УДК 004.584

ИСПОЛЬЗОВАНИЕ БИБЛИОТЕКИ SPACY ДЛЯ АВТОМАТИЗАЦИИ ПРОЦЕССА РАСПРЕДЕЛЕНИЯ ЗАЯВОК ГРАЖДАН С САЙТА АДМИНИСТРАЦИИ Г.КЕМЕРОВО

Алексеев В.С., студент гр. ИТб-191, IV курс
Ванеев О.Н., доцент (к.н.) кафедры ИиАПС
Кузбасский государственный технический университет
имени Т. Ф. Горбачева
г. Кемерово

При помощи виртуальной приёмной любой гражданин может отправить своё обращение в виде электронного документа в администрацию города Кемерово, которое в дальнейшем будет рассмотрено главой города, либо же его заместителями по различным специализированным направлениям.

В настоящее время все исходящие заявления рассматривается специальным отделом в ручном режиме, что требует весомых временных затрат. Основная проблема такого подхода состоит в том, что пока обращение не пройдёт стадию обработки, содержимое документа не может попасть напрямую до получателя, что значительно замедляет ответную реакцию администрации на проблемы жителей города Кемерово.

Для автоматизации данного процесса разрабатывается система с подхватыванием документов с сайта виртуальной приемной, которая заменит человека, поскольку содержимое будет проверяться с помощью технологий обработки текста, что позволит выявлять потенциальный спам, а также своевременно доставлять проверенные файлы до адресатов.

Обработка текстов на естественном языке (Natural Language Processing, NLP) — общее направление искусственного интеллекта и математической лингвистики. Оно изучает проблемы компьютерного анализа и синтеза текстов на естественных языках. Применительно к искусственному интеллекту анализ означает понимание языка, а синтез — генерацию грамотного текста.

NLTK (Natural Language Toolkit) — представляет собой важную библиотеку, поддерживающую такие задачи, как классификация, стемминг, маркировка, синтаксический анализ и семантическое рассуждение в Python. Это основной инструмент для обработки естественного языка и машинного обучения.

NLTK (Natural Language Toolkit) в основном работает с человеческим языком, а не с компьютерным, чтобы применять обработку естественного языка (NLP). Он содержит библиотеки обработки текста, с помощью которых можно выполнять токенизацию, парсинг, классификацию, выделение, тегирование и семантическое обоснование данных, представленная на рис. 1.

