reddit NLP

DSI-7 : Project 3

Arielle Miro
'Data Scientist'

Data Science Problem

We (reddit) have been experiencing some classification difficulties in our platform. Specifically, subreddits that are similar in nature are being misclassified.

How can we use Data Science to solve this?

➢ Connect to the reddit API
➢ Build a NLP model to predict subreddits based on title and selftext

Subreddit 1 : Life Pro Tips

r/LifeProTips

15.7m
Subscribers

5.6k
Online

Tips that improve your life in one way or another.

Posted by u/TakeMeToFatmandu 56 minutes ago

LPT: If you're calling up a company and you're angry, rather than screaming at the advisor then calmly explain you're angry and why. More often than not, they'll provide you a lot more help and bend over backwards for you. If you scream then you'll just get the minimum

💬 3 Comments   ⭐ Give Award   ➤ Share   🔖 Save   ⋯

Posted by u/humanwithfoodname 4 hours ago

11

LPT: Invest in 2 good 32oz water bottles. Fill them up at home and carry them to work/school. You'll stay hydrated throughout the day and won't have to worry about finding a place to recycle a plastic water bottle

Miscellaneous

💬 2 Comments   ⭐ Give Award   ➤ Share   🔖 Save   ⋯

Posted by u/SpikeyHairedOrphan 1 hour ago

LPT It's almost impossible to truly waste a banana. Food & Drink

If you have bananas that you didn't get to eat before they were more ripe than you prefer you can peel them, put them in a freezer bag, mash them, and use them for baking things like banana cake or banana bread for up to 6 months after you freeze them.

💬 5 Comments   ⭐ Give Award   ➤ Share   🔖 Save   ⋯

Subreddit 2: Explain Like I'm Five

r/explainlikeimfive

16.6m
Subscribers

9.0k
Online

Explain Like I'm Five is the best forum and archive on the internet for layperson-friendly explanations.   Don't Panic!

Posted by u/stormyjsin 18 minutes ago

Chemistry **ELI5: How does cool sculpting work to freeze away fat**

I've been seeing a lot of advertisements for different methods of freezing your fat away. i.e - cool sculpting. Can't be that easy to get rid of fat.. can it?

2 Comments   Give Award   Share   Save   ...

Posted by u/toddypoddy 2 hours ago

5

Biology **ELI5: Why does rubbing your head after hitting it on something make it hurt less?**

I'm a tall guy and sometimes I hit my head on things. Walking in people's basements for example. If I rub my head where it was hit it will usually make it feel better quicker.

1 Comment   Give Award   Share   Save   ...

Posted by u/existentialism91342 5 hours ago

6

Other **ELI5: Why are purchases instantly withdrawn from my account but refunds can take days?**

