

Извлечение ключевых слов из набора текстовых документов с помощью программного пакета визуального программирования Orange

Голубева Евгения Павловна

Приамурский государственный университет имени Шолом-Алейхема

Студент

Аннотация

Цель данной статьи – выполнить извлечение ключевых слов из набора текстовых документов. Для выполнения извлечения ключевых слов был использован программный пакет визуального программирования на основе компонентов для визуализации данных Orange и подготовленный текст. С помощью средств визуализации Orange выполнили извлечение ключевых слов из набора текстовых документов и получили итоговую схему.

Ключевые слова: Orange, виджет, слова, извлечение.

Extracting keywords from a set of text documents using the Orange visual programming software package

Golubeva Evgeniya Pavlovna

Sholom-Aleichem Priamursky State University

Student

Abstract

The purpose of this article is to extract keywords from a set of text documents. To extract keywords, we used the Orange visual programming package based on data visualization components and prepared text. Using the Orange visualization tools, we extracted keywords from a set of text documents and obtained the final diagram.

Keywords: Orange, widget, words, extraction.

1 Введение

1.1 Актуальность

Извлечение ключевых слов из набора текстовых документов является фундаментальной задачей в области обработки естественного языка (NLP) и анализа текстовых данных, сохраняющей высокую актуальность в эпоху информационного перегруза. Автоматическое выявление наиболее релевантных и репрезентативных терминов позволяет эффективно сжимать большие объемы неструктурированного текста, выявлять основные темы и тенденции, классифицировать документы, улучшать поисковые системы и мониторить общественное мнение. В контексте научных исследований, анализа социальных медиа, управления знаниями в организациях или обработки клиентских отзывов способность быстро и точно выделить ключевые слова из корпуса документов критически важна для понимания

содержания, выявления скрытых закономерностей и поддержки принятия решений на основе данных. Традиционные ручные методы не масштабируются на большие объемы данных, что подчеркивает необходимость в надежных и доступных автоматизированных решениях.

Использование визуального программирования в Orange для решения этой задачи представляет собой особенно актуальное направление, так как оно существенно снижает барьер входа в сложную область текстовой аналитики. Orange, как платформа для интерактивного анализа данных с открытым исходным кодом, предлагает интуитивно понятный подход к построению конвейеров обработки данных через соединение визуальных виджетов. Для извлечения ключевых слов это означает, что исследователи и аналитики, даже без глубоких знаний программирования, могут легко загружать текстовые корпуса, выполнять предобработку, применять различные статистические и алгоритмические методы извлечения ключевых слов, визуализировать результаты и сравнивать эффективность разных подходов.

1.2 Обзор исследований

Ю.А. Сидоренко, О.М. Атаева, В.А. Серебряков, Д.А. Малахов описывают решение проблемы семантического поиска по текстам документов [1]. Рассматривал возможность семантического поиска информации в Internet и информационных системах В.С. Вороньков [2]. Д. В. Гринченков, Ф. Х. Нгуен, Т. Т. Нгуен, Д. А. Горбушин выполнили краткий обзор и сравнительный анализ возможностей алгоритмов, используемых для интеллектуального анализа данных [3]. Продемонстрировали подходы к поиску информации на основе семантических технологий Б.Н. Нгуен, А.Ф. Тузовский [4]. Л. Нэй, М.Х. Каунг анализируют один из способов семантического информационного поиска [5].

1.3 Цель исследования

Цель исследования - выполнить извлечение ключевых слов из набора текстовых документов.

2 Материалы и методы

Для выполнения извлечение ключевых слов из набора текстовых документов используется программа Orange. Работа будет происходить на готовом наборе данных состоящий из 4 документов, скачать которые можно по ссылке:

<https://drive.google.com/drive/folders/1SORL4HnezlblRJ5q4Nxiq4no5WXgcbRy?usp=sharing>

3 Результаты и обсуждения

Перед началом работы требуется скачать Orange с официального сайта и установить.

Создадим новый файл (см.рис.1).

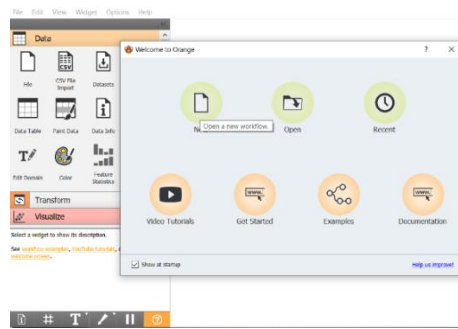


Рисунок 1. Создание нового файла

Для извлечения ключевых слов из набора текстовых документов необходимо установить дополнение Text. Для того, чтобы скачать дополнение, необходимо перейти в Options, далее в Add-ons, в появившемся окне выбираем Text (см.рис.2).

