Sentiment Analysis and Topic Modelling

**Website used:** <https://ca.trustpilot.com/>

**Company used:** <https://ca.trustpilot.com/review/snowbirdadvisorinsurance.ca>

**Reviews Scraped:** 2000+

**Section 1: Web-scarping and preprocessing the data.**

**Steps for web scraping:**

1. Setting Up Selenium WebDriver

2. Navigating and Scrolling for Dynamic Content

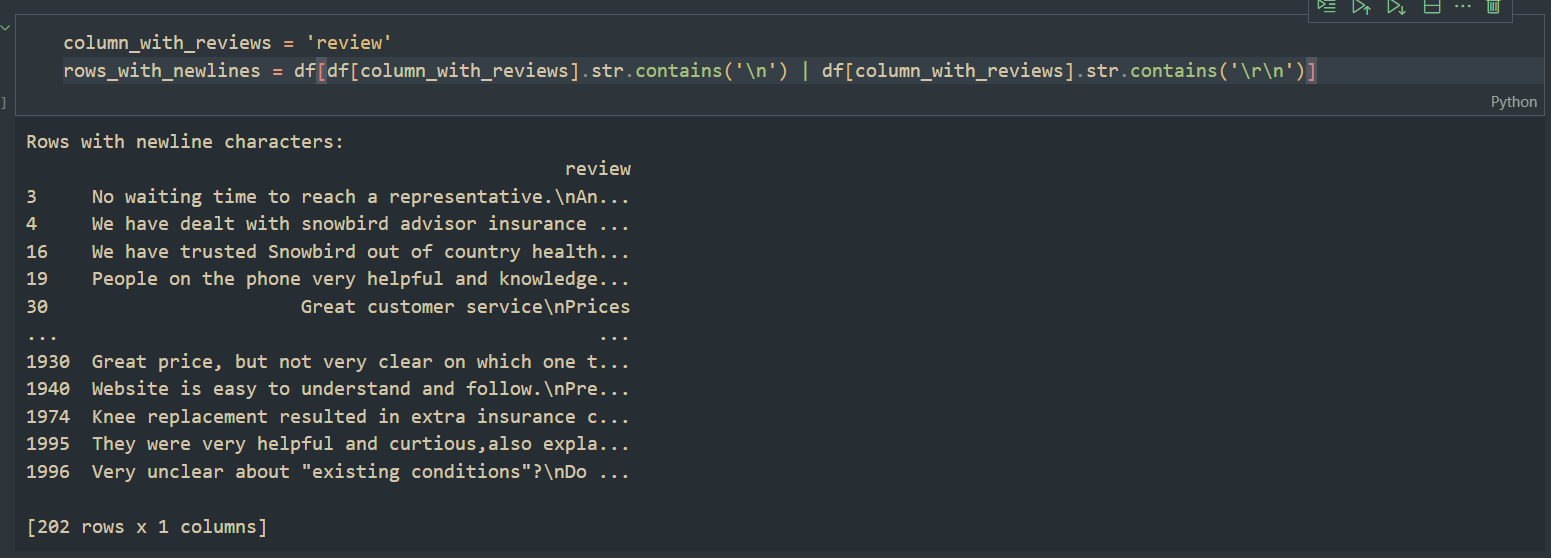
3. Extracting Data by finding the location of Needed Data in HTML

4. Mimicking Human Behaviour to Avoid Detection

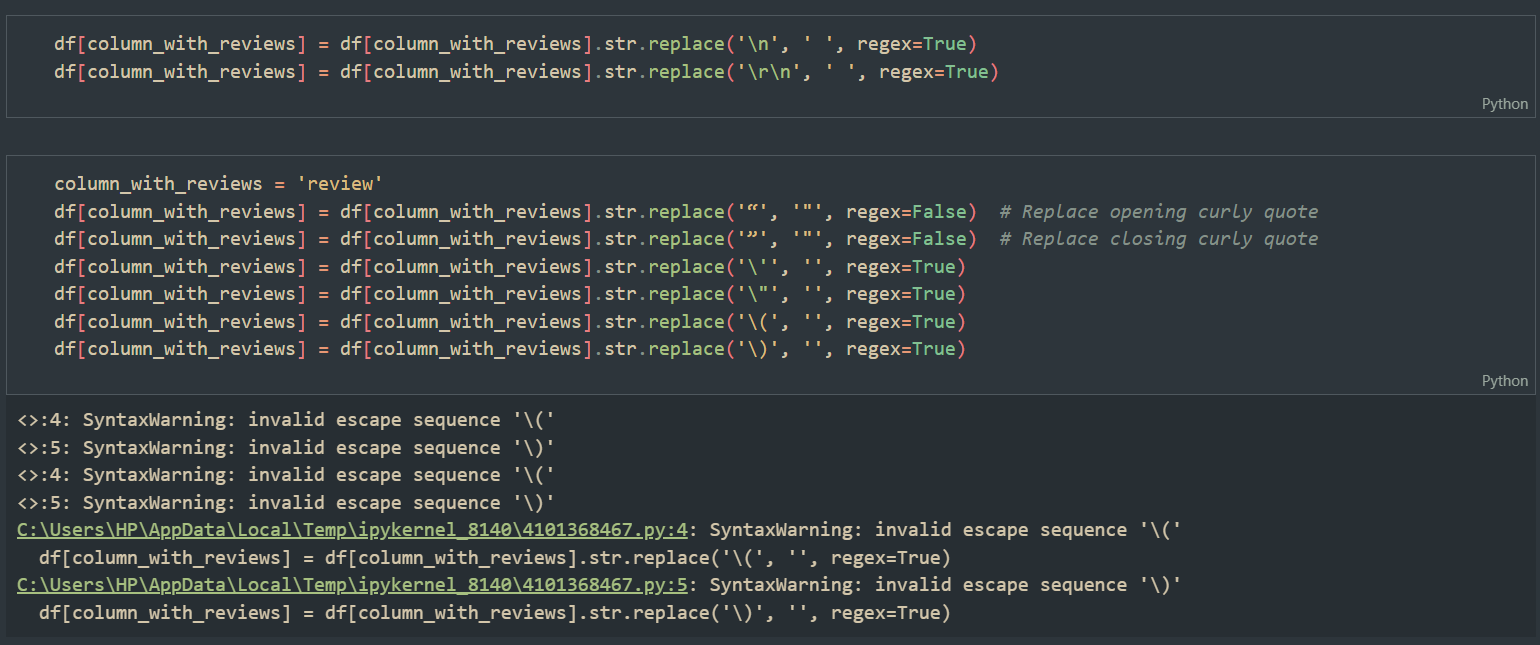
5. Saving Data to CSV & Handling Pagination

