Natural Language Processing

reddit

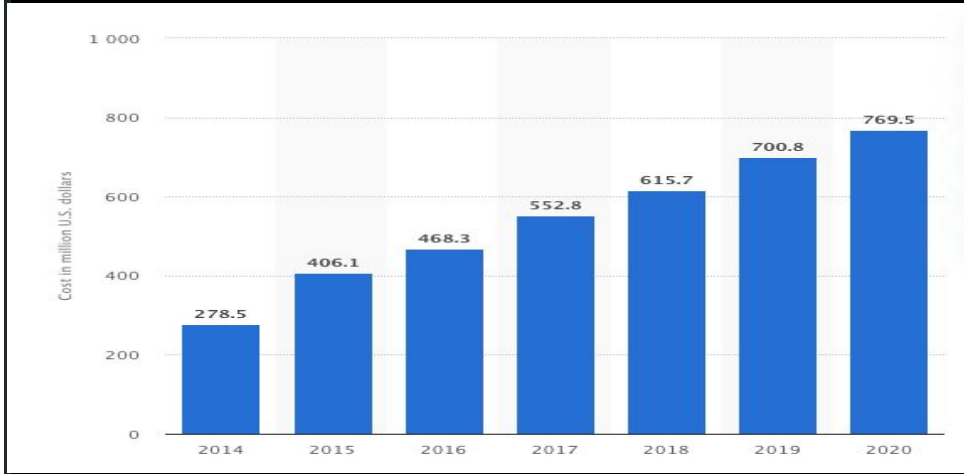
James Opacich
Data Scientist

Problem Statement

Can Machine Learning and NLP distinguish between Homebrewers and Winemakers?

How will it help your marketing objectives?

Constellation Total Ad Spend Per Year



Problem

SOLUTION

- Determine Marketing Cues and Insights
- Inform better marketing and advertising decisions

Metrics And Objectives

- **Balanced Accuracy Score**
  - Accounts for **correct positive** label identification.
  - Also accounts for **correct negative** label identification.

- **Actionable Takeaways**
  - Key words
  - Interesting Findings

The Corpus

2 Subreddits

r/Winemaking

- 1935 **posts.**
- 74,984 **words**
- 1264 **unique authors**

r/Homebrewing

- 2367 **posts**
- 107, 486 **words**
- 1914 **unique authors**

What's The Point?

Data Collection → Data Cleaning → Preprocessing / Exploratory Data Analysis → Marketing Team Utilization

Constellation Brands

Why Even Do The Modeling?

scikit learn

MODEL

NLP Modeling Process

Data Collection [Pushshift]

