УДК 004

Решение задачи кластеризации набора данных различный слов с помощью программного пакета визуального программирования Orange

Голубева Евгения Павловна
Приамурский государственный университет имени Шолом-Алейхема
Студент

Аннотация
Цель данной статьи – решить задачу кластеризации набора данных состоящий из различных слов. Для решения задачи кластеризации был использован программный пакет визуального программирования на основе компонентов для визуализации данных Orange и набор данных различный слов. С помощью средств визуализации Orange решили задачу кластеризации набора данных состоящий из различных слов.
Ключевые слова: Orange, виджет, слова, кластеризация.

Solving the problem of clustering a dataset of different words using the Orange visual programming software package

Golubeva Evgeniya Pavlovna
Sholom-Aleichem Priamursky State University
Student

Abstract
The purpose of this article is to solve the problem of clustering a dataset of different words. To solve the clustering problem, a visual programming software package based on Orange data visualization components and a dataset of various words were used. Using Orange visualization tools, we solved the problem of clustering a dataset of different words.
Keywords: Orange, widget, words, clustering.

1 Введение

1.1 Актуальность

Кластеризация является важным методом анализа данных, позволяющим группировать объекты на основе их сходства. Применение кластеризации к набору данных различных слов может привести к появлению новых знаний и пониманию сходств и различий между этими словами.

Программный пакет визуального программирования Orange предоставляет удобный и интуитивно понятный интерфейс для решения задач анализа данных, включая кластеризацию. Благодаря этому пакету исследователи и специалисты в области анализа данных могут легко применять методы кластеризации к наборам данных, в том числе и к наборам из различных слов.

1.2 Обзор исследований

Е.А. Долгих, Т.А. Першина, Л.А. Давлетшина показали методы многомерного анализа данных: кластерный анализ, дерево решений и случайный лес, анализ изображений [1]. Рассмотрел программные комплексы, включающие в свою функциональность проведение какого-либо вида кластерного анализа, изучил их особенности, перечислил преимущества и недостатки А.В. Гладилин [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]. В статье рассмотрел использование методов кластеризации в программе Orange на основе реальной базы данных. Н. Юсупов [4]. А. В. Леонов в статье рассматривал основные алгоритмы кластеризации категориальных данных применительно к различным типам пользовательских интерфейсов, определяются их достоинства и недостатки [5].

1.3 Цель исследования

Цель исследования - решить задачу кластеризации набора данных состоящий из различных слов.

2 Материалы и методы

Для решения задачи кластеризации используется программа Orange. Работа будет происходить на готовом наборе данных состоящий из различных слов, скачать которые можно по ссылке: https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Ffile.biolab.si%2Fdatasets%2Fwords.xlsx&wdOrigin=BROWSELINK

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 кластеризации 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).

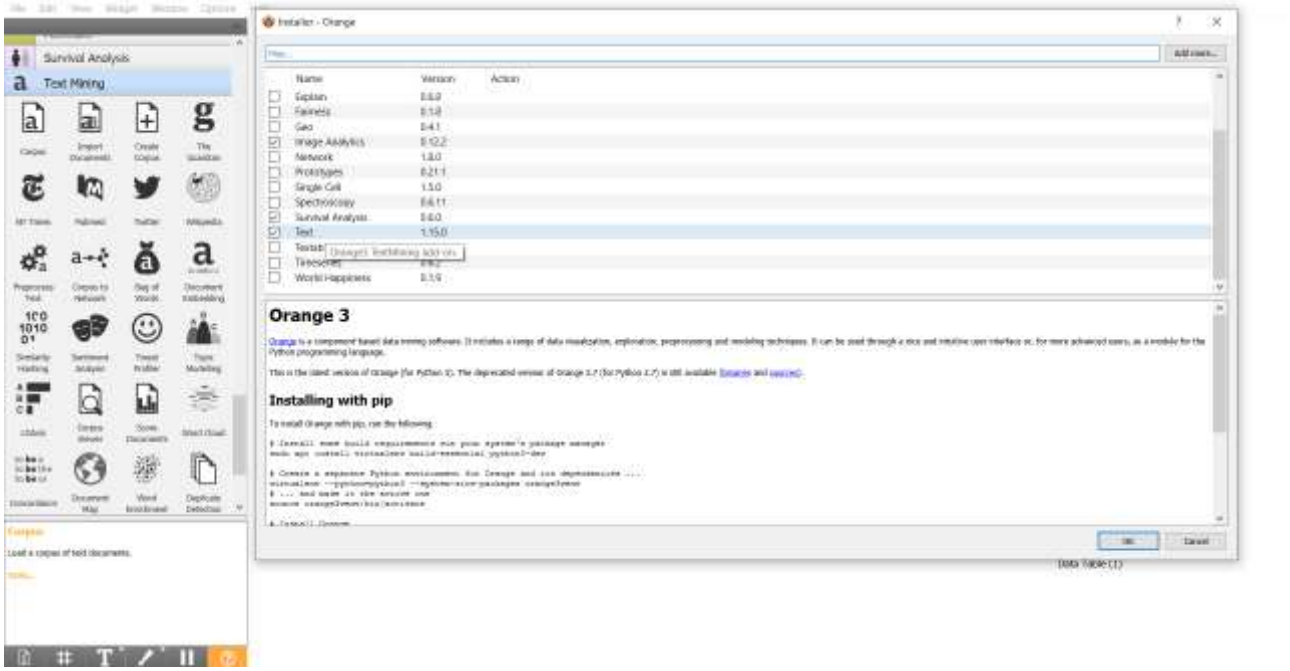

Рисунок 2 - Установка дополнения Text

Для того, чтобы загрузить набор данных состоящий из различных слов, необходимо из раздела Text Mining выбрать виджет Corpus и перенести его на холст (см.рис.3).

