

Анализ и извлечение ключевых фраз из текстов с использованием SpaCy*Андриенко Иван Сергеевич**Приамурский государственный университет имени Шолом-Алейхема**Студент***Аннотация**

В данной статье описан процесс разработки инструмента для автоматического анализа и извлечения ключевых фраз из текстов с использованием библиотеки SpaCy на языке программирования python. Рассмотрены ключевые аспекты работы с текстовыми данными, включая предобработку текста, извлечение именных групп и фильтрацию по значимости. В качестве примера приведена реализация программы для анализа текстов на русском языке, которая извлекает ключевые фразы, основываясь на синтаксическом разборе текста. Показано, как с помощью SpaCy можно эффективно анализировать текстовые данные и выделять ключевые понятия.

Ключевые слова: текстовый анализ, SpaCy, python, извлечение ключевых фраз, обработка текста, синтаксический разбор.

Analyzing and extracting keywords from texts using Space*Andrienko Ivan Sergeevich**Sholom-Aleichem Priamursky State University**Student***Abstract**

This article describes the process of developing a tool for automatic analysis and extraction of key phrases from texts using the SpaCy library in the python programming language. The key aspects of working with text data are considered, including text preprocessing, extraction of name groups and filtering by significance. As an example, the implementation of a program for analyzing texts in Russian is given, which extracts key phrases based on the syntactic analysis of the text. It is shown how SpaCy can be used to effectively analyze text data and identify key concepts.

Keywords: text analysis, SpaCy, python, keyword extraction, text processing, parsing.

1 Введение**1.1 Актуальность**

В условиях постоянно растущего объема текстовой информации, автоматическое извлечение ключевых фраз становится важной задачей для анализа и обработки текстов. Ключевые фразы представляют собой основные

концепты и идеи, содержащиеся в тексте, и их правильное извлечение позволяет существенно улучшить понимание и последующую обработку текстовых данных. В этом контексте использование современных библиотек для обработки естественного языка, таких как SpaCy, предоставляет мощные инструменты для автоматического анализа текстов, позволяя эффективно выделять ключевые фразы и улучшать процессы информационного поиска, классификации текстов и создания кратких аннотаций. Интеграция таких методов становится все более актуальной в различных областях, включая обработку новостных данных, анализ социальных сетей и управление контентом.

1.2 Обзор исследований

Н.П. Орлянская и М.В. Авджян в своей работе исследуют обработку естественного языка на примере англоязычных текстов с использованием библиотеки SpaCy. Авторы подробно рассматривают различные аспекты применения SpaCy для анализа текстов, включая выделение именованных сущностей и разбор предложений [1]. В работе Д.Е. Соколовского рассматривается использование инструмента SpaCy для извлечения собственных именованных сущностей из медицинских наборов данных. Автор анализирует эффективность применения SpaCy для обработки специализированных текстов, отмечая его высокую точность и производительность [2]. М.Л. Малоземова и П.А. Ломов исследуют дистрибутивный подход к анализу естественных языковых текстов для автоматизации пополнения тематического тезауруса. Авторы рассматривают преимущества использования данного подхода в контексте автоматизации обработки больших объемов данных [3]. П.С. Муродов в своей работе проводит анализ библиотек на языке Python для автоматической обработки текстов. Автор акцентирует внимание на возможностях использования данных библиотек для решения различных задач обработки естественного языка [4]. В работе М.А. Фесенко исследуются методы обработки языковых ситуаций с использованием нейронных методов. Автор предлагает подходы к использованию нейросетевых моделей для анализа и интерпретации текстов [5].

1.3 Цель исследования

Цель исследования – разработать эффективный подход к использованию библиотеки SpaCy для автоматического анализа и извлечения ключевых фраз из текстов на русском языке. Это включает в себя настройку модели для работы с русским языком, извлечение именных групп и фильтрацию фраз по частоте встречаемости в тексте.

2 Материалы и методы

Для исследования использовались Python, библиотека SpaCy и дополнительная модель для русского языка. Для работы с библиотекой были

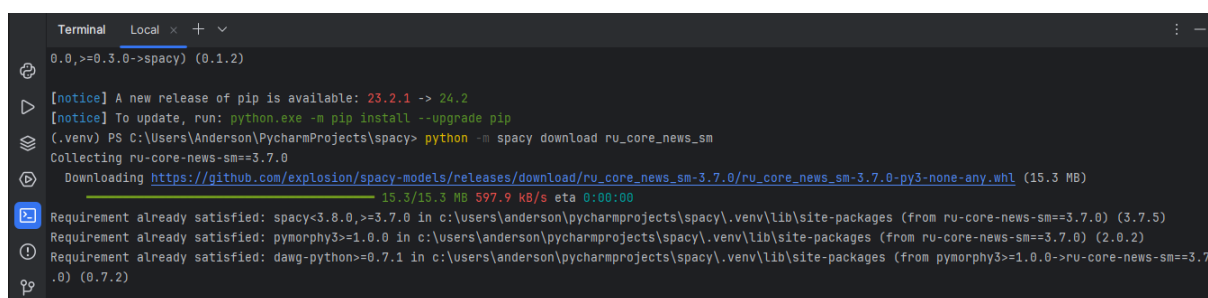
созданы функции для предобработки текста, извлечения ключевых фраз и их фильтрации по частотности.