Data Cleaning

pandas

NLP

python NLTK

seaborn

Initial Modeling

spaCy Preprocessing matplotlib

Exploratory Data Analysis

python

Fine Tuning

scikit learn

Production Model

Voting Classifier

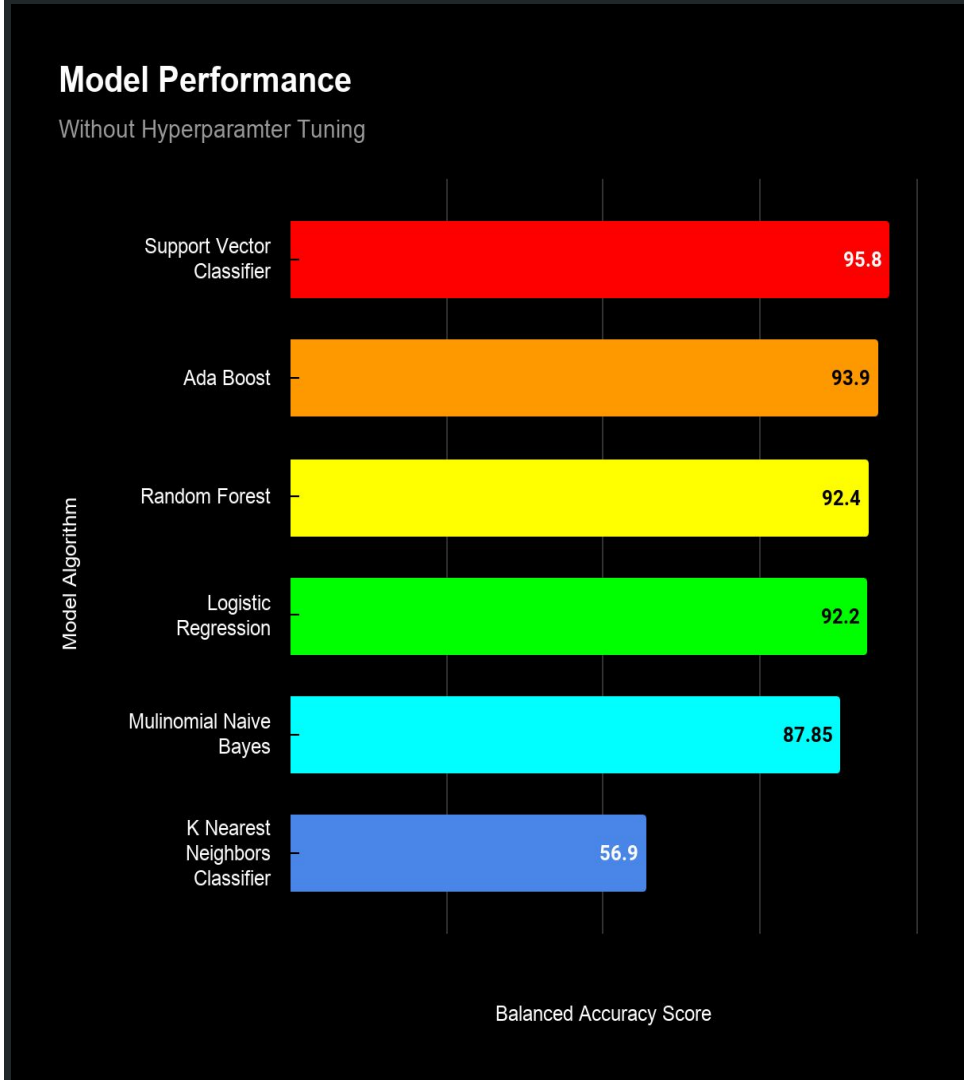
BALANCED ACCURACY SCORE

91.8

Constellation Brands

Marketing Team Utilization

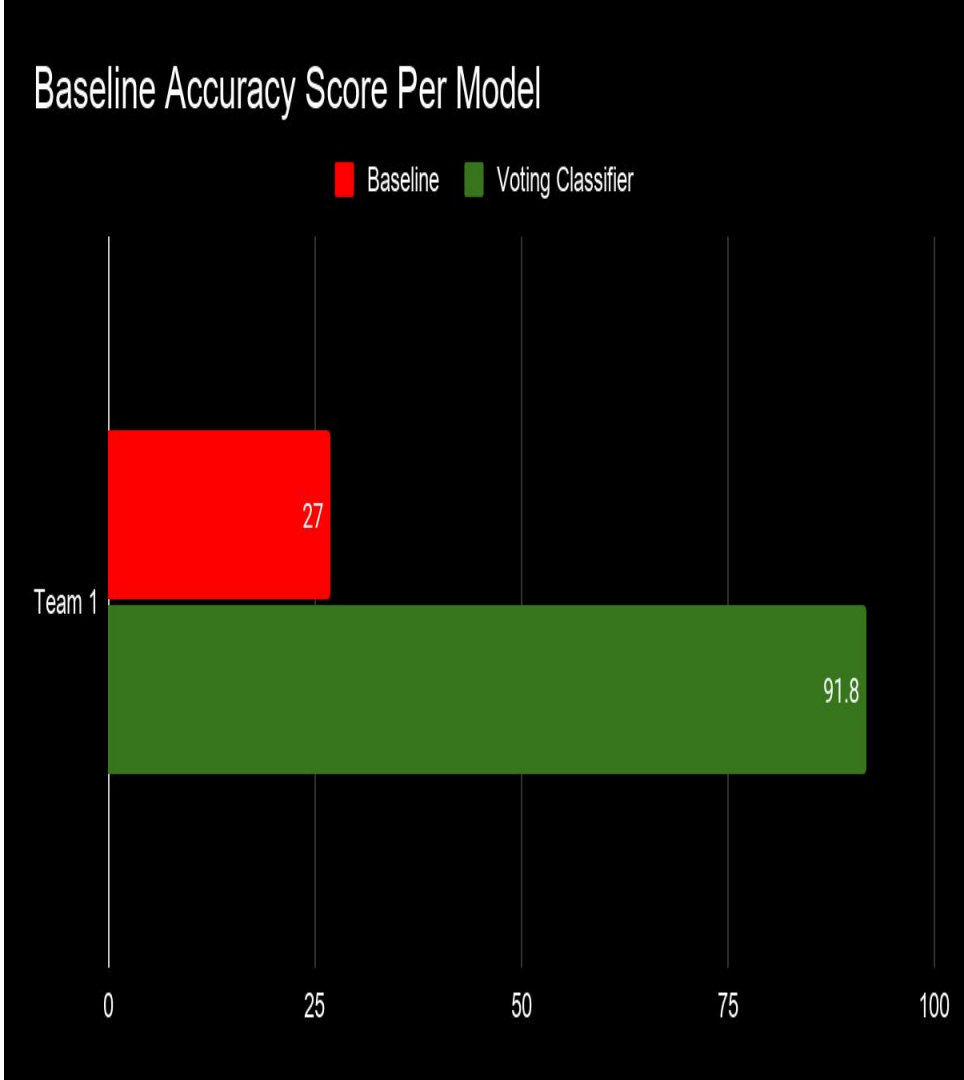
Model Performance



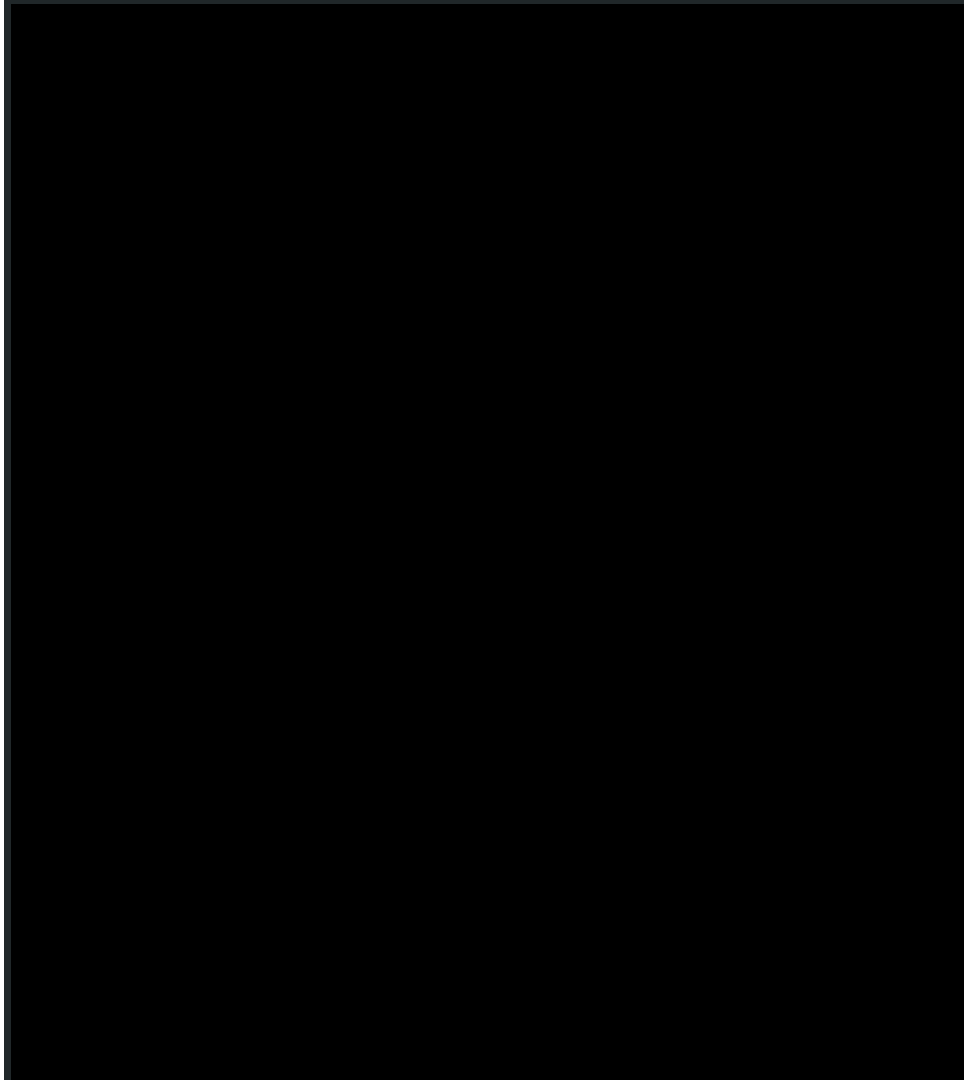