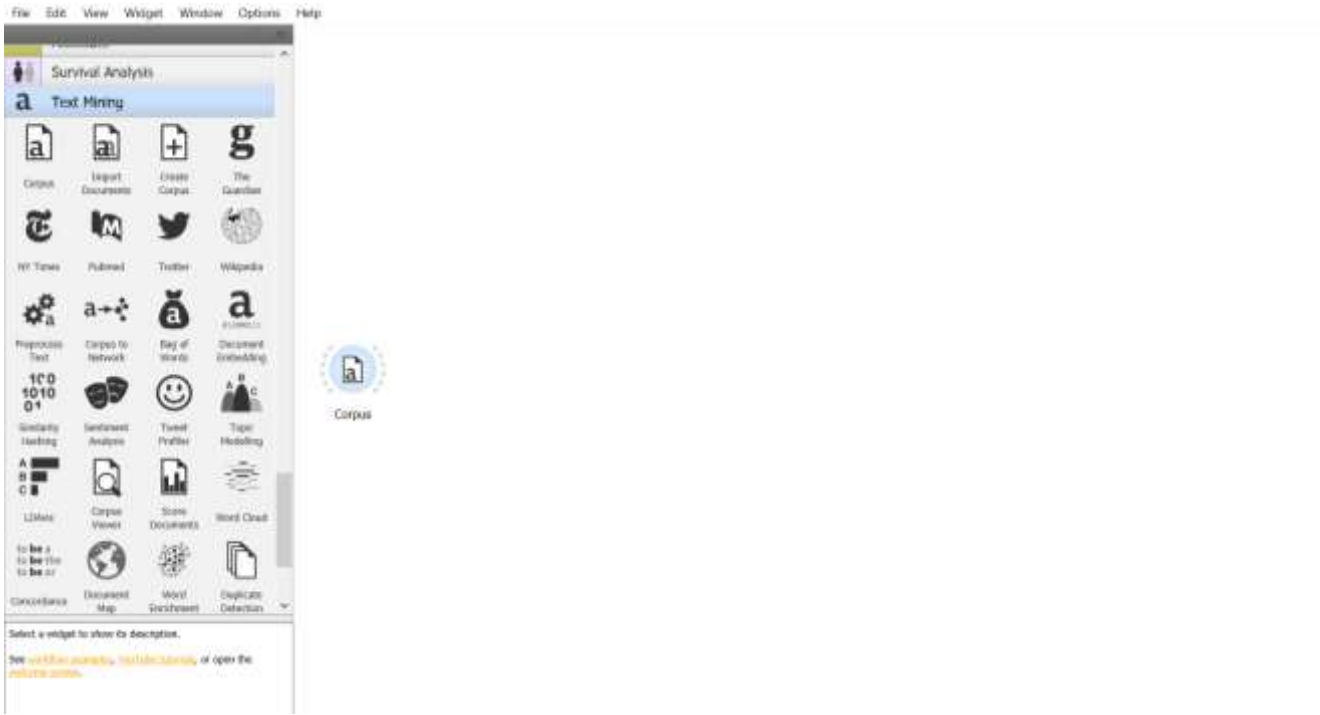

Рисунок 3 - Добавление виджета Corpus на холст

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

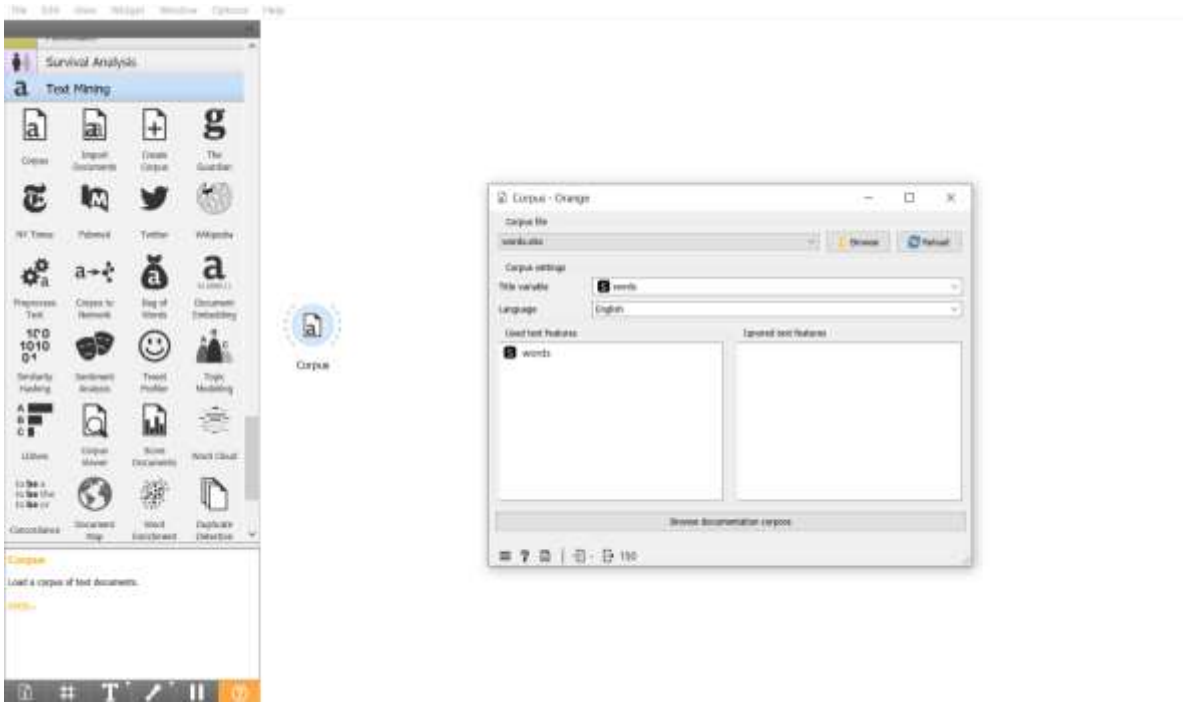

Рисунок 4 - Добавление набора данных words.xlsx

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

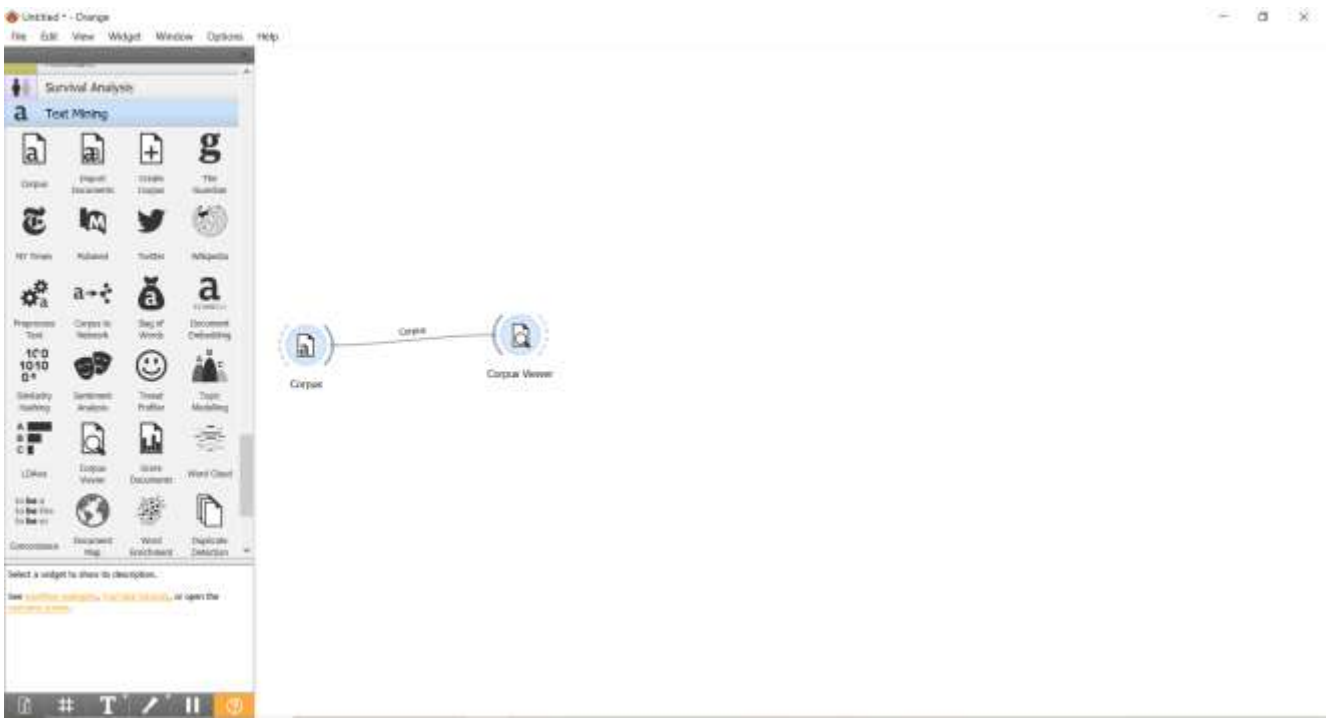

