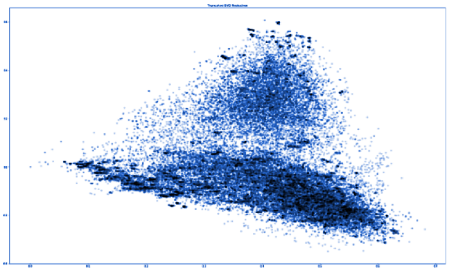
TV CAPTIONS - UNSUPERVISED MACHINE LEARNING AND NLP

SCOPE:

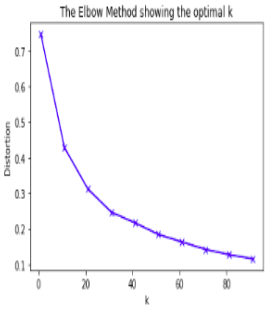
From opinion polls and product reviews to shaping market strategy, NLP is a tool with the ability to transform businesses. The scope of this project is to create an unsupervised classification model and provide brand awareness and sentiment analysis to top brands based on placement during TV broadcasts. With this information companies will be better informed of latent brand perception and take action if necessary.

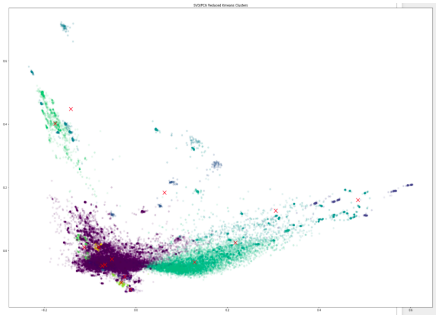


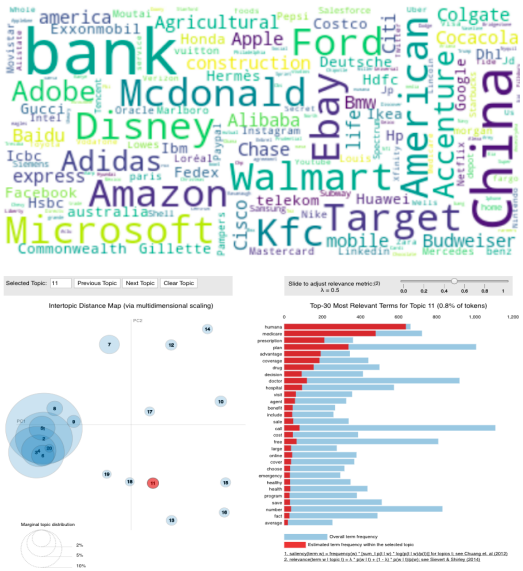
TECH-STACK:

APPROACH:

Closed caption text was reduced in dimension (SVD) then modeled via KMeans and Latent Dirichlet Allocation. K was identified by plotting a range of clusters along with the calculated sum of squared errors (SSE), the K with a reduced SSE was selected for this model. LDA topic modeling parameters were optimized by GridSearch.







SENTIMENT ANALYSIS:

Brand entities were identified using spaCy to captured unique mentions in each text. Brands were then scored based on the polarity score of adjacent adjective-noun combinations. Positive and negative sentiment scores and valence for each brand were generated using NLTK's VADER library.

TAKEAWAY:

Closed caption data is unstructured and unclassified, this model assists in classification and can be further applied to recommender engines, viewership prediction, and brand marketing.

https://www.linkedin.com/in/albertolovell/

albertolovell/closed_caption_classifier

https://www.youtube.com/watch?v=my8ZuVVUFoI