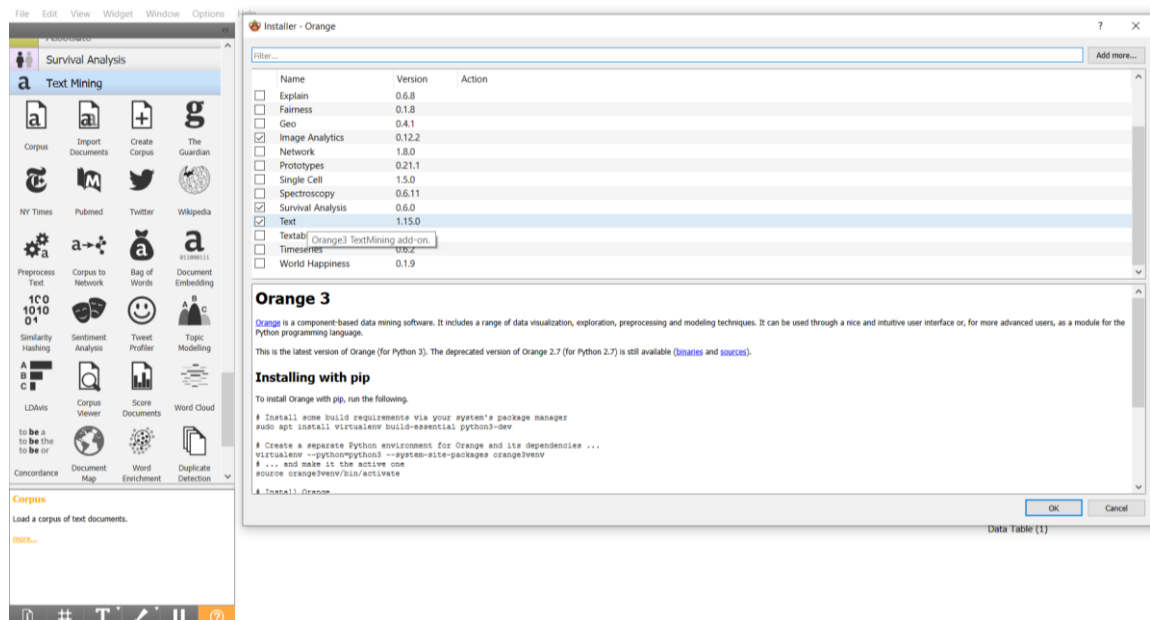


Рисунок 2. Установка дополнения Text

Чтобы загрузить папку с документами выбираем виджет Import Documents из раздела Text Mining и добавляем на холст (см. рис. 3).

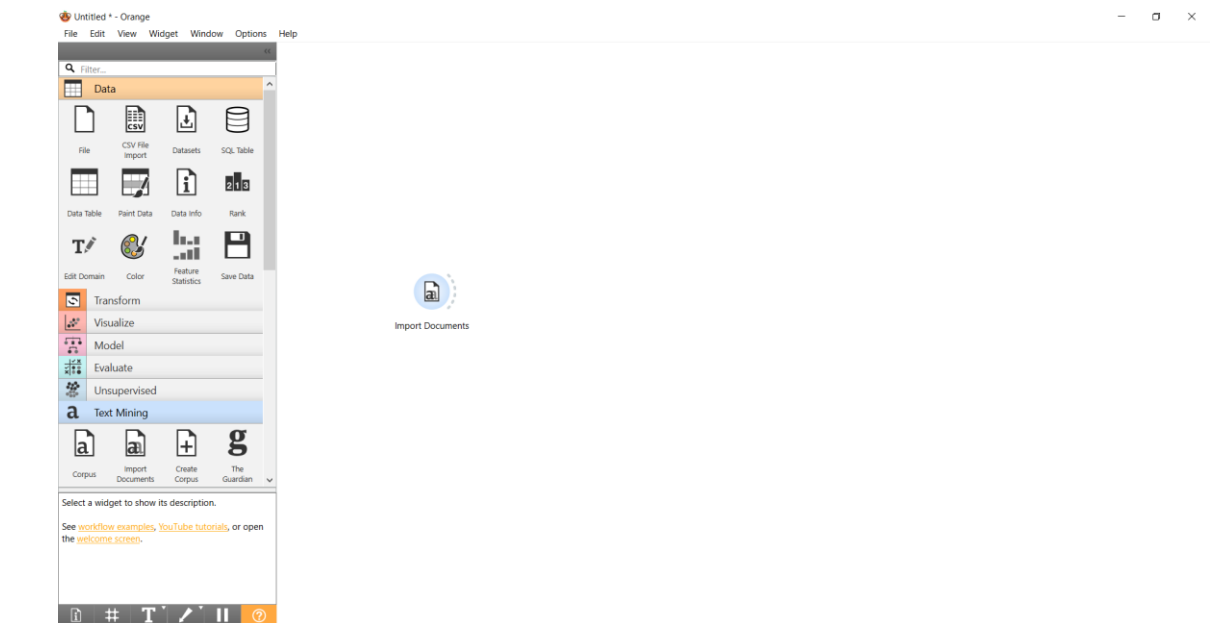


Рисунок 3. Добавление виджета Import Documents на холст

Открываем виджет Import Documents, добавляем папку документы с текстами и с файлом с кастомными стоп-словами, и выбираем русский язык (см.рис.4).