Рисунок 5 - Добавление виджета Corpus Viewer на холст

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

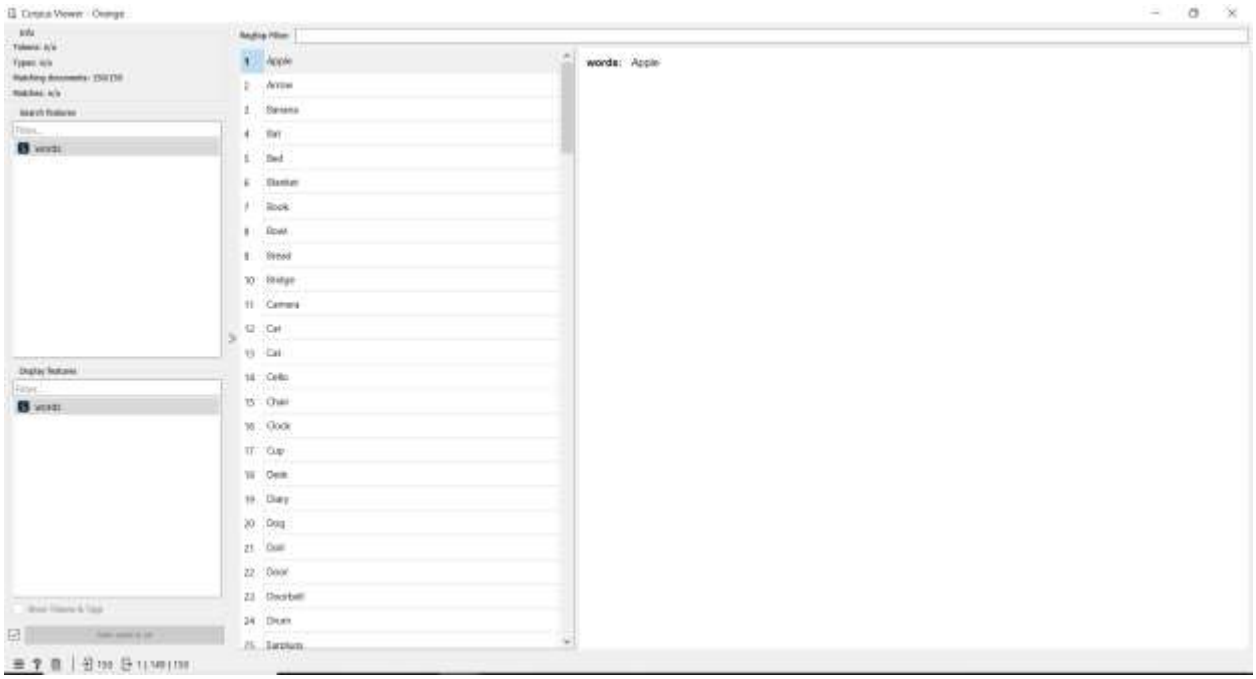

Рисунок 6 - Просмотр набора данных

Далее добавляем виджет Document Embedding на холст, и соединяем с виджетом Corpus. 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).

