РЕАЛИЗАЦИЯ ПРОЦЕССА СБОРА ДАННЫХ ДЛЯ ОБУЧАЮЩЕЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ

Литвинец Д. Н.

Белорусский государственный университет информатики и радиоэлектроники г.Минск, Республика Беларусь

Стержанов М.В. – канд. техн. наук, доцент

Аннотация: В работе рассмотрен процесс сбора текстовых данных с различных веб-ресурсов, выделение лингвистических характеристик, ранжирование данных и их структуризация для последующего применения в обучающих системах и моделях машинного обучения

Проектирование и разработка обучающей информационной системы требует предварительного сбора и анализа данных, качество и структура которых определяет последующее развитие системы. В качестве старта будет подготовлен модуль изучения английского языка, как наиболее популярного в мире и, соответственно, имеющего больше возможностей к анализу.

Первым этапом в создании обучающей информационной системы стал сбор данных, которые в будущем информационная система будет использовать для генерации и рекомендации заданий для пользователей путем выделения наиболее популярных слов и фраз, разделение данных на группы по различным критериям или комбинирование данных для генерации нового материала.

Технологией, с помощью которой будет реализован автоматический сбор данных с веб-сайтов, парсинг HTML-документов и JSON-файлов для обучающей системы и, впоследствии, некоторые манипуляции над этими данными, был выбран язык программирования Python а конкретно библиотека Beautifulsoup4 [1], так как она является несложной для освоения и последующего использования. Для работы с базами данных была выбрана документо-ориентированная NoSQL база данных MongoDB. В нашей работе, в связи с выбором языка программирования Python для работы с базой данных MongoDB будет применяться драйвер PyMongo [2].

Основными данными, которые требовалось собрать, были слова различной частоты использования на английском языке, а также их переводы на русский язык, характеристики данных слов, среди которых часть речи, некоторые особые характеристики, свойственные определенным частям речи (род и число для имён существительных), их категории по значению и примеры употребления данных слов.

Сначала осуществлялся сбор пар слов на русском и английском языках и их категории использования, например, “названия фруктов”, “семья” и так далее. Небольшой трудностью послужило малое количество ресурсов, где можно было в свободном доступе в автоматическом режиме получить изначально систематизированный по группам использования и подготовленный для работы список слов. Еще одним нюансом стало то, что отдельные собранные слова имеют различные значения, что также требует пристального внимания по мере их обработки.

После успешного создания прототипа базы данных, состоящего только из слов, переводов и их категорий был начат второй этап, а именно – этап поиска отсутствующих характеристик, которые не получилось извлечь на первом этапе, во время которого при помощи интернет-ресурса Linguee [3] были получены оставшиеся характеристики. Метод сбора отсутствующих характеристик аналогичен первому этапу: на ресурс отправлялся запрос, а затем информация собиралась с HTML-страницы.

Приведем один из примеров использования собранных данных. С помощью полученных данных мы можем подсчитать количество фраз, которые может выучить пользователь, если выучит N наиболее популярных слов из нашей базы данных при определенном параметре, который мы назвали толерантностью. Толерантность – параметр, показывающий, каким количеством слов в неизвестной фразе мы можем пренебречь и считать эту фразу изученной. Как мы можем видеть, при более высоком параметре толерантности показатель количества изучаемых фраз значительно растет. Данный параметр мы считаем очень важным, так как он также имеет свою важность в реальной жизни. При изучении новых слов/фраз на неизвестном языке человек может понять значение ранее неизвестного ему слова благодаря контексту фразы или ситуации. Именно поэтому, искусственным подобием этой возможности человека подобрать значение исходя из обстоятельств мы можем считать параметр толерантности.

По итогу сбора информации в наше распоряжение была получена база данных, состоящая из 967 слов и их характеристик, а также 2062 примеров использования данных слов. Все слова с

их характеристиками были по итогу отсортированы по частоте использования на основании тех 2062 фраз, а сами фразы в свою очередь были отсортированы по количеству слов, которых нет в нашей базе данных.

