A Tale of Two Subreddits

Web APIs and NLP

———

By L.A. Randrup

03-11-2021

Let's say (for exploration)

You're an entrepreneur with very limited marketing budget, looking to advertise your new coffee/tea brewer product...

Which subreddit should you advertise?

The two subreddits

r/Coffee

a place to discuss any and all things coffee.

Members: 596K

r/Tea

discussion of beverages made from soaking camellia sinensis leaves (or twigs) in water, and, to a lesser extent, herbal infusions, yerba mate, and other tisanes.

Members: 363K

The problem statement and challenge

What are the similarities and differences between the two subreddits?
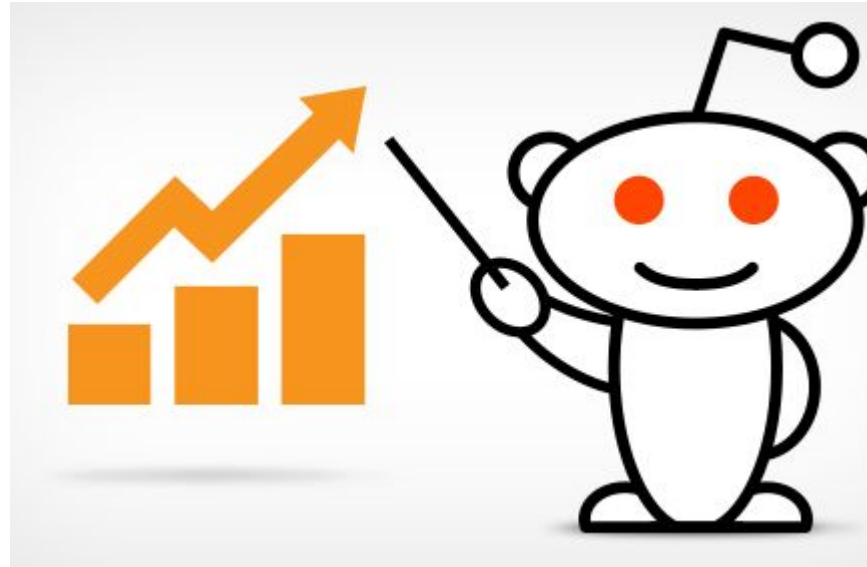
If given a word or set of words, using machine learning classifiers, predict which subreddit it should be posted on.

This could help you get an insight for writing your ad copy!

Research

- Understanding your Target Customers
- Gathering Data using API
- Data Analysis and Exploration
- Exploring the subreddit pages

Goal is to get audience insights

for both subreddits!

Observation

Posted by u/blakjack9 8 hours ago

32

It it fine to store beans in a glass container if

So I got super lucky, and a local coffee shop opened up
house, and they sell in house roasted beans. I buy 2 ba
since some days they don't have beans to sell, so I get i
air tight glass mason jar, since I don't like those paper li
Is that ok? The glass has a slight tint to it now, is that ju
make sure I'm not waisting good beans. Also those 2 ba
weeks, so I was also wondering if I go through it quick e
different in quality.

💬 **16 Comments**   ↱ **Share**   🔖 **Save**   ⋯

Posted by u/jazzb54 8 hours ago

31

Watermelon?

I let my coffee get a bit cold while working, and while I
cools - this is the first time I tasted watermelon. It's one
subscription coffees, and the card says I should taste "c

Has anybody else ever tasted watermelon as a primary
palate broken this morning?

**GUYS LOOK HOW COOL (honey jar was from a
found at value village)**

💬 **1 Comment**   ↱ **Share**   🔖 **Save**   ⋯

Posted by u/Adventurous_Matter_3 33 minutes ago

Vote   The tea wall is coming along (⭒ᵕ⭒)   Photo

Findings

From 3,500 posts...

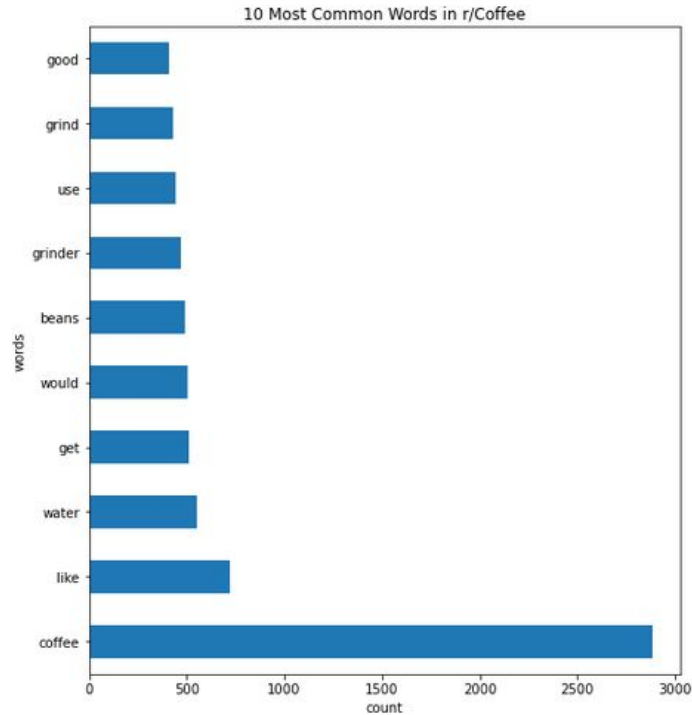|	r/Coffee	r/Tea
Unique Users	2026	1287
Max Post of a User	60	10
Min Post of a User	1	1
Average Post per User	1.149	1.26
Average Word Counts per Post	115.54	84.43

Diving into texts…

Understanding the audience and culture of the page through text.

