УДК 004

Семантический поиск слов с помощью программного пакета визуального программирования Orange

Голубева Евгения Павловна
Приамурский государственный университет имени Шолом-Алейхема
Студент

Аннотация
Цель данной статьи – выполнить семантических поиск слов. Для выполнения семантического поиска слов был использован программный пакет визуального программирования на основе компонентов для визуализации данных Orange и набор данных различный слов. С помощью средств визуализации Orange выполнили семантических поиск слов и получили итоговую схему.
Ключевые слова: Orange, виджет, слова, семантический.

Semantic word search using the Orange visual programming software package

Golubeva Evgeniya Pavlovna
Sholom-Aleichem Priamursky State University
Student

Abstract
The purpose of this article is to perform semantic word searches. To perform semantic word search, a visual programming software package based on Orange data visualization components and a dataset of different words were used. Using Orange visualization tools, we performed semantic word searches and obtained the final scheme.
Keywords: Orange, widget, words, semantic.

1 Введение

1.1 Актуальность

Семантический поиск позволяет находить слова и понятия, связанные по смыслу, что имеет важное практическое применение в таких областях, как информационный поиск, анализ тональности текста, категоризация документов и т.д.

Кроме того, семантический поиск становится все более востребованным в связи с быстрым ростом объемов текстовой информации в цифровом формате. Традиционные методы поиска, основанные на ключевых словах, часто оказываются недостаточно эффективными, поскольку не учитывают контекст и смысловые связи между понятиями. Использование семантических методов позволяет повысить точность и релевантность

поиска, что имеет большое значение для широкого круга приложений, от информационных систем до систем искусственного интеллекта.

Использование визуального программирования в Orange упрощает разработку и применение таких семантических алгоритмов, делая их доступными для широкого круга пользователей, не обладающих глубокими навыками программирования.

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 семантических поиск слов.

2 Материалы и методы

Для выполнить семантического поиска слов используется программа Orange. Работа будет происходить на готовом наборе данных состоящий из различных слов, скачать которые можно по ссылке:

https://docs.google.com/spreadsheets/d/1ZBCg9WHKwWI-9mOEF6CJJ_AO7Fk_rut_/edit?usp=sharing&ouid=104272149632818699735&rtpof=true&sd=true

3 Результаты и обсуждения

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

Создадим новый файл (см.рис.1).

Рисунок-1 Создание нового файла

Для решения задачи классификации необходимо установить дополнение Text. Для того, чтобы скачать дополнение, необходимо перейти в Options, далее в Add-ons, в появившимся окне выбираем Text (см.рис.2).

Рисунок-2 Установка дополнения Text

Чтобы загрузить набор данных, содержащий различные слова, выбраем виджет Corpus из раздела Text Mining и добавляем на холст (см. рис. 3).

Рисунок- 3 Добавление виджета Corpus на холст

Открываем виджет Corpus и добавляем набор данных Words.xlsx (см.рис.4).

Рисунок-4 Добавление набора данных Words.xlsx

Далее добавляем виджет Corpus Viewer на холст, и соединяем с виджетом Corpus (см.рис.5).

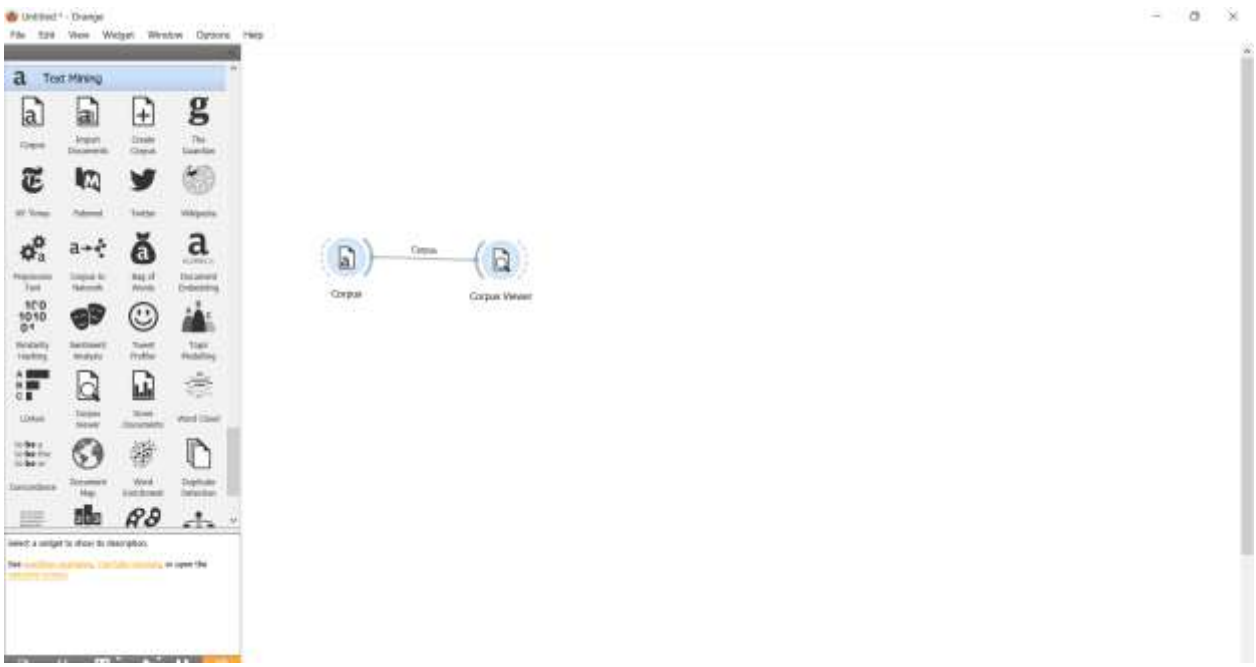

Рисунок-5 Добавление виджета Corpus Viewer на холст

Открываем окно виджета Corpus Viewer. В открывшемся окне можно увидеть, что набор данных содержит 150 различных слов (см.рис.6)

