ST3001 –PYTHON ASSIGNMENT

INTRODUCTION

The purpose of this report is to review the data visualization packages available to use through the Python programming language, and how effective they are in their visualization capabilities. To analyse the visualization capabilities of each package, we will be using Twitter data that was scraped in February 2015. This data set is a sentiment analysis carried out on thousands of tweets for each of the major U.S. airlines. This data was classified in terms of positive, neutral and negative tweets, further categorized by negative sentiment, such as if the customer experienced a late flight or a booking problem.

The visualization packages used were:

- Matplotlib package – found at https://matplotlib.org/
- Plotly Package – found at https://plot.ly/python/getting-started/
- Bokeh Package – found at https://bokeh.pydata.org/en/latest/

Data visualization of the airline sentiment analysis was created out in each package to analyse and compare. Each visualization package had their individual areas where they performed well in, and these will be explored more in depth below.

ANALYSING THE DATA

The data was read into the Spyder IDE and analysed using the Pandas package. Pandas was chosen to analyse the data as it is best suited for Tabular data. It is also capable of handling large datasets, keeping it's processing speeds at a high rate.

MATPLOTLIB

Matplotlib is a Python 2D plotting visualization package which produces graphs and plots, using Python scripts. It is a useful complement to Pandas. To use Matplotlib with the Pandas dataframe we imported it from the matplotlib module. Further details of the matplotlib library can be found at the link above.

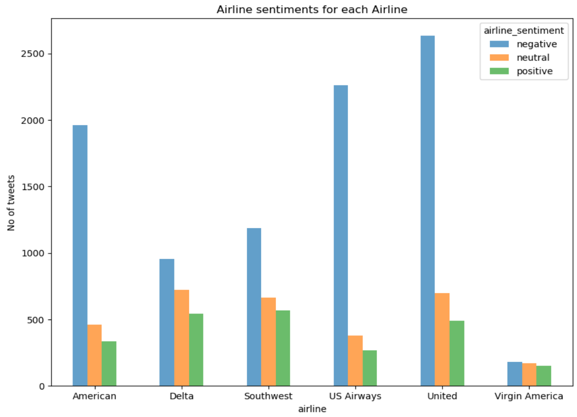
DIFFICULTY / EASE OF USE

It is relatively easy to use Matplotlib in conjunction with Pandas dataframes for the sentiment analysis. As Pandas comes equipped with useful wrappers around several matplotlib plotting routines, dataframes can be plotted quickly and easily with just a few lines of code. Plots can also be customised with the manipulation of the dataframe. There are also numerous examples of plots generated from matplotlib's online documentation.

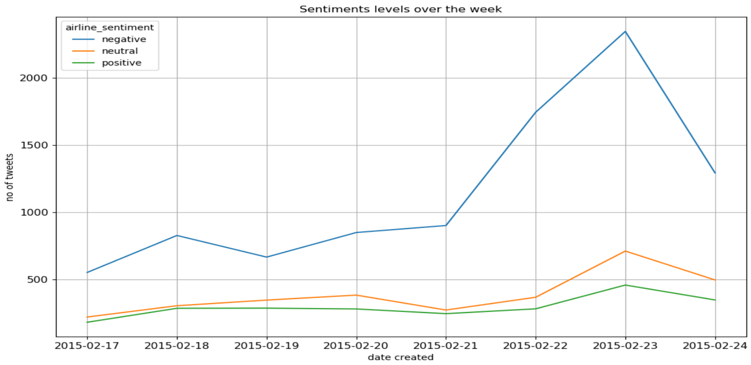
CAPABILITIES

Matplotlib can be used to create a wide range of plots such as bar-charts, line plots and boxplots. It can produce plots quickly in hardcopy formats for publication, and also in interactive formats. Using Matplotlib, you can also easily manipulate plot properties like the font, axes, and line styles. Matplotlib also allows you to zoom into plots to inspect them further.