**Data preprocessing techniques:**

1. Removing new line characters in the string formatted csv:



1. Removing all the symbols which would hamper parsing.



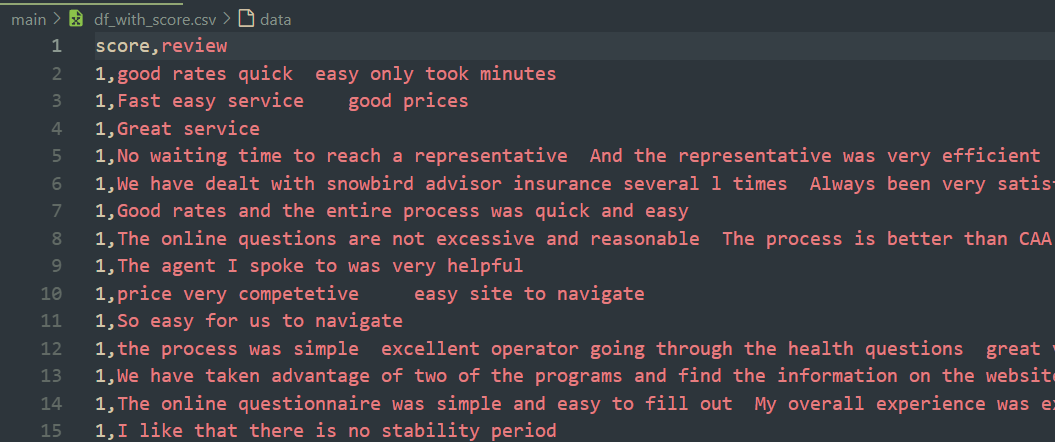
**Section 2: Performing sentimental analysis using VADER and categorizing it.**

**Using VADER for sentiment analysis:**

1. VADER library was used to classify the review into positive, negative and neutral.



We now have a final csv file with clean data with their sentimental scores

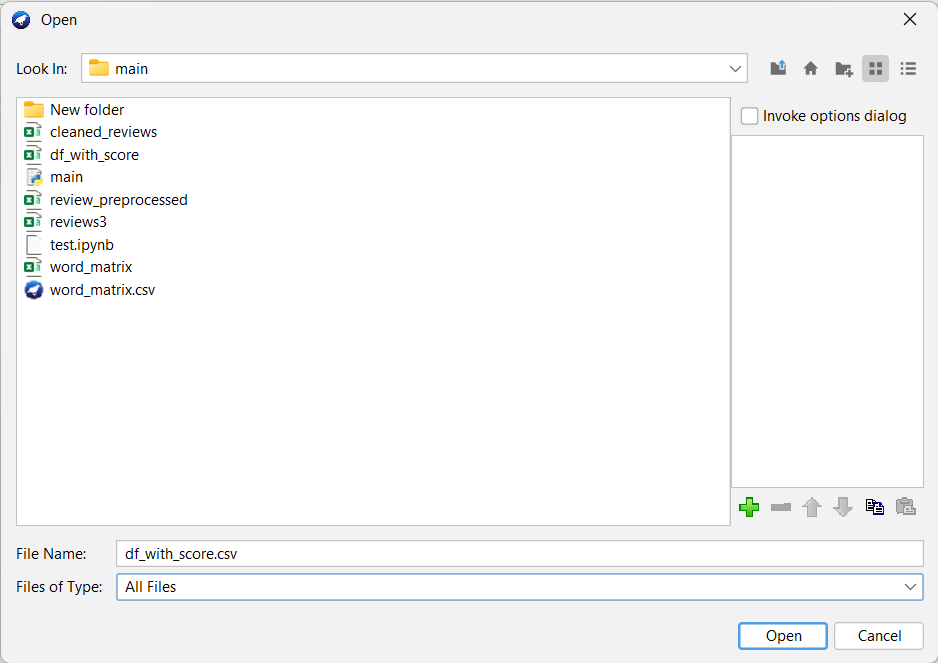


**Section 3: Using WEKA for machine learning analysis.**

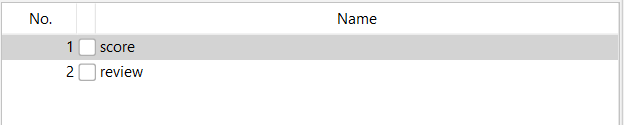
Using Weka for applying a machine learning model on this dataset:

Steps:

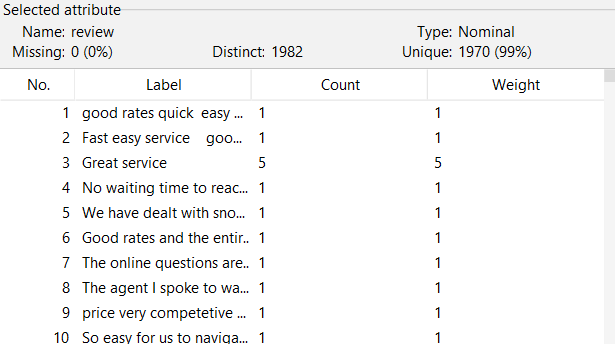
1. Load the data in weka.



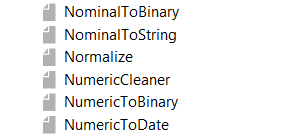
1. Now we have loaded the csv file in weka now we need to do feature engineering on this dataset.



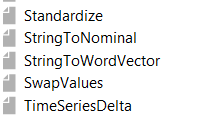
1. We can observe that the review column is nominal we need to convert it to string.



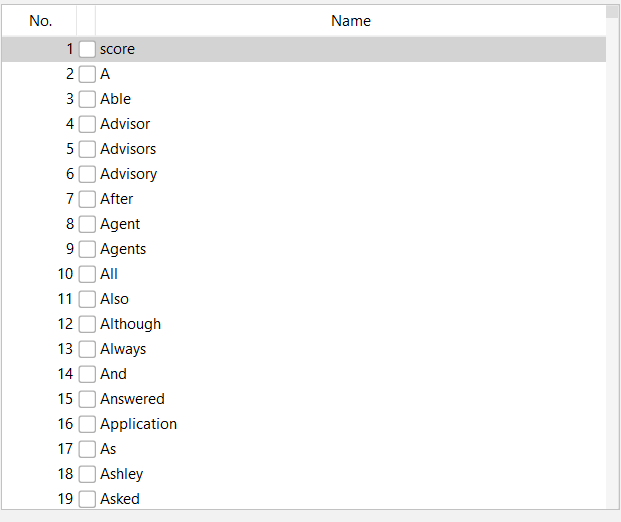
1. We will use the function “NomialToString”.