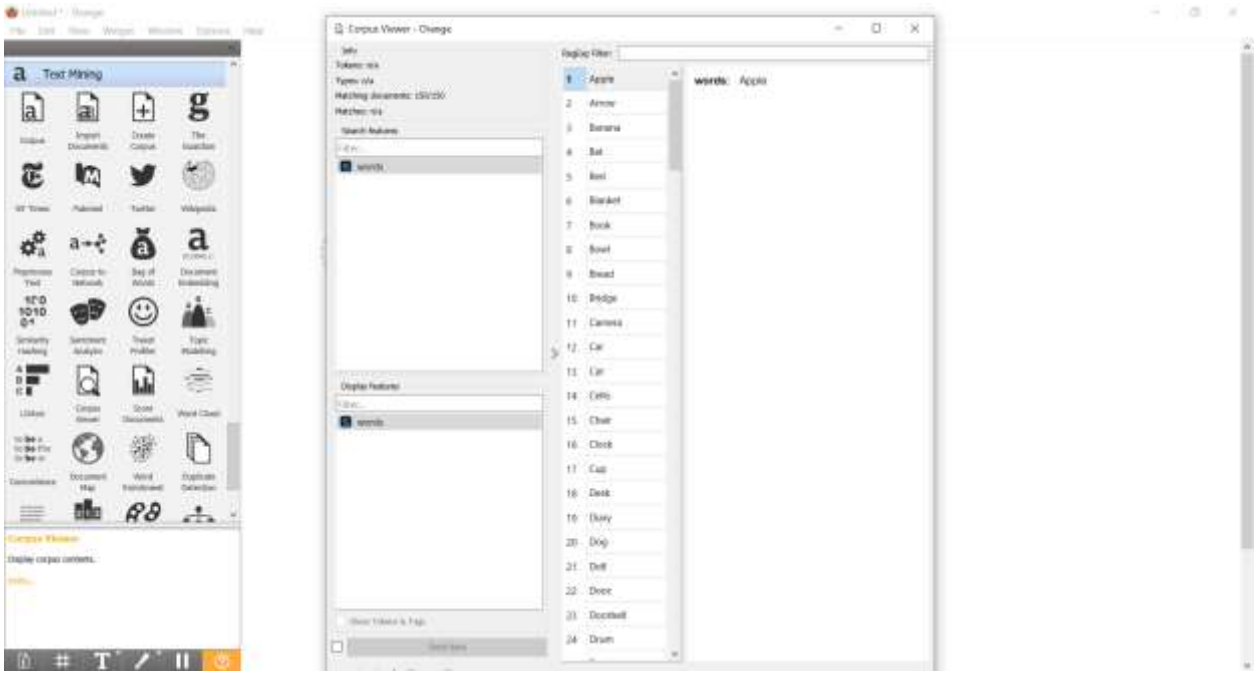

Рисунок-6 Просмотр набора данных

Далее создадим процесс семантического поиска слов. Для этого добавим виджет Create Corpus на холст (см.рис.7). Виджет Create Corpus будет отвечать за ввод слов, которые будут использоваться для семантического поиска.

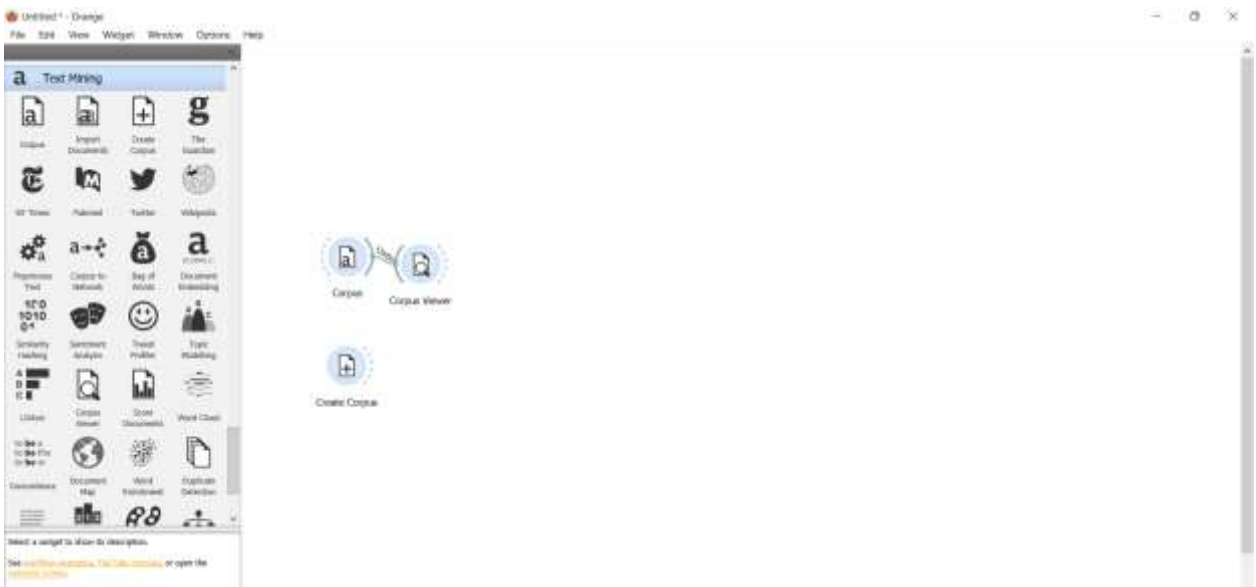

Рисунок - 7 Добавление виджета Create Corpus

Открываем виджет Create Corpus, выбираем английский язык, в названии файла пишем New text, и в окне ввода напишем учеба на английском (см.рис.8).

