# Задание для первого студента.

Парсер сайта https://bloknot-volgograd.ru  


Рисунок 1 - кнопка запуска парсера

Описание принципа работы парсера новостей

1. Загрузить содержимое страницы с новостями.
   1. Ссылка типа «https://bloknot-volgograd.ru/?PAGEN\_1=#n#», где #n# – номер страницы. На каждой странице содержатся 10 новостей.
2. Найти по тегам блок с новостями и получить ссылки на новости.
   1. Найти все объекты с именем «li» родительского объекта «ul» с классом «bigline».
   2. Взять значения атрибутов «href» дочерних объектов «a».
3. Загрузить содержимое страницы с новостью.
   1. Ссылка типа «https://bloknot-volgograd.ru/news/#header#», где #header# – текст заголовка новости транслитом.
4. Найти по тегам содержимое новости (заголовок, дата создания, текст новости, ссылки на видеоролики, если таковые имеются).
   1. Заголовок новости – объект «article», содержимое дочернего объекта «h1».
   2. Дата создания новости – содержимое объекта «span» с классом «news-date-time».
   3. Текст новости – содержимое всех объектов «p» с классом «news-text».
   4. Ссылки на видео – содержимое всех объектов «iframe».
5. Сформировать из найденных атрибутов объект новости и сохранить его в базу данных. Если при попытке записи новости в базу новость с таким адресом уже существует, то обновляется текст новости, дата создания новости и количество комментариев к новости.

Описание базы данных и коллекций

Система управления базами данных: MongoDb.

Описание коллекций и их полей:

1. news – коллекция с текстами новостей
   1. id
   2. name\_news – заголовок новости
   3. date\_news – дата новости
   4. link\_news – ссылка на новость
   5. text\_news – текст новости
   6. link\_video – ссылки на видео (если есть)
   7. comments\_count – количество комментариев
2. person
   1. id
   2. name\_object – имя персоны/места
   3. num\_object – номер персоны/места
3. sentence
   1. id
   2. sentence – предложение с персоной/местом
4. tonality
   1. id
   2. sentence – предложение с персоной/местом
   3. tonality – тональность данного предложения