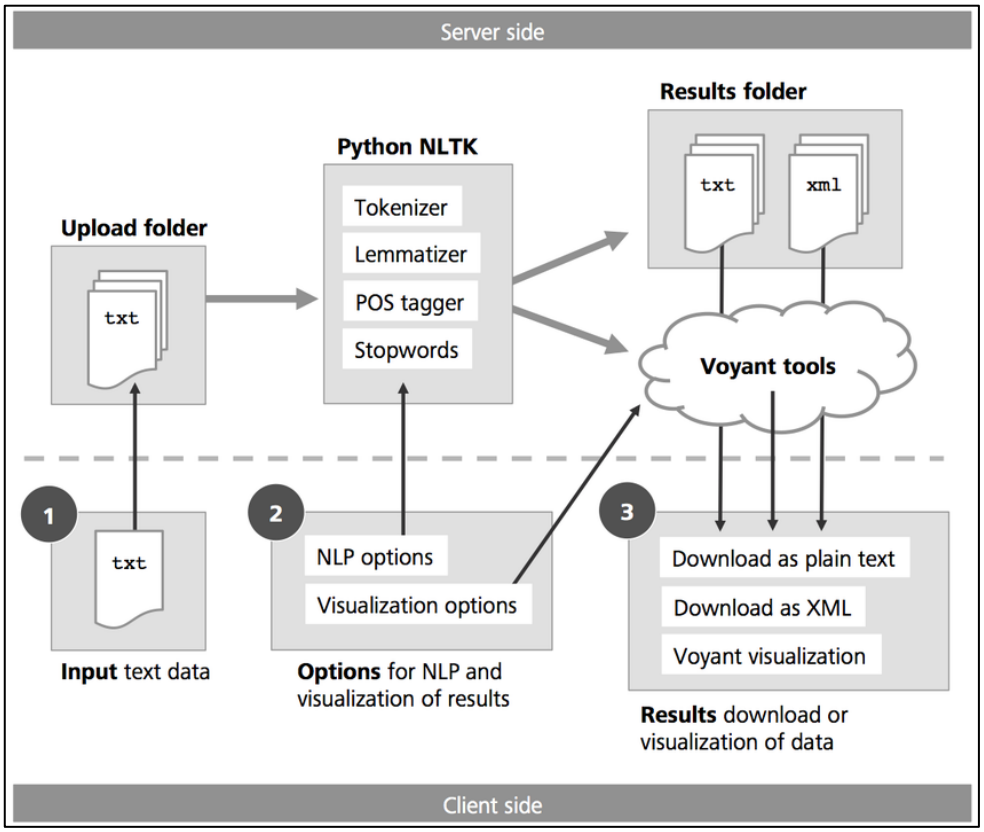

Рис. 1. Схема сети, обеспечивающая защищённость данных

Библиотека довольно универсальна, однако её трудно использовать для обработки естественного языка. NLTK может быть довольно медленным и не соответствовать требованиям быстро развивающегося производственного использования.

Сегодня библиотека NLTK служит образовательной основой для разработчиков Python, которые только приступают к изучению NLP и машинного обучения.

spaCy относительно молодая библиотека, предназначенная для производственного использования. Вот почему она гораздо доступнее других NLP-библиотек Python, таких как NLTK.

К преимуществам данной библиотеки можно отнести:

- Поддержка 72+ языков;
- Многозадачное обучение с предварительно обученными преобразователями;
- Предварительно мотивированная токенизация
- Компоненты для распознавания именованых сущностей, тегирования частей речи, анализа зависимостей, сегментации предложений, морфологического анализа, связывания сущностей и т.д.;
- Легко расширяемая с помощью настраиваемых компонентов и атрибутов;
- Простая упаковка модели, развертывание и управление рабочим процессом;
- Надежная, тщательно оцененная точность.

Центральными структурами данных в spaCy являются Doc и Vocab. Объект Doc хранит последовательности токенов и все их аннотации. Объект Vocab хранит набор справочных таблиц, что делает общую информацию доступной для всех документов. При централизованном хранении строк, векторов слов и лексических атрибутов отсутствует необходимость хранения нескольких копий этих данных. Это экономит память и обеспечивает единый источник правды.

Объект Doc владеет данными, а Span и Token являются представлениями, что позволяет Spacy работать быстро, без лишних копирований, указывающими на них. Объект Doc создается объектом Tokenizer, а затем модифицируется in-place компонентами pipeline. Объект Language координирует эти компоненты. Он берет необработанный текст и отправляет его по pipeline, возвращая аннотированный документ.