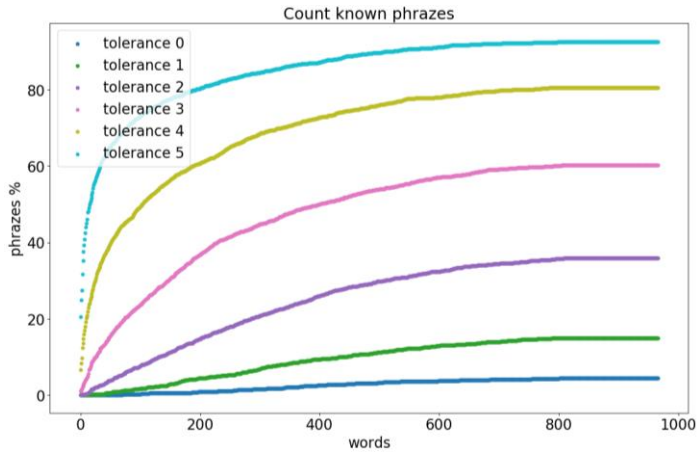

Рисунок 1 – график зависимости известных фраз от количества известных слов при заданном параметре толерантности

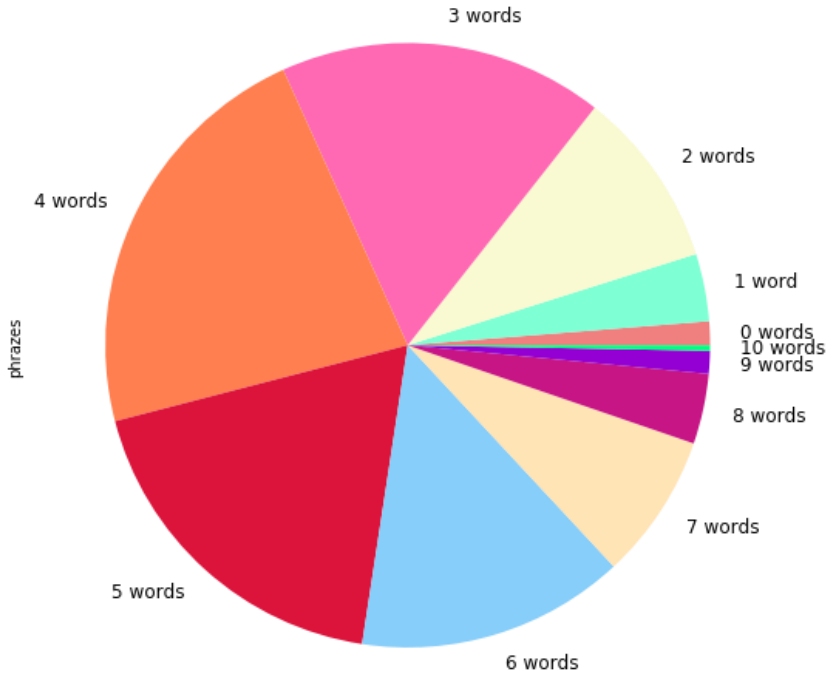

Рисунок 2 – диаграмма фраз по количеству неизвестных слов

На Рисунке 1 представлена зависимость количества доступных для изучения фраз от количества изученных ранее пользователем слов. На Рисунке 2 можем наблюдать график диаграмму количества фраз и количества неизвестных в ней слов.

Как мы можем наблюдать, самые многочисленные группы фраз содержат от 3 до 6 слов, отсутствующих в нашей базе, что составляет около 72% от общего числа фраз, а группы фраз, содержащие от 0 до 2 и более 7 неизвестных слов составляют 14% и 12% соответственно. После небольшой проверки также было выяснено, что 13% слов не имеют примеров использования в нашей базе, что показывает, что требуется дальнейший поиск ресурсов для сбора отсутствующей информации, например, путем ручного ввода данных или путем скраппинга статей или книг.

Таким образом, по итогам проделанной работы собран минимально необходимый объем данных для практического применения при создании обучающей модели или программы, проведено выделение как общих характеристик, так и искусственных, связанных с постановкой исходной задачи, например, метрика толерантности. В дальнейшем планируется расширять набор и улучшать качество данных, продолжать реализацию обучающей системы.

Список использованных источников:

[1]. Beautiful Soup: We called him Tortoise because he taught us [Электронный ресурс]. – 2022. – Режим доступа: https://www.crummy.com/software/BeautifulSoup/

[2]. PyMongo – MongoDB drivers [Электронный ресурс]. – 2022. – Режим доступа: https://www.mongodb.com/docs/drivers/pymongo/

[3]. Linguee [Электронный ресурс]. – 2022. – Режим доступа : https://www.linguee.ru