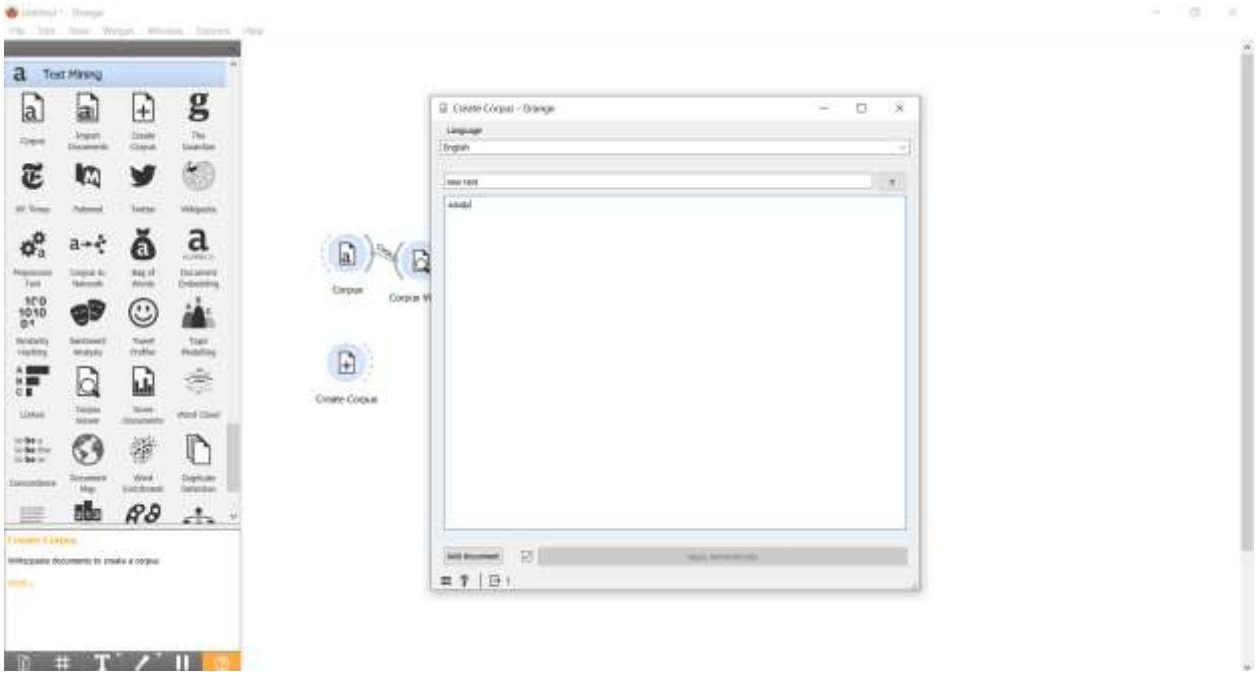

Рисунок - 8 Ввод текста в виджет Create Corpus

Для того, чтобы посмотреть какую информацию содержит виджет Create Corpus, добавим виджет Corpus Viewer, и соединяем с Create Corpus (см.рис.9).

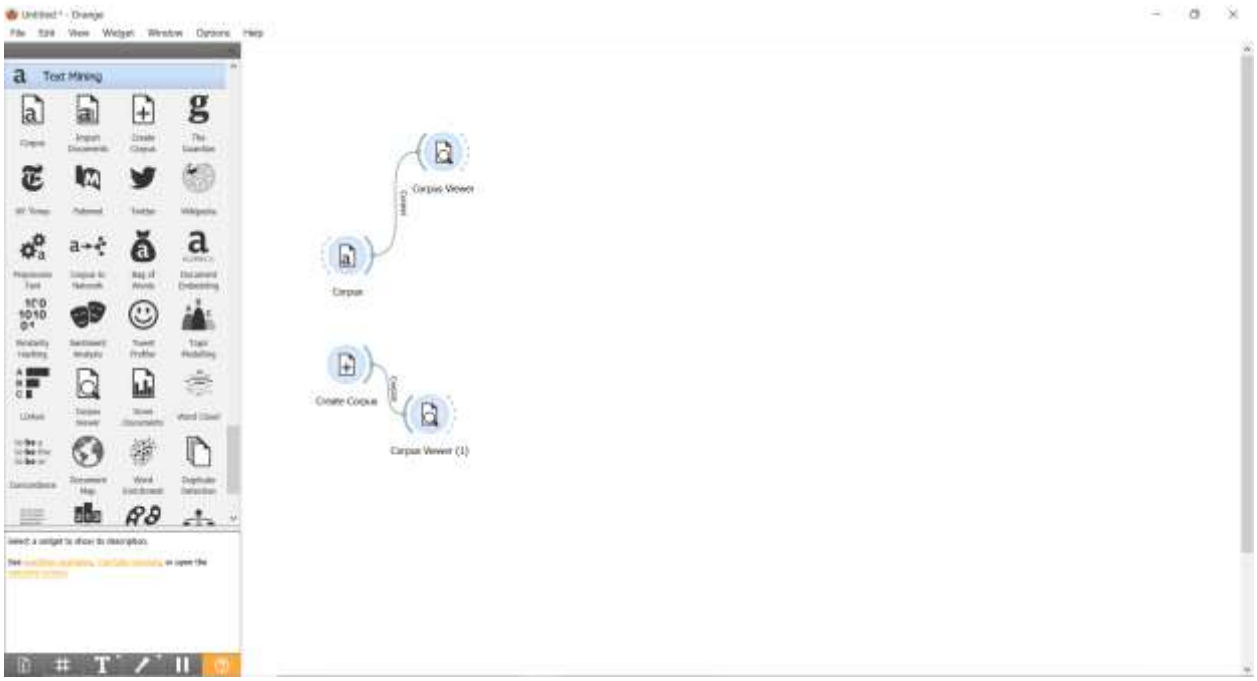

Рисунок - 9 Добавление виджета Corpus Viever

Открываем виджет Corpus Viever (1), и можем увидеть, что виджет содержит один документ (см.рис.10).

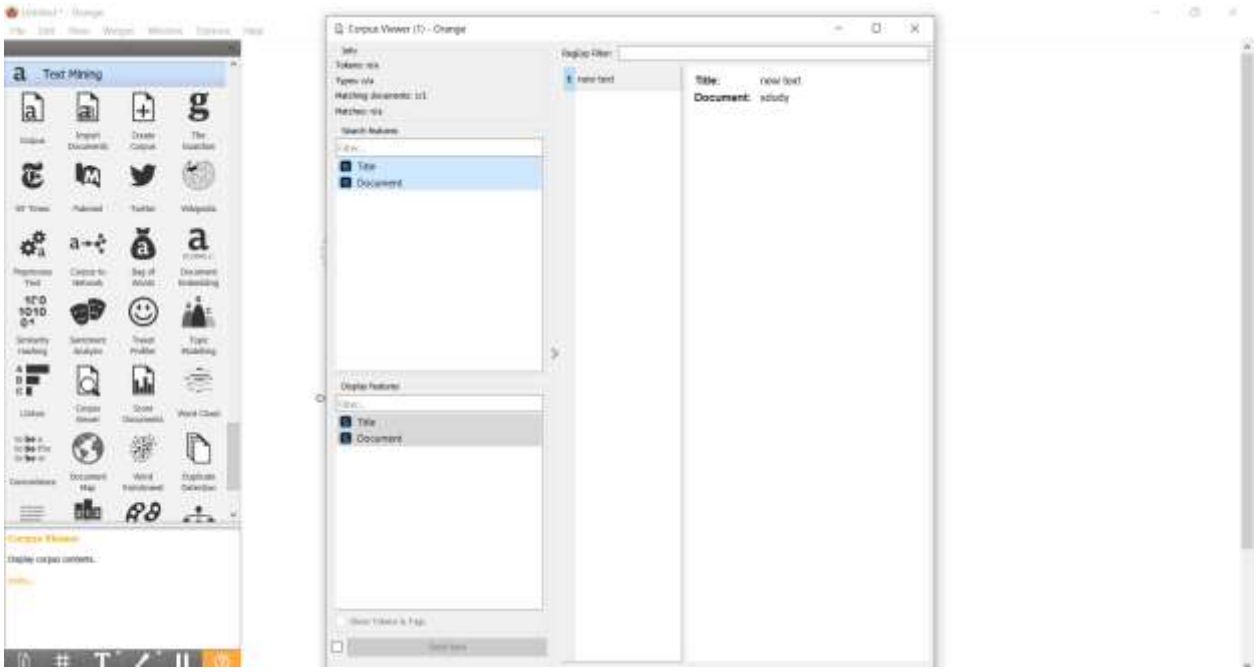

Рисунок - 10 Просмотр данных виджета Corpus Viever (1)

Для того, чтобы найти более близкие слова из набора данных по схожим признакам слов из Creat Corpus, добавим виджет Document Embedding на холст, и соединяем с виджетом Corpus (см.рис.11).