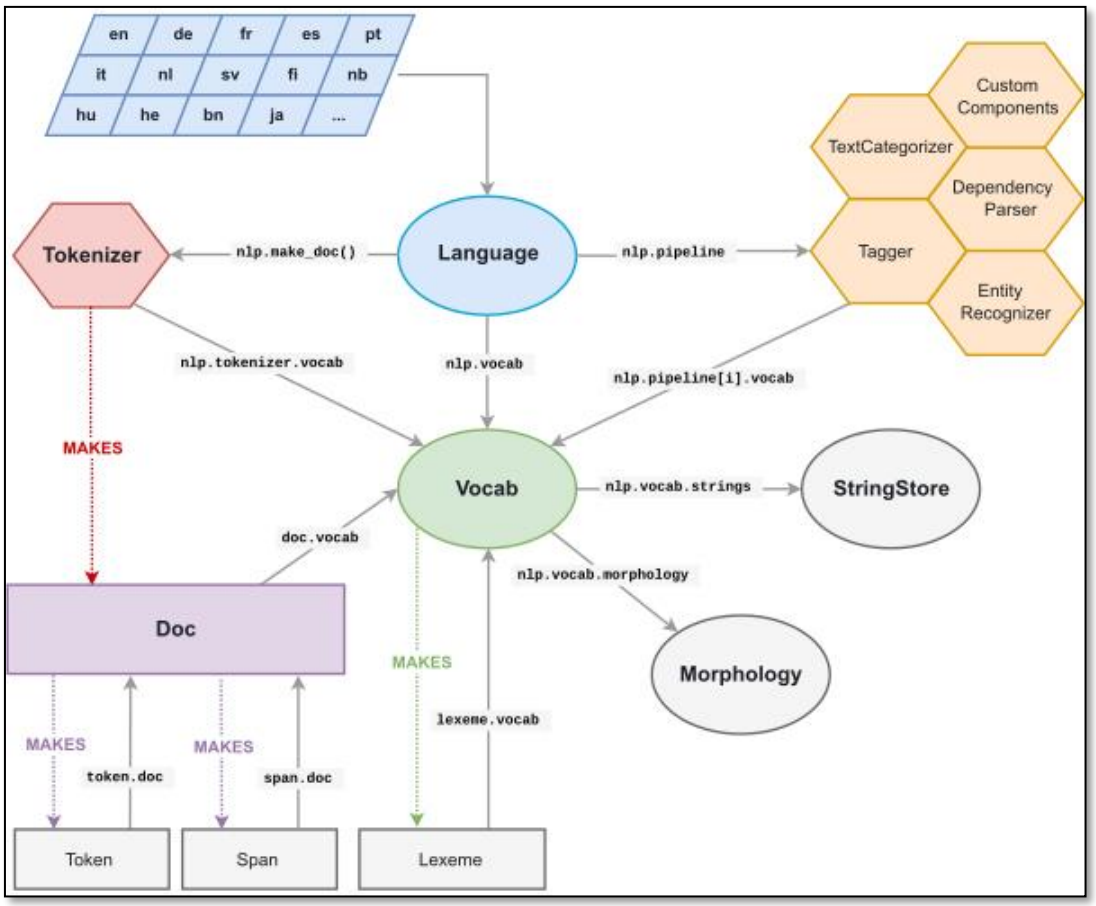

Рис. 2. Архитектура spaCy

Если кратко, то spaCy может примерно всё то же самое, что и NLTK и их аналоги, но быстрее и точнее.