13 Comments   Give Award   Share   Save   ...
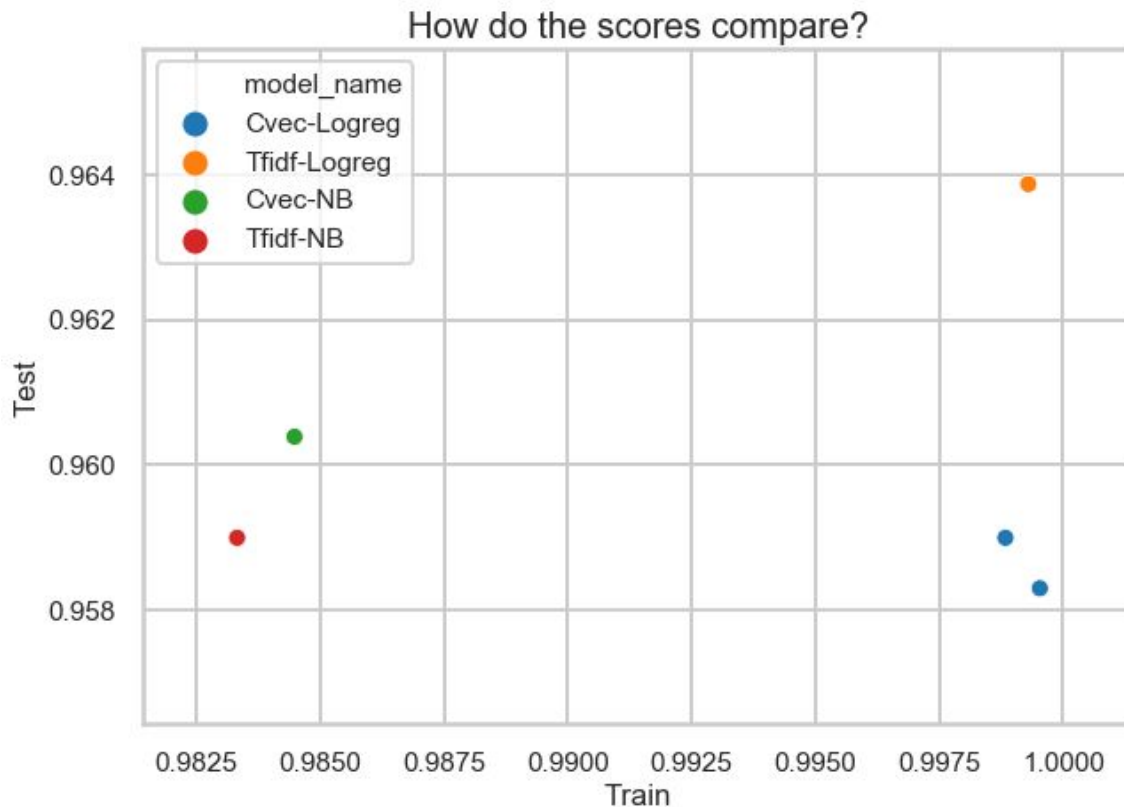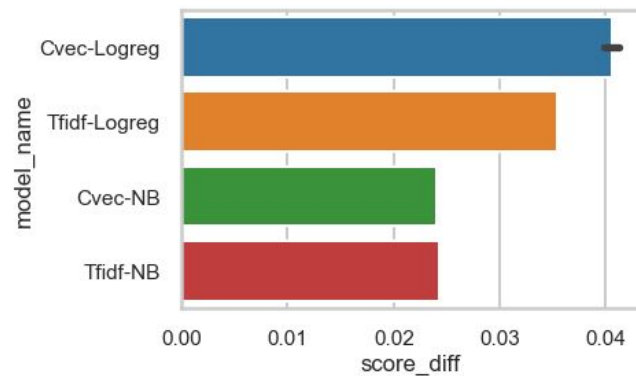
Methodology

PART I

➢ API Connection

➢ Webscraping

PART II

➢ Data Cleaning

PART III

1. Baseline

2. Cvec - Logreg (default params)

3. Cvec - Logreg

4. Tfidf - Logreg

5. Cvec NB MultiNomial

6. Tfidf NB MultiNomial

Model Scores

Final Model

LogReg - TFIDF

➢ *slightly* overfit

➢ "Scaled" with Tf-Idf vectorizer

➢ Tuned Hyperparameters

➢ *Highest* train and test scores

Best Params

➢ logreg__C: 10

➢ logreg__n_jobs: -2

➢ logreg__penalty: l2

➢ tfidf__max_features: 15,000

➢ tfidf__ngram_range: (1, 2)

➢ tfidf__stop_words: None

Confusion Matrix

➢ Accuracy: 96.25 %

➢ Sensitivity: 96.48 %

➢ Specificity: 96.04 %

	Predicted Negative	Predicted Positive
Actual Negative	728	30
Actual Positive	24	657