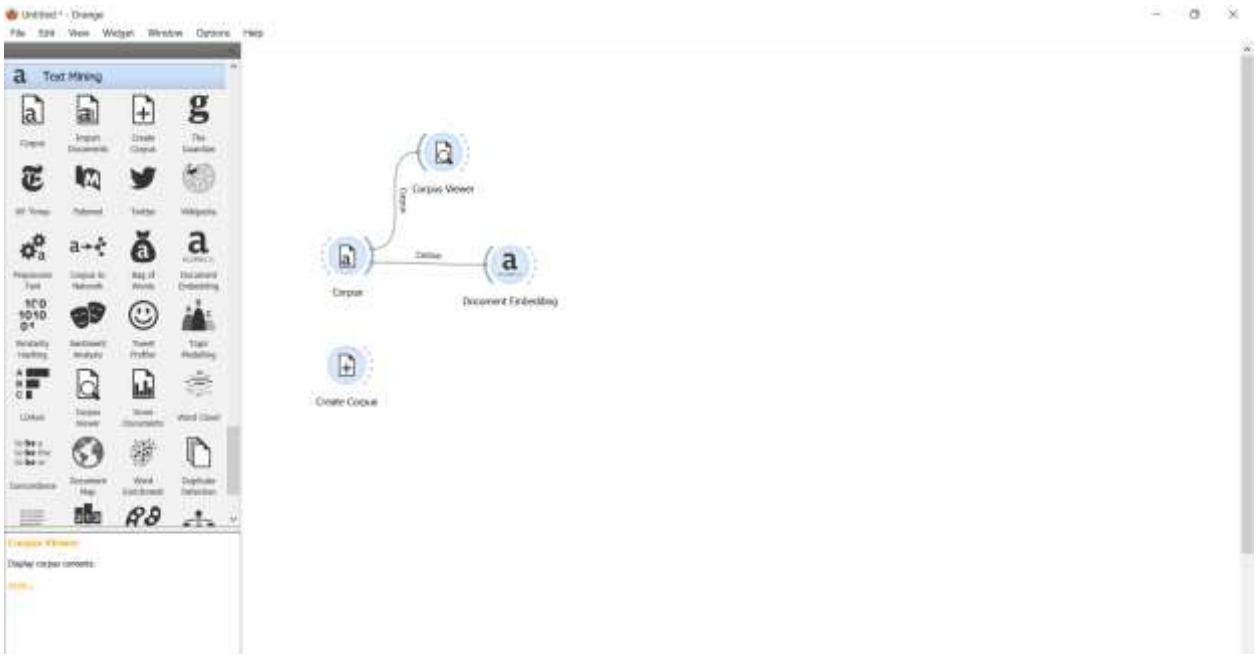

Рисунок - 11 Добавление виджета Document Embedding на холст

Перепроверим данные виджета Document Embedding, для этого добавим виджет Data Table на холст и соединим с Document Embedding (см.рис.12).

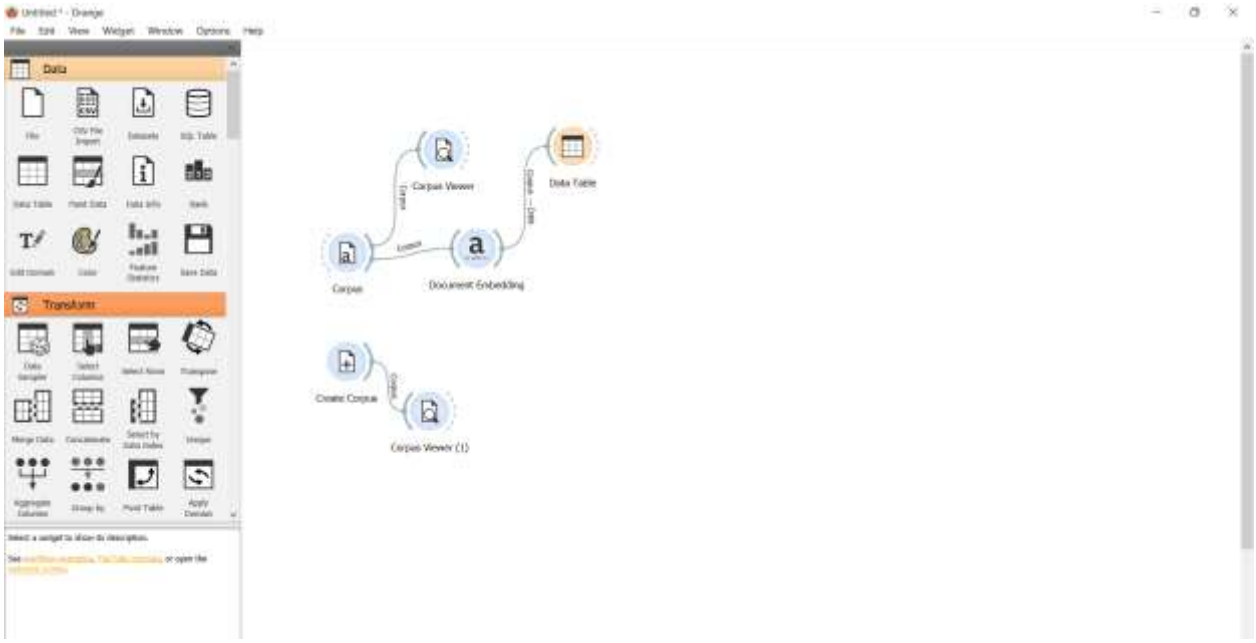

Рисунок - 12 Добавление виджета Data Table на холст

Откроем виджет Data Table. С помощью таблицы можем увитеть, что для каждого слова добавились 300 дополнительных признаков (см.рис.13).