Пример заполнения базы данных

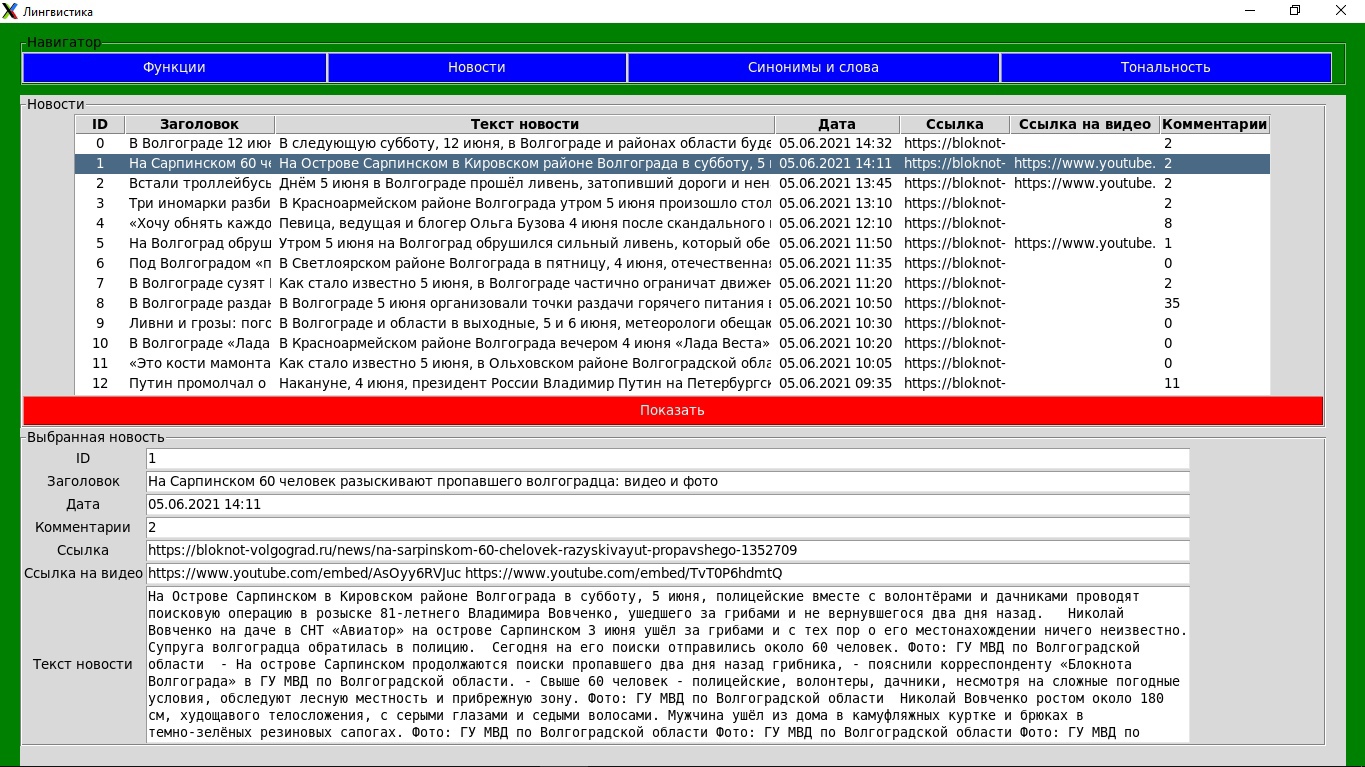


Рисунок 2 - новость в базе данных

# Задание для второго студента.

Описание работы томита-парсера

Для дальнейшего анализа новостей необходимо найти предложения, в которых упоминаются персоны и места из заданного списка. Для этого при помощи томиты находятся такие предложения и записываются в базу данных без изменений. После этого персоны заменяются на их условное обозначение и предложения с замененными персонами перезаписываются.

Описание работы Word2Vec

Для получения контекстных синонимов для персон и мест была обучена модель w2v на 1500 новостях с сайта <https://bloknot-volgograd.ru>, которые были получены при помощи парсера. После обучения модели в интерфейсе программы можно посмотреть контекстные синонимы во вкладке “Синонимы и слова” по нажатию на кнопку “Контекстные синонимы”.

Описание работы Natasha

Для получения слов окружения была использована библиотека Natasha для python. Для ее обучения были использованы предложения, в которых участвуют важные для Волгограда персоны и места. Чтобы просмотреть слова окружения, нужно зайти во вкладку “Синонимы и слова”, выбрать персону и нажать на кнопку “Слова окружения”.