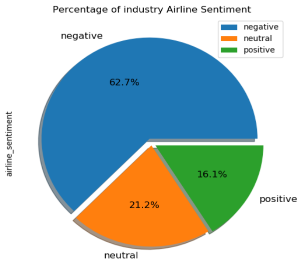
VISUALIZATIONS

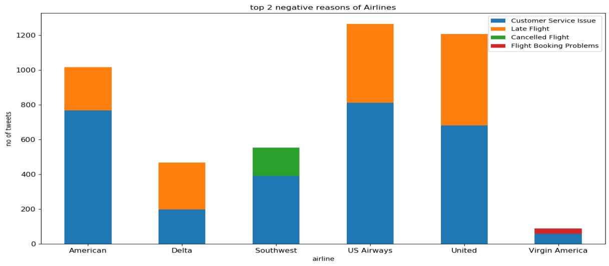


In this visualization, using Matplotlib we have plotted a sentiment summary for each of the airlines contained in the dataset in bar chart form. From manipulating the dataframe to extract the data needed, the plot was generated by applying Matplotlib's plot function to the resulting dataframe. From this visualisation, one can simply analyse and observe the number of tweets associated with either of the sentiments for each airline. It was easy to visualize this data in terms of a bar chart using the "kind=bar" parameter. Matplotlib automatically colour-coded the sentiment categories, allowing for straightforward interpretation of the data. It was also easy to set the size of the plot generated by using the figsize parameter.



Here we have a visualization of the levels of airline sentiment categories for the industry over the week the twitter data was scraped from. Initially, this was plotted using a bar-chart, but was transform into a line chart using "kind=line". The x-axis label was automatically generated as the dates were passed as the index of the dataframe, while the y-axis label was added by using the plt.ylabel parameter. As seen above, this plot contains a grid which can be added effortlessly by passing "grid=True" as a parameter. It is straightforward to save plots to your computer, by using the savefig() function or by clicking the save button on the plot generated. From the two plots above, it is clear Matplotlib produces plots that are clear and easy to digest. More examples of plots generated by Matplotlib are shown below.





PLOTLY

Plotly is an online collaborative data analysis and graphing tool that can be used to create dashboards and data powered applications. The Python API allows access to all of Plotly's functionality. Plotly figures are shared, tracked and edited all online through your Plotly account. Plotly can be installed on your machine using pip installer.

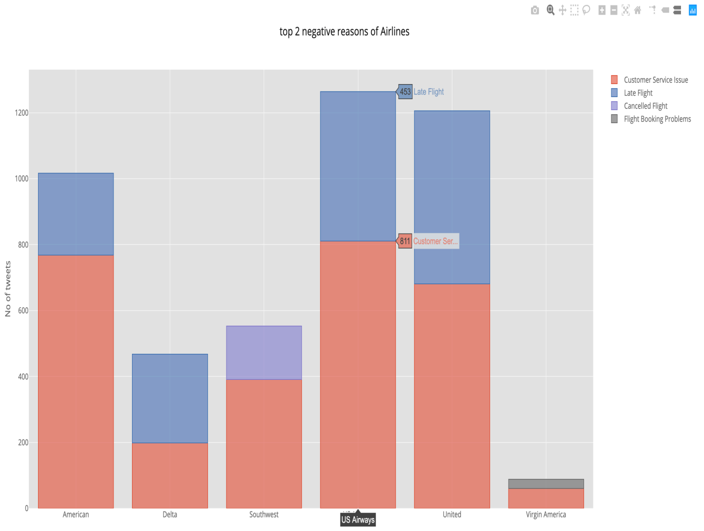
DIFFICULTY / EASE OF USE

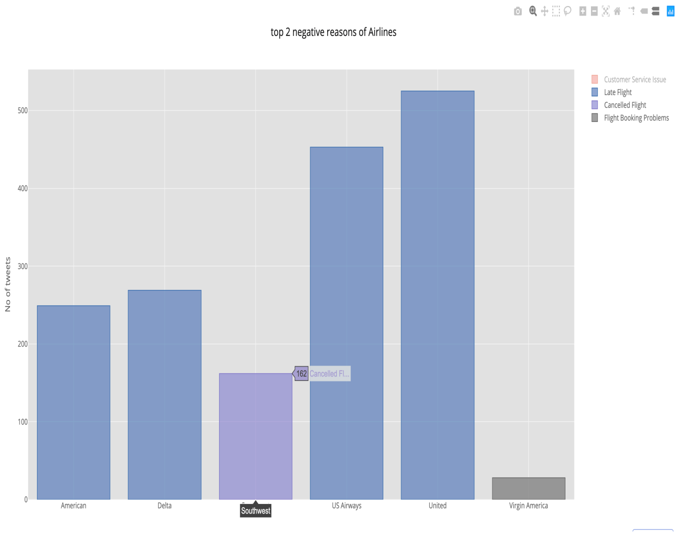
Plotly is not difficult to use when working with data. Generating visualisations with Plotly is explained through its online library with numerous examples. By installing Cufflinks (also installable through pip installer), you can bind Plotly directly to Pandas dataframes – making it intuitive to visualise data when analysing it through Pandas.