1. Then we will convert the review column to word matrix for applying machine learning model.



1. Now we have a word matrix, can create a data frame out of it if we want.

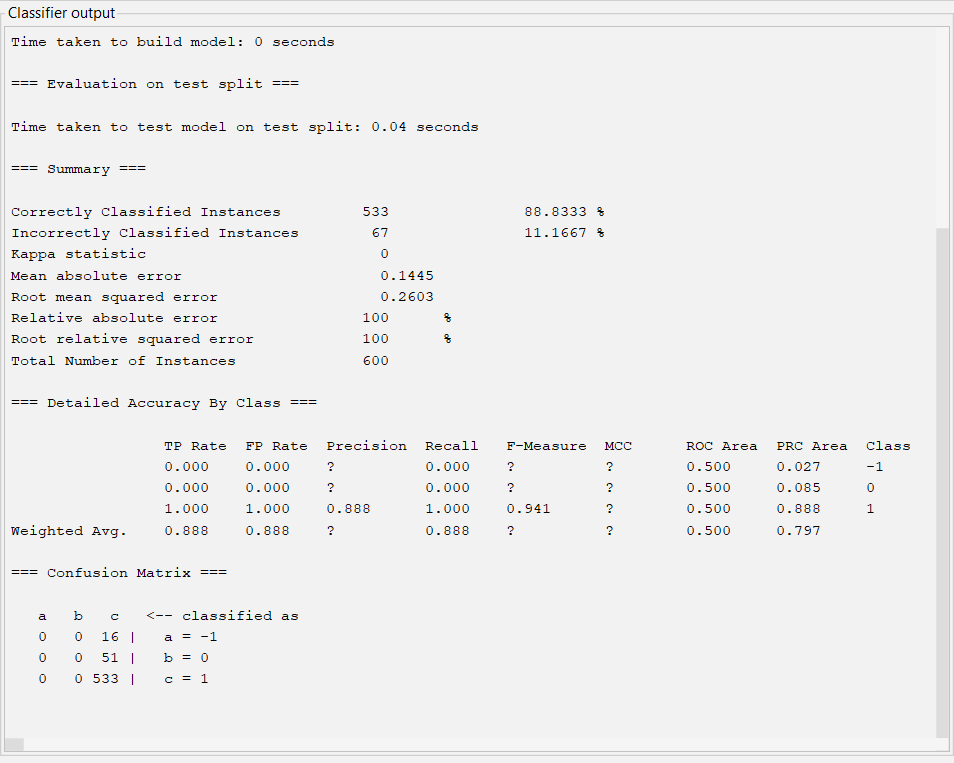


1. Now we have to convert score attribute to a class variable and make it nominal.

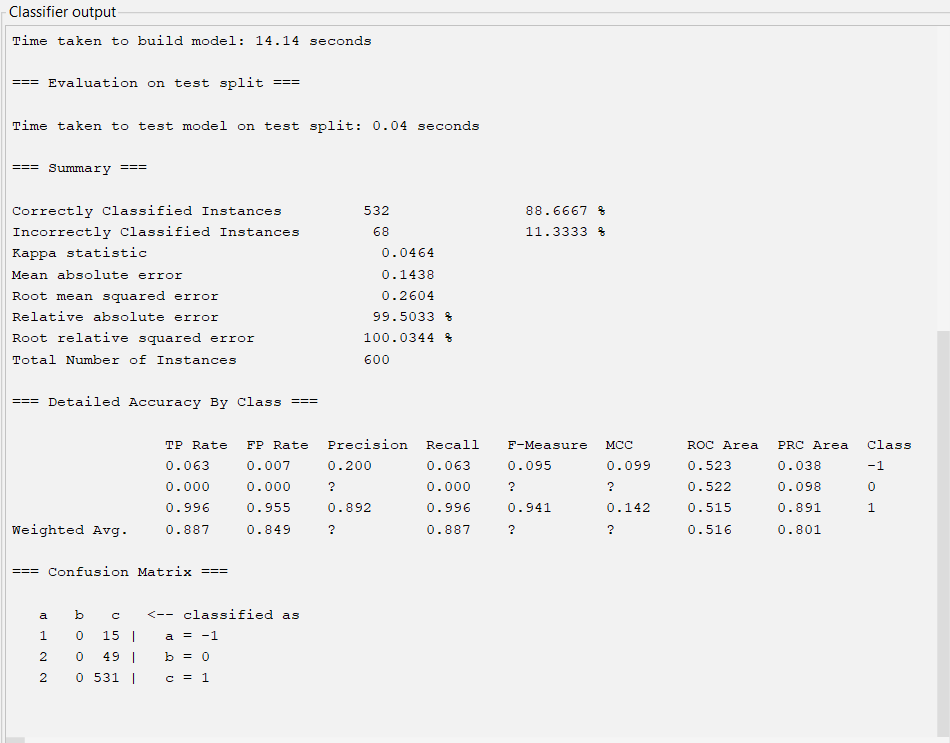
**Classification:**

I have used various classification techniques:

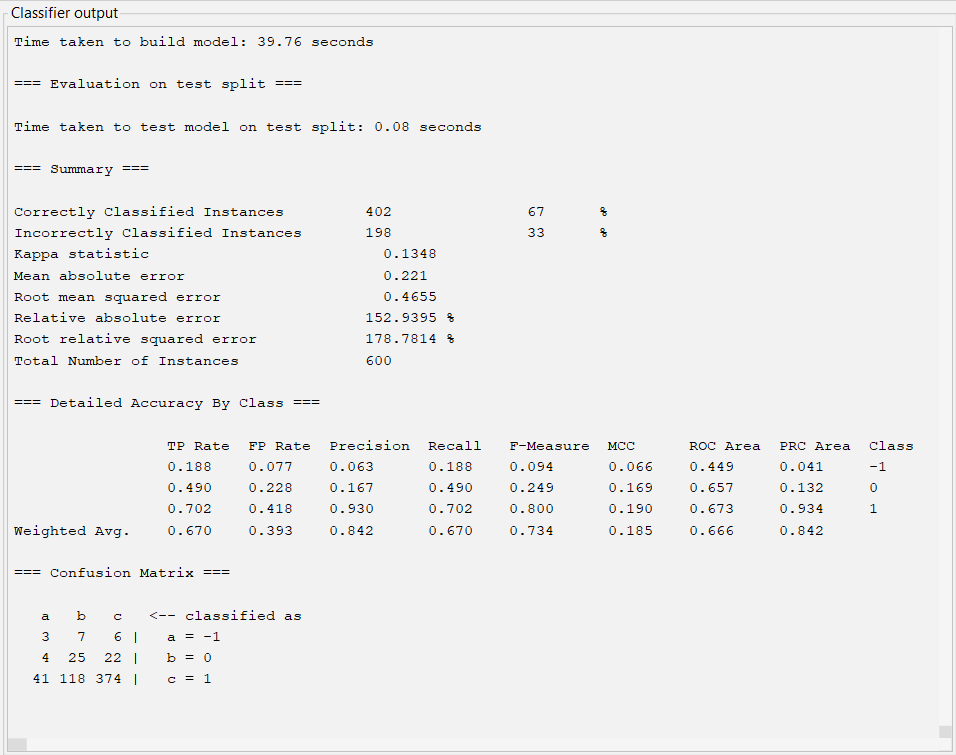
1. Zero Rules:



1. Decision Tree:



1. Logistic Regression:

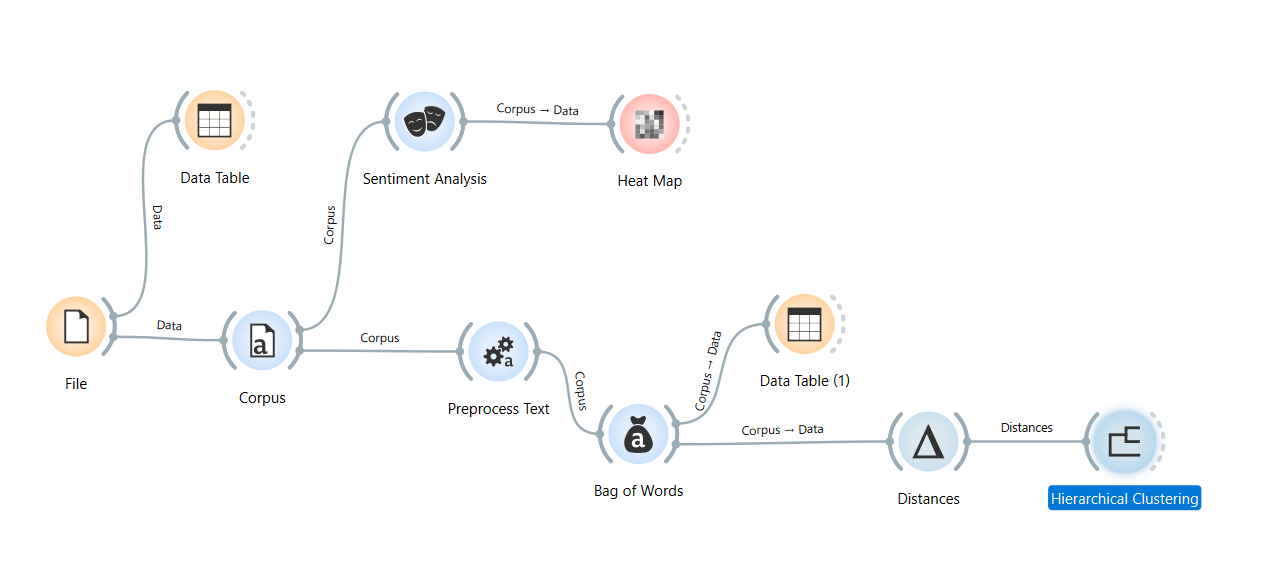


**Section 4: Using orange for machine learning analysis.**

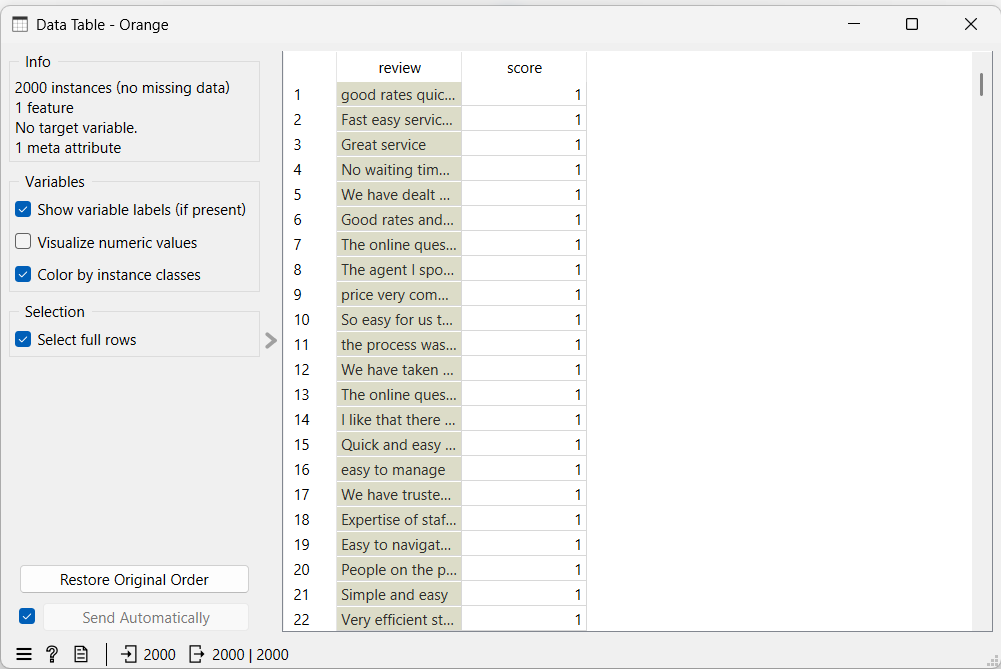
**Steps:**

* First create a file widget and import the data in it.
* Then view it in a data table.
* Then use corpus to use strings from the dataset.
* Then preprocess the text
* Then use “Bag of Words” to vectorize the text.
* Then use the distances and hirerical clustering to cluster the data.
* Now once we are clear with understanding the data we can apply sentimental analysis widget on it.
* Then finally we connect the heatmap with it to check out the results.