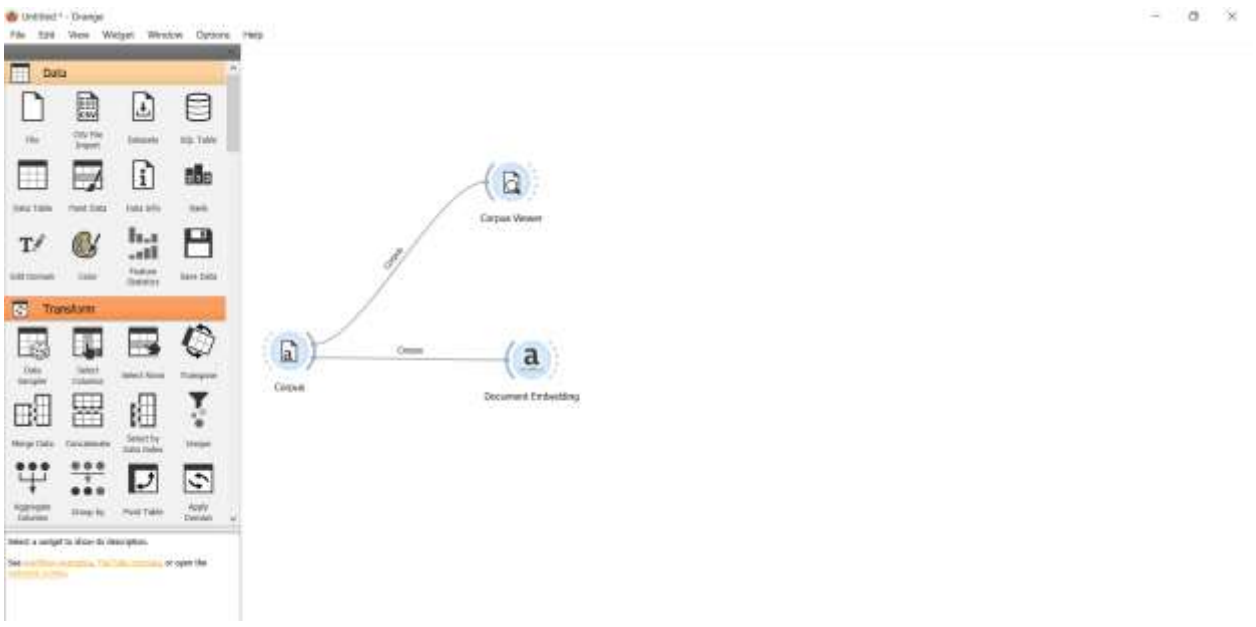

Рисунок 7 - Добавление виджета Document Embedding на холст

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

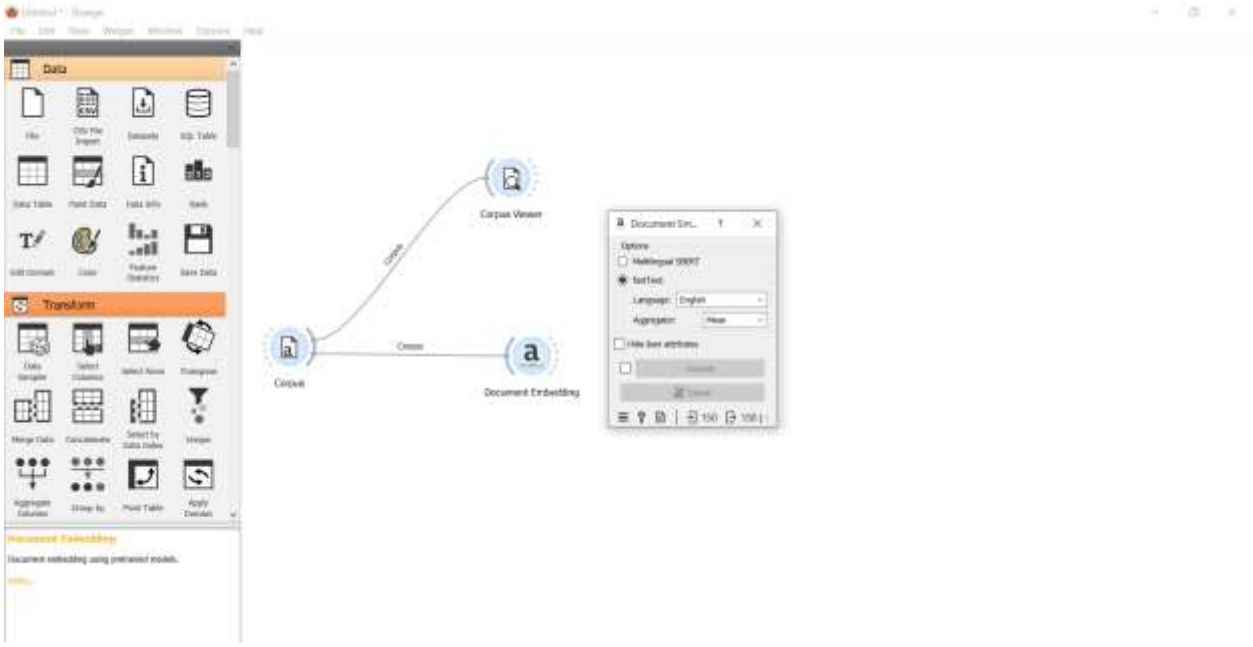

Рисунок 8 - Изменение настроек виджета Document Embedding

Добавим виджет Data Table на холст, и соединим с виджетом Document Embedding, для того чтобы посмотреть данные виджета Document Embedding (см.рис.9).

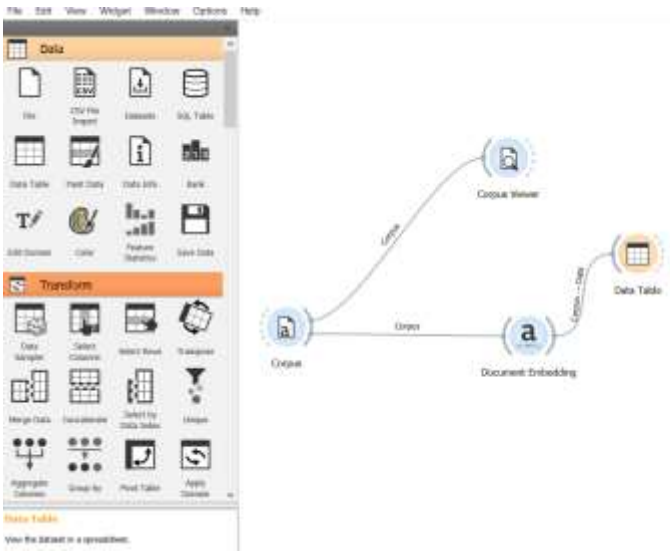

Рисунок 9 - Добавление виджета Data Table

Открываем виджет Data Table. С помощью таблицы можем увитеть, что для каждого слова добавились 300 дополнительных признаков (см.рис.10).

