Reddit Subreddits Classification

PROJECT 3

GA SG DSI 26

KWEK JUN HONG

Introduction

reddit

- Reddit is a social news aggregation, web content rating, and discussion website founded in 2005.

Ranked 9th most popular social media app in US

430 million monthly active users

Over 100,000 active communities

- Users can upload posts or comments in the subreddits facilitating a discussion. Users can also upvote or downvote the posts as a form of ratings.

Problem Statement

- Employee at a PC building company in Singapore
- Increasing volume of customer enquires for PC building and for after sales tech support
- No database of past customers enquiries
- Objective is to build a classification model to identify incoming enquires
- Will be using posts from two relevant subreddits in place of real customer enquires to train and test the model

Subreddits

- The subreddits identified are r/buildapc and r/techsupport.

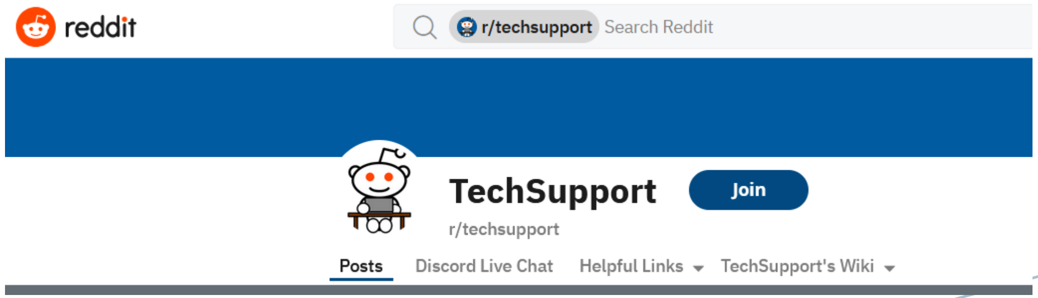
r/buildapc

- 4.8 million members
- Created on 11th Apr 2010
- Community-driven subreddit dedicated to custom PC assembly

r/techsupport

- 1.6 million members
- Created on 10th Jun 2008
- Providing answers or advice to a tech problem

/r/buildapc - Planning on building a computer but need some advice? This is the place to ask!



Methodology

- Web scraping for subreddit posts
- EDA
- Data Cleaning, Lemmatization & Feature Engineering
- Model Preparation
- Model Testing
- Model Results
- Recommendations and Limitations

