**https://pete-lda-dash-app.herokuapp.com/**

**Project Description**

The main data sources were 10-K reports collected from the SEC EDGAR portal through web scraping. A list of companies downloaded from NASDAQ.com was used to direct the scraper and provide light background info on each company. Not every company had an easily scrapable 10-K so some were skipped over. From these pages raw text data was harvested, normalized, and fed into an LDA topic model (genism). The outputs of this model were then transformed and loaded into mlab for access by the main app.

The main app is an interactive dashboard built with Plotly Dash. There are two tabs; an exploration of the topics present in the 10-K corpus and a company-centric page. The first page is driven by a scatter plot of the topics, which maps the topic space (the coordinate determinations were based on the source code from pyldavis). Like topics are near each other and there are noticeable themes in the four quadrants. This scatter cross-filters two bar graphs, showing the words present in the chosen topic and the companies whose reports belong to the chosen topic. The second page allow the user to see the topic break down by company. This breakdown, done with a pie chart, controls an excerpt from the 10-K, highlighting the words most relevant to the chosen topic.

Sales professionals are the intended audience of this app. The front page is intended to be exploratory in nature, giving a broad market overview and allowing the user to find companies who have a need for the user’s product or service based. The topics scatter can be shaded by market cap, allowing the user to find rich sales opportunities, or sector, letting them find trends in particular industries. The second page, because of the 10-K excerpt, is very versatile. It could be used as a quick reference while contacting a lead or conducting in depth research.

**Scraping:** CompanyScraper.py

**Topic Model:** All\_word.py, LDA.py

**Front End:** topic\_dash.py

**User Guide**

*Tab 1*

**Scatter Plo**t: The color scale can be controlled by the dropdown on the far left. By default it is set to market cap. Hover over marks to see the topic name, which is generated by choosing the most salient word from that topic. Click bubbles to filter the two bar graphs.

Words Bar Graph: Each word belonging to a topic is scored using a formula that takes the frequency within that topic and the frequency overall as inputs. The words are sorted by this score

Company Bar Graph: The topics related to each company sum to 100%. The companies shown here are those related to the chosen topic.

*Tab 2*

**Pie Chart:** This controls the highlighting on the 10-K excerpt

**10-K Excerpt:** Words with a rank (based on the score defined in the Tab 1 description) of 30 or less within the chosen topic are bolded.

**Sources**

<https://github.com/bmabey/pyLDAvis>

<https://radimrehurek.com/gensim/#>

Various Dash and Plotly tutorials