Model Performance

Baseline
Vs.
Production Model

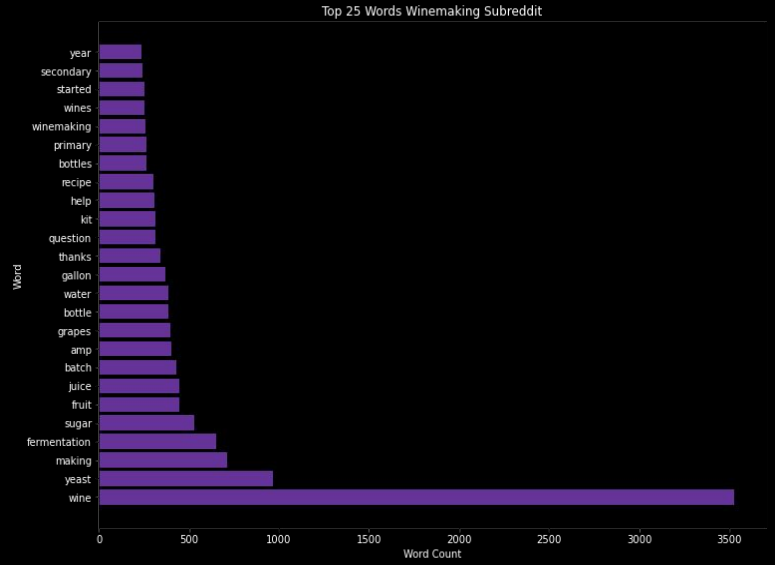
Voting Classifier had Logistic Regression, Random Forest, Ada Boost and SVC in the estimators hyperparameter