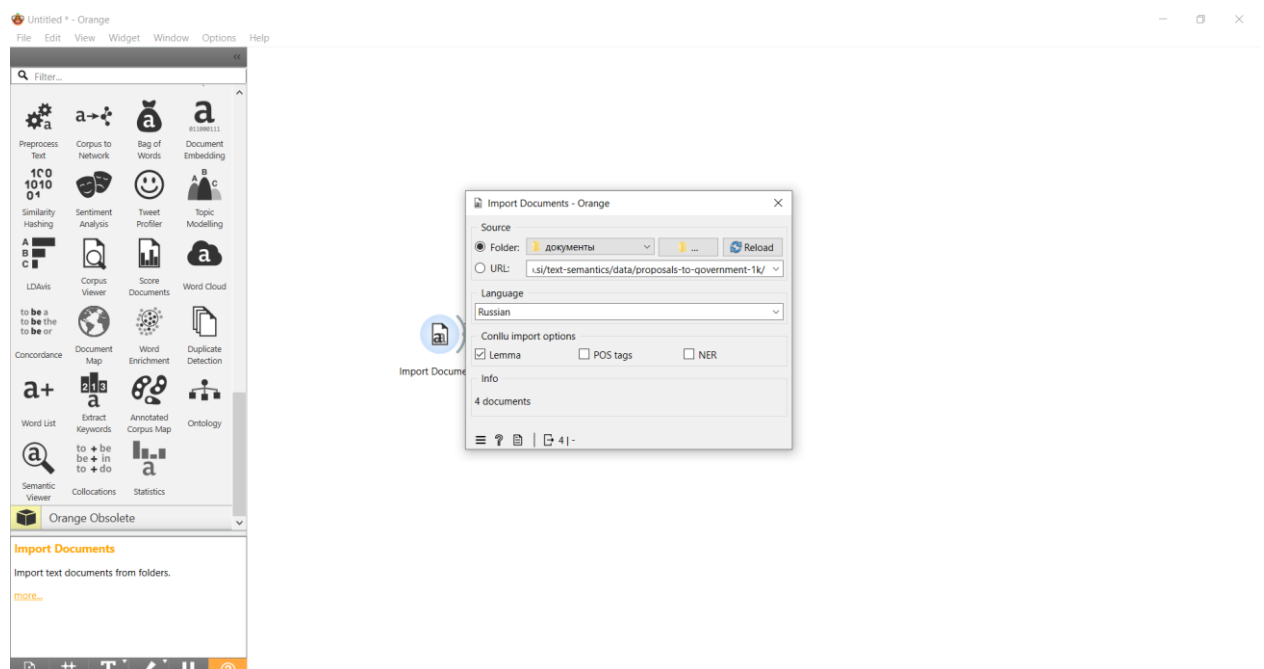


Рисунок 4. Добавление папки с файлами

Далее добавляем виджет Preprocess Text на холст, и соединяем с виджетом Import Documents (см.рис.5).



Рисунок 5. Добавление виджета Preprocess Text на холст

Открываем окно виджета Preprocess Text. В открывшемся окне необходимо выбрать настройки обработки текста. Также надо добавить файл с кастомными стоп-словами, для этого в Stopwords выбираем русский язык, и добавляем файл custom_stopwords.txt (см.рис.6).

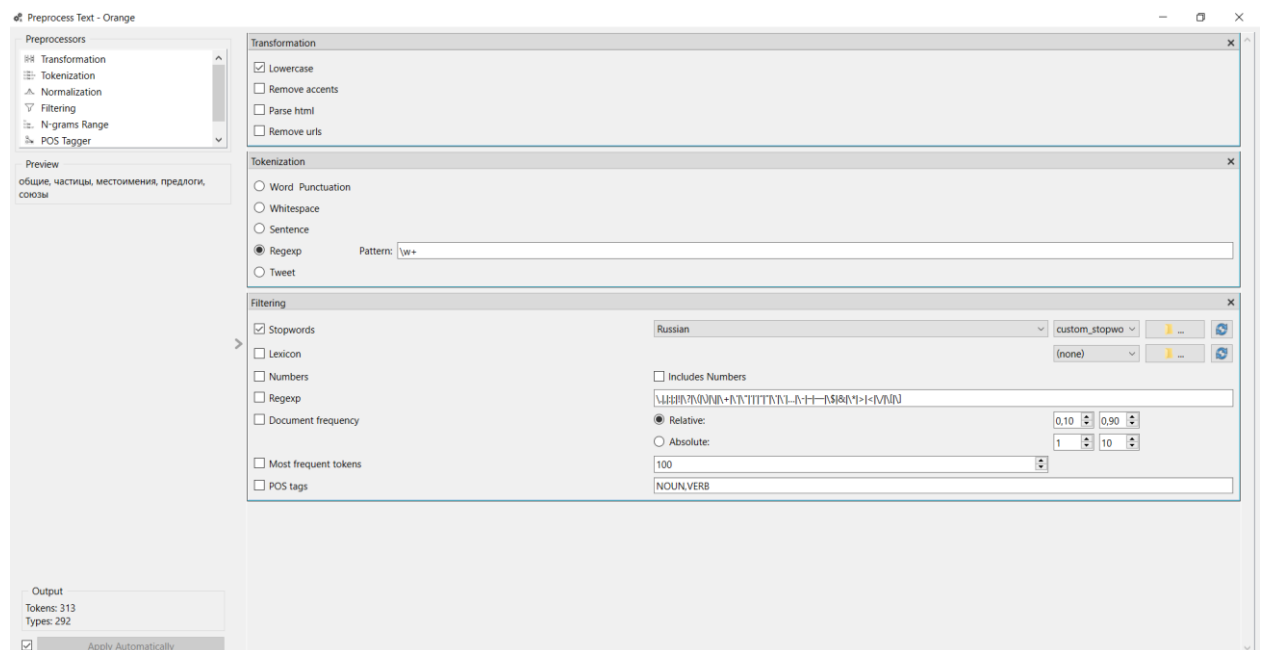


Рисунок 6. Предварительная обработка текста

Далее добавляем виджет Document Embedding на холст, и соединяем с виджетом Preprocess Text. Виджет Document Embedding представляет слова в многомерном пространстве таким образом, что слова со схожими значениями имеют сходное вложение. Это означает, что каждое слово сопоставляется с вектором вещественных чисел, представляющих слово (см.рис.7).