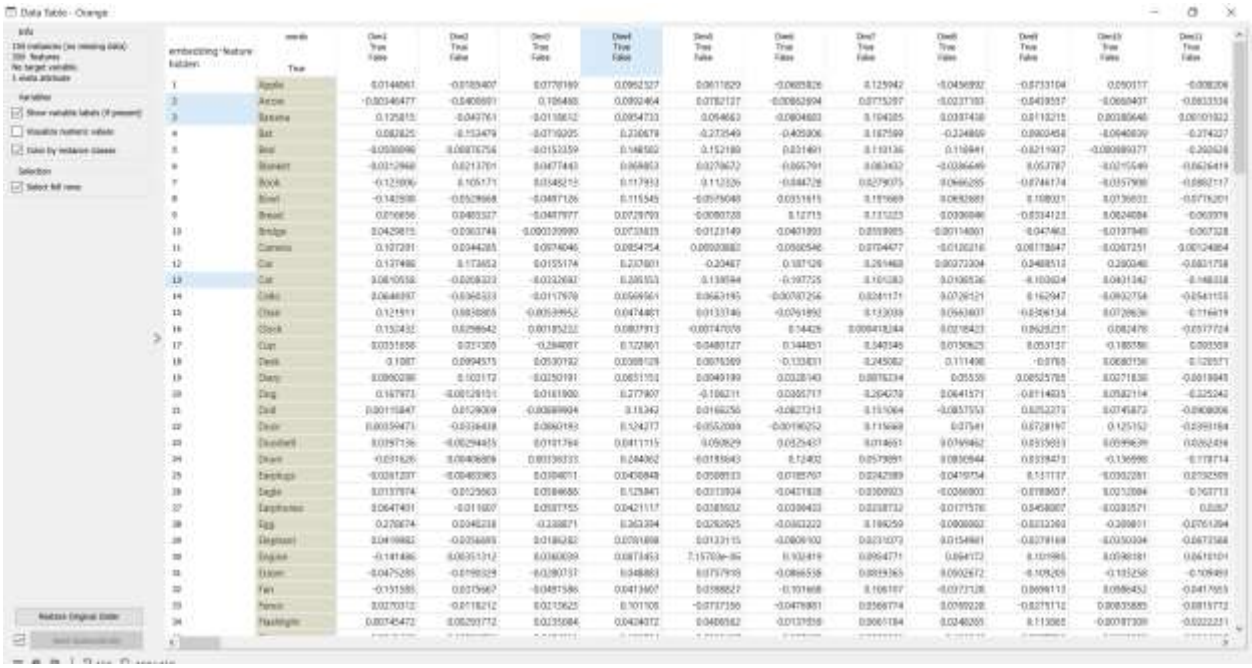

Рисунок 10 - Просмотр данных виджета Document Embedding

Далее добавляем виджет Distances на холст, и соединим с виджетом Document Embedding. Откроем виджет Distances, и выберем метрику расстояния cosine (см.рис.11).

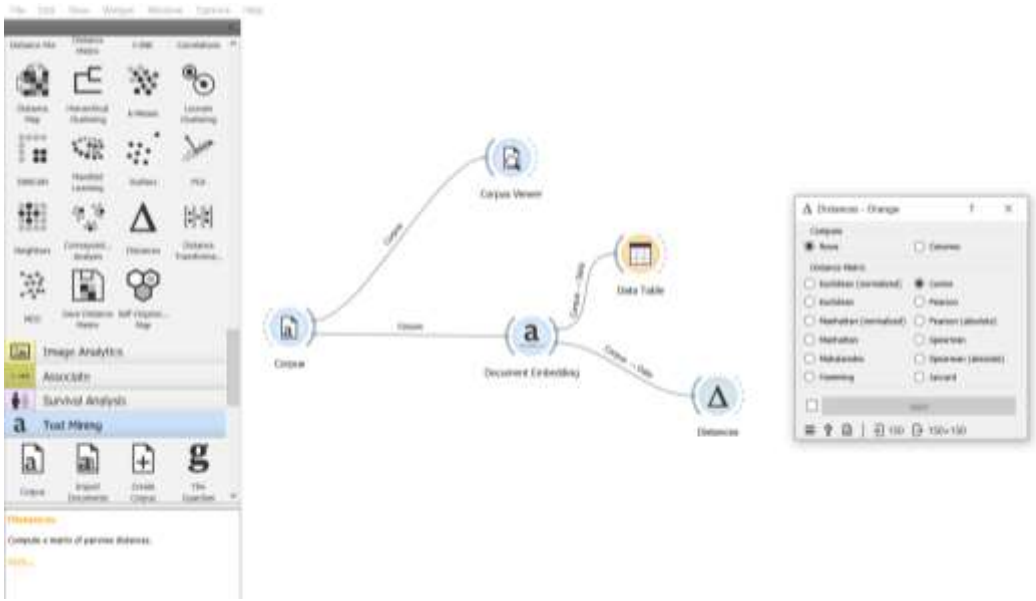

Рисунок 11 - Добавление виджета Distances на холст

Для того чтобы посмотреть данные результата виджета Distances, добавим виджет Distance Matrix и соединим с виджетом Distances (см.рис.12)

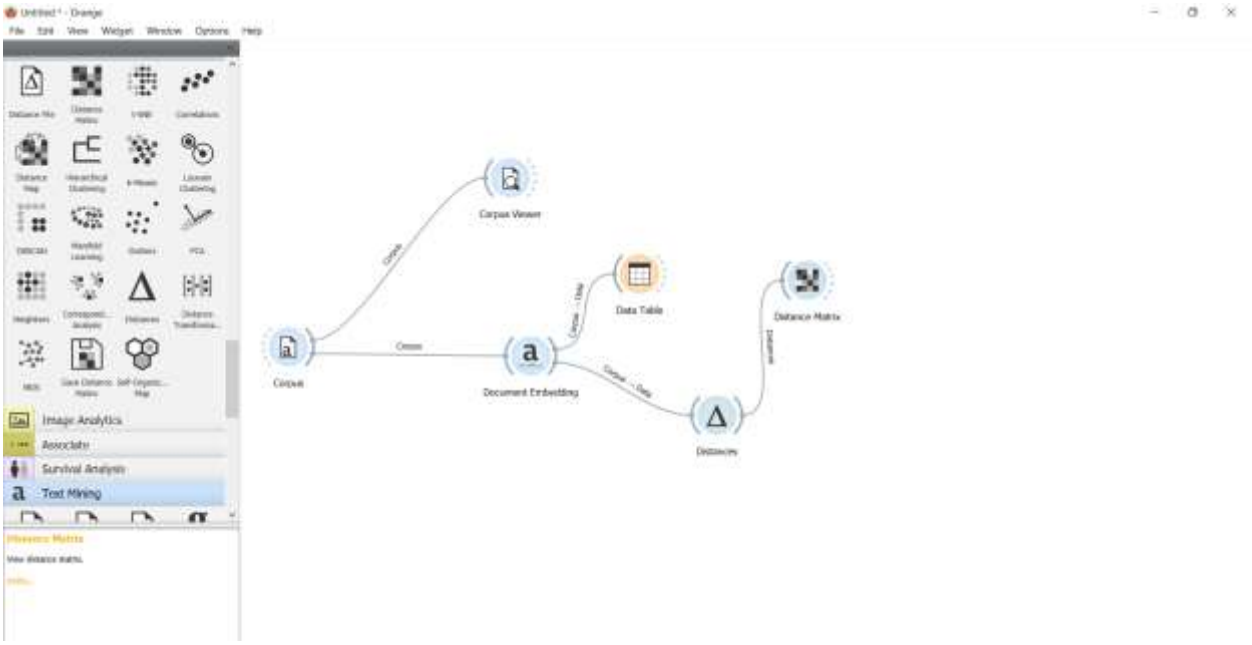

Рисунок 12 - Добавление виджета Distance Matrix на холст

На таблице матрица расстояния можем определить, какое расстояние между словами. По матрице можно определить, чем меньше значение между словами, тем больше они схоже по признакам (см.рис.13).