CAPABILITIES

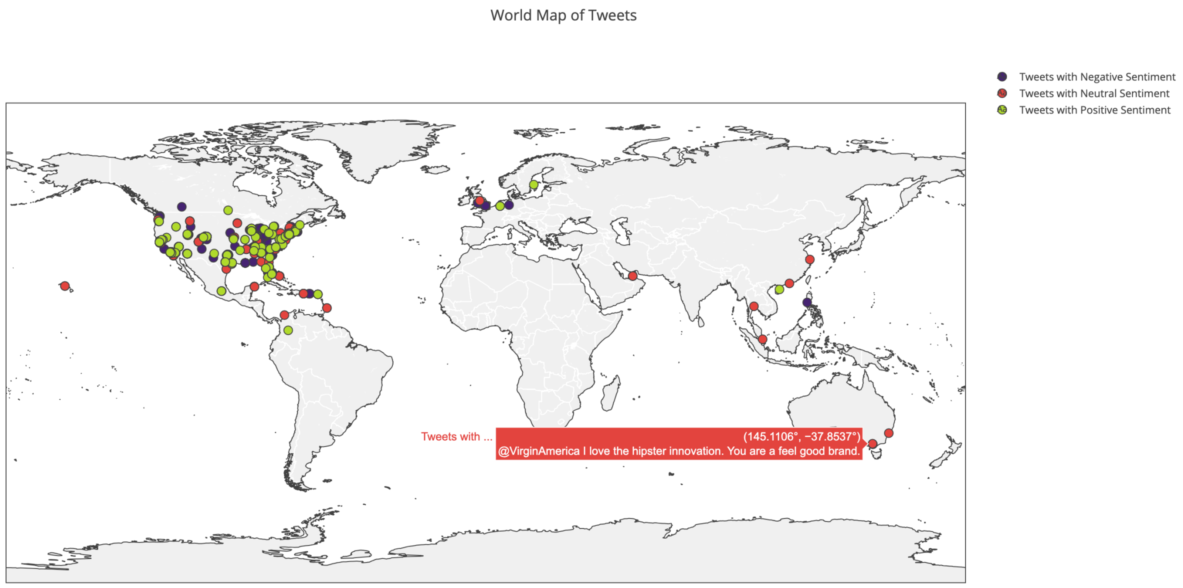
Like Matplotlib, Plotly allows you to generate data visualizations through a wide range of plots. However, compared to Matplotlib – Plotly excels through producing more interactive visualizations. Plotly also allows users to visualize data through a range of different programming languages - for example R, Perl, and Matlab. Users can also generate offline graphs from the environment they are programming in.