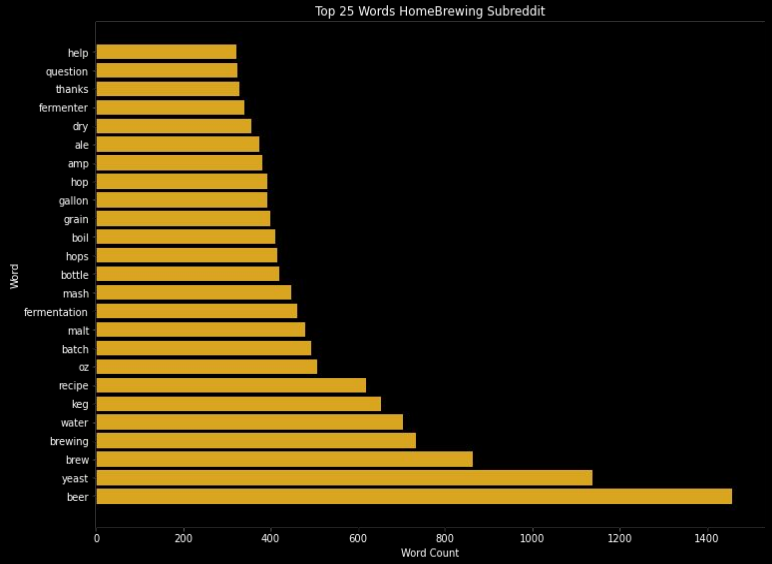


Brief Insights

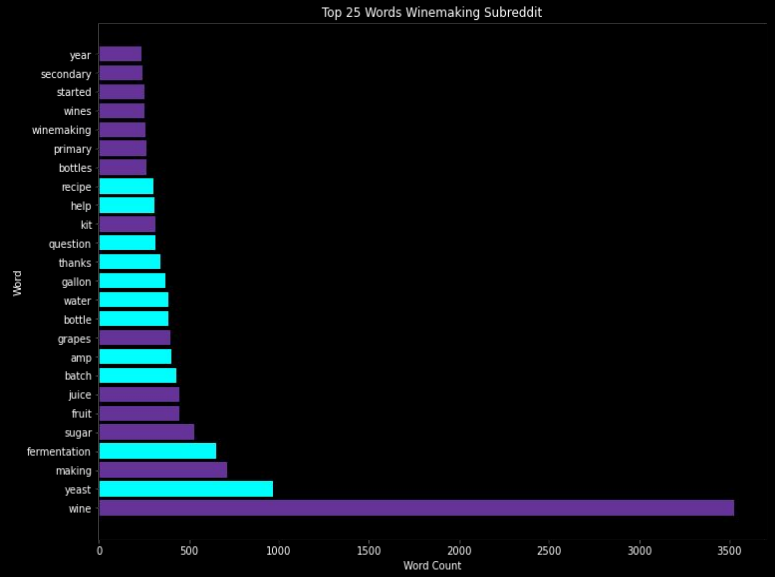


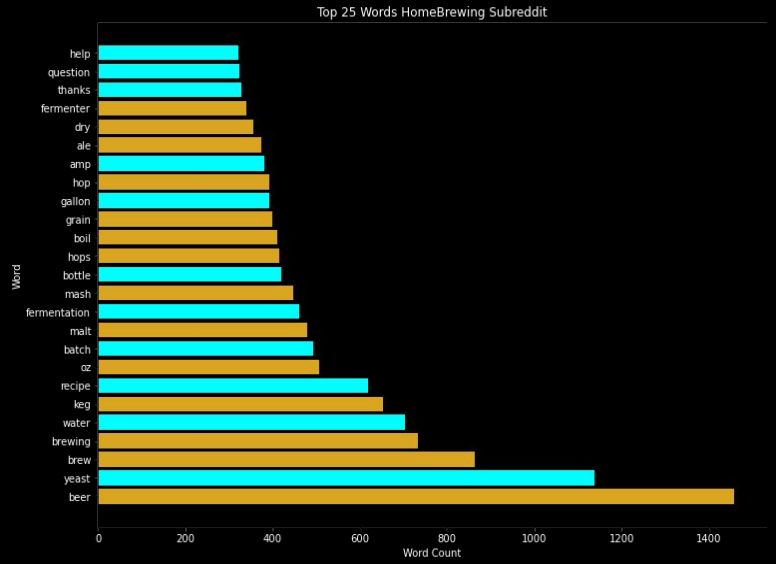
Top 25 Words


Top 25 Words Winemaking Subreddit
Word
year
secondary
started
wines
winemaking
primary
bottles
recipe
help
kit
question
thanks
gallon
water
bottle
grapes
amp
batch
juice
fruit
sugar
fermentation
making
yeast
wine
0
500
1000
1500
2000
2500
3000
3500
Word Count


Top 25 Words HomeBrewing Subreddit
Word
help
question
thanks
fermenter
dry
ale
amp
hop
gallon
grain
boil
hops
bottle
mash
fermentation
malt
batch
oz
recipe
keg
water
brewing
brew
yeast
beer
0
200
400
600
800
1000
1200
1400
Word Count

