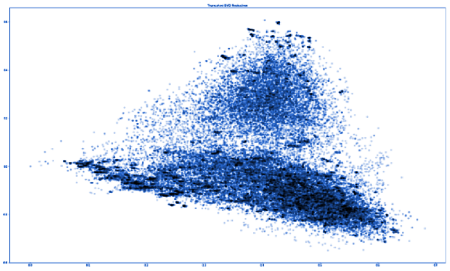
TV CAPTIONS - UNSUPERVISED MACHINE LEARNING AND NLP

SCOPE:

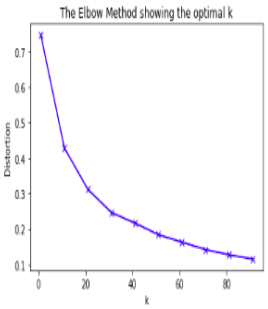
From opinion polls and product reviews to shaping market strategy, NLP is a tool with the ability to transform businesses. The scope of this project is to create an unsupervised classification model and provide brand awareness and sentiment analysis to top brands based on placement during TV broadcasts.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

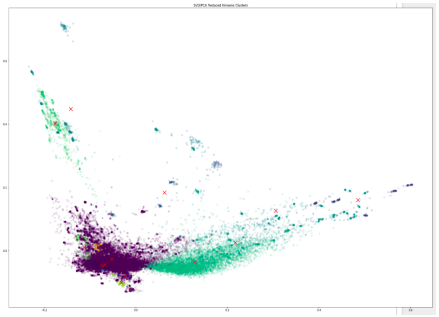


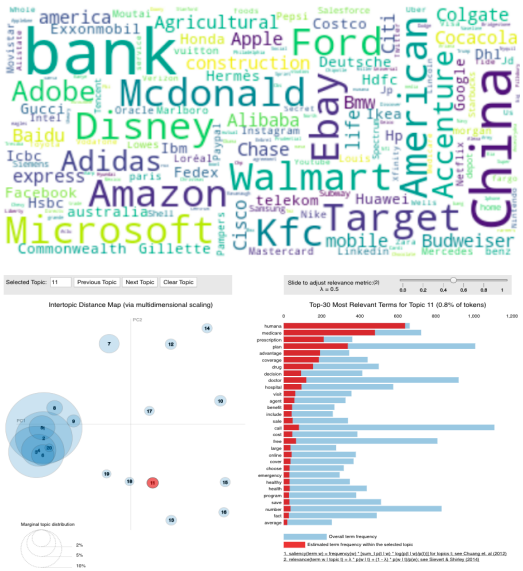
TECH-STACK:

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

Alberto Lovell

San Francisco, CA (925)752-1903 | alberto.lovell@gmail.com | linkedin.com/in/albertolovell | github.com/albertolovell

EXECUTIVE SUMMARY

- Data scientist with 8+ years quantitative and analytical experience in Healthcare and Genetics. Experienced in Tech, Entertainment, and Business data analytics
- Key qualifications: applying Python, SQL, statistics, data mining, and machine learning algorithms (regression, classification, ensemble methods, clustering, time series, NLP, recommender systems) to identify useful insights from structured and unstructured data, and communicating those insights to both technical and non-technical audiences

EXPERIENCE

GALVANIZE, SAN FRANCISCO, CA 2018
DATA SCIENCE IMMERSIVE

- Partnered with Gracenote/Nielsen to create an unsupervised classification model using KMeans, dimensionality reduction, LDA topic modeling, and pos/neg polarity sentiment scores to identify latent topics and provide brand awareness to top brands and TV broadcast stations via closed caption content
- Deployed a web-based model to prevent profit loss by flagging potential fraud events for a popular events ticketing website using NLP, Naive Bayes, and Gradient Boosting algorithms
- Created a KNN based user churn model to predict ride-share customer churn and reduce expenditure
- Built a movie recommendation engine via Collaborative Filtering and SVD dimensionality reduction

STANFORD UNIVERSITY, PALO ALTO, CA 2014 – 2018
LIFE SCIENCE RESEARCH PROFESSIONAL

- Led genomic services generating 25% annual revenue and implemented key organizational changes to decrease expenditure
- Built analytical models on gene expression datasets to identify potential targets of interest
- Established a client-friendly reproducible linear regression workflow which can be easily implemented on any gene expression profile to validate Next Generation Sequencing data
- Developed a CRISPR gRNA identification method to target molecules with optimized gene editing specificity and reduced off-target effects
- Managed genomics QC team and spearheaded client engagement, project management, and PHP database management

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO (UCSF), CA 2011 -2014
RESEARCH ASSOCIATE

- Developed an image classifier for abnormal CT/X-ray images via open-source Linux packages, authored and published results in peer-reviewed journals
- Created a statistical classification and scoring system to investigate correlation between T2 Diabetes and intervertebral discs

SKILLS

Machine Learning
- Supervised and Unsupervised
- Classification, regression, clustering
- Time series, dimensionality reduction
- Natural language processing (NLP)

Programming / Scripting / Libraries
- Python, sklearn, Numpy, Pandas
- SQL, PostgreSQL, MongoDB
- PsychoPG2, BeautifulSoup
- Spark, AWS, Git, Flask

Analysis
- Statistics (Frequentist + Bayesian)
- Hypothesis testing, A/B Testing
- Experimental design
- Visualization with matplotlib, seaborn

EDUCATION

Galvanize Data Science Immersive Program, San Francisco
2018

B.S. in Biochemistry and Molecular Biology, **University of California, Davis**
2010