Рисунок 13 - Матрица расстояния

Далее проведем кластеризацию слов. Для этого добавим виджет Hierarchical Clustering на холст и соединим с виджетом Distances (см.рис.14).

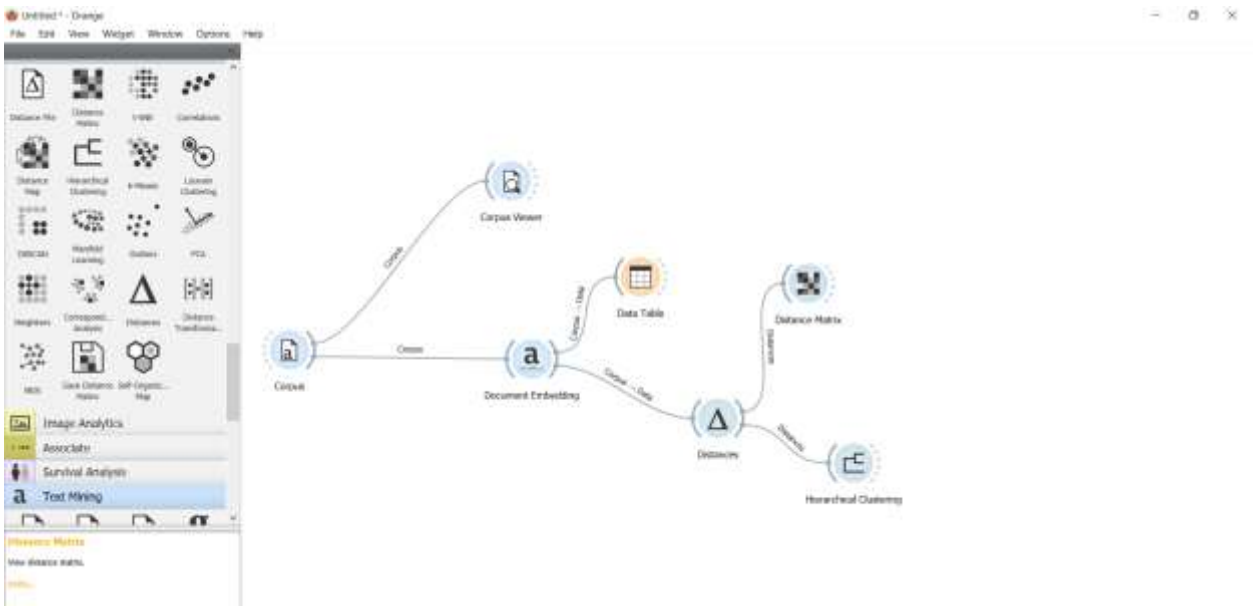

Рисунок 14 - Добавление виджета Hierarchical Clustering на холст

Открываем виджет Hierarchical Clustering. В появившемся окне можно увидеть, как слова распределились на 5 кластеров. Например, кластер c1 можно отнести к группе музыкальные инструменты, кластер c2 можно отнести к группе вещи (см.рис.15).

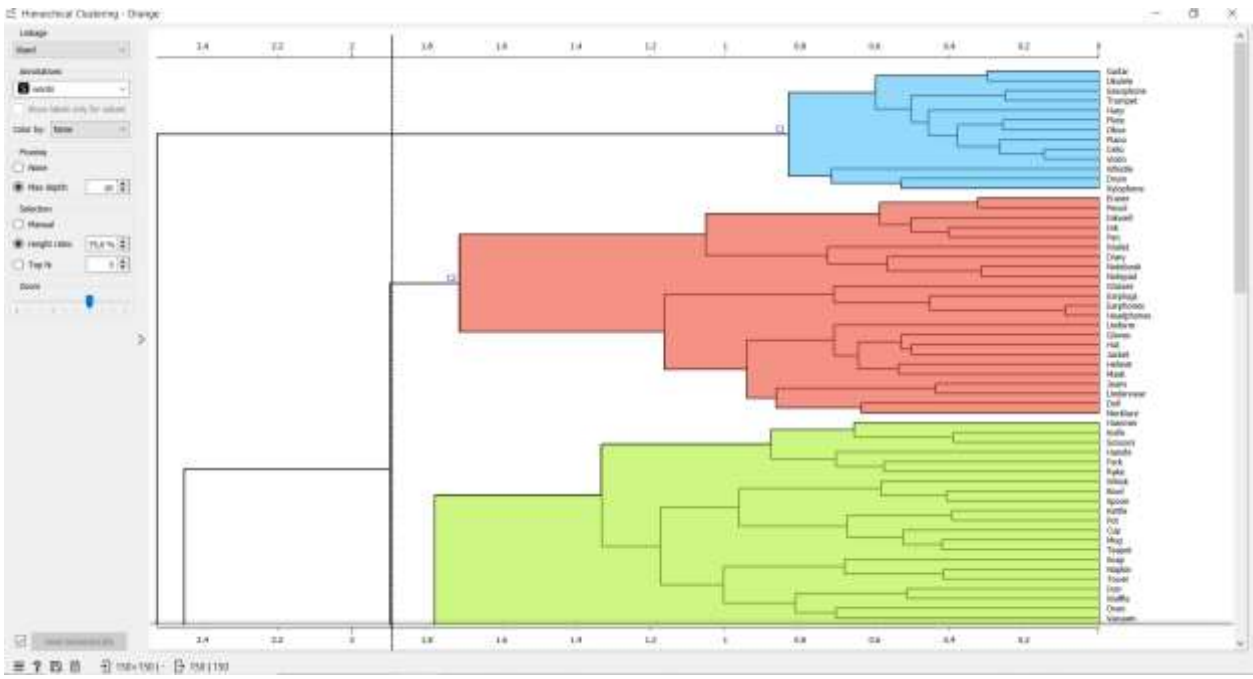

Рисунок 15 - Распределение слов по кластерам

Также можно визуализировать кластерный анализ. Для этого добавляем виджет t-SNE на холст, и соединяем с виджетами Document Embedding и Hierarchical Clustering (см.рис.16).

