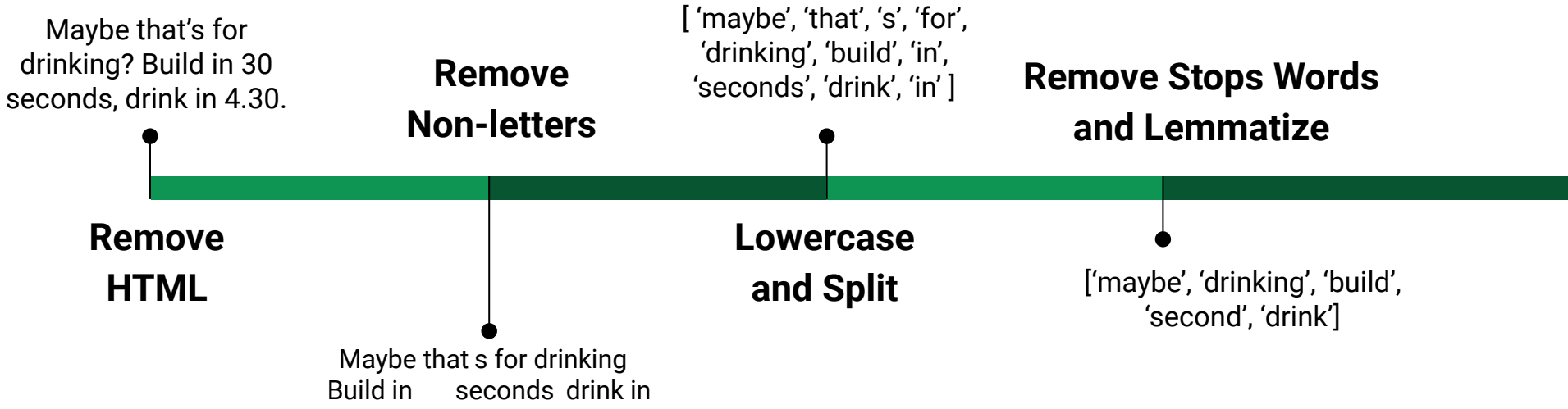
Wine? or Cocktail?

Jobeth Muncy

Collecting the Data

- Where is the data? https://api.pushshift.io/reddit/search/comment
- Get 50K Comments from the Subreddit Wine and Subreddit Cocktails
- Clean the Text Data
- **Can the model correctly classify comment as wine or cocktail?**

<html><body> Maybe that’s for drinking? Build in 30 seconds, drink in 4.30. </body></html>