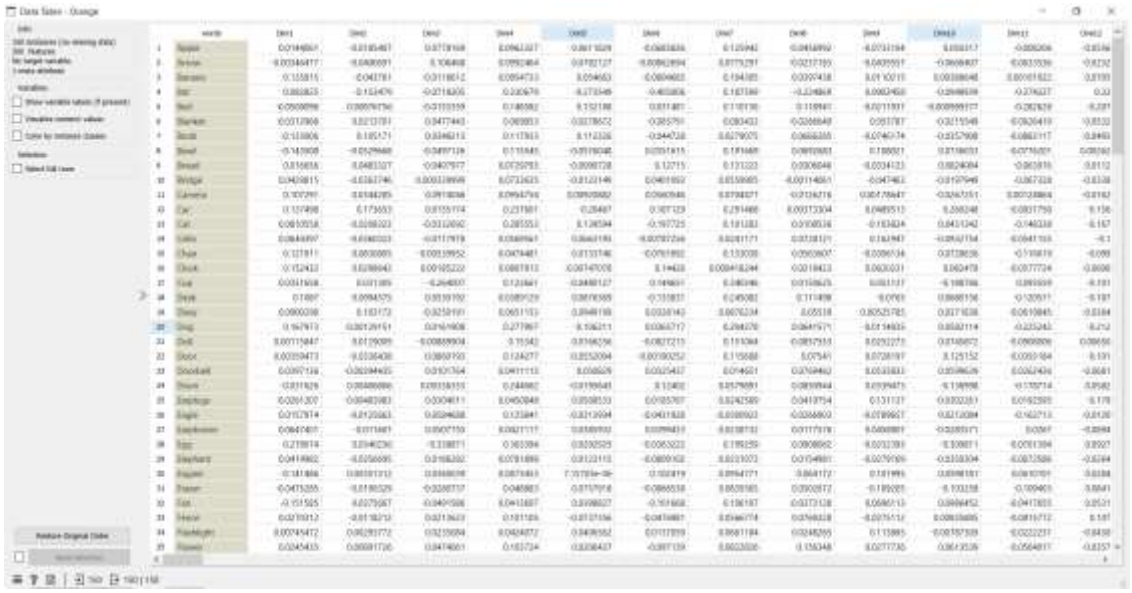

Рисунок - 13 Просмотр данных виджета Data Table

Также аналогично подключим виджет Document Embedding к Create Corpus. Далее подключим к Document Embedding (1) виджет Data Table для просмотра информации Document Embedding (1) (см.рис.14).

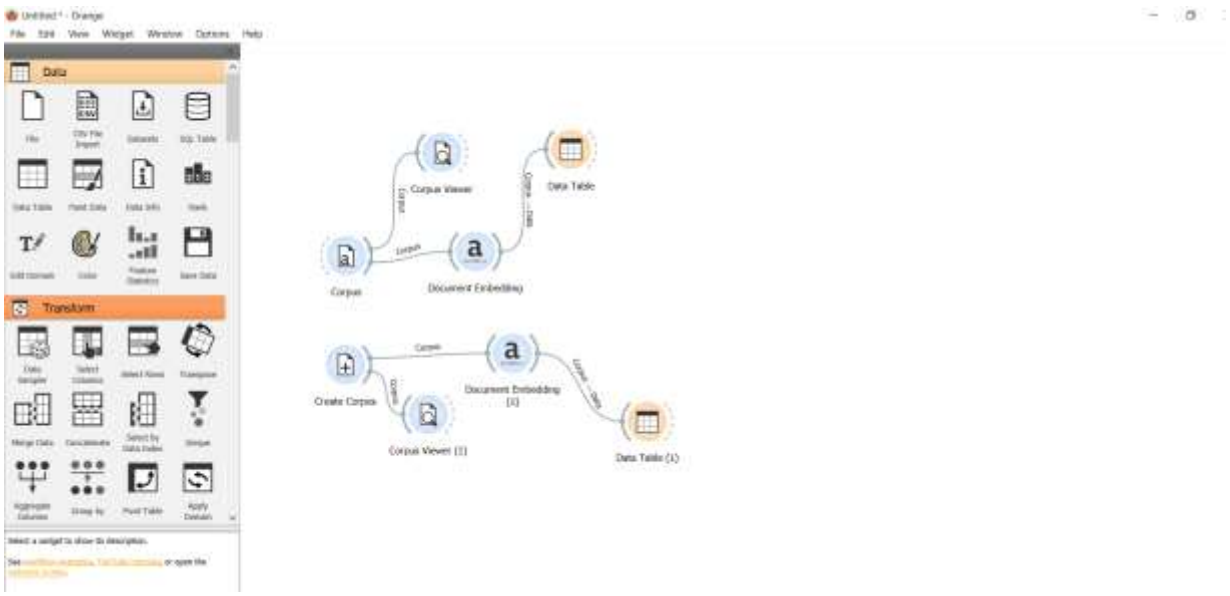

Рисунок - 14 Добавление виджетов Document Embedding и Data Table

Открываем виджет Data Table (1). С помощью таблицы можно увидеть, что у слова study добавилось 300 дополнительных признаков (см.рис.15).

Рисунок - 15 Просмотр данных виджета Data Table (1)

Далее добавляем виджет Neihbors, и к нему соединяем виджеты Document Embedding. Виджет Neihbors вычисляет ближайших соседей в данных в соответствии с ссылкой. Document Embedding будет передавать слова из набора данных, а Document Embedding (1) будет передавать слова, которые будем вводить для поиска похожих слов (см.рис.16)