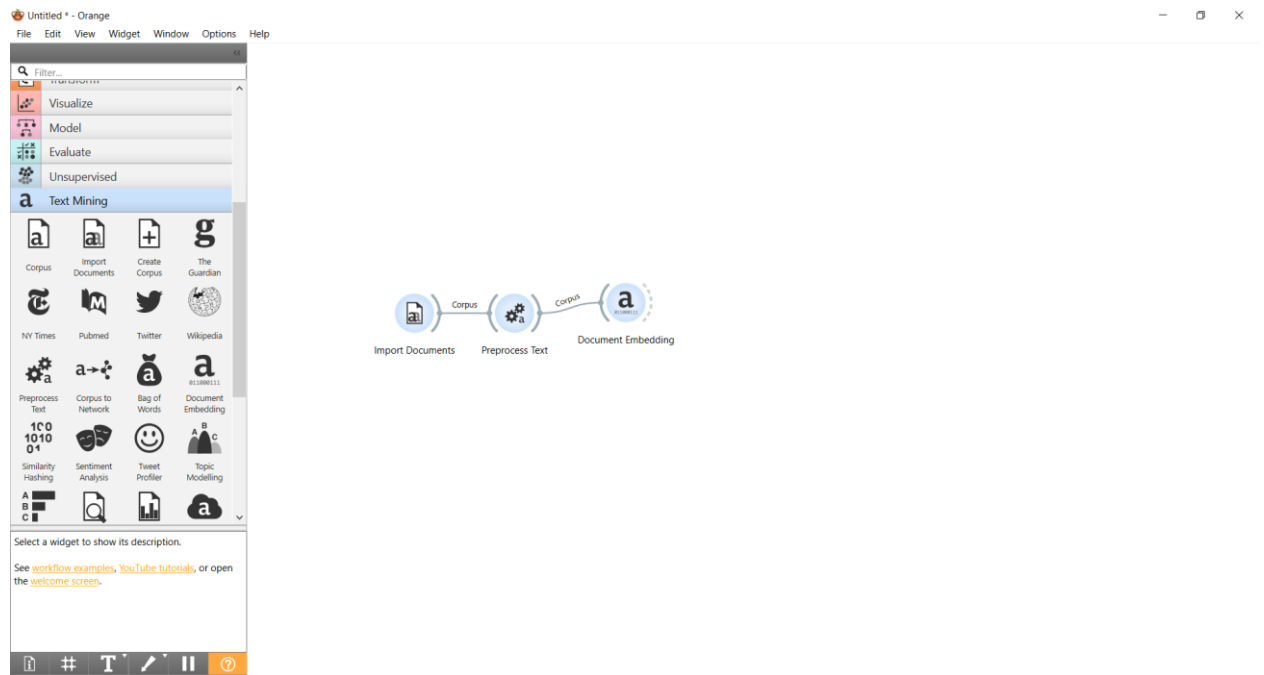


Рисунок 7. Добавление виджета Document Embedding

Открываем виджет Document Embedding, и в появившемся окне выбираем Multilingual SBERT (см.рис.8).

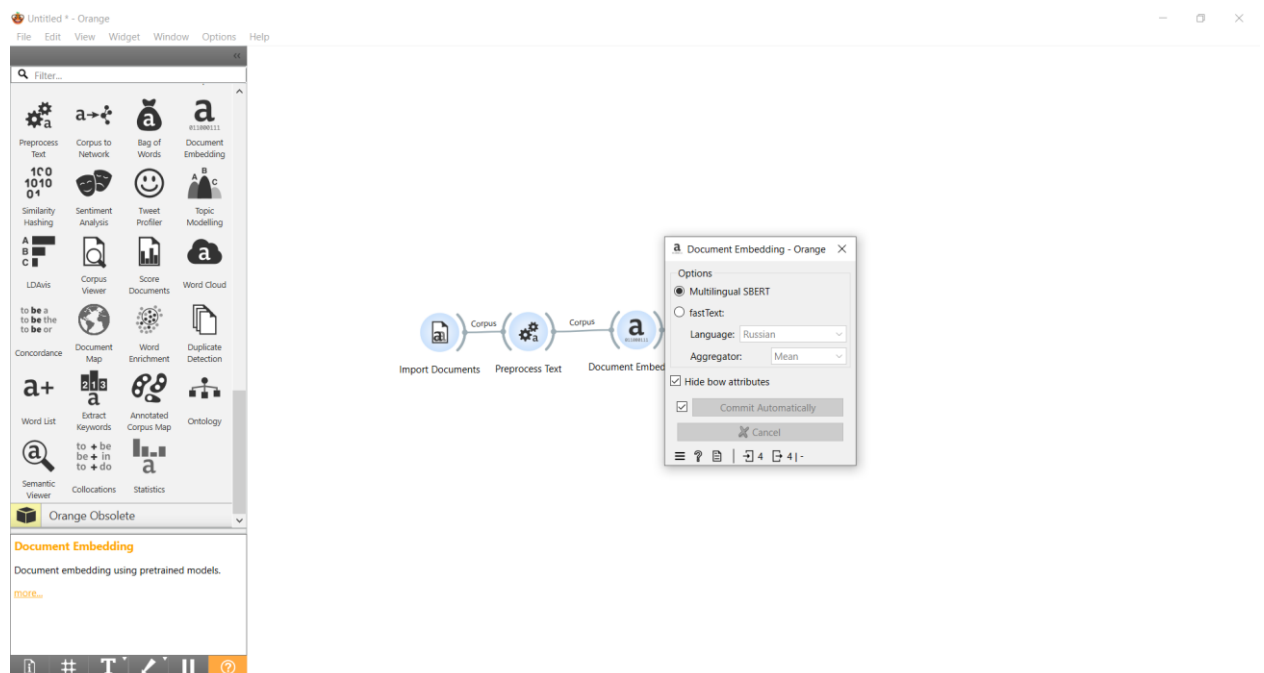


Рисунок 8. Изменение настроек виджета Document Embedding

Далее необходимо добавить виджет t-SNE на холст для визуализации, и соединить с виджетом Document Embedding (см.рис.9).