VISUALIZATIONS



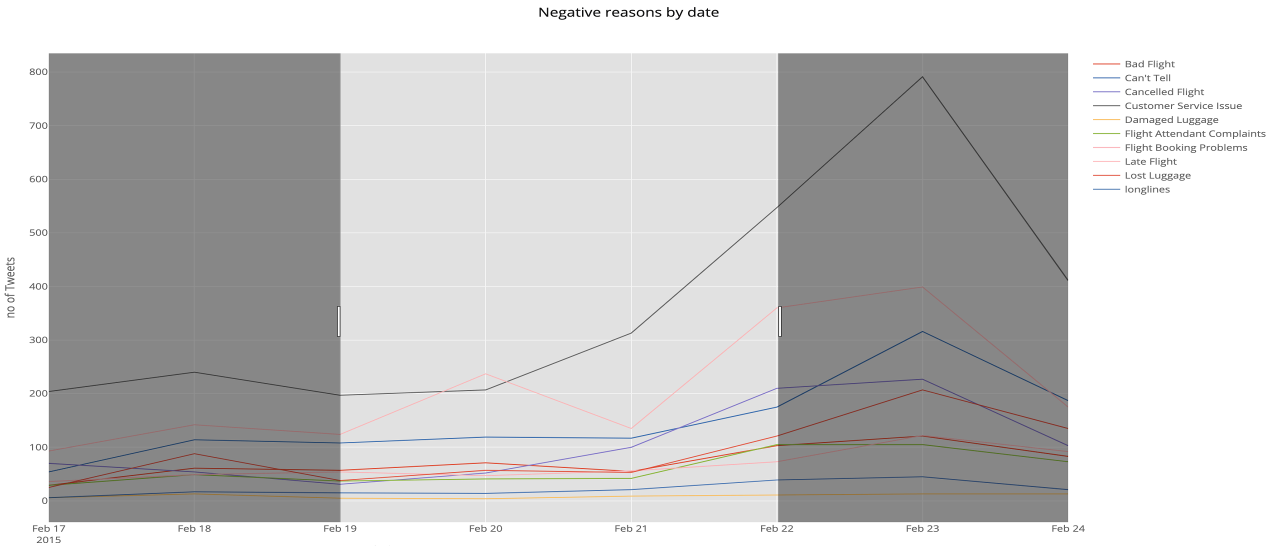


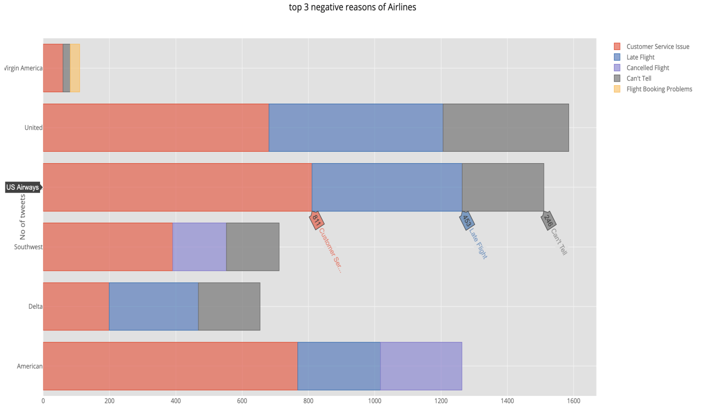
The left stacked bar chart shows the top two negative reasons tweeted for each airline. It was plotted in just a few lines of code as importing cufflinks allows the binds Plotly directly to the dataframe. As it can be observed from the graph, when the user's mouse moves over the visualization, Plotly displays the contents of each bar and the number of tweets related to each reason is displayed on the screen. The user can also easily select what type of data they want displayed. For example, the bar chart on the right only shows the data where Late Flight, Cancelled Flight, and Flight Booking Problems are selected from the legend.

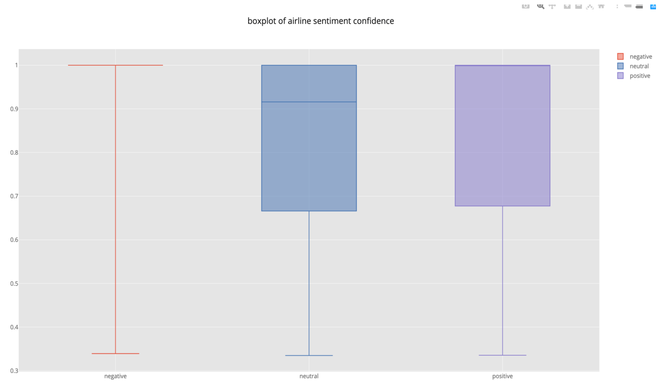

World Map of Tweets

Tweets with Negative Sentiment
Tweets with Neutral Sentiment
Tweets with Positive Sentiment
Tweets with ...
(145.1106°, −37.8537°)
@VirginAmerica I love the hipster innovation. You are a feel good brand.

