TU 257 – Fundamentals of Data Science

Data Analytics

Lab 11 – Text Mining

Brendan Tierney

Agenda

- Demo Notebook 1 – Introduction to Text Mining
    - Part 1 – Example from Notes
    - Part 2 – Text Mining & Word Cloud for a website
- Demo Notebook 2 – Text Mining & Classification Machine Learning

Demo Notebook 1 – Introduction to Text Mining

Demo Notebook 1 – Introduction to Text Mining

- Part 1 – Example from Notes
  - Takes the simple example used in the notes to illustrate each step of Text Mining
  - Run the cells for this example, follow what is being done at each step

  - Exercise: Change the text to something else, rerun the cells and see what happens.
  - Exercise: Change the stop words list, rerun and see what happens

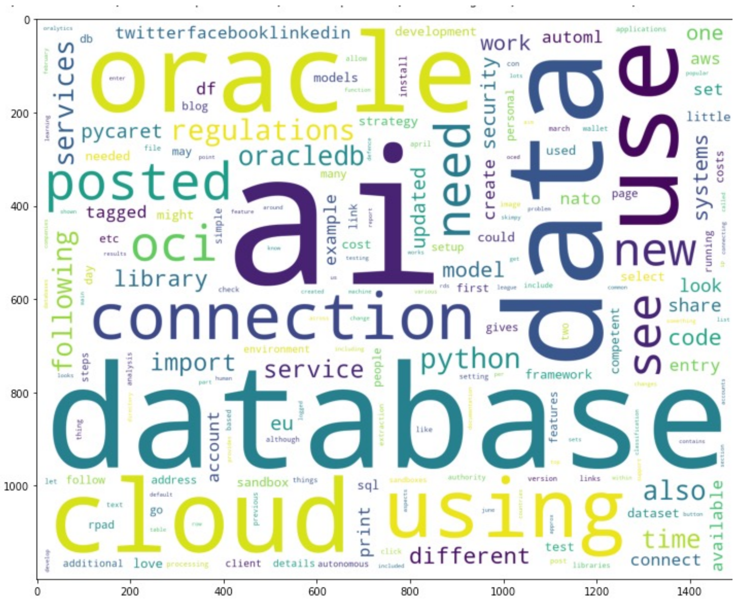
Demo Notebook 1 – Introduction to Text Mining

- Part 2 – Text Mining & Word Cloud for a website. (Part of same/previous Notebook)
  - See blog post – Creating a WordCloud using Python
  - https://oralytics.com/2018/05/21/creating-a-word-cloud-using-python/

  - This is one example of using Text Mining to analyse text
  - 2 different examples are given

  - Run the cells for this example and examine what happens

- Exercise: Use a different website, rerun and examine outputs



Demo Notebook 1 – Introduction to Text Mining

- [Optional]

- Use Text Mining to examine the Election Manifestos from the last two general elections

- The following two blog posts contain links to the Manifestos for each Party

- And Text Mining code used to process these data and produce WordClouds
  - **#GE2020 Analysing Party Manifestos using Python**
  - **Comparing Party Manifestos to 2016**

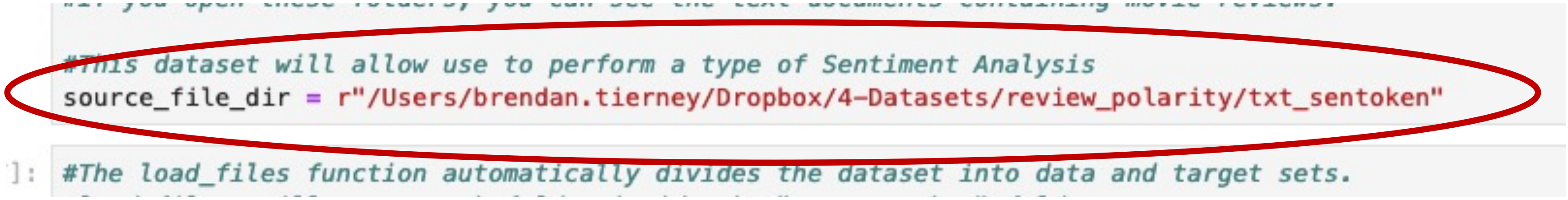
- These WordClouds help to identify what are the comment themes in each manifestos, their relative importance and by comparing between 2016 and 2020 elections how their manifestos have changed.

Demo Notebook 2 – Text Mining & Classification Machine Learning

Demo Notebook 2 – Text Mining & Classification Machine Learning

- Combining Text Mining with Classification (Machine Learning
- This Notebook is based on this blog post
- https://oralytics.com/2021/11/01/combining-nlp-and-machine-learning-for-document-classification/

- Download the Notebook and Dataset from the module webpage.
  - Unzipp the dataset
  - Change the location of the dataset to where you have unzipped it.



```
#This dataset will allow use to perform a type of Sentiment Analysis
source_file_dir = r"/Users/brendan.tierney/Dropbox/4-Datasets/review_polarity/txt_sentoken"
```

```
]: #The load_files function automatically divides the dataset into data and target sets.
```

  - Run all the cells and examine what is done in each cell, particularly the Training and Testing the model

Any Questions ?

What Now/Next ?

Complete **all** Lab Exercises before <u>Next Week</u>