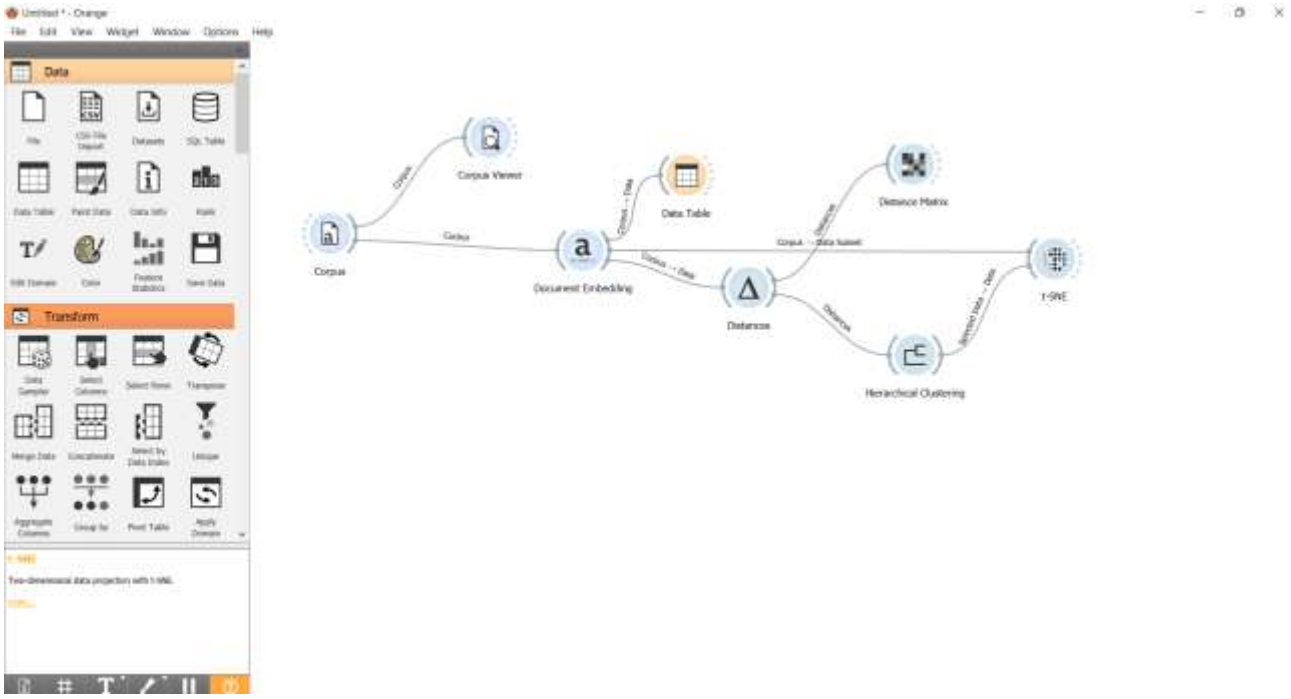

Рисунок 16 - Добавление виджета t-SNE на холст

Открываем виджет t-SNE. Виджет покажет 2D-карту, где образцы с похожими профилями экспрессии генов будут расположены близко друг к другу (см.рис.17).

Рисунок 17 - Диалоговое окно t-SNE

Для того чтобы убедится, что расположенные рядом точки схожи по значению добавляем на холст виджет Corpus Viewer и соединяем с виджет t-SNE (см.рис.18).

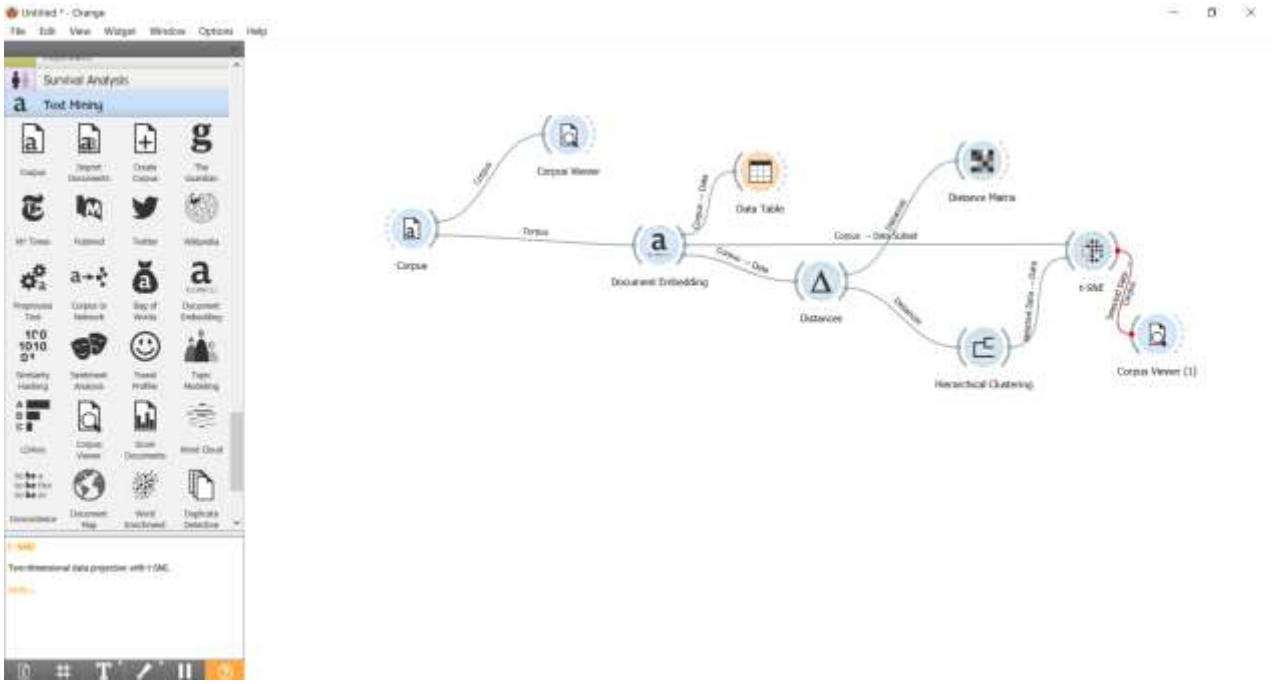

Рисунок 18 - Добавление виджета Corpus Viewer на холст

Открываем виджеты t-SNE и Corpus Viewer. Для того чтобы проверить правильность значений в виджете t-SNE выделяем группу точек, которые расположены близко к друг другу. В окне Corpus Viewer появились значение выделенных точек. Можем сделать вывод, что точки расположены близко к друг другу относятся к категории музыкальные инструменты (см.рис.19).