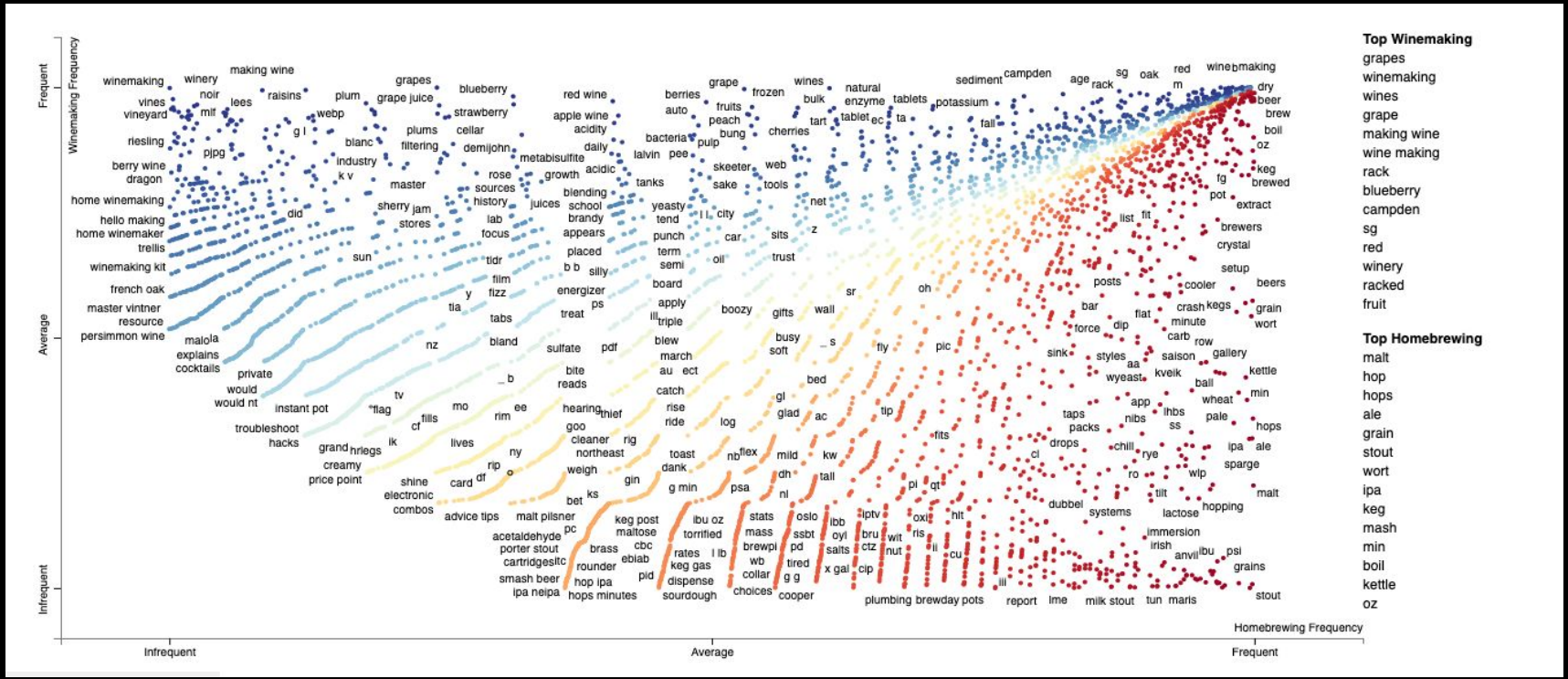
Top 25 Words



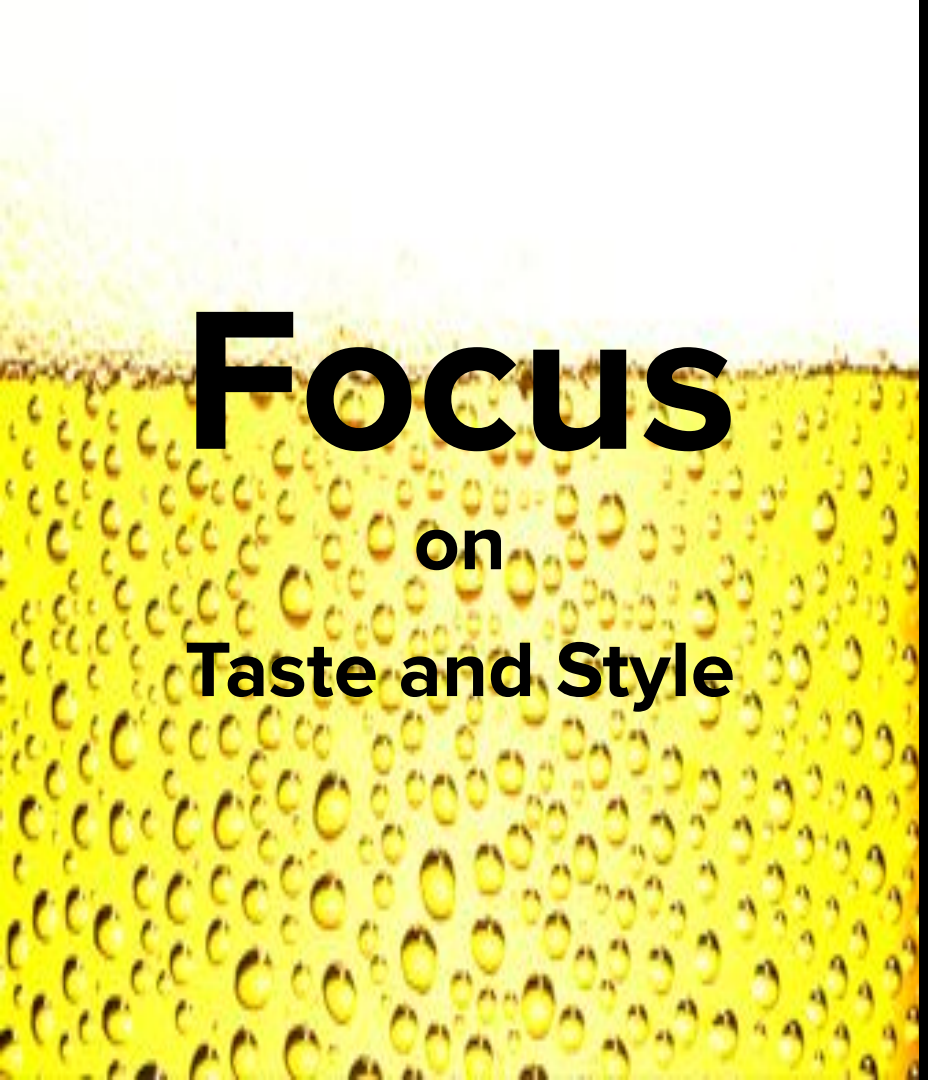


The light blue bars are words shared by both corpus’

spaCy Scattertext



Visualization of Balanced F-Scores and Word Frequency Divided by Subreddit



Focus

on

Taste and Style

‘Taste’
74 per 25,000 terms
79 per 1,000 docs
218 posts

‘Taste’
47 per 25,000 terms
68 per 1,000 docs
196 posts

‘dry’
43 per 25,000 terms
53 per 1,000 docs
127 posts

‘dry’
85 per 25,000 terms
93 per 1,000 docs
356 posts

Focus

Inputs and Ingredients

‘Oak’
40 per 25,000 terms
32 per 1,000 docs
118 posts

‘Keg’
107 per 25,000 terms
95 per 1,000 docs
448 posts

‘Grapes’
136 per 25,000 terms
113 per 1,000 docs
398 posts

‘Malt’
115 per 25,000 terms
98 per 1,000 docs
479 posts

Focus on The Process

‘Ferment’
61 per 25,000 terms
71 per 1,000 docs
178 posts

‘Ferment’
46 per 25,000 terms
65 per 1,000 docs
191 posts

‘Secondary’
82 per 25,000 terms
86 per 1,000 docs
240 posts

‘Secondary’
44 per 25,000 terms
51 per 1,000 docs
183 posts

Conclusion

NLP Can ...

1. Identify Distinguishing Characteristics

Which Will Help You...

2. Streamline and compound marketing efforts

Take Aways from Reddit NLP Results

- Focus on Keywords to gain insights
- See relationships with visualizations
- Compound on your creativity

What is next?

- Try more Spacy Visualizations
- Attempt to incorporate more powerful modeling techniques and deploy useful tools like lead generation funnels, recommendations engines and forecasting models based on sentiment analysis.
-
- Web Deployment of Scattertext model via Streamlit or Flask with Heroku
- Analyze model coefficients (word importance)
- Entity visualization to find mentions of key figures and places in the data
- Incorporating Corpora across several social media sources.
- Post comment relationships
- Post score relationships
- Sentiment Analysis