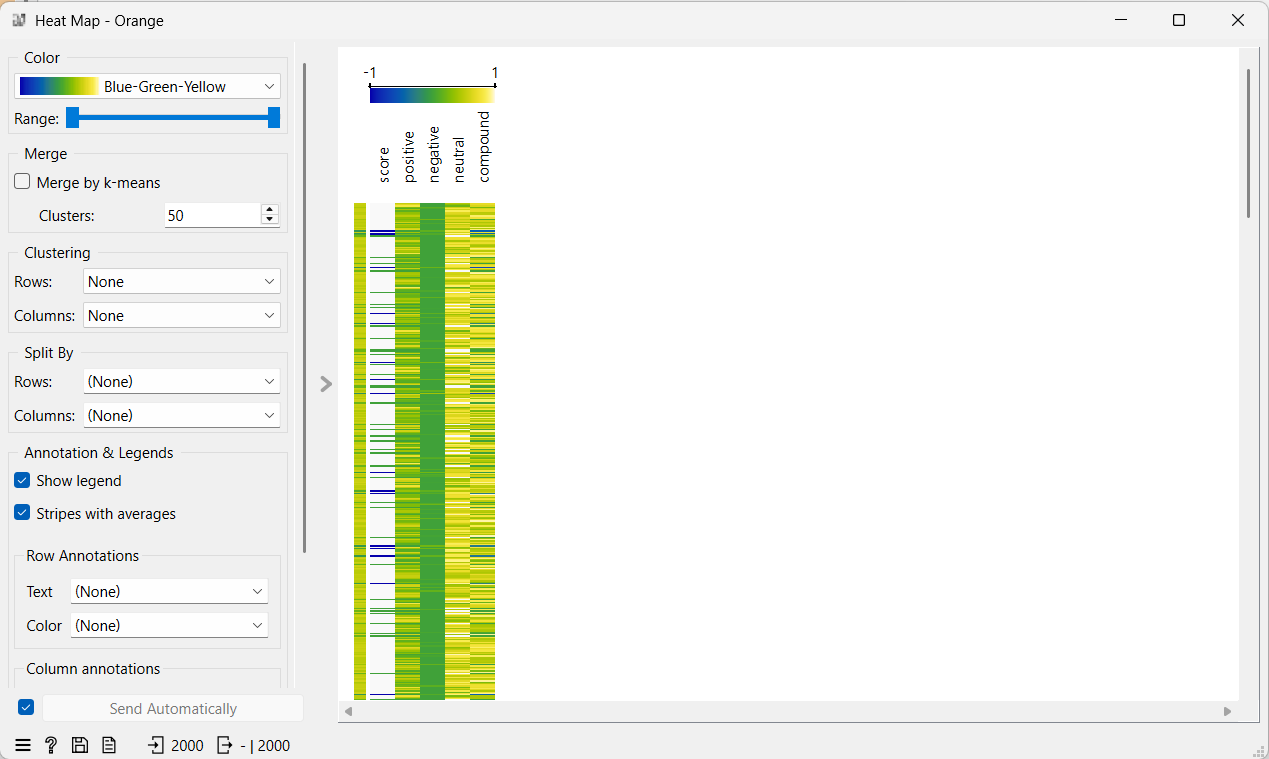
**Orange representation:**



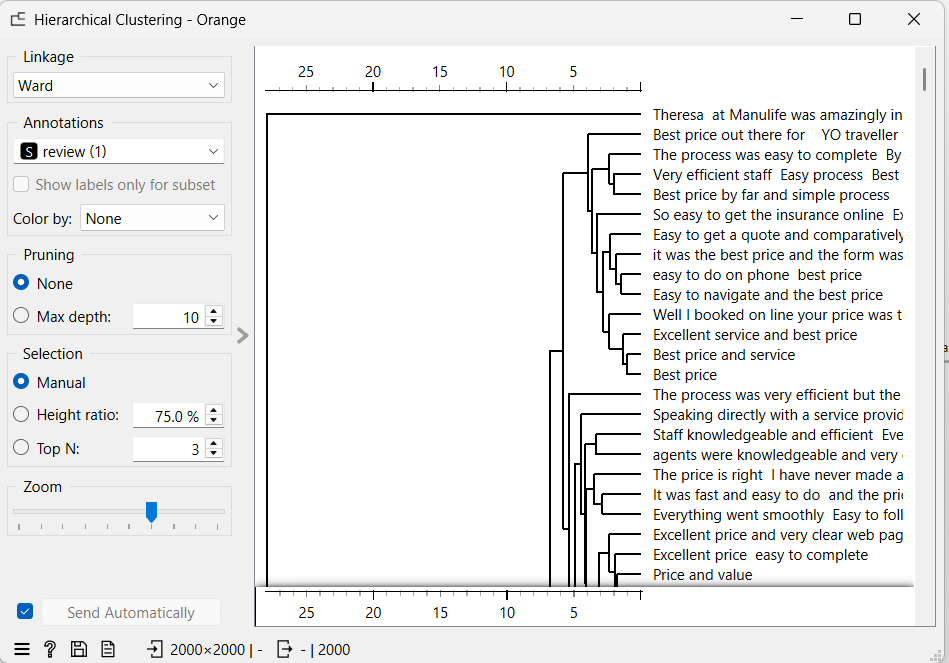
Data Table View:



Heat Map View:



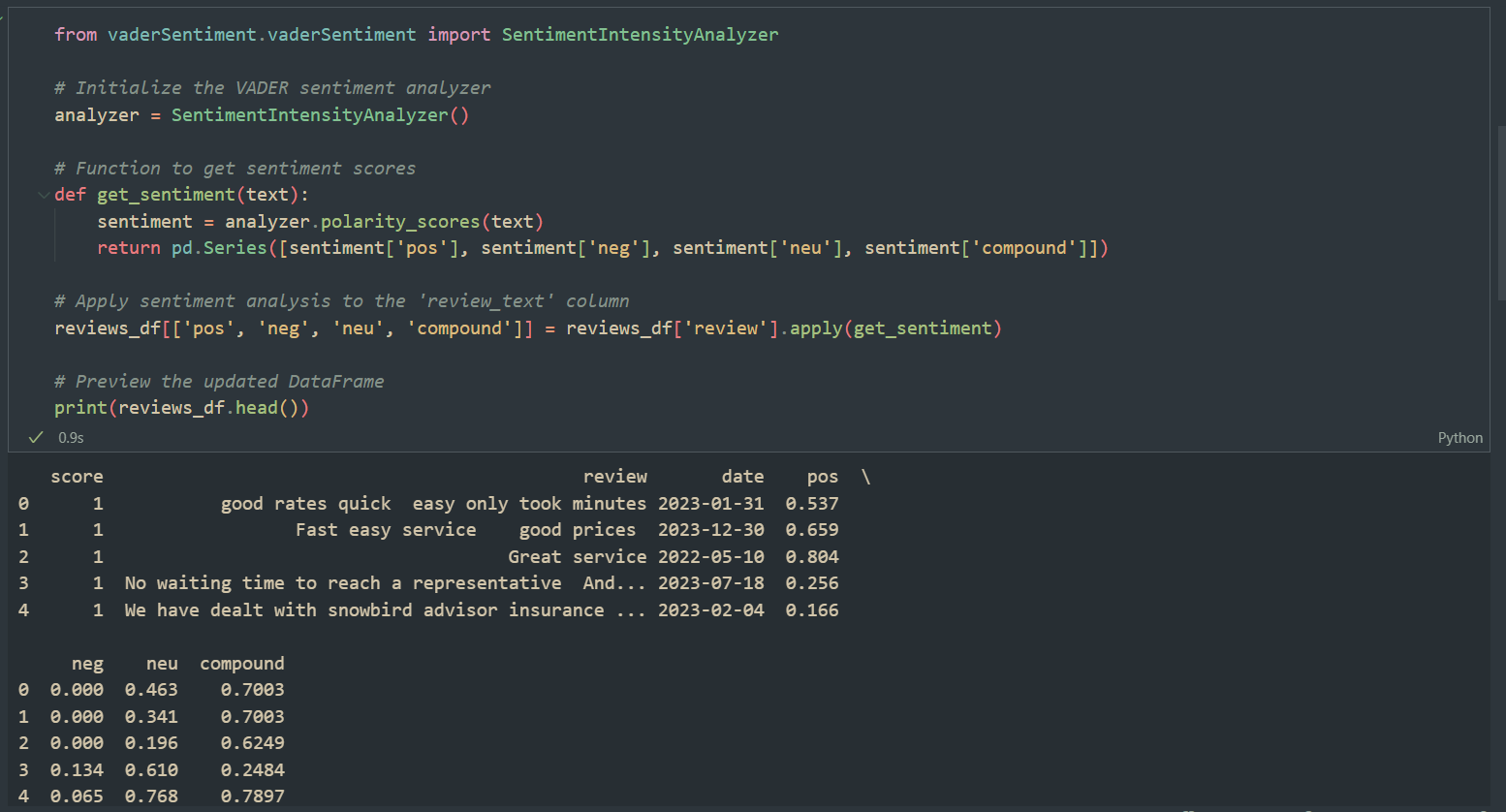
Clustering View:



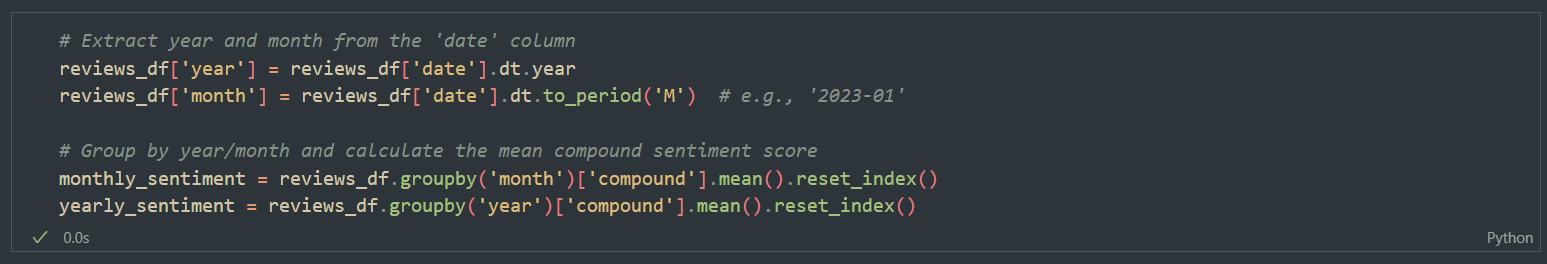
**Section 5: Time framed Sentimental Analysis**