The capability of Plotly’s visualization and interactive use can be seen by the plot placed above. The world map was plotted using Plotly’s Scattergeo function. The points on the map were plotted by passing tweet coordinates into the function. The points were customizable by colour, size, and shape. Using the mouse, the user can drag and change position on the map. When the user hovers over one of the points on the map, the tweet content and coordinates are shown. Plotly produces clear, visually appealing, highly descriptive plots with great interactivity. More examples can be seen below.


Negative reasons by date


top 3 negative reasons of Airlines


boxplot of airline sentiment confidence

BOKEH

Bokeh is an interactive visualization library that targets modern web browsers for presentation. It is used to create dashboards and data applications, with high-performance capabilities over very large or streaming datasets. Bokeh can be installed using pip installer. Visualizations are generated through python scripts.

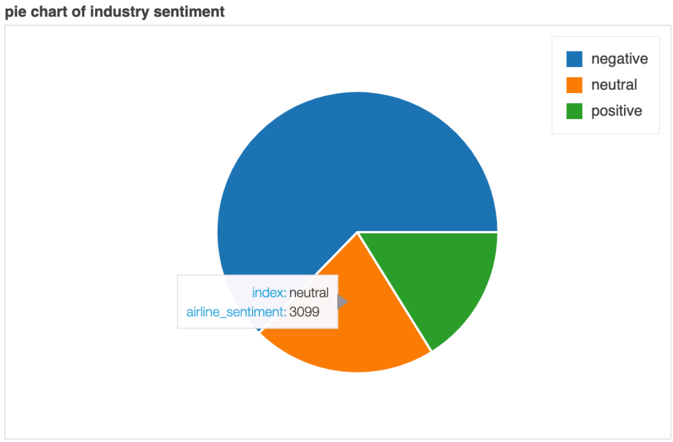
DIFFICULTY / EASE OF USE

Compared to Matplotlib and Plotly, Bokeh is not as easy/intuitive to use. Installing Pandas_bokeh allows for a more complementary use between bokeh and Pandas. However, it is still less integrated with Panda than Plotly and Matplotlib.

CAPABILITIES

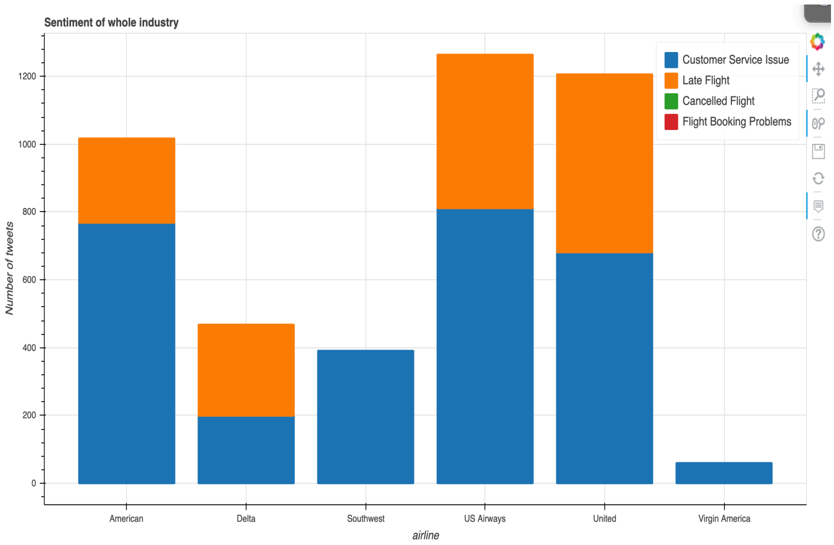
Like the two previously mentioned data visualization, Bokeh can visualize a wide range of plots ranging from bar charts to line plots. Bokeh is more interactive than Matplotlib, but not as interactive right out of the box compared to Plotly. Greater interactivity can be created in Bokeh, but it will require the user to have a high level of data analysis skill.