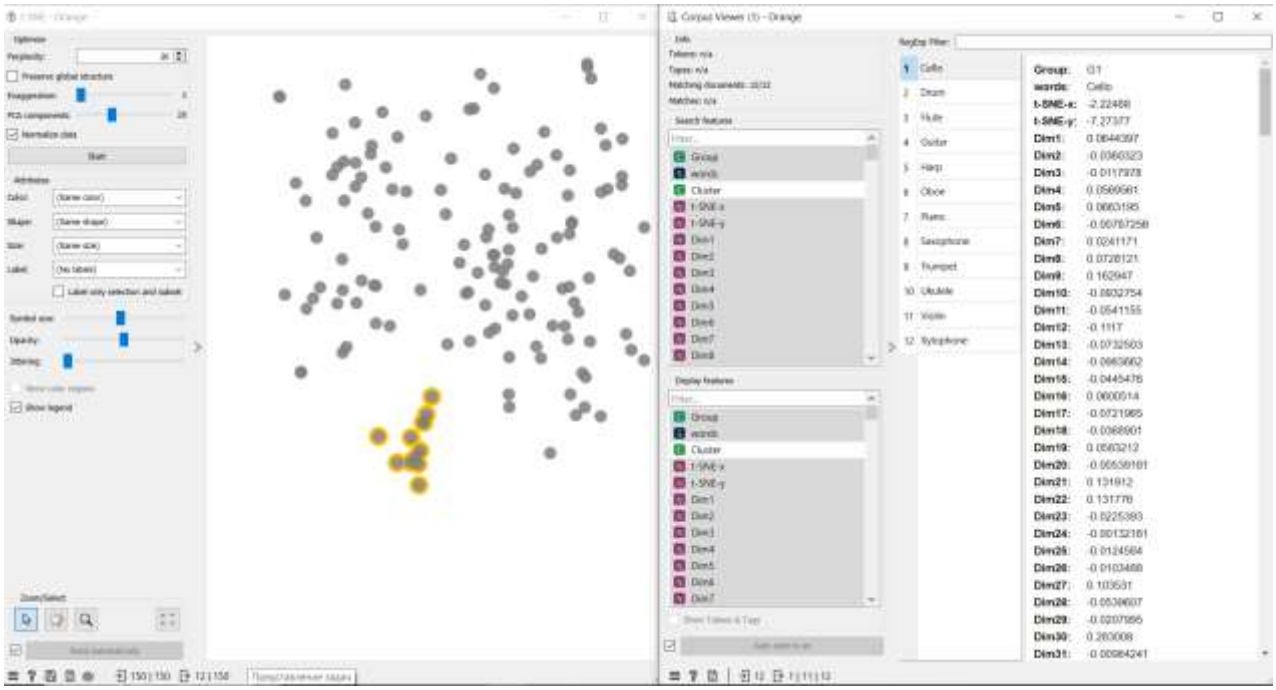

Рисунок 19 - Проверка выделенных точек

В итоге получилась готовая схема, с помощью которой можно решить задачу кластеризации различных слов (см.рис.20).

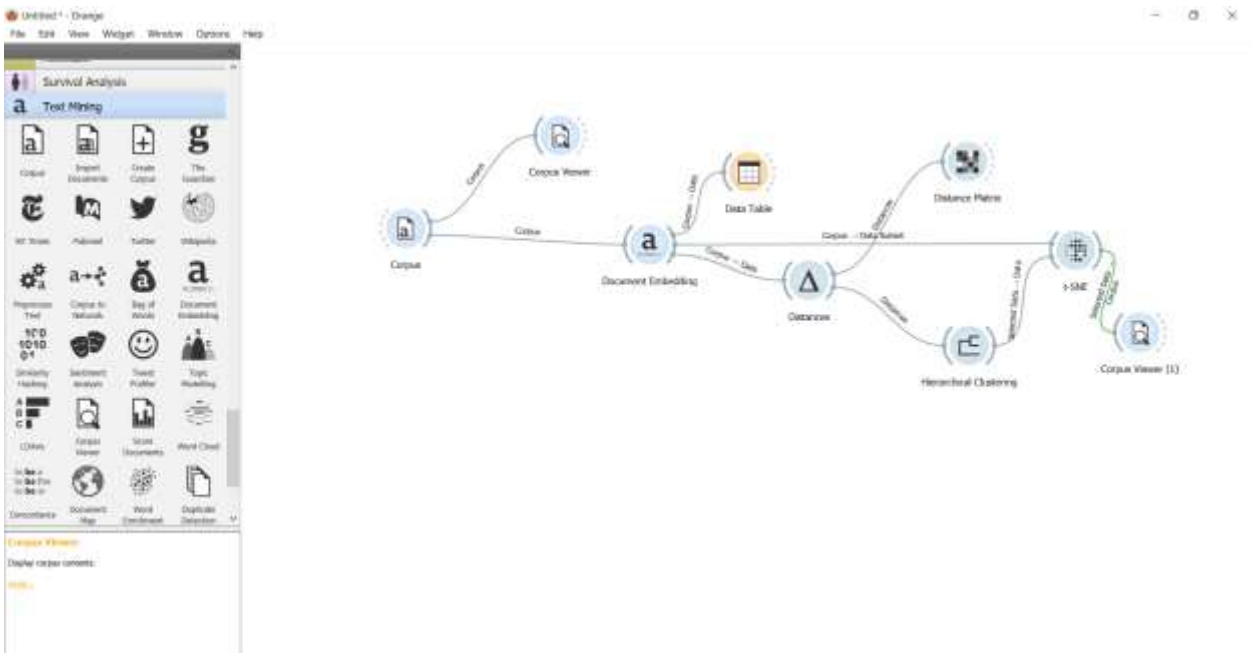

Рисунок 20 - Итоговая схема

Выводы

В данной работе была выполнена задача кластеризации набора данных различный слов с помощью программного пакета визуального программирования на основе компонентов для визуализации данных Orange. С помощью виджетов Corpus, Corpus Viewer, Document Embedding, Data Table, Distances, Hierarchical Clustering, t-SNE, Distance Matrix выполнили кластеризацию различных слов и получили итоговую схему.

Библиографический список

1. Гладилин А. В. Обзор существующих программных комплексов для проведения кластерного анализа //Новые информационные технологии и системы. 2017. С. 224-226.
2. Моисеенко Г. А. и др. Классификация и распознавание изображений живой и неживой природы //Оптический журнал. 2015. Т. 82. №. 10. С. 53-64.
3. Юсупов Н. Исследование методов кластеризации в программе Orange //Молодежная школа-семинар по проблемам управления в технических системах имени АА Вавилова. 2020. Т. 1. С. 35-37.
4. Клименко А. В., Слащев И. С. кластерный анализ данных //Вестник науки. 2019. Т. 1. №. 1. С. 159-163.
5. Гринченков Д. В. и др. Сравнительный анализ алгоритмов интеллектуального анализа данных //Моделирование. Теория, методы и средства. 2016. С. 263-266.
6. Мастевной С. С., Петрова А. Н. Data mining: обзор методов и области их применения // Наука, инновации и технологии: от идей к внедрению. 2022. С. 38-40.