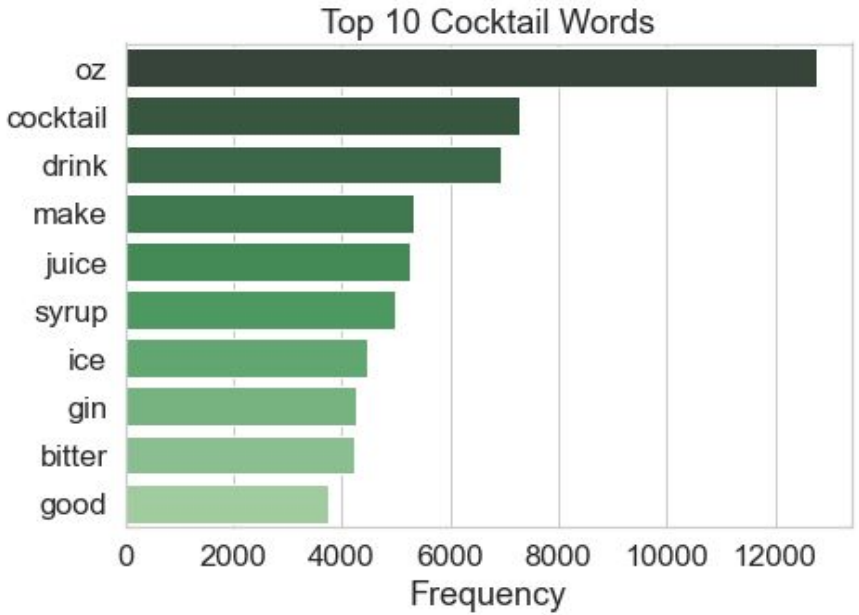


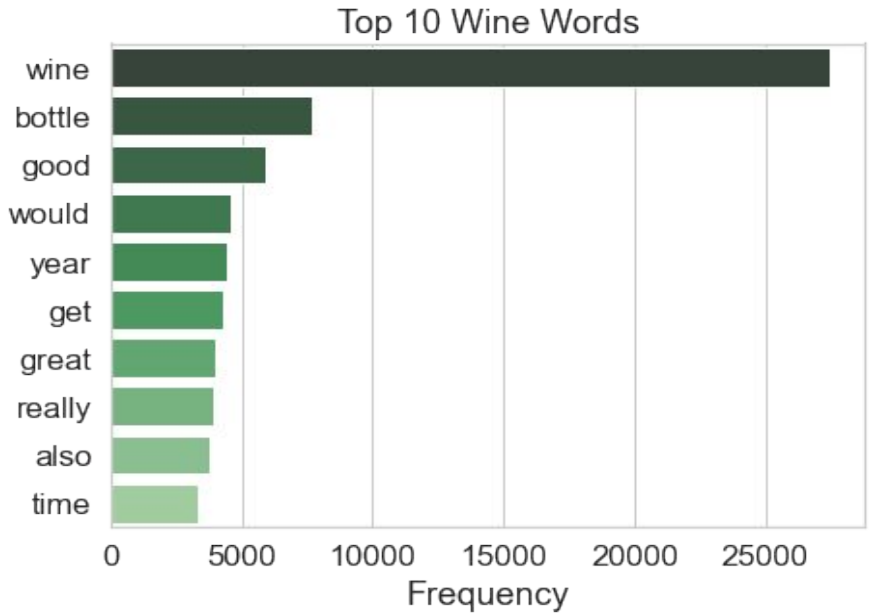
What Changed?

- ➔ HTML
- ➔ Non-letters
- ➔ Stopwords
- ➔ Lemmatized Words

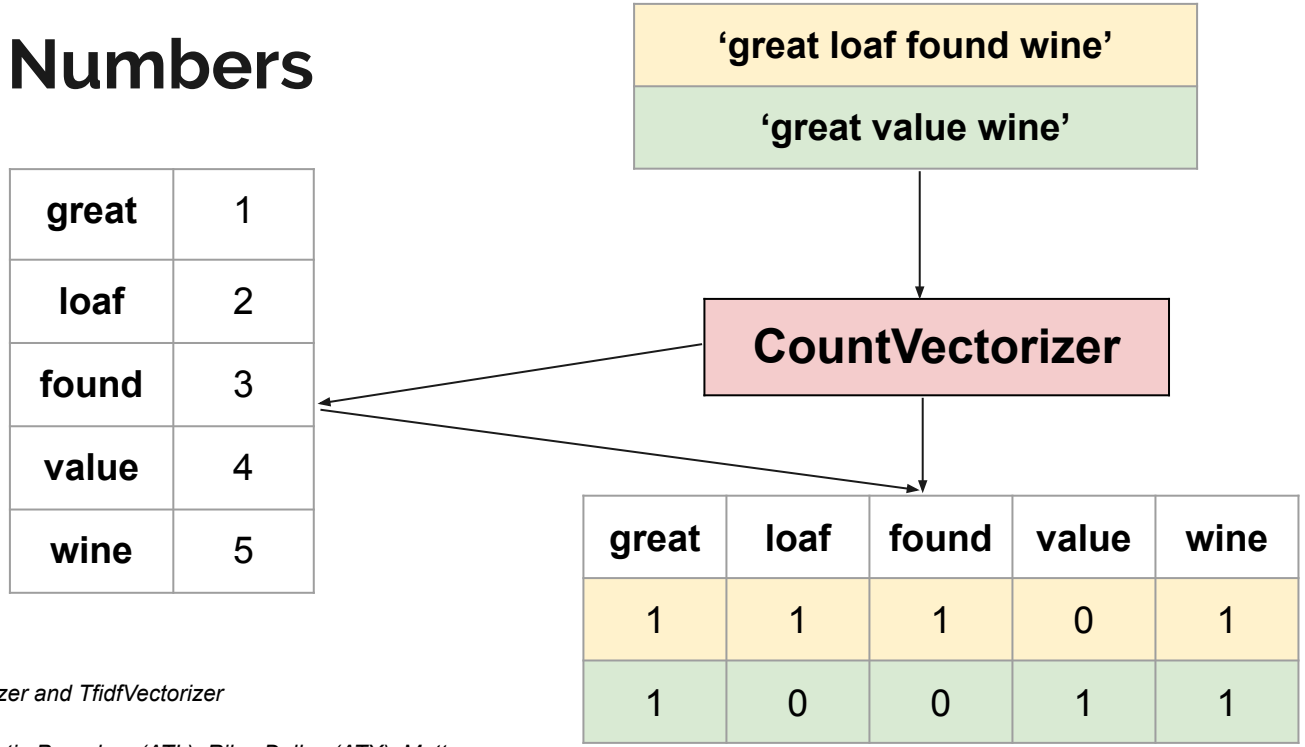
	Original Word Count	Clean Word Count	Percent Removed	Missing Values
Cocktails	1,460,871	818,864	43%	309
Wine	1,718,265	906,371	47%	357

What are the Most Common Words?





Text to Numbers



Adapted from NLPII: CountVectorizer and TfidfVectorizer

Authors: Dave Yerrington (SF), Justin Pounders (ATL), Riley Dallas (ATX), Matt Brems (DC)

Best Model

CountVectorizer

Naive Bayes Classifier

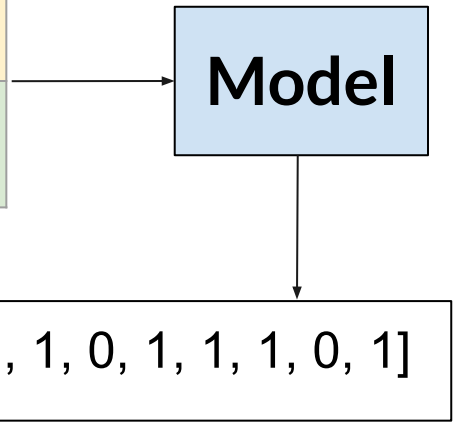
The model predicts what subreddit each comment belongs to with 86% accuracy

1. Asked 3 friends for a few comments on wine or cocktails
2. Ran their comments through the process
3. Checked the model predictions

Clean the text:

1. ‘No, it’s not a snow globe, that’s a Pet Nat. Hello sediment!’
2. ‘Who says you can’t drink your artichokes? Say hello to Cynar.’

1. ‘snow globe pet nat hello sediment’
2. ‘say drink artichoke say hello cynar’



Wine: 1
Cocktail: 0

Text	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Predict	0	0	1	0	0	1	0	1	1	0	1	1	1	0	1
Actual	0	0	1	0	0	1	0	1	0	0	1	1	1	0	1

Summary

1. Collected data from Subreddit Wine and Subreddit Cocktail using API
2. Cleaned the text data
3. Changed text data to numerical data
4. Found the best model at 87% accuracy
5. Predicted scores from 15 new comments with 93% accuracy

https://www.reddit.com/
https://api.pushshift.io/reddit/search/comment
https://pixabay.com/photos/frogs-beverages-bottles-alcohol-1639476/

Contributors:
Sara Thomson
Ali Thorburn
Zoe Nystrom