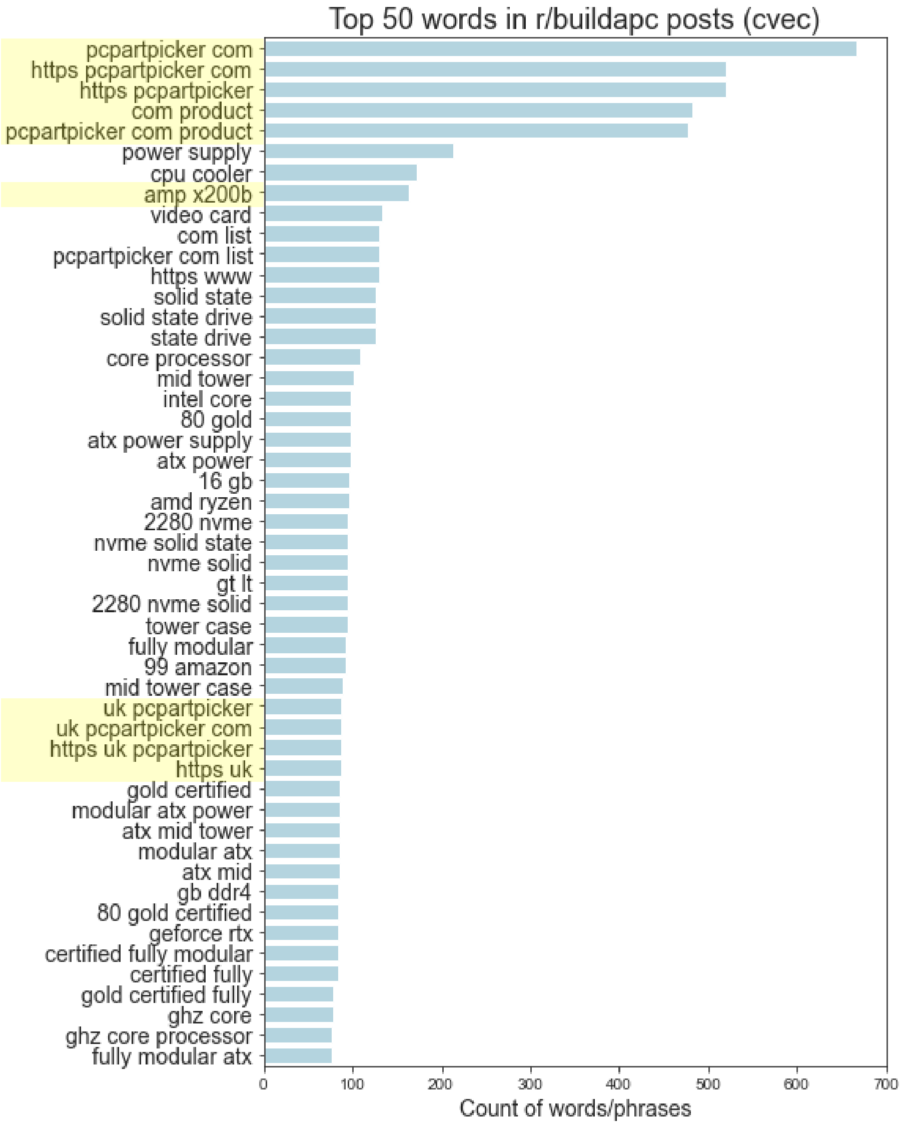
Web scraping for subreddit posts

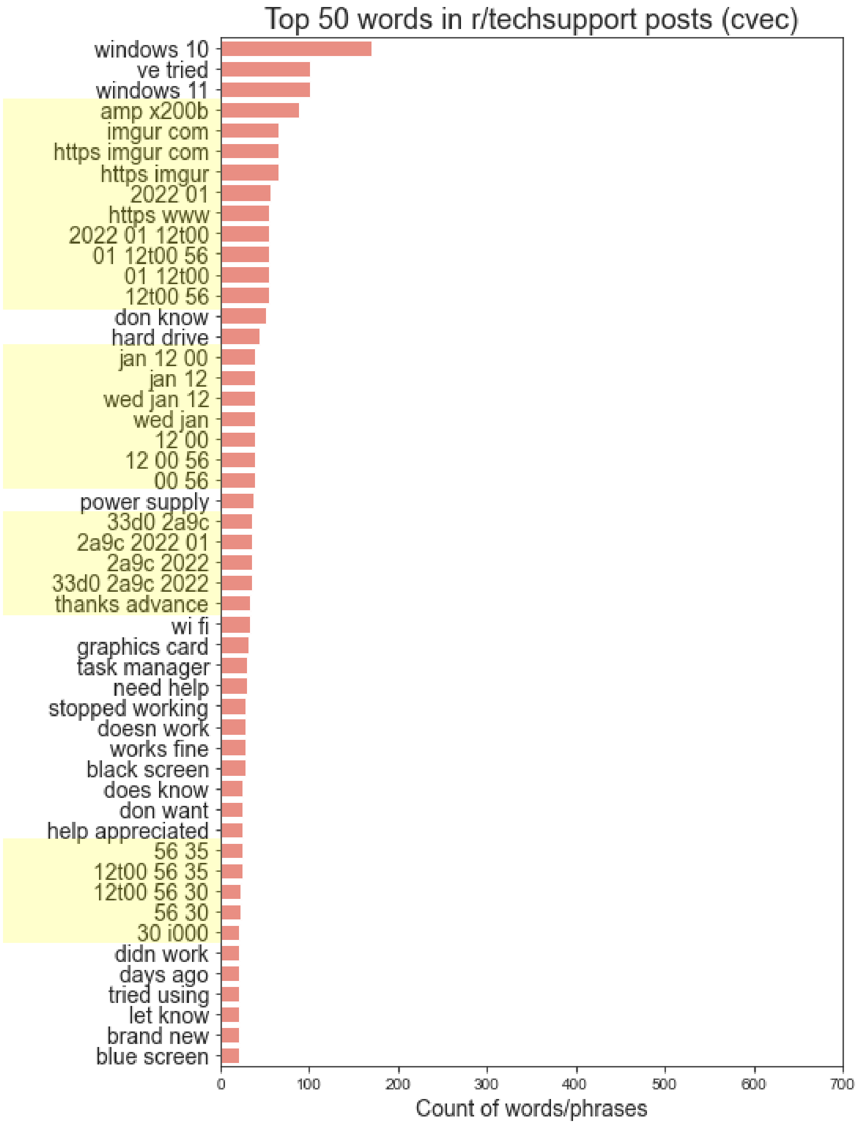
- Using Push API to scrape the posts from the subreddits
- Posts were scraped on 13th January 2022
- Obtain 1000 posts from each subreddits