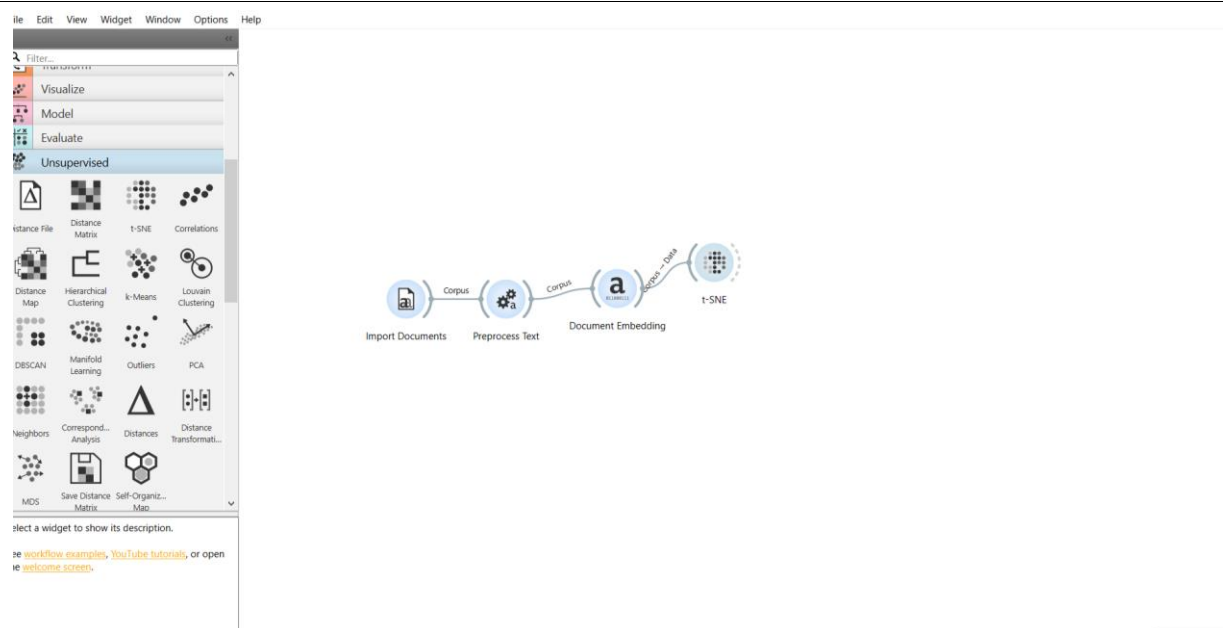


Рисунок 9. Добавление виджета t-SNE

Добавляем виджет Corpus Viewer для просмотра содержимого файлов, и соединяем с виджетом t-SNE (см.рис.10).

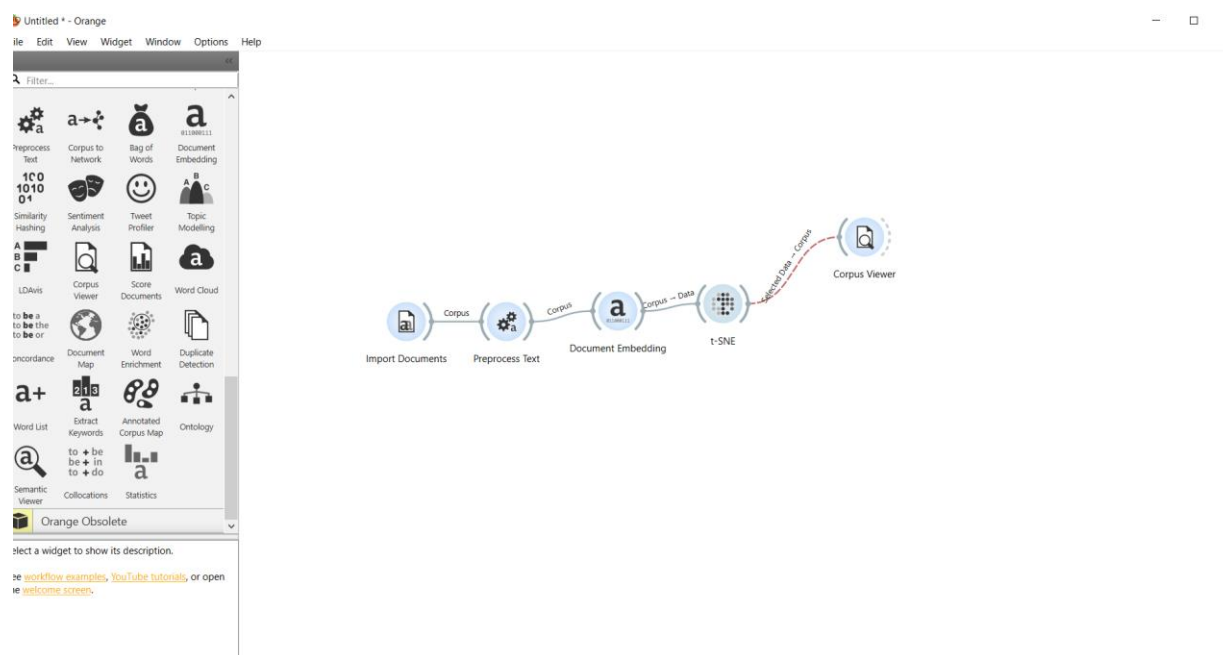


Рисунок 10. Добавление виджета Corpus Viewer

Для извлечения ключевых слов добавляем виджет Extract Keywords на холст, и соединяем с виджетом t-SNE (см.рис.11).

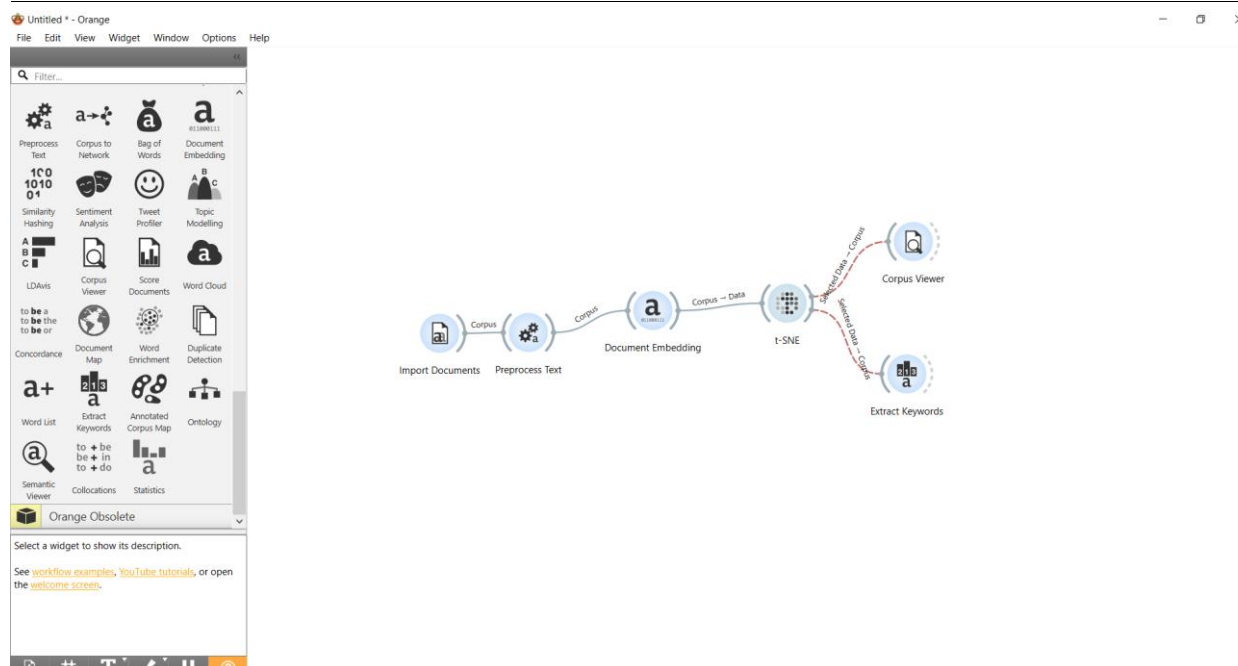


Рисунок 11. Добавление виджета Extract Keywords на холст

Далее для того, чтобы просмотреть какие были извлечены ключевые слова открываем виджет t-SNE, и выбираем любую точку файла, но только не выбираем файл с кастомными стоп-словами. На изображении видно, что был выбран файл с текстом про здоровое питание, и были извлечены ключевые слова «жизни», «здоровое», «питание» (см.рис.12).