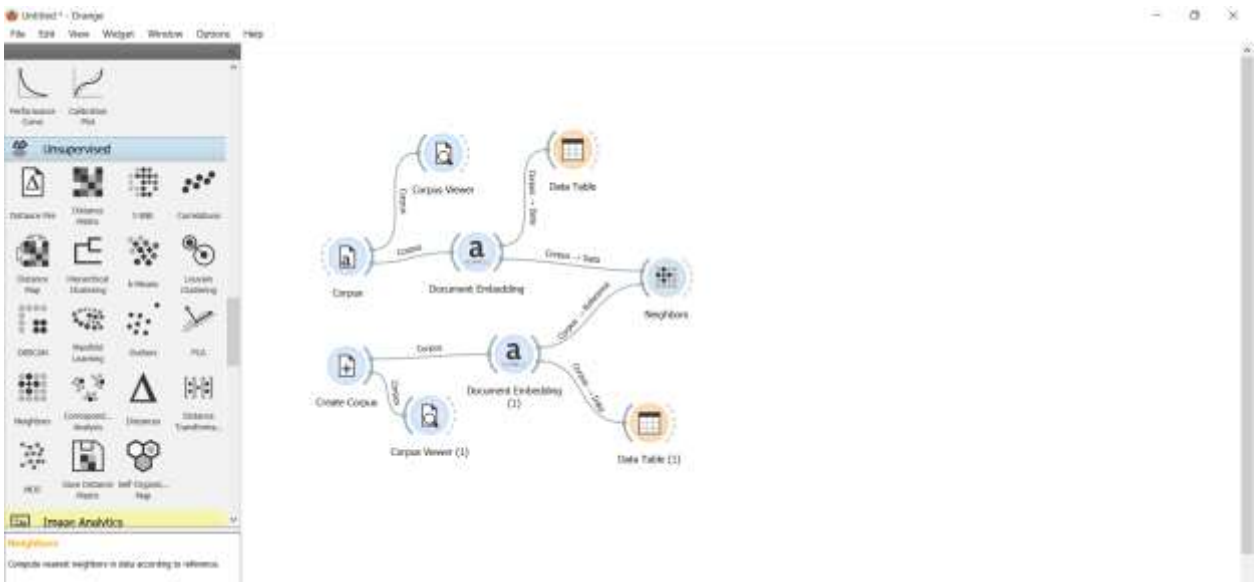

Рисунок - 16 Добавление виджета Neighbors

Открываем виджет Neighbors. В появившемся окне выбираем расстояние Cosine, и выбираем трех соседей (см.рис.17).

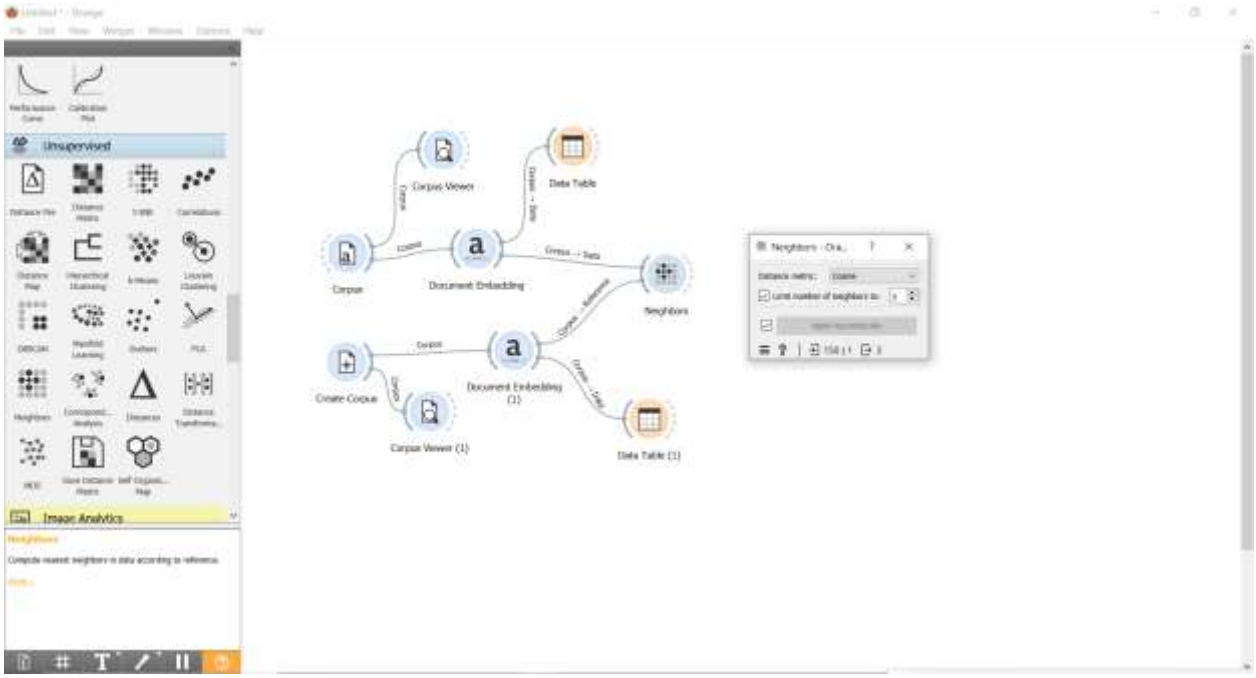

Рисунок - 17 Настройка виджета Neighbors

Для того, чтобы просмотреть как виджет Neighbors нашел схожие слова с набора данных со словом sdudy, добавим виджет Corpus Views и соединяем с виджетом Neighbors (см.рис.18).