3 Результаты и обсуждения

SpaCy – это библиотека для обработки естественного языка, которая предоставляет мощные инструменты для анализа текстов, включая синтаксический разбор и извлечение именных групп. В данной работе библиотека использовалась для анализа текстов на русском языке, что включает в себя настройку модели и использование специальных функций для выделения ключевых фраз. Основные компоненты программы:

- Предобработка текста – удаление стоп-слов, нормализация текста и разбор на леммы для последующего анализа.
- Извлечение именных групп – определение и выделение существительных и связанных с ними прилагательных, что позволяет формировать ключевые фразы, представляющие основные идеи текста.
- Фильтрация ключевых фраз – подсчет частотности фраз и отбор наиболее значимых из них.

Для реализации автоматического анализа и извлечения ключевых фраз из текстов с использованием библиотеки SpaCy необходимо установить соответствующие программные зависимости и языковые модели. Основным инструментом для анализа текста в данном исследовании является библиотека SpaCy для обработки естественного языка. Для корректной работы с русскоязычными текстами необходимо выполнить команды в терминале. Команда `pip install spacy` установит саму библиотеку SpaCy, а команда `python -m spacy download ru_core_news_sm` скачает языковую модель для обработки русскоязычных текстов. Эта модель включает в себя лексические и синтаксические особенности русского языка, что позволяет эффективно проводить анализ текстов и извлечение ключевых фраз (рис. 1).



```
Terminal Local x + v
0.0.>=0.3.0->spacy) (0.1.2)

[notice] A new release of pip is available: 23.2.1 -> 24.2
[notice] To update, run: python.exe -m pip install --upgrade pip
(.venv) PS C:\Users\Anderson\PycharmProjects\spacy> python -m spacy download ru_core_news_sm
Collecting ru-core-news-sm==3.7.0
  Downloading https://github.com/explosion/spacy-models/releases/download/ru_core_news_sm-3.7.0/ru_core_news_sm-3.7.0-py3-none-any.whl (15.3 MB)
    15.3/15.3 MB 597.9 kB/s eta 0:00:00
Requirement already satisfied: spacy<3.8.0,>=3.7.0 in c:\users\anderson\pycharmprojects\spacy\.venv\lib\site-packages (from ru-core-news-sm==3.7.0) (3.7.5)
Requirement already satisfied: pymorphy3>=1.0.0 in c:\users\anderson\pycharmprojects\spacy\.venv\lib\site-packages (from ru-core-news-sm==3.7.0) (2.0.2)
Requirement already satisfied: dawg-python>=0.7.1 in c:\users\anderson\pycharmprojects\spacy\.venv\lib\site-packages (from pymorphy3>=1.0.0->ru-core-news-sm==3.7.0) (0.7.2)
```

Рис. 1. Установка библиотеки и модели

Далее необходимо импортировать библиотеку SpaCy. После импортируем класс Counter, который используется для подсчета и хранения количества различных элементов (в данном случае слов) в коллекции. Загружаем предварительно обученную модель русского языка `ru_core_news_sm`, которая включает в себя алгоритмы для анализа русского текста (рис. 2).

```
1 import spacy
2 from collections import Counter
3
4 nlp = spacy.load('ru_core_news_sm')
5
```

Рис. 2. Загрузка модели

Создадим функцию `extract_key_phrases`, которая принимает текст и количество ключевых фраз, которые нужно извлечь. Сначала текст анализируется с помощью библиотеки SpaCy, которая разбивает его на токены. Затем для каждого токена проверяется, является ли он существительным или прилагательным, и не относится ли он к стоп-словам. Если условия выполняются, создается ключевая фраза, включающая леммы связанных токенов, таких как прилагательные или составные части, которые объединяются с леммой основного токена. Полученные фразы приводятся к нижнему регистру и добавляются в список ключевых фраз. После этого используется `Counter` для подсчета частоты появления каждой ключевой фразы в тексте. В конце функция возвращает список из наиболее частых ключевых фраз в соответствии с указанным количеством `top_n` (рис. 3).

```
1 usage
2
3 def extract_key_phrases(text, top_n=10):
4
5     doc = nlp(text)
6
7     key_phrases = []
8     for token in doc:
9
10         if token.pos_ in {"NOUN", "ADJ"} and not token.is_stop:
11
12             phrase = " ".join([child.lemma_ for child in token.children if child.dep_ in {"amod", "compound"}])
13             phrase = f"{phrase} {token.lemma_}".strip()
14             key_phrases.append(phrase.lower())
15
16     phrase_freq = Counter(key_phrases)
17
18     return phrase_freq.most_common(top_n)
```

