Reddit E-Cigarette Data Mining

By Tyler Trimble

Description of Methodology

- Data Collection
  - Subreddits: r/vaping, r/e_cigarette, r/electronic_cigarette, r/vapes
  - For each subreddit I collected the top 250 posts all time and their top 15 comments
- Obtained a total of 8,662 comments
- Data Cleaning
  - I converted non UTF-8 characters to strings and replaced newline and special characters with a space
  - I removed posts containing a URL
  - I removed all posts with less than three words
  - I removed all posts containing words related to CBD or marijuana
  - I removed stopwords, including vape, vaping, and ecig
- Finished with 6,801 comments

Details of Data Mining Process

- Topic Modeling
  - I began by vectorizing the posts using the bag of words vectorizer
  - Applied LDA topic modeling to all of the posts
    - Did not give me good results, regardless of the number of topics I used, ranging from 3 to 10
- Clustering
  - Used bag of words vectorizer to vectorize the posts and then applied K-Means clustering
    - I set hyperparameter K to 20 and got an inter-cluster distance of 68
    - These clustering techniques best helped me analyze the data and generate conclusions

Sentimental Analysis

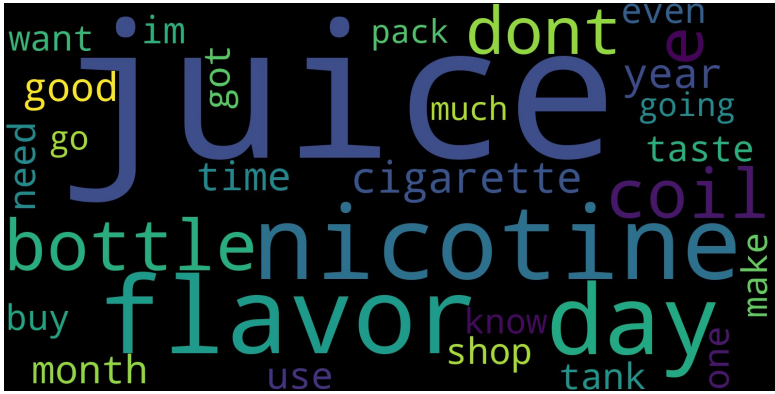
- I applied a sentimental analysis to evaluate the public opinion surrounding electronic cigarette dialogue
- Results:
  - 6,801 samples scored
  - 320 **very bad** samples → sample_score <= –0.75
  - 1753 **bad** samples → –0.75 < sample_score <= –0.1
  - 5874 **neutral** samples → –0.1 < sample_score < 0.1
  - 3425 **good** samples → 0.1 <= sample score < 0.75
  - 927 **very good** samples → 0.75 <= sample_score

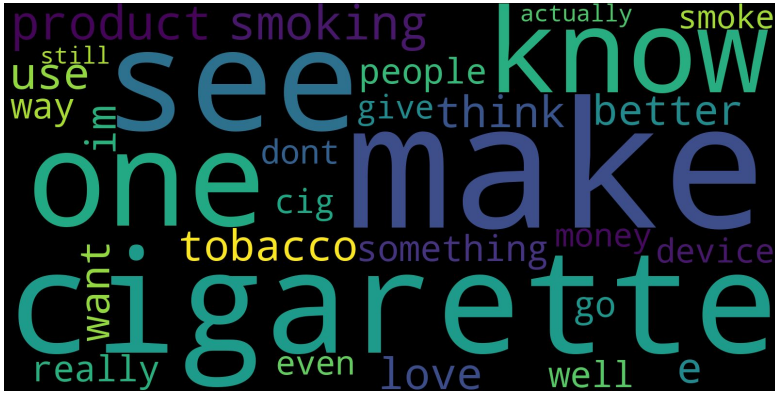
Iterative Development

- Subreddits
  - Began with subreddits focused on quitting smoking and quitting vaping
  - Results were skewed as topics were very specific and didn't represent the whole electronic cigarette community
- Preprocessing the posts
  - All comments on one line vs one comment on one line
- Vectorizers
  - Bag of Words vs TF IDF

What I Found

- Flavored vapes
  - “I gotta be serious when I say that their mango juice is something special. Such a good flavor”
- Financial inconsistency between users
  - “Blew a whole lot of money up front looking for my perfect setup”
  - “Buy the vape because it’s usually substantially cheaper and it looks like juice will be the same way”
- Healthier alternative to smoking
  - “I saved my lungs from 1000s of cigarettes and feel much healthier each morning”
  - “For me, nothing worked for over two decades until I started using e cigarettes. This quite likely saved my life”





What I Learned

- Each comment or post should be its own data point
- The chosen source of the data can be biased
  - Source selection depends on the topic that I am researching
  - If I want to persuade somebody to stop smoking, I should select sources that reflect the pros of my argument
- Preprocessing the data well will improve the reliability of the analysis
- I should have continued to iterate the data mining process to gather more insights, but was restricted by time