Steps:

1. Include the timestamps of the review in the data frame and perform sentimental analysis.

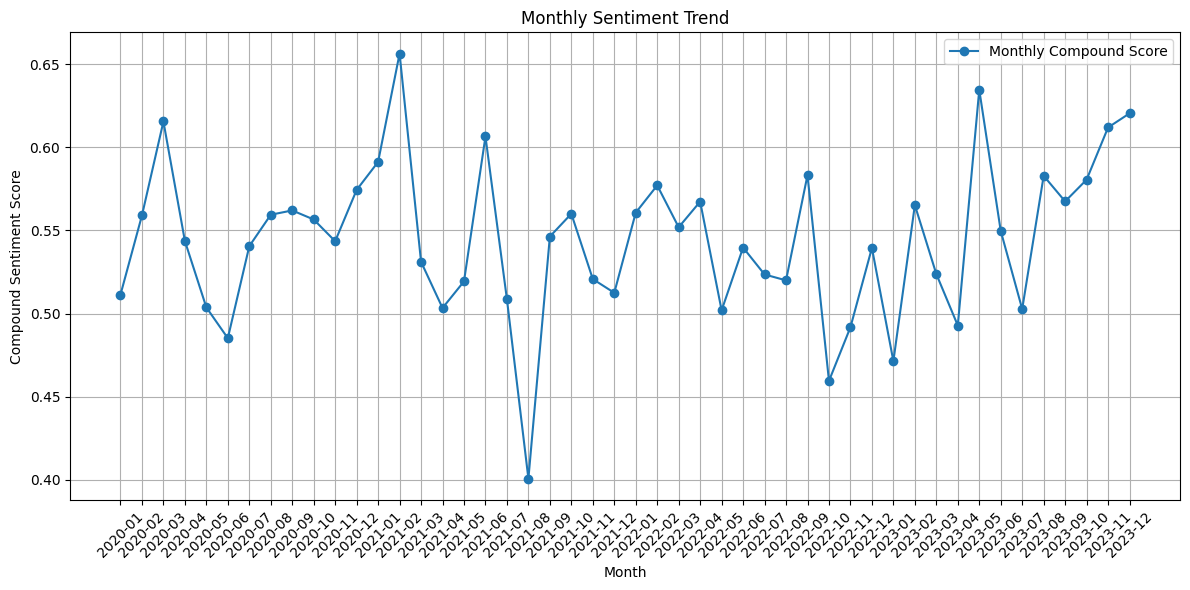


1. Divide the time series into month and years

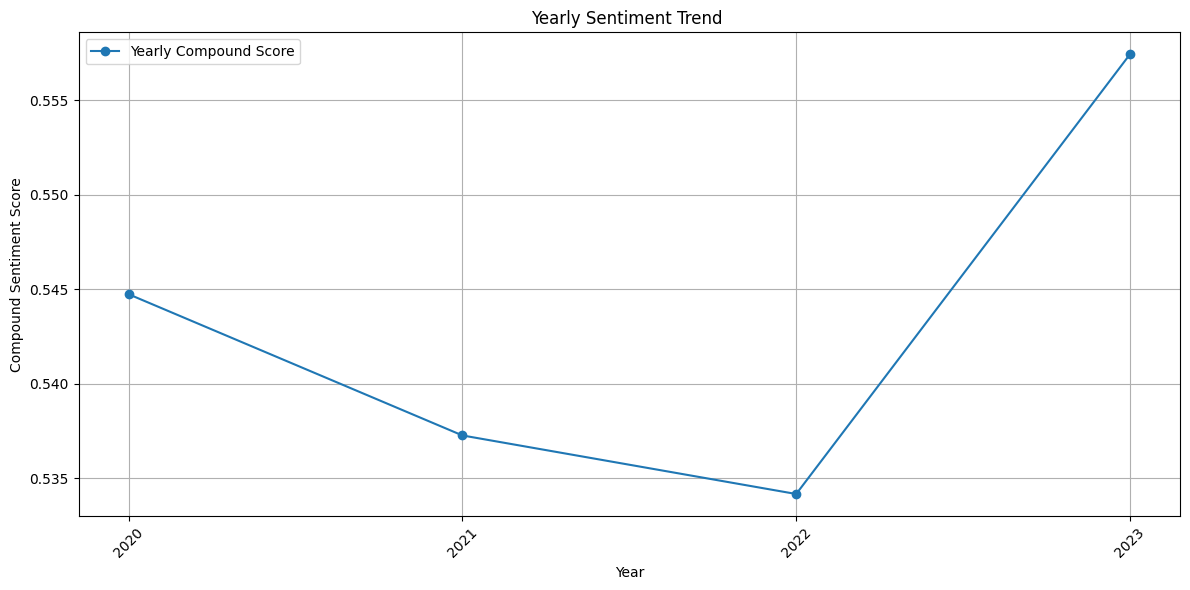


1. Use matplotlib to visualize the trends:

Monthly Trends:



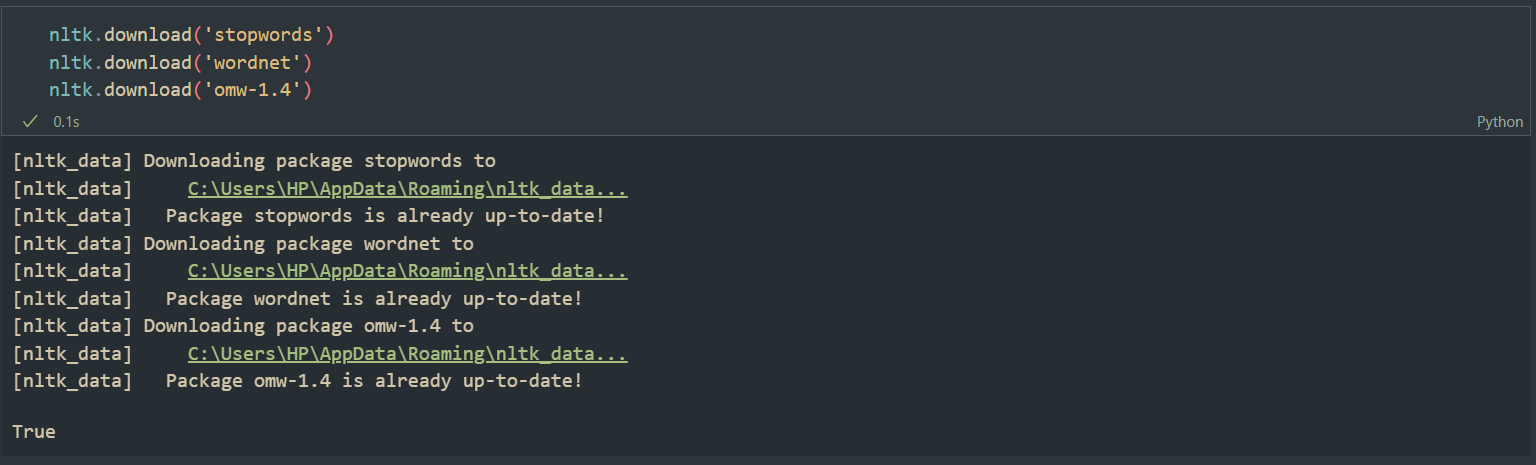
Yearly Trends:



**Section 6: Using LDA Topic Modelling Algo for the most reviewed topics.**

Steps:

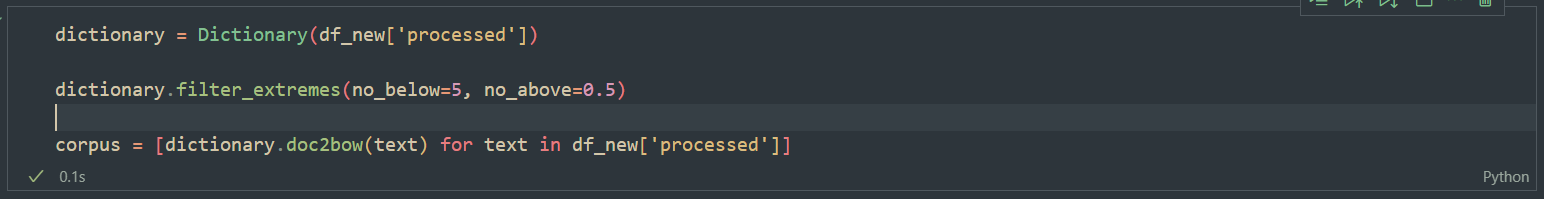
1. Use NLTK library to download the stop words and word net.



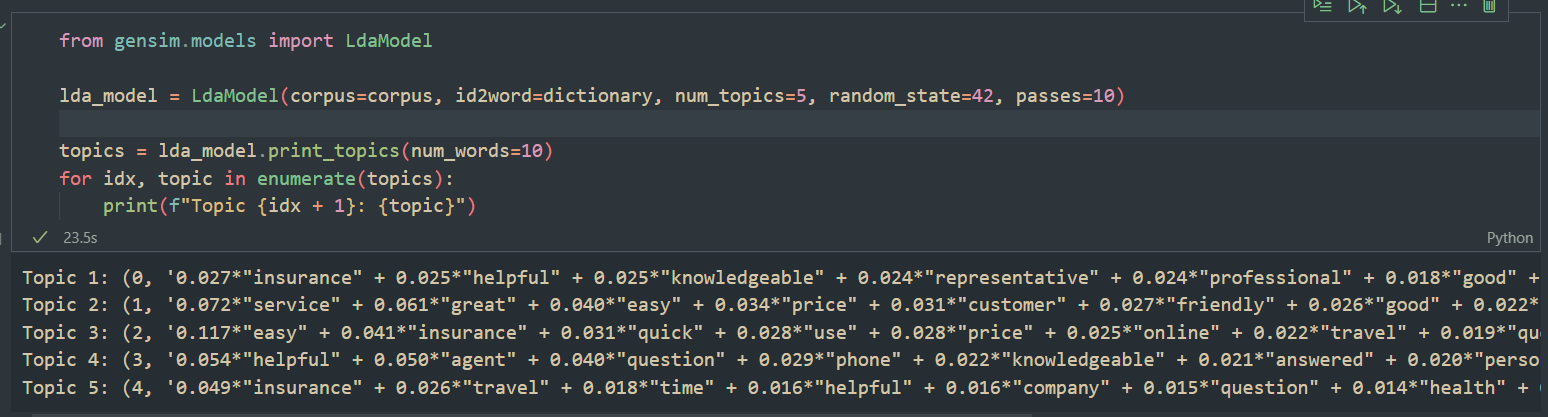
1. Create a new column and tokenize the reviews.