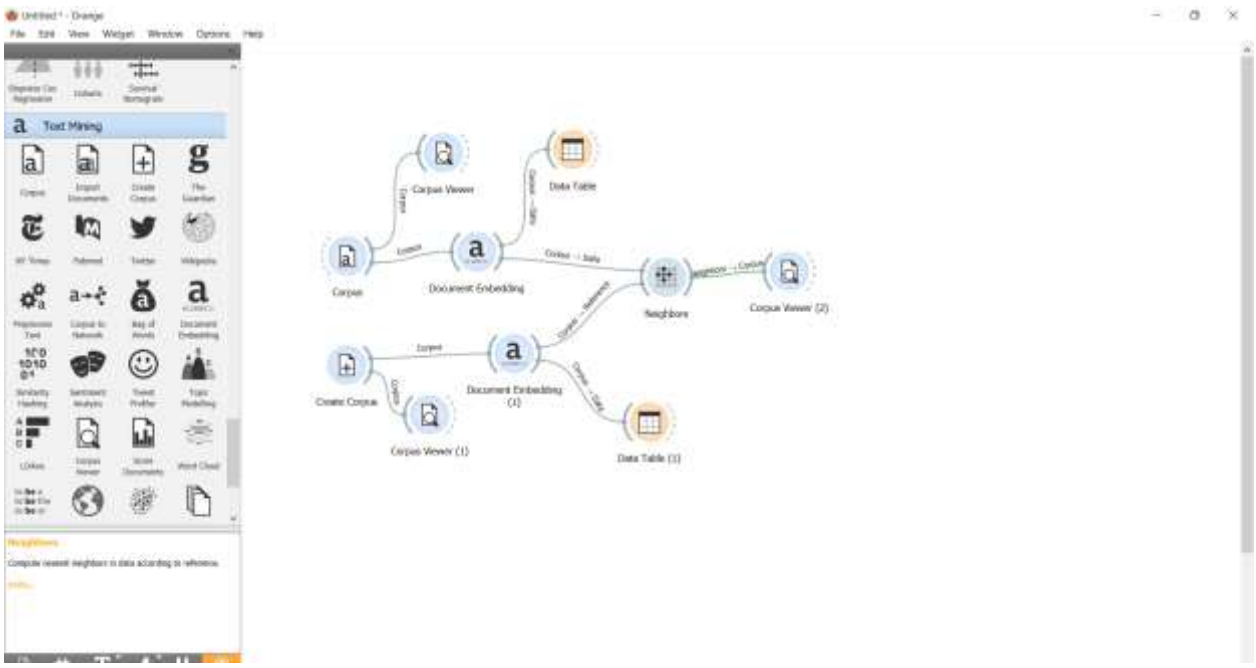

Рисунок - 18 Добавление виджета Corpus Viewer на холст

Открываем виджет Corpus Viewer (2). В появившемся окне можем увидеть слова анкета, книга и дневник, которые по смыслу схожие со словом учеба (см.рис.19).

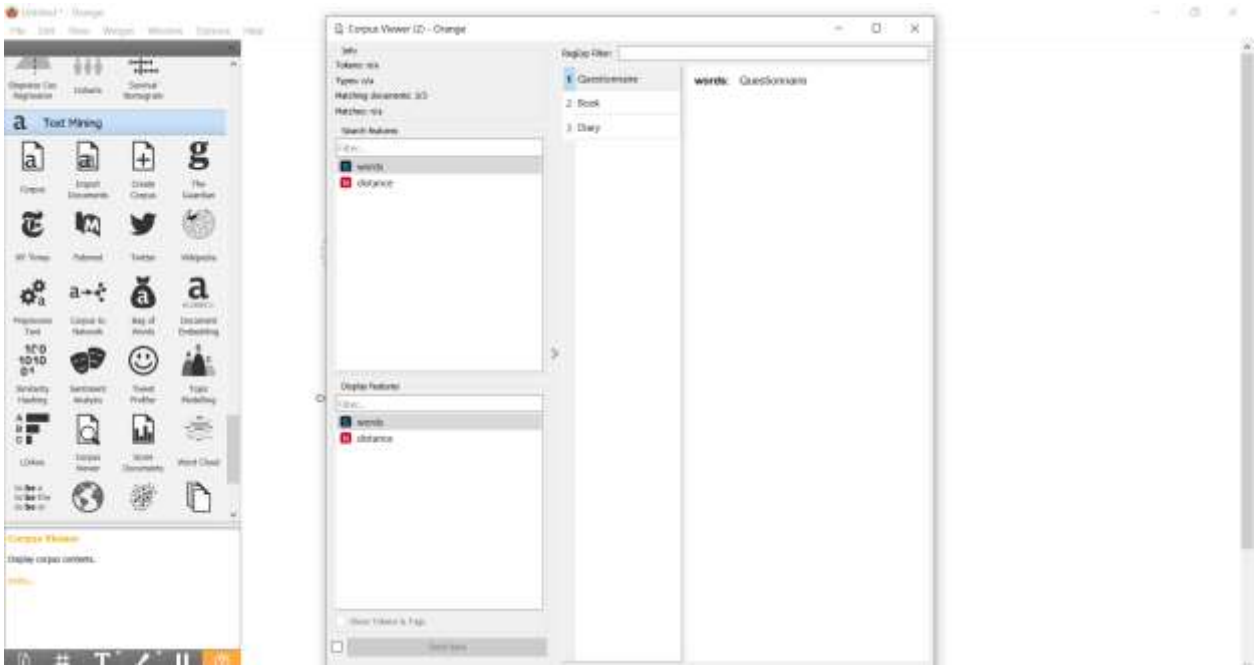

Рисунок - 19 Просмотр данных виджета Corpus Viewer (2)

Далее найдем похожие слова по смыслу для слова транспорт. Для этого введем слово Transport в виджет Create Corpus, и подождем пока похожие слова отображаться в окне виджета Corpus Viewer (2). Можем увидеть, что в виджете Corpus Viewer (2) появились слова автомобиль, мост и трактор, которые относятся к транспорту (см.рис.20).

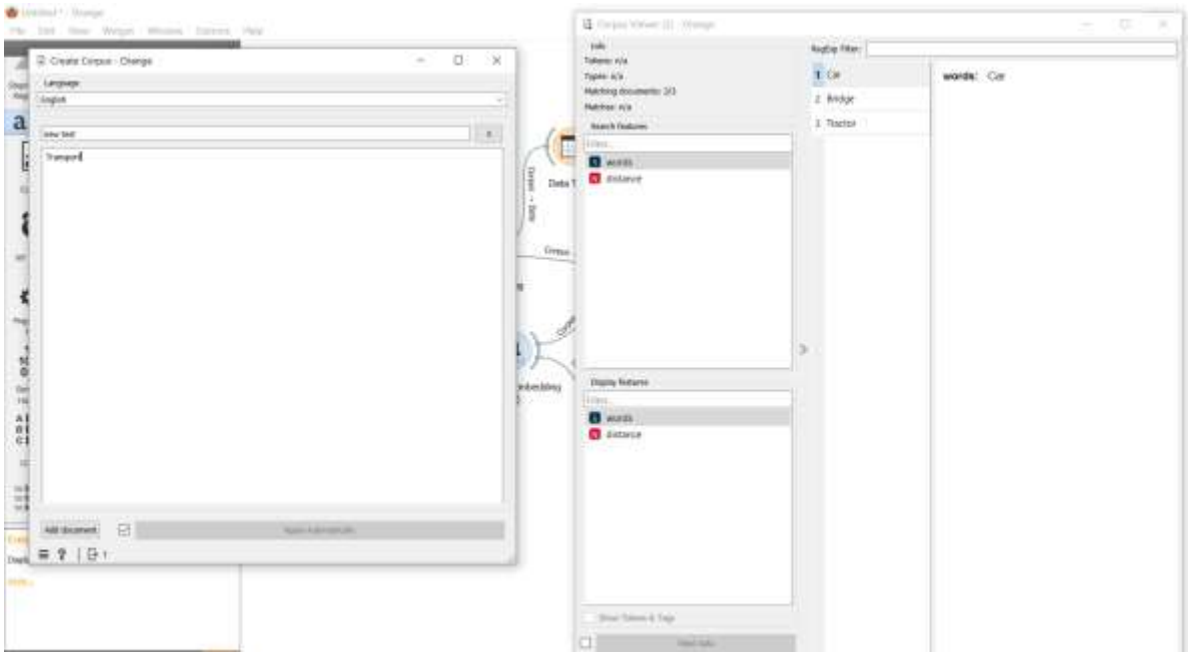

Рисунок - 20 Вывод семантический слов

Так же проверим семантический поиск слов, с помощью текста «играть на музыкальных инструментах». Для этого напишем данный текст в виджет Create Corpus, и подождем пока обработается текст. После обработки текста в виджете Corpus Viewer (2) появились слова пианино, скрипка и гитара которые являются примерами музыкальных инструментов, связанных с действием "играть". Это подтверждает, что семантический поиск успешно идентифицировал ключевые слова, связанные с заданным контекстом (см.рис.21).