Misclassified Subreddits

Predicted	Actual	Full Text (title & subtext without keywords)
LPT	ELI5	before traveling to a foreign country learn about its currency notes and how to spot counterfeit bills tourists who are unfamiliar with the local currency can be easy marks
ELI5	LPT	apnea disease 90 suffering from snoring even their partners apnea is becoming a big problem or disease how do you guys stop this comment it will share the solution too
ELI5	LPT	instead of rubbing your eyeballs rub circling around them it does the job and doesn't harm your eyes

Do the to words make a difference?
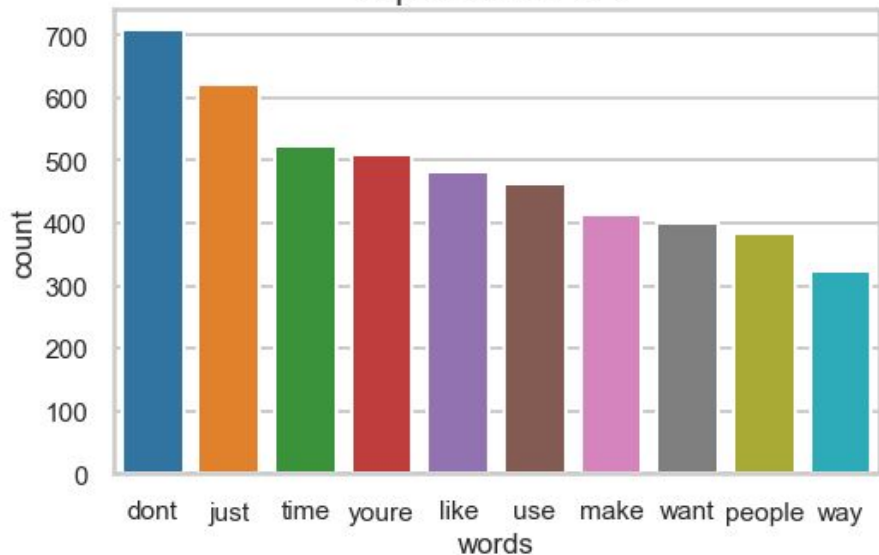
Top words for LPT / Top words for ELI5
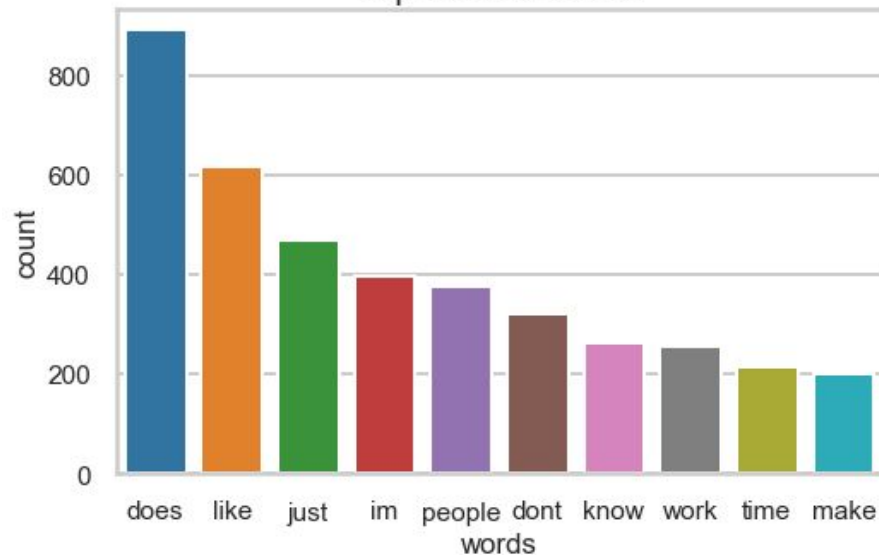
Similar Words: don't, just, time, like, time, make

Next Steps

➢ Try a Random Forests model

➢ Try a Voting Classifier model

➢ Test the current model on smaller and more specific subreddits

➢ Test the current model on other subreddits that don't have too many words in common