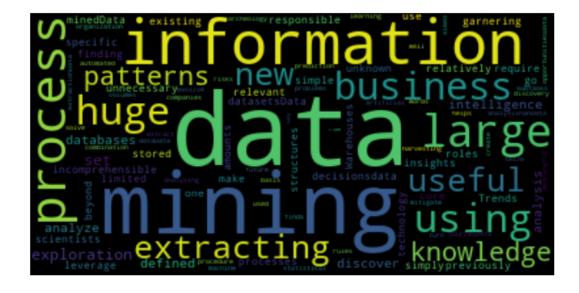
Programming Assignment 1 - Data Mining

January 25, 2020

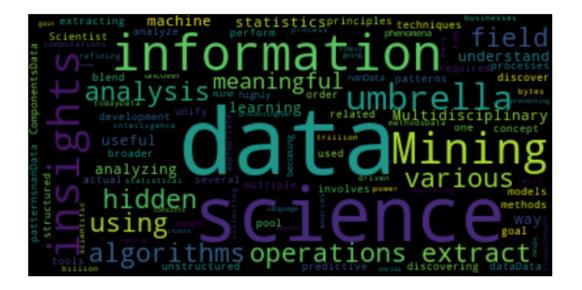
1 Programming Assignment 1 - Data Mining

Author: Skylar Trendley Course: CS5402 Date: January 25, 2020

```
In [31]: #Imported for data management (dataframes)
         import pandas as pd
         #Imported to create word clouds
         from wordcloud import WordCloud
         #Imported to allow for the display of word clouds
         import matplotlib.pyplot as plt
         #read in the .csv file
         df_mining=pd.read_csv("data_mining.csv",dtype=str)
         #turn any instances of floats into strings
         df_mining = df_mining.applymap(str)
         #join the columns of text to make a "word glob"
         str_mining=''.join(list(df_mining['Text']))
         #generate the word cloud counting common occurences
         wordcloud_mining=WordCloud().generate(str_mining)
         #display the word cloud
         plt.figure(figsize=(10,8))
         plt.axis('off')
         plt.imshow(wordcloud_mining,interpolation='bilinear')
Out[31]: <matplotlib.image.AxesImage at 0x7f744eec08d0>
```



```
In [30]: #Imported for data management (dataframes)
         import pandas as pd
         #Imported to create word clouds
         from wordcloud import WordCloud
         #Imported to allow for the display of word clouds
         import matplotlib.pyplot as plt
         #read in the .csv file
         df_science=pd.read_csv("data_science.csv",dtype=str)
         #turn any occurences of floats into strings
         df_science = df_science.applymap(str)
         #join the text columns to make a "word glob"
         str_science=''.join(list(df_science['Text'].fillna('')))
         #generate the word cloud using most word patterns
         wordcloud_science=WordCloud().generate(str_science)
         #display the word cloud
         plt.figure(figsize=(10,8))
         plt.axis('off')
         plt.imshow(wordcloud_science,interpolation='bilinear')
Out[30]: <matplotlib.image.AxesImage at 0x7f744ef067b8>
```



1.0.1 Concept Description:

In this homework, our goal is to determine the difference in the language between data mining and data science as reported by fourteen randomly recorded websites.

1.0.2 Data Collection:

Using google, fourteen example sites that return results for both "What is Data Mining" and "What is Data Science." were chosen to record language patterns. The results of "Data Mining is..." or "Data Science is..." in the form of a text block was lifted from the website. Pertinent record data such as category, pull date, URL, and source site were also recorded to ensure search result integrity.

1.0.3 Example Description:

Category: The term which was searched for. Either "Data Mining" or "Data Science". Pull Date: The date in which the data was retrieved from the website. URL: The URL of the website where the data was taken. Source Cite: The owner of the URL which the data was posted. Text: The textblock of data akin to the search results for data mining or data science.

1.0.4 Data Import and Wrangling:

Three python3 library imports were used in the analysis of this dataset. The first, being pandas, allowed for reading the data in as a .csv for data processing. The second, being wordcloud, allowed for the data to be turned into a "word blob", which is essentially just a large string of data. This allowed for the word blob to be analyzed for patterns, or commonly occurring words. The third, matplotlib, allowed for displaying the word cloud in a visual format.

1.0.5 Exploratory Data Analysis:

N/A

1.0.6 Mining or Analytics:

Wordcloud and matplotlib allowed for the data to be transformed into a visual format. In the visualization, the largest words are the words that appear the most, and as they appear smaller in size, their occurences in the follow a linear correlation.

1.0.7 Evaluation:

N/A

1.0.8 Results:

The language used for data mining and data science appear to be very similar. Common occurences in both include words such as "information, extracting/extraction, useful, analysis, and statistics". An interesting trend in the data science category is that it often mentioned data mining in its definition, whereas data mining hardly mentions data science. This could be due to the fact that data mining is a field of data science. Data science was often described to be multidisciplinary and an umbrella term for subcategories in its definition, whereas data mining was often more specific in what it actually meant.

1.0.9 References:

Websites used to collect Data: https://www.datasciencecentral.com/profiles/blogs/differenceof-data-science-machine-learning-and-data-mining https://data-flair.training/blogs/datamining-and-data-science/ https://www.datasciencegraduateprograms.com/datamining/ https://www.houseofbots.com/news-detail/11973-1-clarifying-differencesbetween-data-analysis-data-mining-data-science-machine-learning,-and-big-data https://www.edureka.co/blog/whathttps://www.newtechdojo.com/what-is-data-mining/ is-data-science/ https://www.coursehero.com/file/p5p65iv/What-is-Data-Science-Data-Science-is-a-blend-of-various-tools-algorithms-and/ https://medium.com/tech-in-200-words/what-is-data-mining-understandingin-200-words-cd6ddf4fcf11 https://www.talend.com/resources/whatis-data-mining/ https://www.educba.com/advantages-of-data-mining/ https://pinformatics.tamhsc.edu/phpm631/ppt/lec1_ds.pdf https://365careers.net/datascience-explained-predictive-modeling-data-mining-data-warehousing-querying/ https://becominghuman.ai/8-key-differences-between-data-science-and-data-mining-674f09599df2 https://www.microstrategy.com/us/resources/introductory-guides/data-miningexplained Imports used to shape and analyze data: https://pandas.pydata.org/https://matplotlib.org/

In []:

https://github.com/amueller/word_cloud