1. Import genism library to create a “Bag of Words” and find label frequent words.

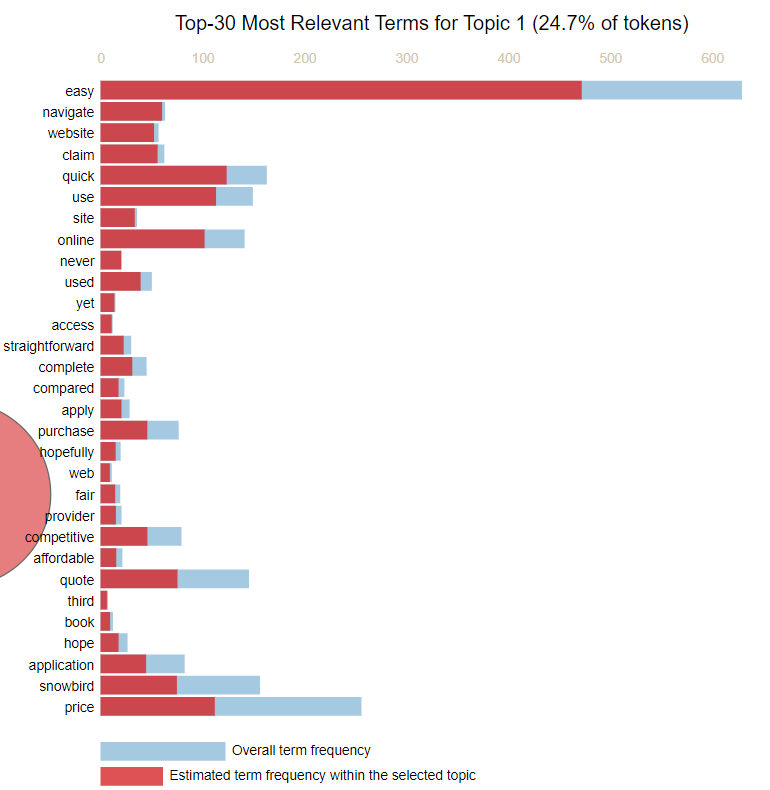


1. Train a LDA Model.

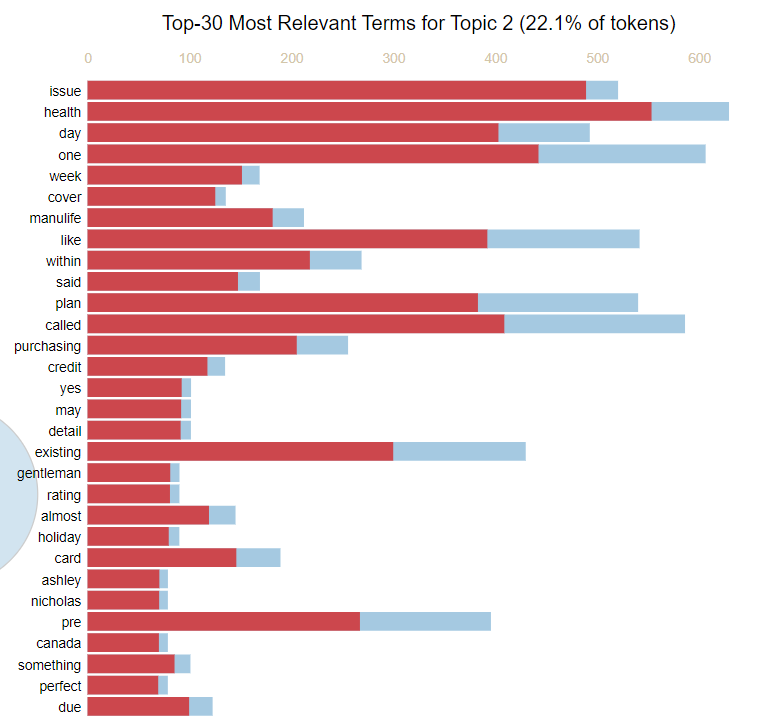


1. Import pyLDAvis and visualize the LDA model.

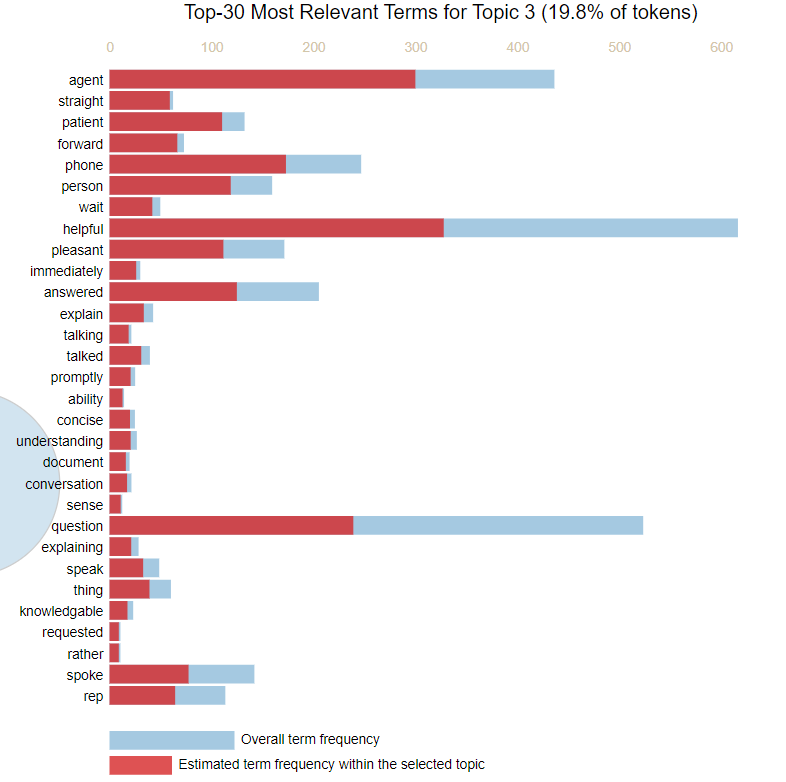
Topic 1:



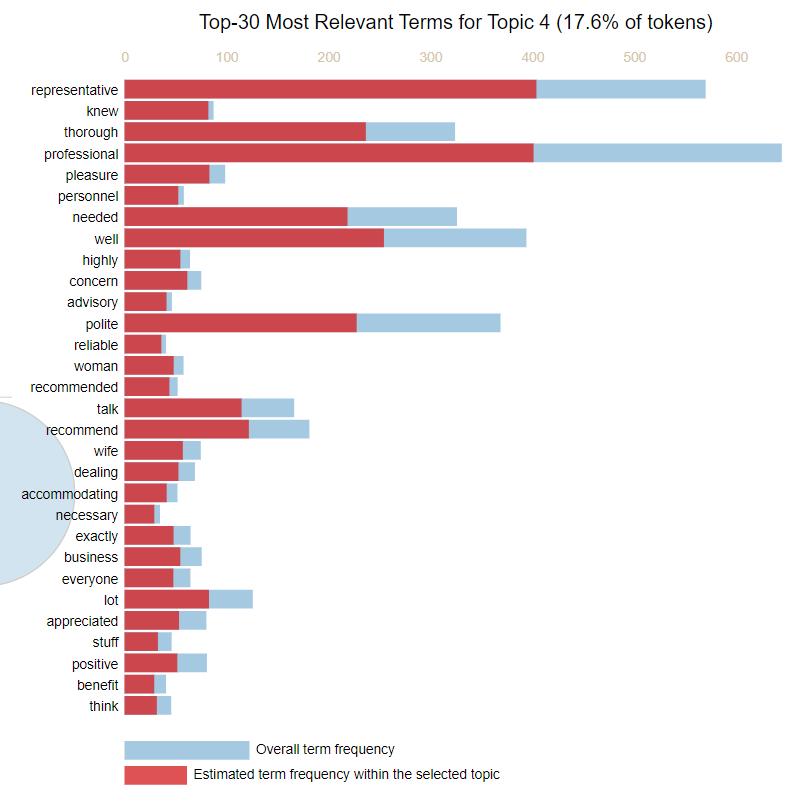
Topic 2:



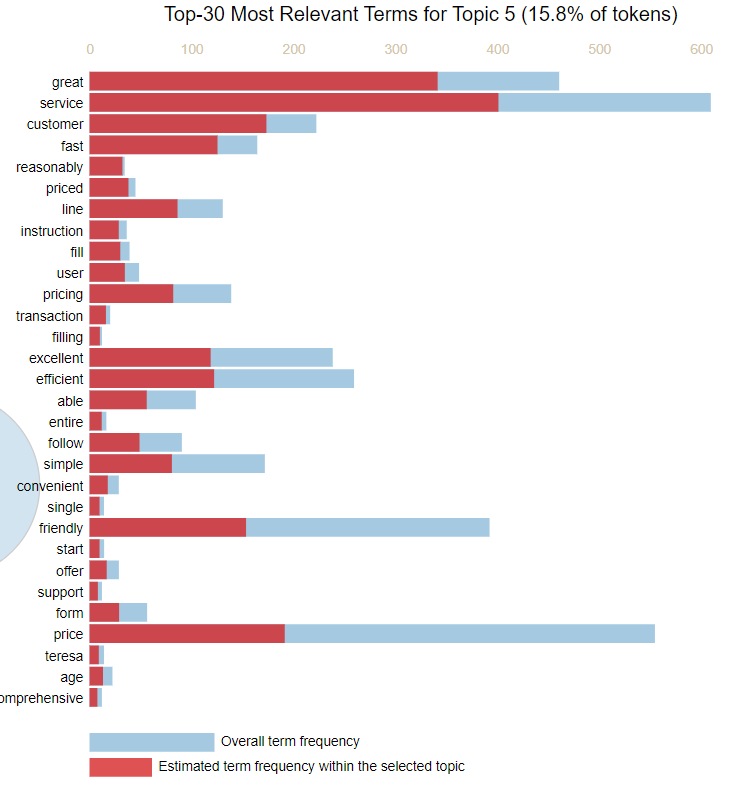
Topic 3:



Topic 4:



Topic 5:



Distance Map:

