-курсов платформы Stepik.

В процессе обзора предметной области были определены основные термины рассматриваемой области, дано описание подходов выделения ключевых слов, проанализированы области их применения, временные затраты на применение, наличие требований использования специфических средств для работы и особенности подходов. Проведен выбор двух подходов, реализуемых в дальнейшем в веб-приложении. Выбранными подходами стали статистический и графовый.

На основании обзора предметной области и анализа проблемы, описанной во введении, были выдвинуты требования к разрабатываемому решению. Учитывая данные требования, была достигнута основная задача работы – спроектирован программный продукт для автоматического генерирования оглавления и указателей терминов для русскоязычных массовых открытых онлайн-курсов. Разработанное решение представлено веб-приложением на языке программирования Python, с использованием веб-фреймворка Flask и СУБД PostgreSQL. Полученное программное решение имеет пользовательский интерфейс. Приложение интегрируется платформой Stepik с помощью REST-интерфейса Stepik API, а с хранилищем данных с помощью библиотеки **SQLAlchemy**. Посредством анализа алгоритмами Frequency, TopicRank и TextRank из выбранного раздела МООК выделяются термины, для которых имеются определения.

Было проведено исследование свойств программного решения: исследование времени работы алгоритмов выделения ключевых слов и их точности. По результатам исследования быстродействия алгоритмов самым быстрым оказался алгоритм TopicRank (6,556 с.), а самым медленным – Frequency (8,761 с.). Это объясняется использованием готовых решений для обработки текста, что сильно влияет на скорость работы алгоритмов. В результате исследования точности, ни один алгоритм не показал отличных результатов из-за неподходящего для них сценария реализации этапа предобработки (TopicRank, 50%; TextRank, 20%) или из-за неподходящего проверяемому алгоритму типа текста (Frequency, 50%).

В будущем данную проблем, а также проблему малой скорости работы планируется решить для алгоритмов, потенциально подходящих для задачи выделения терминов из содержимого МООК, то есть для графовых алгоритмов TopicRank и TextRank.

Кратко (на одну – две страницы) описать основные результаты работы, проанализировать их соответствие поставленной цели работы, показать рекомендации по конкретному использованию результатов исследования и перспективы дальнейшего развития работы.

Научная ценность данной работы заключается в улучшении качества результатов проекта автоматизации службы технической поддержки компании ООО ГПП.

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Что такое МООС (Массовые открытые онлайн курсы) [Электронный ресурс]. URL: https://etu.ru/ru/on-line-obuchenie/mooc (дата обращения: 15.05.2020).
2. Чекалина Т. А., Тумандеева Т. В., Максименко Н. В. / Основные направления и перспективы развития онлайн-обучения // Профессиональное образование в России и за рубежом. – 2018. – (№) 3 (31). – С. 44-52. [Электронный ресурс] URL: https://cyberleninka.ru/article/n/osnovnye-napravleniya-i-perspektivy-razvitiya-onlayn-obucheniya/viewer (дата обращения: 15.05.2020).
3. Ванюшкин А. С., Гращенко Л. А. / Методы и алгоритмы извлечения ключевых слов // Новые информационные технологии в автоматизированных системах. – 2016. – (№) 19. – С. 85-93. [Электронный ресурс]. URL: https://cyberleninka.ru/article/n/metody-i-algoritmy-izvlecheniya-klyuchevyh-slov/viewer (дата обращения: 24.05.2020).
4. Manning C. D., Manning C. D., Schütze H. Foundations of statistical natural language processing. – MIT press, 1999. 680p.
5. Лазарева О. Ю., Боломутова М. С. / Методы выделения ключевых слов в контексте электронных обучающих систем // Молодой ученый. – 2016. – (№) 26 (130). С. 143-146. [Электронный ресурс] URL: https://moluch.ru/archive/130/35952/ (дата обращения: 16.05.2020).
6. Лекция. Задачи Data Mining. Классификация и кластеризация // НОУ ИНТУИТ. [Электронный ресурс] URL: https://www.intuit.ru/studies/courses/6/6/lecture/166 (дата обращения: 16.05.2020).
7. Keyword Extraction: A Guide to Finding Keywords in Text // MonkeyLearn. [Электронный ресурс] URL: https://monkeylearn.com/keyword-extraction/ (дата обращения: 17.05.2020).
8. Girish K.P. Keyword Extraction from a Single Document Using Centrality Measures / Pattern Recognition and Machine Intelligence, Kolkata, India, December 18–22, 2007 / Kolkata, 2007, pp. 503–510.
9. Litvak M. Graph-based Keyword Extraction for Single-Document Summarization / Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization, Manchester, United Kingdom, August 2008 / Manchester, 2008, pp. 17–24.
10. Григорьева Е. Г., Клячин В. А., Помельников Ю. В., Попов В. В. / Алгоритм выделения ключевых слов на основе графовой модели лингвистического корпуса // Вестник ВолГУ. Серия 2, Языкознание. – 2017. – Т. 16 № 2. – С. 58-67. [Электронный ресурс] URL: https://cyberleninka.ru/article/n/algoritm-vydeleniya-klyuchevyh-slov-na-osnove-grafovoy-modeli-lingvisticheskogo-korpusa/viewer (дата обращения: 24.05.2020).
11. Лекция. Принципы работы и структура Web-приложений на основе ASP.NET // НОУ ИНТУИТ [Электронный ресурс] URL: https://www.intuit.ru/studies/courses/1139/250/lecture/6422 (дата обращения: 19.05.2020).
12. Welcome to Python.org // Python [Электронный ресурс] URL: https://www.python.org/ (дата обращения: 20.05.2020).
13. О нас // Stepik [Электронный ресурс] URL: https://welcome.stepik.org/ru/about (дата обращения: 20.05.2020).
14. StepikOrg/Stepik-API: API documentation and Examples // GitHub [Электронный ресурс] URL: https://github.com/StepicOrg/Stepik-API (дата обращения: 20.05.2020).
15. Welcome to Flask // Flask Documentation (1.1.x) [Электронный ресурс] URL: http://flask.pocoo.org/ (дата обращения: 20.05.2020).
16. SQLAlchemy – The Database Toolkit for Python // SQLAlchemy [Электронный ресурс] URL: https://www.sqlalchemy.org/ (дата обращения: 20.05.2020).
17. PostgreSQL: The world’s most advanced open source database // PostgreSQL [Электронный ресурс] URL: https://www.postgresql.org/ (дата обращения: 20.05.2020).
18. Collections – Container datatypes // Python 3.8.3 documentation [Электронный ресурс] URL: https://docs.python.org/3/library/collections.html (дата обращения: 21.05.2020).
19. Boudinfl/pke: Python Keyphrase Extraction module // GitHub [Электронный ресурс] URL: https://github.com/boudinfl/pke (дата обращения: 21.05.2020).
20. Pke documentation // pke 1.8 documentation [Электронный ресурс] URL: https://boudinfl.github.io/pke/build/html/index.html (дата обращения: 23.05.2020).
21. Docker Documentation // Docker Documentation. [Электронный ресурс] URL: https://docs.docker.com/ (дата обращения: 20.05.2020).
22. Заглавная страница // Википедия – свободная энциклопедия. [Электронный ресурс] URL: https://ru.wikipedia.org/wiki/Заглавная\_страница (23.05.2020).
23. Русская Википедия // Википедия. [Электронный ресурс] URL: https://ru.wikipedia.org/wiki/Русская\_Википедия (дата обращения: 23.05.2020).
24. Bougouin A. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction / International Joint Conference on Natural Language Processing, Nagoya, Japan, 14-18 October 2013 / Nagoya, 2013, pp 543–551.
25. Unsupervised models (TopicRank) // pke 1.8 documentation [Электронный ресурс] URL: https://boudinfl.github.io/pke/build/html/unsupervised.html#topicrank (дата обращения: 23.05.2020).
26. Mihalcea R. TextRank: Bringing Order into Texts / EMNLP, July 2004 / EMNLP, July 2004, № 32.
27. Unsupervised models (TextRank) // pke 1.8 documentation [Электронный ресурс] URL: https://boudinfl.github.io/pke/build/html/unsupervised.html#textrank (дата обращения: 23.05.2020).
28. Алгоритм PageRank // Data Science [Электронный ресурс] URL: http://datascientist.one/pagerank-algorithm/ (дата обращения: 23.05.2020).
29. Docker Hub // Docker Hub [Электронный ресурс] URL: https://hub.docker.com/repository/docker/vika2108991/vkr (дата обращения 24.05.2020).
30. Moevm/bsc\_tereshchenko // GitHub [Электронный ресурс] URL: https://github.com/moevm/bsc\_tereshchenko (дата обращения: 24.05.2020).
31. Разработка модели алгоритма составления оглавления и указателя терминов для массовых открытых онлайн курсов. / Терещенко В. Н., Заславский М. М. // 73-я научно-техническая конференция профессорско-преподавательского состава СПбГЭТУ «ЛЭТИ», СПб, 5 февраля 2020 г. / СПбГЭТУ «ЛЭТИ», СПб, 2020.
32. 15.3. time – Time access and conversions // Python 2.7.18 documentation. [Электронный ресурс] URL: https://docs.python.org/2/library/time.html (дата обращения: 24.05.2020).