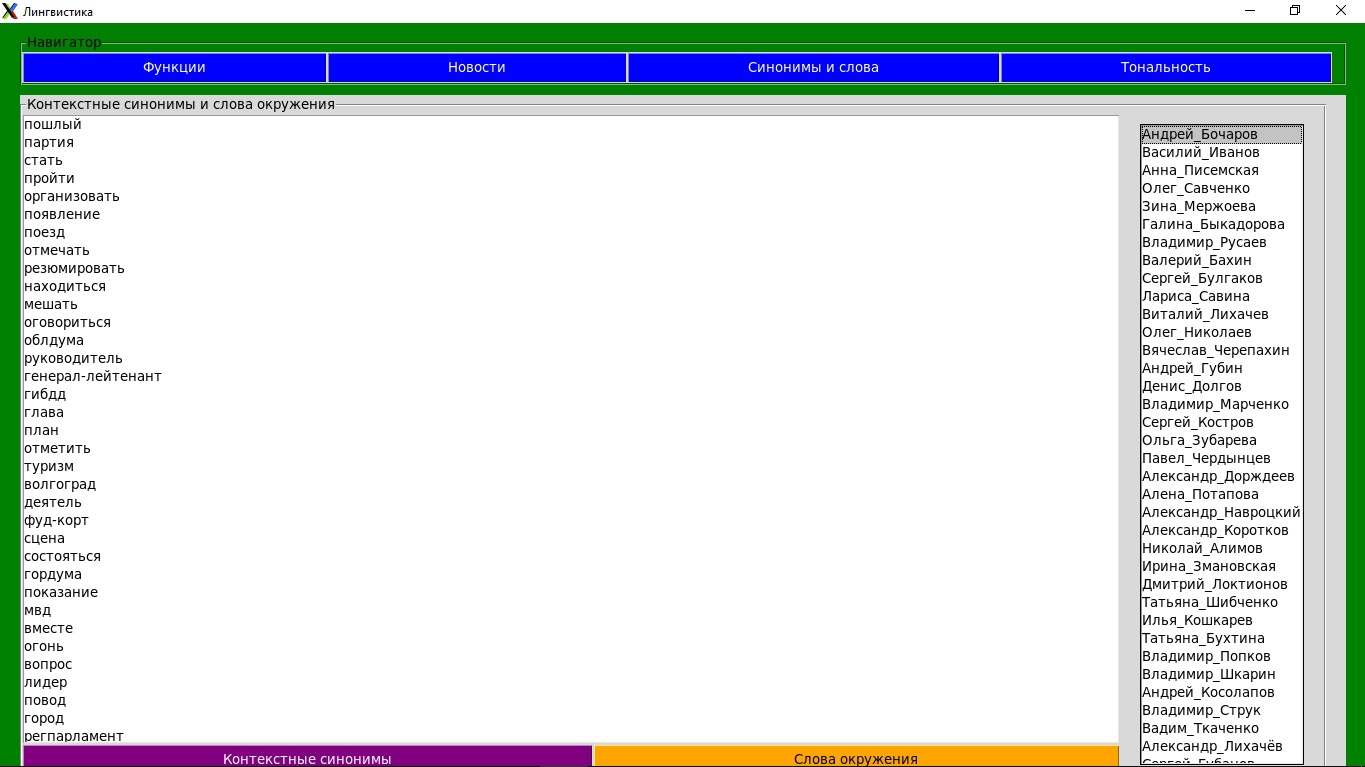


Рисунок 3 - слова окружения для Андрея Бочарова

# Задание для третьего студента.

Для анализа тональности используются предложения, в которых есть персоны или места. Они находятся при помощи томита-парсера. Для определения тональности используется библиотека nltk и датасеты с твитами positive.csv и negative.csv. После обучения на этих датасетах, в модель подаются предложения и результаты анализа записываются в базу данных. Для просмотра необходимо зайти на вкладку “Тональность” и нажать на кнопку “Показать”.

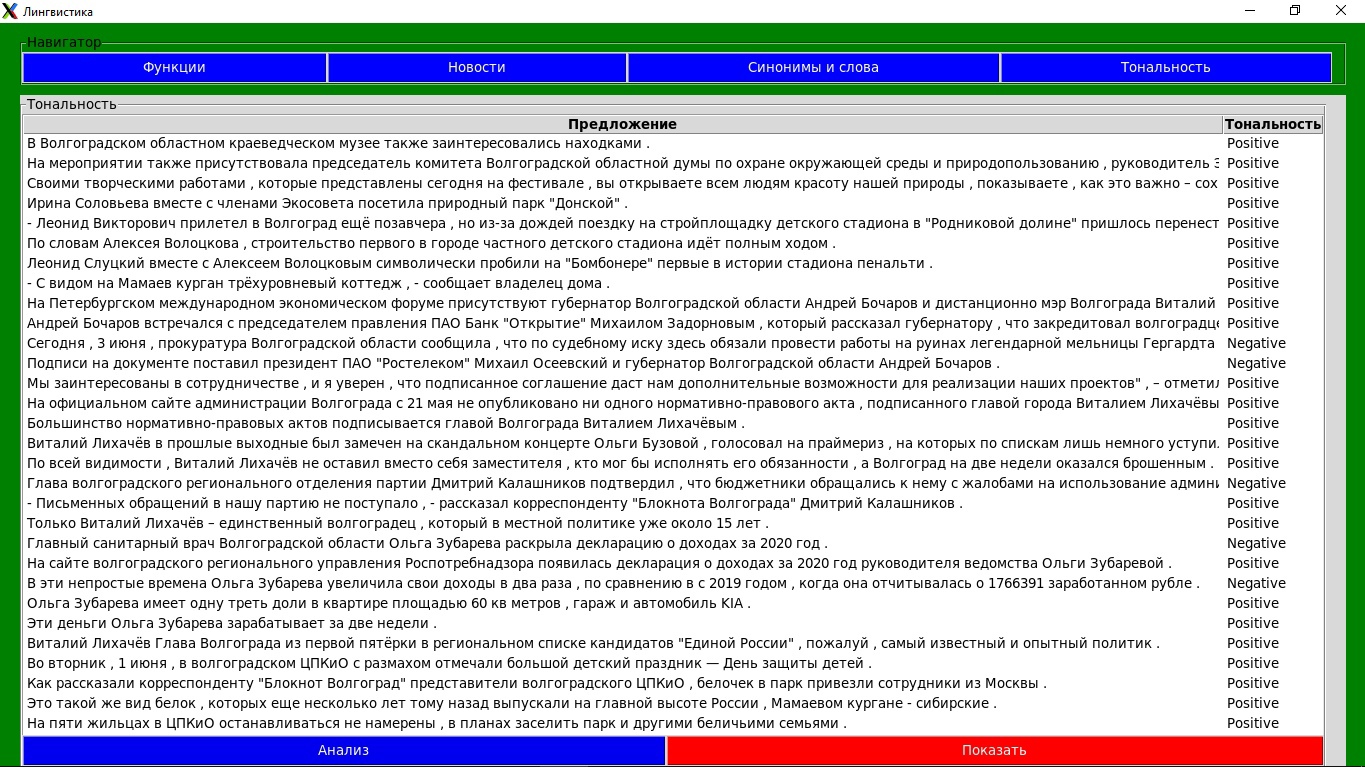


Рисунок 4 - тональность предложений