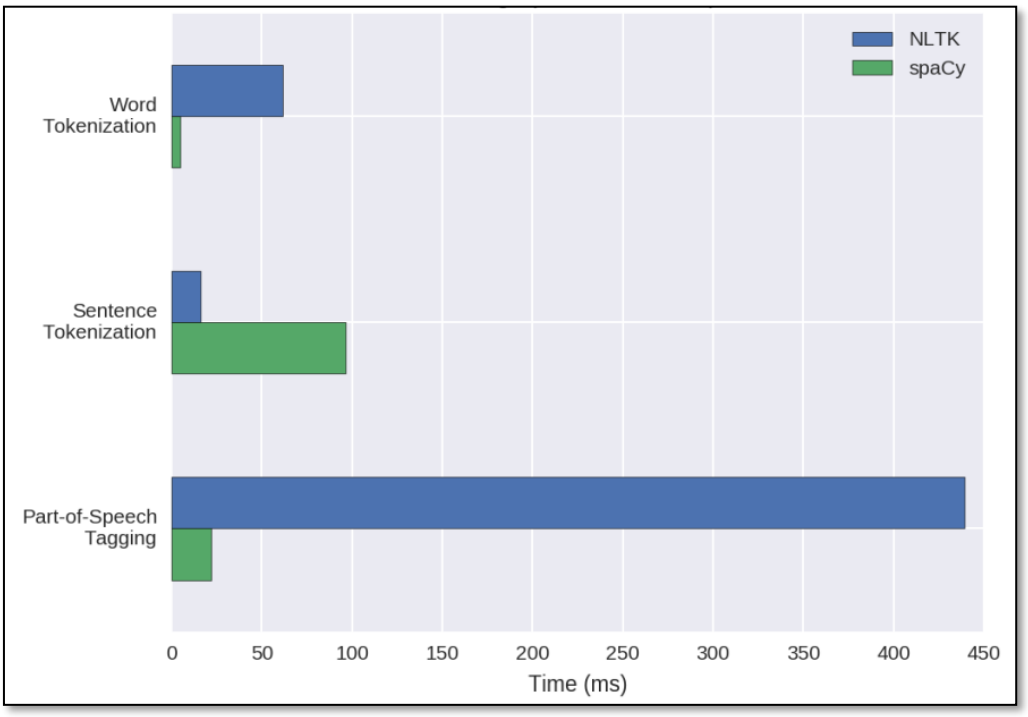

Рис. 3. Тест скорости двух библиотек NLTK и spaCy

Непосредственно перед установкой spaCy необходимо проверить наличие следующих компонентов:

Кроссплатформенная интегрированная среда разработки для языка программирования Python PyCharm – источник для скачивания jetbrains.com/ru-ru/pycharm/download

NLP библиотека, которая фокусируется на том, чтобы сделать естественный человеческий язык пригодным для использования компьютерными программами pypi.org/project/NLP-python/

Все вышеуказанные компоненты являются свободно распространяемыми. Установка каждого компонента не вызывает особых трудностей.

Источник для скачивания spaCy – spacy.io/usage. Установка библиотеки так же не вызывает никаких затруднений.

Основные элементы проекта:

import spacy – Добавление библиотеки;

nlp = space.load(' ') – Добавление языковой группы;

doc = nlp(u' ') – Добавление строковых значений в объект doc;

for token in doc: – Создание цикла с использованием объекта doc;

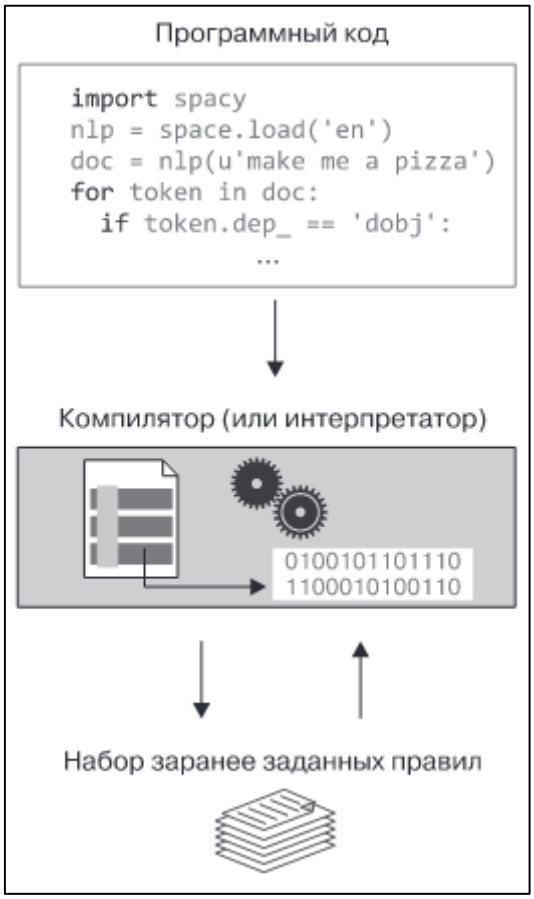

Рис. 4. Упрощенный технологический процесс обработки кода на языке программирования Python с использованием библиотеки spaCy

Благодаря внедрению технологий обработки текста, администрация города Кемерово сможет своевременно получать информацию о поступающих проблемах. Также данная система снизит нагрузку на бюджет, поскольку заменит труд человека в данной области.

Список литературы:

1. Документация по библиотеке NLTK. [Электронный ресурс] URL: https://www.nltk.org/ (дата обращения: 20.03.2023).
2. Документация по библиотеке spaCy. [Электронный ресурс] URL: https://spacy.io/api/doc (дата обращения: 20.03.2023).