Приложение А

Таблица – Время работы токенизаторов униграмм

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Инструменты | Измерения времени, с | | | | | Среднее значение, с |
| 1 | 2 | 3 | 4 | 5 |
| **RegexpTokenizer** | 1,8858 | 1,8439 | 1,8160 | 1,8434 | 1,8104 | 1,8399 |
| WhitespaceTokenizer | 0,0 | 0,0 | 0,0 | 0,0 | 0,0010 | 0,0002 |
| SpaceTokenizer | 0,0 | 0,0 | 0,0010 | 0,0 | 0,0 | 0,0002 |
| ToktokTokenizer | 0,0 | 0,0010 | 0,0010 | 0,0 | 0,0 | 0,0004 |
| TweetTokenizer | 0,0080 | 0,0090 | 0,0090 | 0,0090 | 0,0090 | 0,0088 |
| NLTKWordTokenizer | 0,0080 | 0,0090 | 0,0090 | 0,0100 | 0,0090 | 0,0090 |
| Re | 0,0010 | 0,0010 | 0,0010 | 0,0010 | 0,0010 | 0,0010 |
| SpaCy | 5,7478 | 5,5067 | 5,6582 | 5,5220 | 5,5408 | 5,5951 |
| Pattern | - | - | - | - | - | - |
| Gensim | 0,1318 | 0,1333 | 0,1315 | 0,1230 | 0,1340 | 0,1307 |
| Scikit-learn | 0,0010 | 0,0010 | 0,0020 | 0,0010 | 0,0010 | 0,0012 |
| Keras | 0,0040 | 0,0040 | 0,0040 | 0,0030 | 0,0040 | 0,0038 |
| Rutokenizer | 0,0200 | 0,0203 | 0,0200 | 0,0200 | 0,0190 | 0,0199 |
| TextBlob | 0,1300 | 0,1298 | 0,1290 | 0,1300 | 0,1310 | 0,1300 |

Тестирование времени проводилось на предложении «Не ветер, а какой-то ураган!»

Приложение Б

Таблица Б1 – время работы модулей NLTK для токенизации на 2-граммы (для Таблицы 6 основной части документа)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Инструменты | Измерения времени, с | | | | | Среднее время работы инструмента, с |
| 1 | 2 | 3 | 4 | 5 |
| Bigram | 0.00100 | 0.0 | 0.0 | 0.0 | 0.00104 | 0.00041 |
| Ngrams | 0.00096 | 0.0 | 0.00100 | 0.00105 | 0.0 | 0.00060 |

Таблица Б2 – время работы модулей NLTK для токенизации на 3-граммы (для Таблицы 6 основной части документа)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Инструменты | Измерения времени, с | | | | | Среднее время работы инструмента, с |
| 1 | 2 | 3 | 4 | 5 |
| Trigram | 0.0 | 0.00105 | 0.00053 | 0.00100 | 0.0 | 0.00052 |
| Ngrams | 0.00036 | 0.00080 | 0.0 | 0.00054 | 0.00105 | 0.00055 |

Тестирование времени проводилось на предложении «Пушкин неоднократно писал о своей родословной в стихах и прозе; он видел в своих предках образец истинной «аристократии», древнего рода, честно служившего отечеству, но не снискавшего благосклонности правителей и «гонимого».».