Using NLP preprocessing methods, we can find more insights on the two datasets.

- Removal of special characters

- Tokenize

- Lemmatize

- Stop Word Removal

Top Words in r/Coffee

10 Most Common Words in r/Coffee

As expected, most of the words are coffee related.

COFFEE
LIKE
WATER
BEANS
GRINDER
USE
ESPRESSO
TIME
CUP
BREW

Top Words in r/Tea
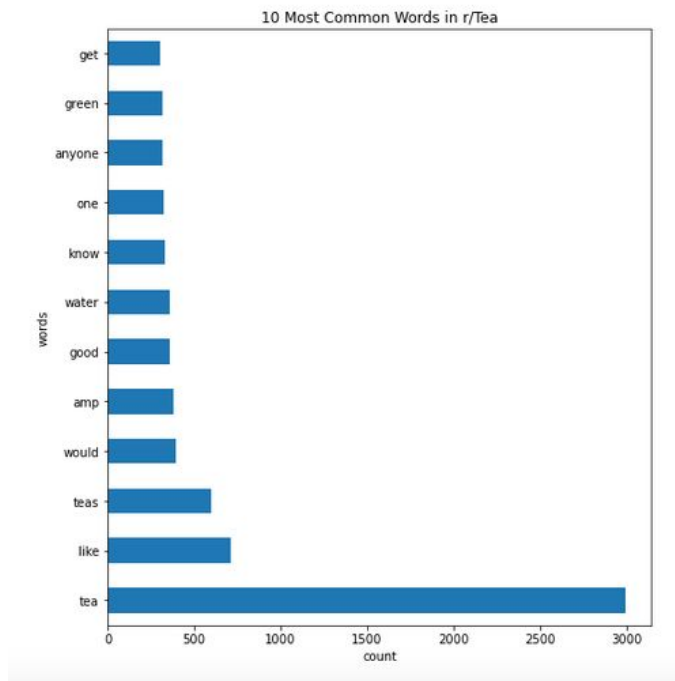
Same with tea...

TEA
WATER
LOOSE-LEAF
TIME
TASTE
TRIED
LOVE
DRINK
USE
THANKS

Common Words in both subreddits

After removing "Coffee", "Beans, "Tea", and "Loose-leaf" words for comparison.

r/Coffee	r/Tea	Both subredddits
water	water	water
beans	loose-leaf	use
grinder	time	time
use	taste	cup
espresso	tried	grinder
time	love	make
cup	drink	brew
machine	use	want
brew	find	looking
make	thanks	taste
want	make	tried
new	different	thanks
pour	even	first
taste	got	new
thanks	brew	got

AhA!

As a brewing device entrepreneur, which subreddit would you advertise?

Both are great market, but there can be only one that prevails…

Aha!
Discoveries

- **r/Coffee members members post more discussions, suggestions, and questions and r/Tea members post more photos.**

- r/Coffee has 206K more members than r/Tea.

- **r/Tea users post emotion words like "love" and "thanks" where r/Coffee users post more on coffee-related words.**

- There are plenty of common words that has similar meaning in both pages including "brew", "looking", and "want".

Conclusion

r/Coffee is a bigger forum with a more discussion-involved audience and r/Tea has a share-a-photo audience. Both pages mainly talks about coffee and tea drinks that includes preparation, suggestions, and stories the difference is r/Tea users uses emotion words on their posts and mainly shares photos while r/Coffee users are posting more discussion posts about coffee.

Based on these findings and keeping all other things constant, I would recommend advertising on r/Coffee.

Machine Learning Models

Model	Training Accuracy Score	Testing Accuracy Score
Random Forrest Classifier	95.01%	94.35%
Extra Trees Classifier	94.12%	92.94%
Multinomial Naive Bayes Classifier	98.38%	95.76%

Prediction Confusion Matrix

This model shows that we can predict which subreddit would words should be posted to, we can use that for insights on your ad copy to match the tone and words of that subreddit.
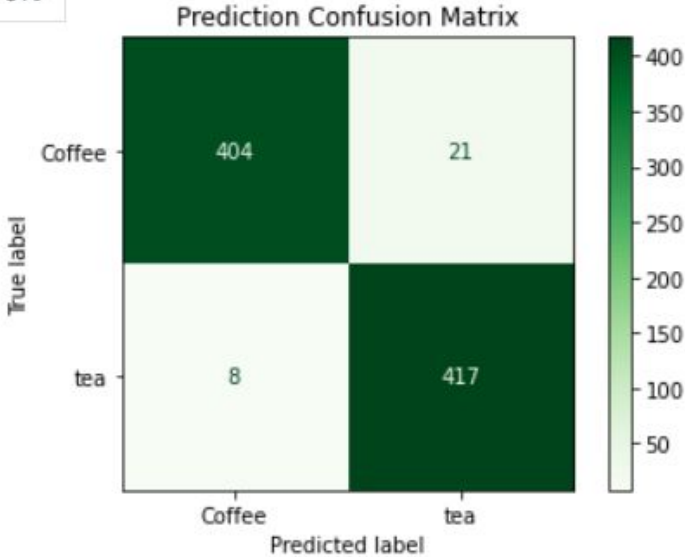
What will I do next?

Looking forward on working with strings (n_grams) to understand context more and setting a more by-the-subreddit-focused stop word dictionary to be able to filter the text better and get more insights.

Thank You!

Questions/Comments?