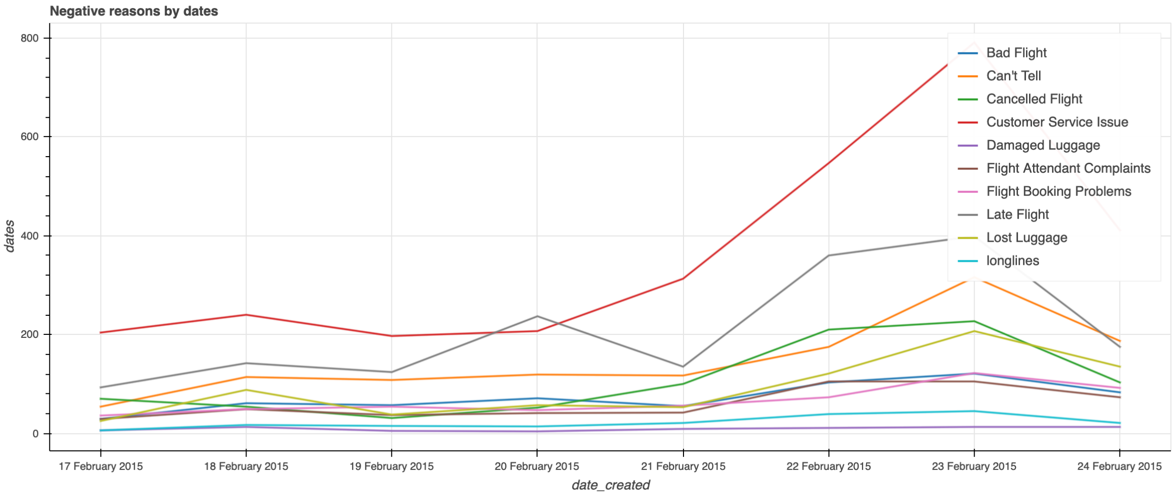
VISUALIZATIONS



As can be seen from the graph, when a pie chart is plotted in bokeh – the values of each slice are only shown when the mouse is hovered over it. Although it is visually appealing and has good interactive features, compared to Matplotlib and Plotly - it doesn't plot the value of each slice onto the graph when the visualization is generated. To plot pie chart values, additional data manipulation is needed.

This stacked bar chart shows the most frequent two negative reasons found for each airline. However, compared to Matplotlib and Plotly – Bokeh plots incomplete visualisations when data is missing (NaN values present). The remaining reasons for Southwest and Virgin America airlines aren't plotted because some unknown values are present in their rows in a crosstab, while as seen in graphs above – Matplotlib and Plotly could plot the data anyway. Additional data manipulation would be required to make sure all data is displayed correctly.





Above a plot is displayed of the negative reasons tweets over the course of the week. It is plotted in a clear and visually appealing way, but the legend is placed over the line data points in the plot. Plotly automatically places this outside of the main plot area to avoid overlapping content, while Matplotlib places the legend in the “best positioned place”. Bokeh produces visually appealing, interactive, easy to interpret graphs, however, because a high level of data analysis is needed in certain situations – it makes it difficult to work with at times.

CONCLUSION

After generating data visualisations of the twitter airline sentiment analysis, it was clear that each data visualization package had their individual characteristics and strengths. Matplotlib is visualization package that works seamlessly with Pandas data frames, allowing users to generate and customise plots quickly and efficiently in several different ways. However, compared to bokeh and Plotly, Matplotlib lacks interactivity.

Bokeh offers more interactive elements to its plots than Matplotlib – but producing customised plots can be a difficult process. This is largely due to its lack of integration with Pandas. Therefore, to produce customised plots to the level seen with Matplotlib and Plotly, it is likely that one would have to be skilled in data analysis. Overall, Bokeh is a data visualization package that is perfectly suited to highly skilled data analysers who want to build customisable, interactive visualizations.

Overall, Plotly is the visualization package that would be most suitable for presenting the data given in the twitter sentiment analysis of airlines. Installing cufflinks allows you to work seamlessly with Pandas data frames. Highly interactive graphs can be produced with a few lines of code, with scope to customize plots. Plotly keeps all plots online in a centralized location, ensuring smooth and easy access when needed.

INSTALLATIONS NEEDED FOR PACKAGES

Running the following commands in the computer terminal will ensure that the necessary components of the packages are installed, so visualization through them is possible.

PLOTLY

```
pip install plotly

pip install cufflinks
```

BOKEH

```
pip install bokeh
```