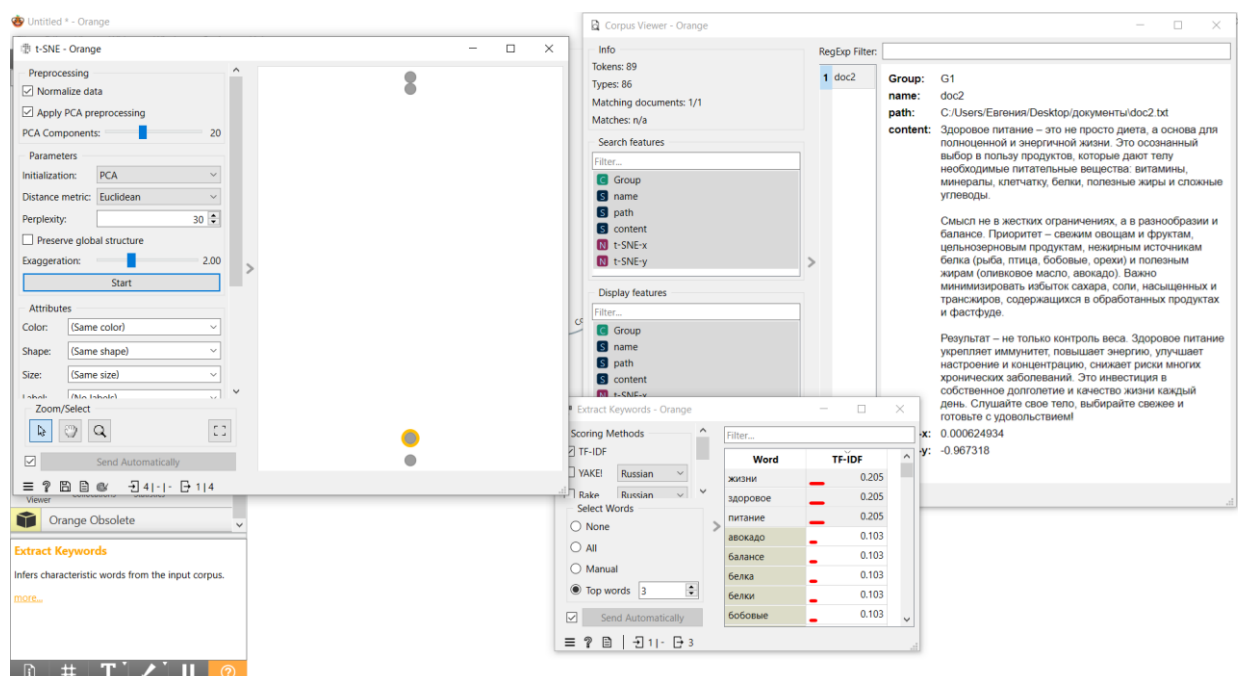


Рисунок 12. Просмотр извлеченных ключевых слов из текста про здоровое питание

Для текста про экотуризм были извлечены ключевые слова «культуру», «экотуризм», «активной» (см.рис.13).

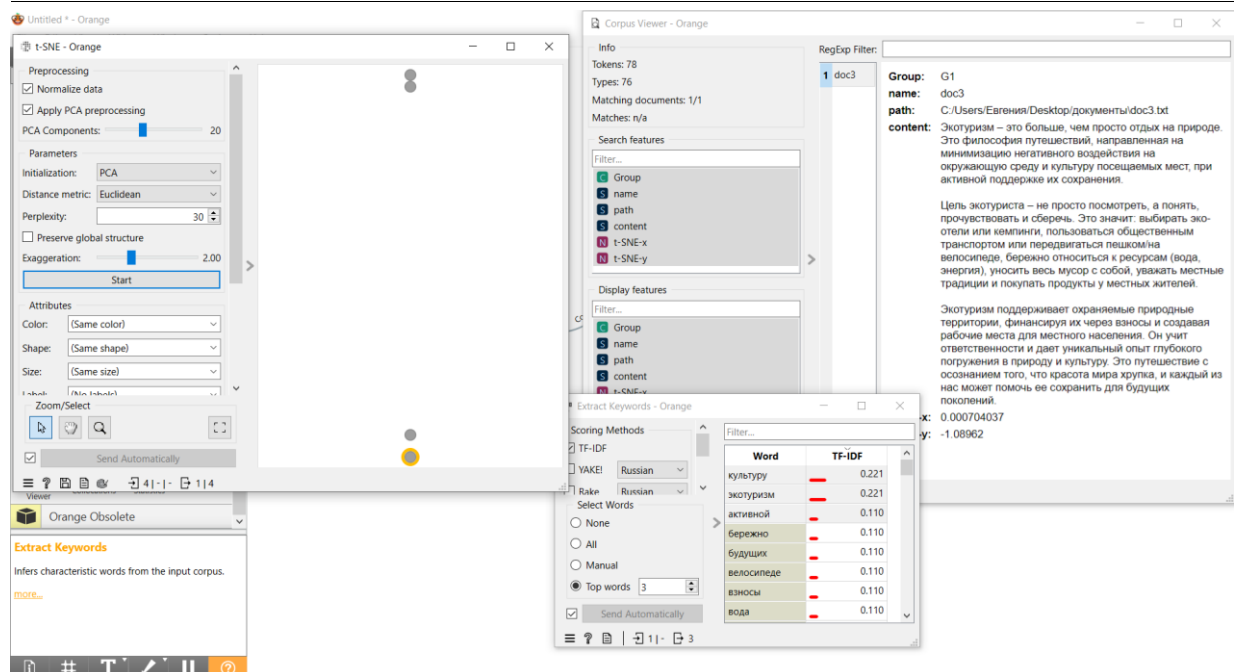


Рисунок 13. Просмотр извлеченных ключевых слов из текста про экотуризм

Для текста про искусственный интеллект были извлечены ключевые слова «ии», «решения», «алгоритмы» (см.рис.14).

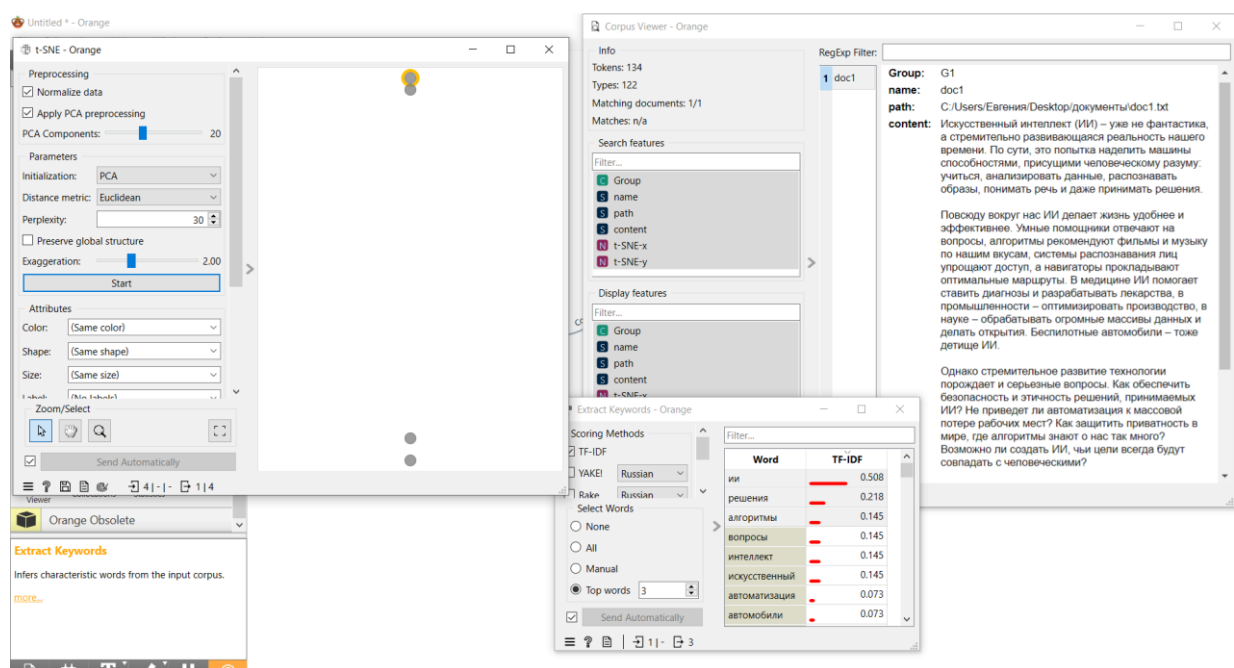


Рисунок 14. Просмотр извлеченных ключевых слов из текста про искусственный интеллект

В итоге получилась итоговая схема, с помощью которой можно извлечь ключевые слова из текста (см.рис.15).

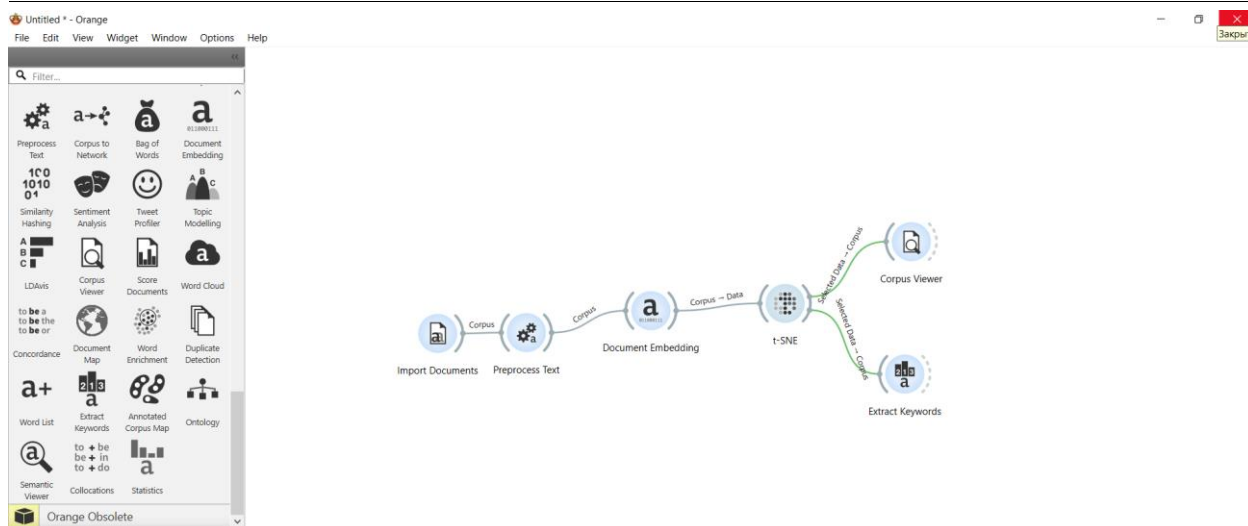


Рисунок 15. Итоговая схема

4 Выводы

В данной работе было выполнено извлечение ключевых слов из набора текстовых документов с помощью программного пакета визуального программирования на основе компонентов для визуализации данных Orange. С помощью виджетов Import Documents, Preprocess Text, Document Embedding, t-SNE, Corpus Viewer, Extract Keywords выполнили извлечение ключевых слов из набора текстовых документов и получили итоговую схему.

Библиографический список

1. Малахов Д. А. и др. Семантический поиск как средство взаимодействия с электронной библиотекой //Труды XVIII Межд. конф. DAMDID/RCDL. – 2016. – С. 85-91.
2. Вороньков В. С. Метод семантического поиска информации в информационных системах //Вестник Воронежского государственного технического университета. 2007. Т. 3. №. 12. С. 122-125.
3. Гринченков Д. В. и др. Сравнительный анализ алгоритмов интеллектуального анализа данных // Моделирование. Теория, методы и средства : Материалы 16-ой Международной научно-практической конференции, посвященной 110-летию Южно-Российского государственного политехнического университета (НПИ) имени М.И. Платова, Новочеркасск, 06–07 декабря 2016 года. Новочеркасск: ООО "Лик", 2016. С. 263.
4. Нгок Н. Б., Тузовский А. Ф. Обзор подходов семантического поиска //Доклады Томского государственного университета систем управления и радиоэлектроники. 2010. №. 2-2 (22). С. 234-237.
5. Нэй Л., Каунг М. Х. Семантический информационный поиск //Проблемы разработки перспективных технологических систем. 2017. С. 100-103.