Приложение В

Таблица В1 – время работы инструментов для токенизации на 2-граммы (для Таблицы 5 основной части документа)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Инструменты | Измерения времени, с | | | | | Среднее время работы инструмента, с |
| 1 | 2 | 3 | 4 | 5 |
| SpaCy | 0.00100 | 0.00100 | 0.0 | 0.00100 | 0.0 | 0.00060 |
| Pattern | 2.33088 | 2.27264 | 2.26600 | 2.24720 | 2.54472 | 2.33229 |
| Scikit-learn | 0.00100 | 0.00102 | 0.00101 | 0.00102 | 0.00100 | 0.00101 |
| TextBlob | 0.00925 | 0.01004 | 0.00901 | 0.01004 | 0.00997 | 0.00966 |

Таблица В2 – время работы инструментов для токенизации на 3-граммы (для Таблицы 5 основной части документа)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Инструменты | Измерения времени, с | | | | | Среднее время работы инструмента, с |
| 1 | 2 | 3 | 4 | 5 |
| SpaCy | 0.0 | 0.00100 | 0.0 | 0.00117 | 0.0 | 0.00043 |
| Pattern | 2.44092 | 2.43213 | 2.56806 | 2.42952 | 2.47671 | 2.46947 |
| Scikit-learn | 0.00184 | 0.00165 | 0.00195 | 0.00184 | 0.00159 | 0.00177 |
| TextBlob | 0.00948 | 0.00933 | 0.00917 | 0.00899 | 0.00777 | 0.00895 |

Тестирование времени проводилось на предложении «Пушкин неоднократно писал о своей родословной в стихах и прозе; он видел в своих предках образец истинной «аристократии», древнего рода, честно служившего отечеству, но не снискавшего благосклонности правителей и «гонимого».».

Приложение Г

Таблица Г1 – время работы модулей NLTK для токенизации на 1-2-3-граммы (для Таблицы 7 основной части документа)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Инструменты | Измерения времени, с | | | | | Среднее время работы инструмента, с |
| 1 | 2 | 3 | 4 | 5 |
| Unigram + Bigram + Trigram | 0.00100 | 0.00093 | 0.00105 | 0.0 | 0.00104 | 0.00080 |
| Ngrams | 0.00066 | 0.00200 | 0.00028 | 0.00078 | 0.00095 | 0.00093 |
| Everygrams | 0.00100 | 0.00101 | 0.00096 | 0.00100 | 0.0 | 0.00079 |

Таблица Г2 – время работы инструментов для токенизации на 1-2-3-граммы (для Таблицы 8 основной части документа)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Инструменты | Измерения времени, с | | | | | Среднее время работы инструмента, с |
| 1 | 2 | 3 | 4 | 5 |
| SpaCy | 0.0 | 0.00105 | 0.00100 | 0.00100 | 0.00100 | 0.00081 |
| Pattern | 2.45792 | 2.42852 | 2.48038 | 2.46185 | 2.44867 | 2.455468 |
| Scikit-learn | 0.00300 | 0.00199 | 0.00100 | 0.00186 | 0.00105 | 0.00178 |
| TextBlob | 0.01099 | 0.01097 | 0.01204 | 0.00996 | 0.00998 | 0.010788 |

Тестирование времени проводилось на предложении «Пушкин неоднократно писал о своей родословной в стихах и прозе; он видел в своих предках образец истинной «аристократии», древнего рода, честно служившего отечеству, но не снискавшего благосклонности правителей и «гонимого».».