CS598 Data Mining Capstone

Khemchandra Persadsingh (kdp3)

Task 7 – Word2Vec Data Explorer

Introduction

The application provides a user interface to query a Word2Vec model. Word2Vec is a two-layer neural network that takes a text corpus as input and outputs a set of vectors. Each vector is a feature vector that represents words in the corpus. The feature vectors produced have the property that similar words are grouped together in the vector space. Word2Vec can make accurate guesses about a word's meaning based on past appearances or cluster documents and classify by topic. Word2Vec is used in feature engineering.

The application is designed to accept a saved Word2Vec model as input and the user can use the HTML front end to preform queries and visualize the data.

Application Overview

Intended Users

The intended users are data scientists who are interested in exploring the dataset and extracting features. The public can use the tool to visualize a dataset to view word groupings.

Features

The application has three features:

- Auto Complete – this accepts a word or word segment and displays words in the model that start or end with the word or segment.
- Close Words – this accepts two words and shows the words that are closer to the first word than the second word in the vector space. The first word can be a query (described below).
- Explore – this accepts a word or a query and shows the words that are like it. There is an option to cluster the results and view the words at the centroids of the clusters

A query is made up of words connected by AND or AND NOT. For example, "pasta AND tomato" finds words that are similar to pasta and tomato, "pasta AND tomato AND NOT cheese" finds words that are similar to pasta and tomato but not cheese. A query can be a single word as well.

Figure 1 shows the page when the application is launched. The features above can be accessed using the top menu.

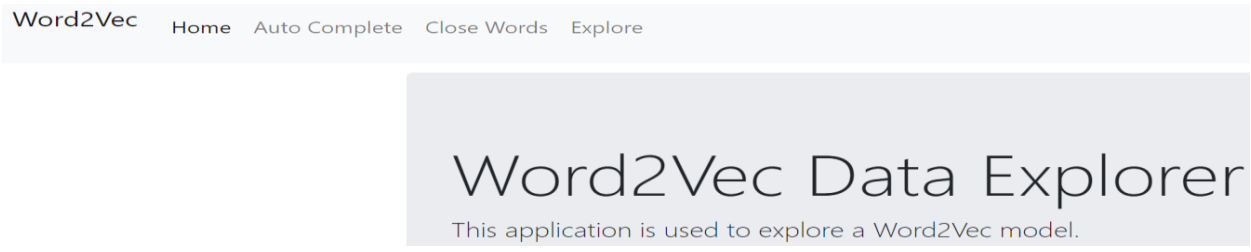

Figure 1

Auto Complete

Figure 2 shows the auto complete form.

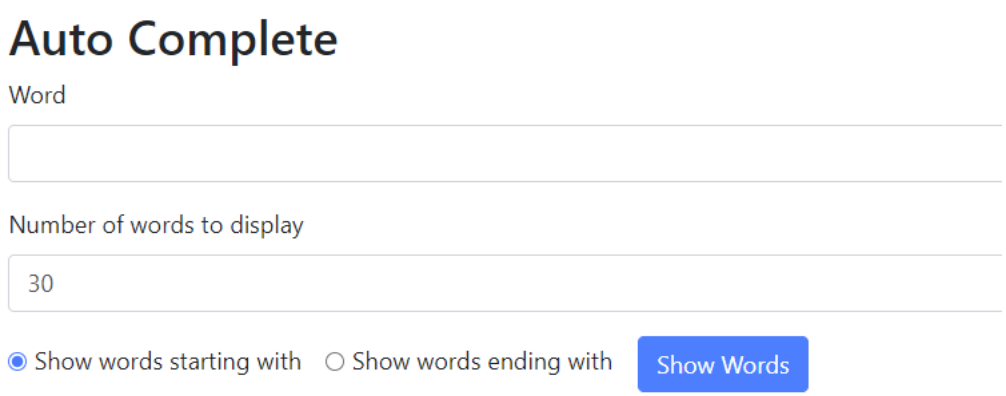

Figure 2

The word or segment is entered in the “Word” field and the number of words to be displayed is entered in the “Number of words to display” field. The “Show Words” button is used to display the words that start or end (depending on radio button chosen) with the entered word. The words are displayed below the form.

Close Words

Figure 3 shows the close words form.

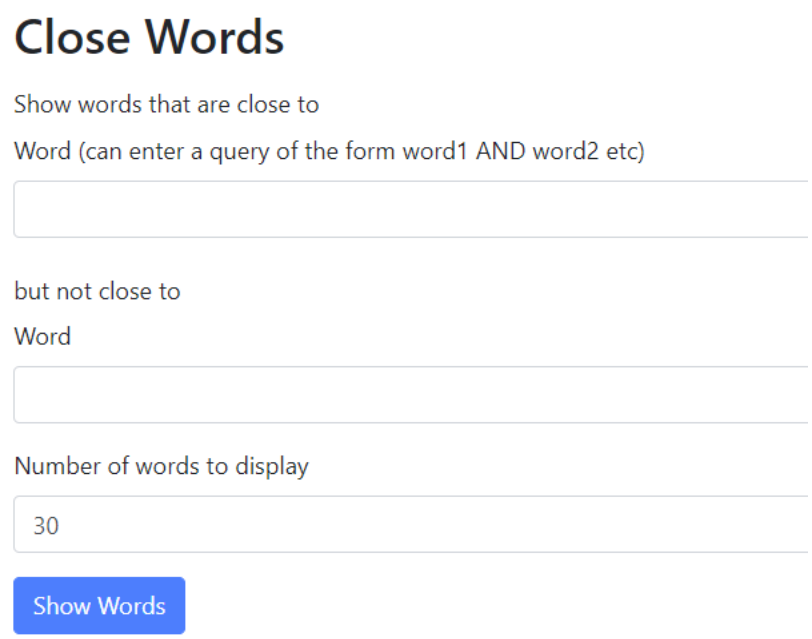

Figure 3

The word or query is entered in the first “Word” field and the second word is entered in the second “Word” field. The number of words to be returned is entered in the “Number of words to display” field. The list of words is shown below the form.

Explore

The explore form is shown in figure 4.

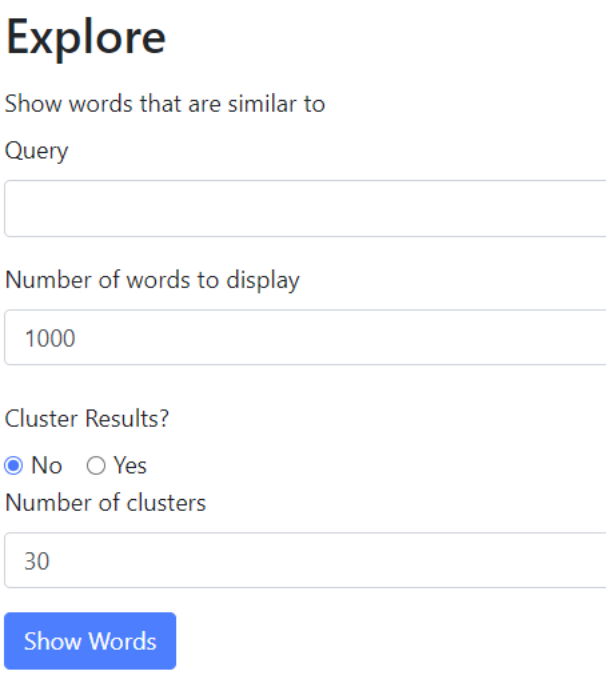

Figure 4

The query is entered in the “Query” field. The “Number of words to display” specifies the number of words to be returned by the query. The “Cluster Results?” radio button indicates if the system should cluster the results of the query and the “Number of clusters” field specifies the number of clusters to cluster the words into.

The results of the query shown in figure 5 is shown in figure 6.

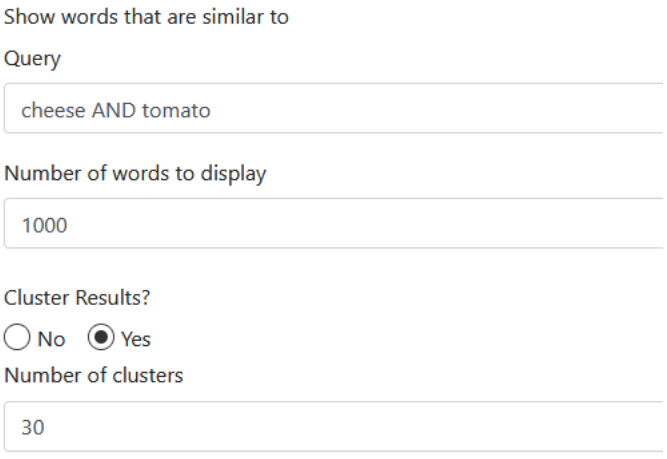

Figure 5

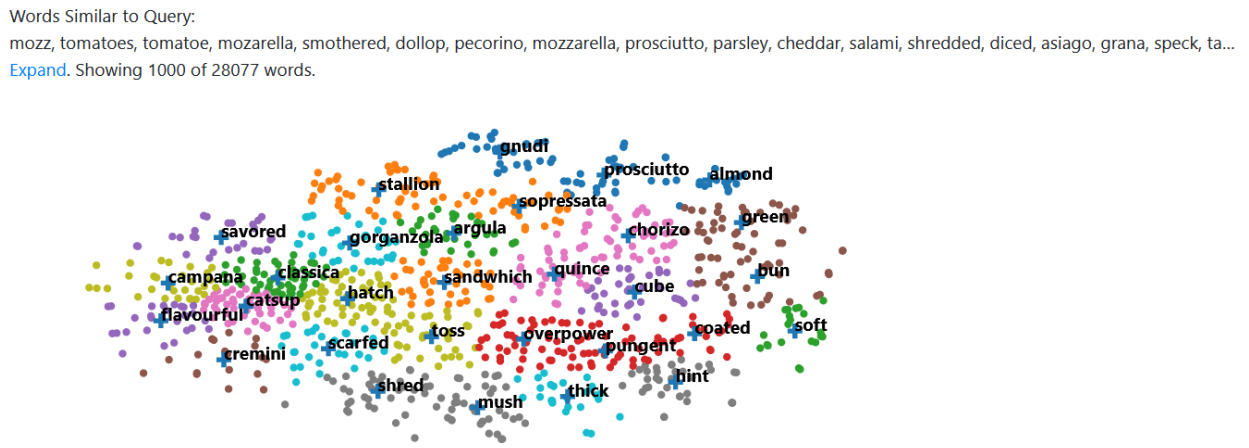

Figure 6

The first line of the words returned is shown. The "Expand" link is used to display all the words. Next to the "Expand" link is the statistics of the model. In this case, the model has 28077 words. The clusters are then shown. The centroids are indicated with a cross symbol and the word at the centroid is displayed.

The words at the centroids can be used in the explore form to view the words that are similar to it.

Technical Details

The gensim package is used to query the Word2Vec model. TSNE from the sklearn package was used to reduce the dimensions of the word vectors to two and KMeans from the sklean package was used to find the clusters in the reduced dimensional dataset.

Flask was used to create a web application. The application is built in a REST like manner. It accepts requests and returns the results in JSON format.

The front end is a separate HTML page and uses jQuery to perform the requests and d3 to draw the charts.

Accessing the Application

The application is deployed with a Word2Vec model created from the Yelp Dataset. The reviews related to business categories "Restaurants" and "Italian" were extracted from the dataset. This was used to train the Word2Vec model and this saved model is loaded by the application.

Repository for the application - https://github.com/khemnavet/CS598DM

The front end can be accessed from https://khemnavet.github.io/CS598DM/front/index.html

Example of Usage

For task 3 of the capstone, the application can be used to enhance the cuisines mined. The explore option can be used with the topics mined using SegPhrase. The similar words can then be used to enhance the label file used by AutoPhrase to get the cuisine list.