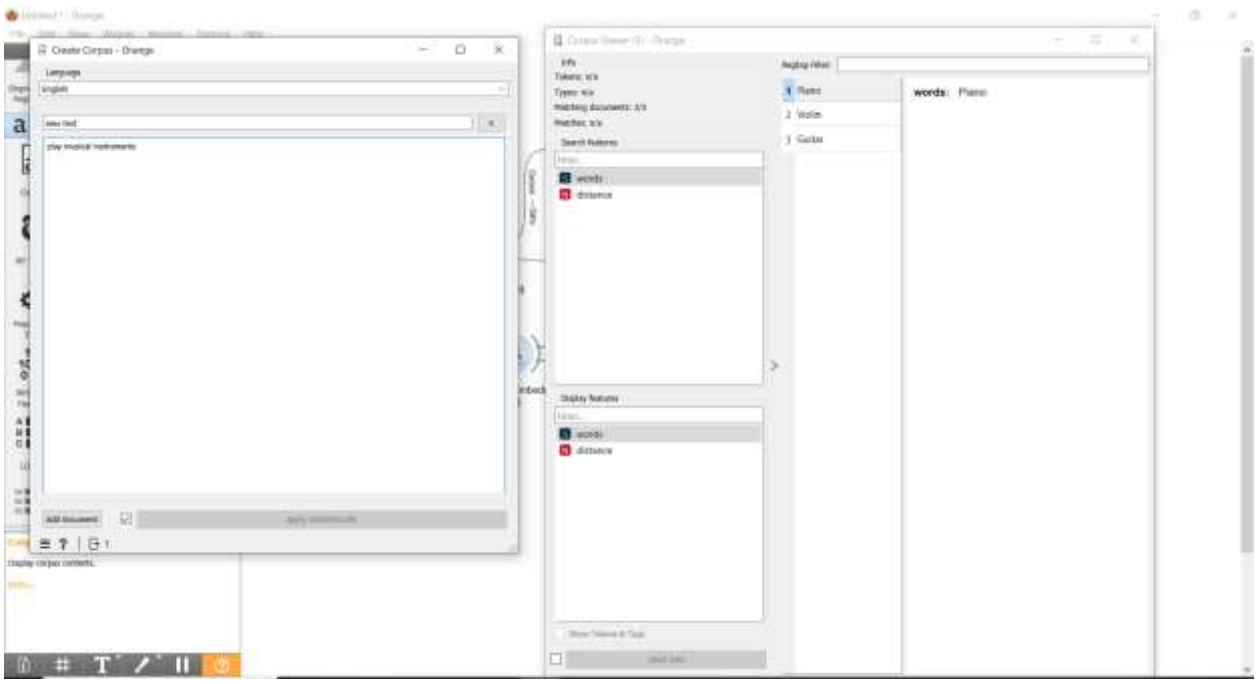

Рисунок - 21 Вывод семантический слов

В итоге получилась итоговая схема, с помощью которой можно выполнить семантического поиска слов (см.рис.22)

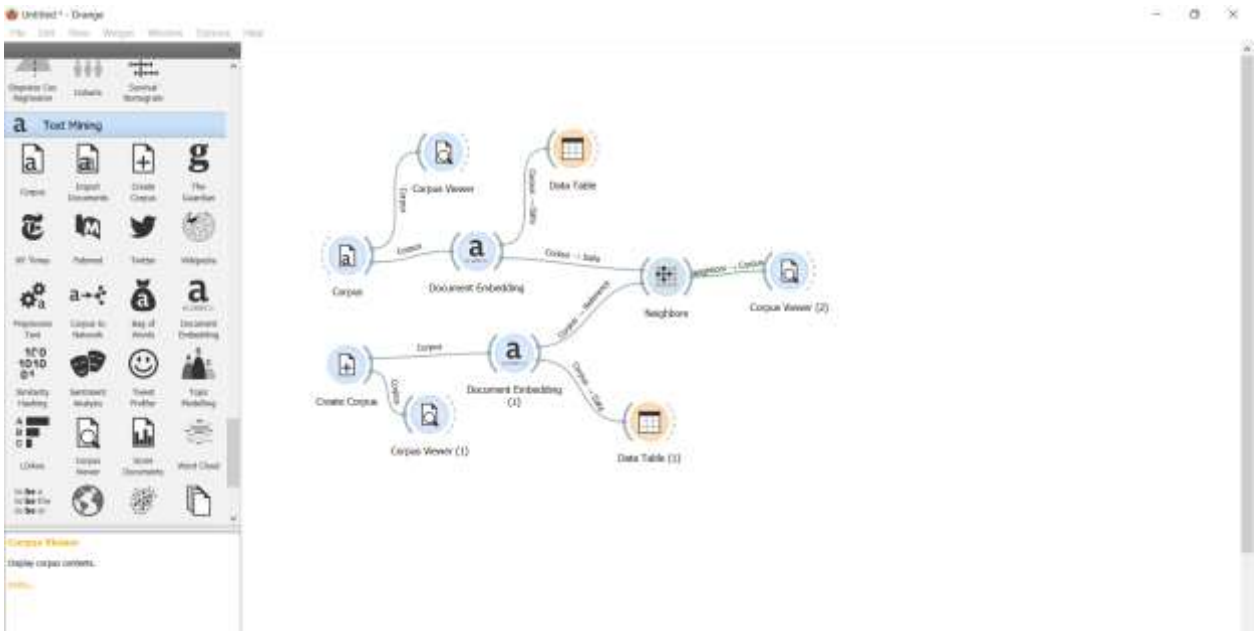

Рисунок - 22 Итоговая схема

4 Выводы

В данной работе был выполнен семантический поиск слов с помощью программного пакета визуального программирования на основе компонентов для визуализации данных Orange. С помощью виджетов Corpus, Data Table, Document Embedding, Neighbors, Create Corpus, Corpus Viewer выполнили семантический поиск слов 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, 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