Рис. 3. Создание функции `extract_key_phrases`

Зададим текстовый блок. В нем будет отрывок на тему искусственного интеллекта и обработки естественного языка. Этот текст будет служить входными данными для анализа. Затем текст передается в функцию `extract_key_phrases`, которая выполняет обработку текста, чтобы извлечь ключевые фразы. Эта функция анализирует текст, выделяя наиболее важные словосочетания, и возвращает список ключевых фраз, отсортированных по частоте их появления. После того как ключевые фразы извлечены, они выводятся на экран. Для каждой ключевой фразы указывается сама фраза и частота её появления в тексте. Этот вывод позволяет увидеть, какие фразы были определены как наиболее значимые в данном тексте (рис. 4).

```
25 if __name__ == "__main__":  
26     text = """Искусственный интеллект (ИИ) стремительно трансформирует индустрии,  
27     позволяя машинам выполнять задачи, которые обычно требуют человеческого интеллекта.  
28     Обработка естественного языка (NLP), являющаяся частью ИИ,  
29     позволяет компьютерам понимать, интерпретировать и генерировать человеческий язык."""  
30  
31     key_phrases = extract_key_phrases(text)  
32     print("Ключевые фразы:")  
33     for phrase, freq in key_phrases:  
34         print(f"{phrase} (Частота: {freq})")  
35
```

Рис. 4. Передача текста в функцию

Программа автоматически выделяет фразы, которые содержат основную информацию текста, что помогает лучше понять его ключевые идеи. В результате работы программы были извлечены такие ключевые фразы, как "искусственный интеллект", "человеческий интеллект" и другие. Эти результаты показывают, как использование SpaCy может эффективно способствовать анализу и обработке текстовых данных, позволяя быстро выделять основные идеи и понятия (рис. 5).

```
↓  
⇐  
⇓  
🖨  
🗑  
Ключевые фразы:  
человеческий (Частота: 2)  
искусственный (Частота: 1)  
искусственный интеллект (Частота: 1)  
индустрия (Частота: 1)  
машина (Частота: 1)  
задача (Частота: 1)  
человеческий интеллект (Частота: 1)  
обработка (Частота: 1)  
естественный (Частота: 1)  
естественный язык (Частота: 1)  
  
Process finished with exit code 0
```

Рис. 5. Результат программы

Выводы

В данной статье представлен процесс разработки инструмента для автоматического анализа и извлечения ключевых фраз из текстов с использованием библиотеки SpaCy на языке программирования python. Рассмотрены ключевые аспекты, включая предобработку текста, извлечение именных групп и фильтрацию фраз по частотности. Приведены примеры реализации, которые демонстрируют, как библиотека SpaCy упрощает анализ текстовых данных. Использование SpaCy обеспечивает надежное и эффективное выделение ключевых фраз, что позволяет разработчикам легко анализировать и управлять текстовыми данными. Библиотека предоставляет

мощный инструмент для анализа текстов на естественном языке, способный эффективно справляться с задачами обработки текстов на русском языке.

Библиографический список

1. Орлянская Н.П., Авджян М.В. Обработка естественного языка на примере англоязычных текстов с помощью библиотеки SpaCy // Электронный сетевой политематический журнал "Научные труды КубГТУ". 2023. № 1. С. 71-85.
2. Соколовский Д.Е. Использование инструмента SpaCy для извлечения собственных именованных сущностей из медицинских наборов данных // Сборник избранных статей научной сессии ТУСУР. 2023. № 1-2. С. 217-220.
3. Малоземова М.Л., Ломов П.А. Опыт применения дистрибутивного подхода к анализу естественных языковых текстов для автоматизации пополнения тематического тезауруса // В сборнике: Теория и практика системной динамики. Материалы конференции VIII Всероссийской конференции (с международным участием). Ответственный редактор: А.Г. Олейник. 2019. С. 93-94.
4. Муродов П.С. Библиотеки на языке Python для автоматической обработки текстов // Вестник Филиала Московского государственного университета имени М.В. Ломоносова в городе Душанбе. 2023. Т. 1. № 4 (35). С. 72-85.
5. Фесенко М.А. Processing of Language Situations Using Neural Methods // В сборнике: Современные проблемы науки, общества и культуры. Сборник статей VII Международной научно-практической конференции. Омск, 2022. С. 137-140.