HTML
HTML Websites
Web Scraping
Data

Preliminary EDA

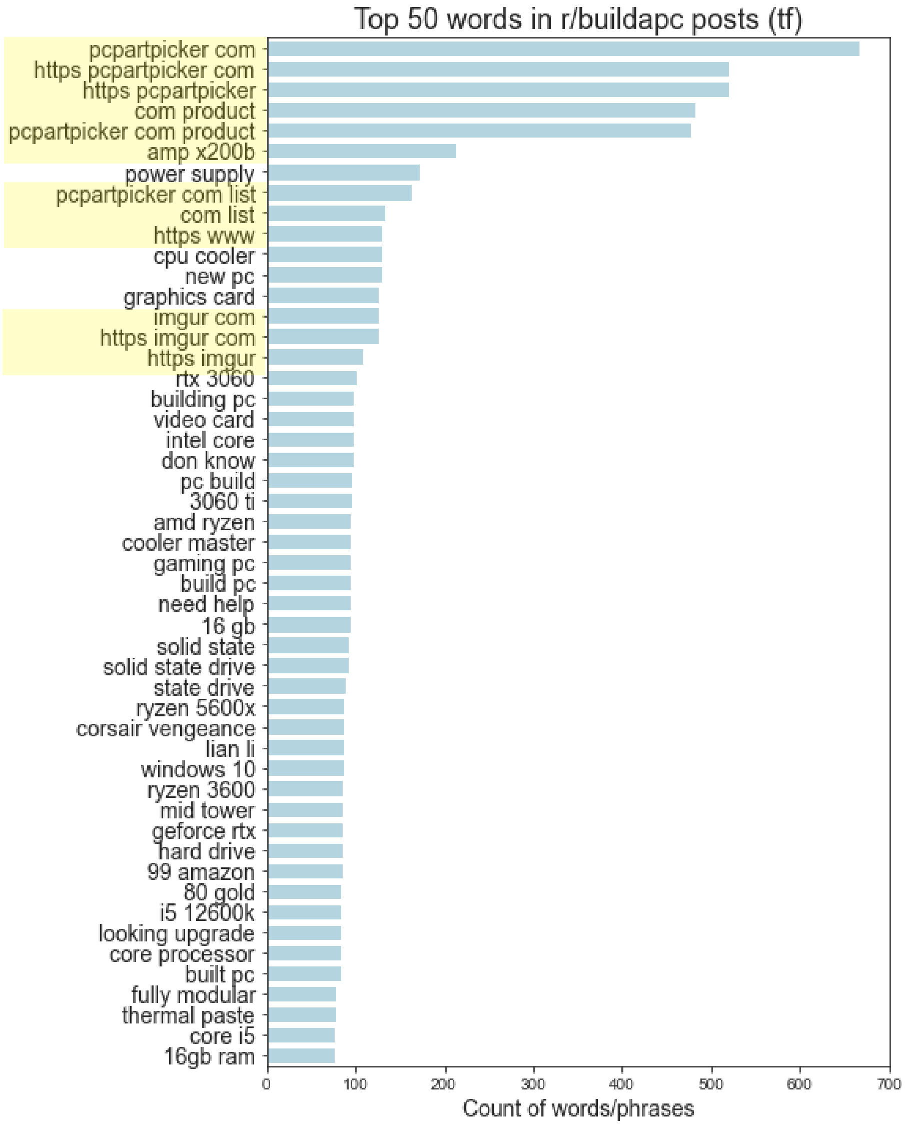
Using CountVectorizer



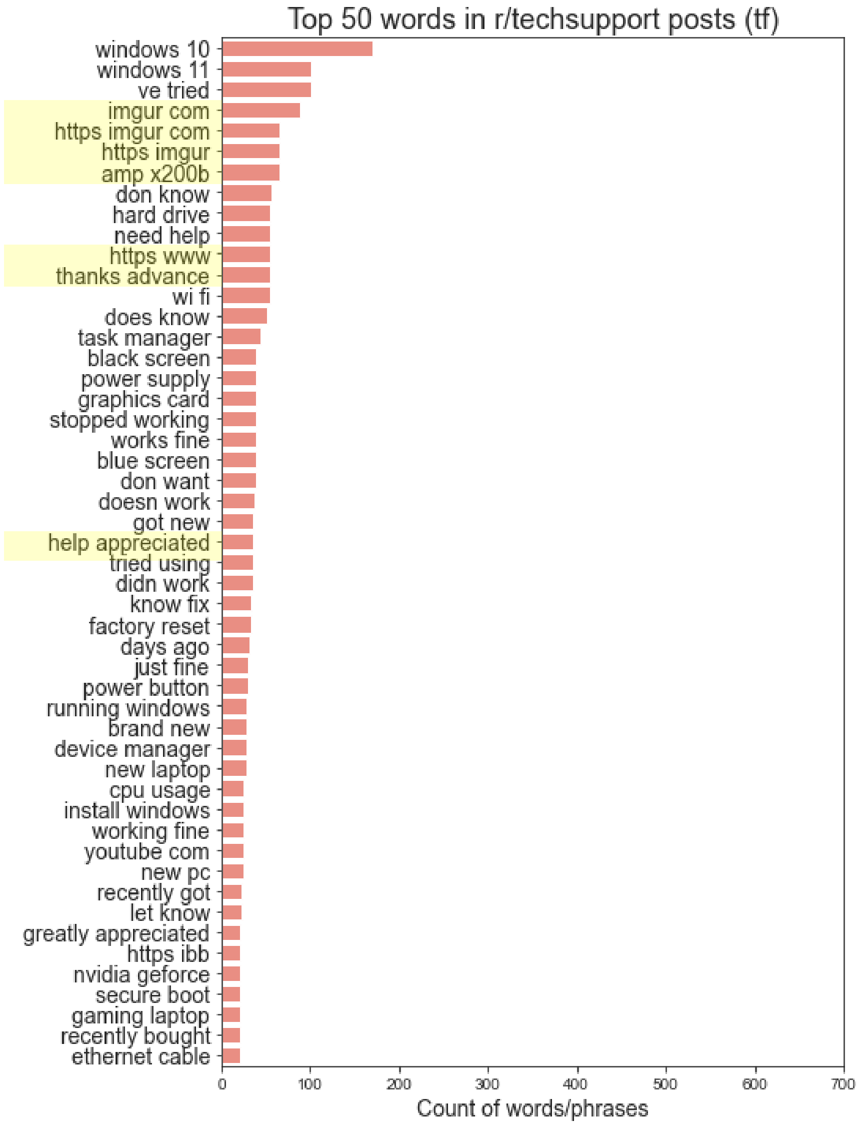


Preliminary EDA

Using TfidfVectorizer



Top 50 words in r/buildapc posts (tf)



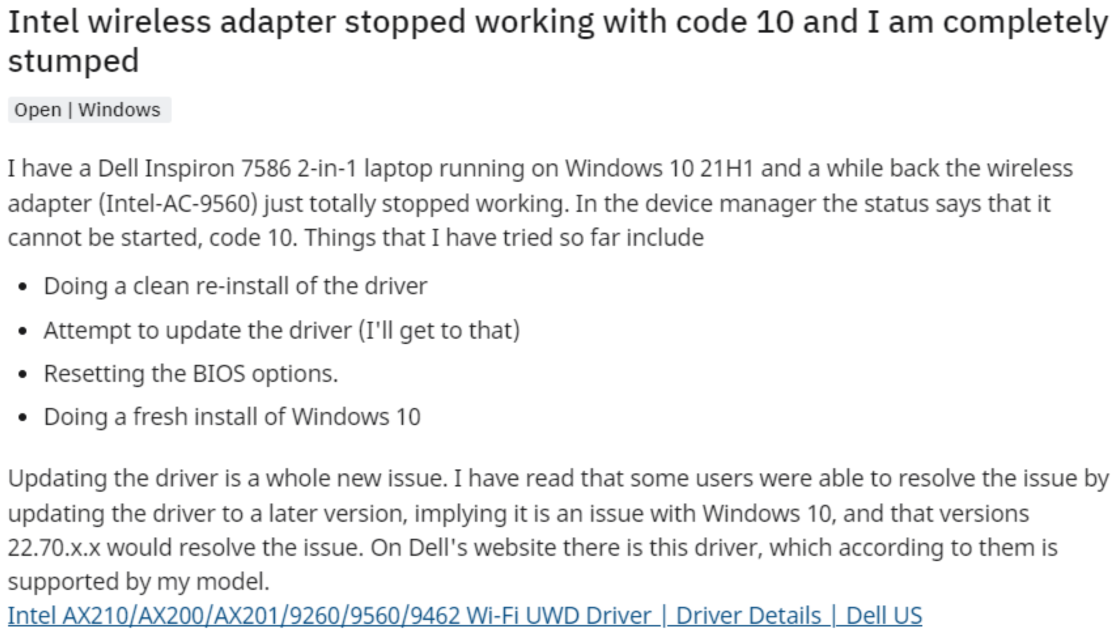
Top 50 words in r/techsupport posts (tf)

Data Cleaning

- Check for null values in the columns that is needed (title and selftext)
    - Null values were not in targeted columns
- Check for duplicates in title and selftext columns
    - r/buildapc
        - Rows that are duplicated in ‘title’ columns (double posts) → Rows dropped
        - Rows with blank field in ‘selftext’ column but with valid titles → Rows kept
        - Rows with [removed] in ‘selftext’ column but with irrelevant titles → Rows dropped
    - r/techsupport
        - Rows that are duplicated in ‘title’ columns (double posts) → Rows dropped
        - Rows with [removed] in ‘selftext’ column but with irrelevant titles → Rows dropped

Data Cleaning

- Remove website links (words like https, www) in both subreddits
- Remove ‘amp x200b’ character from both subreddits
- Remove the date figures in r/techsupport



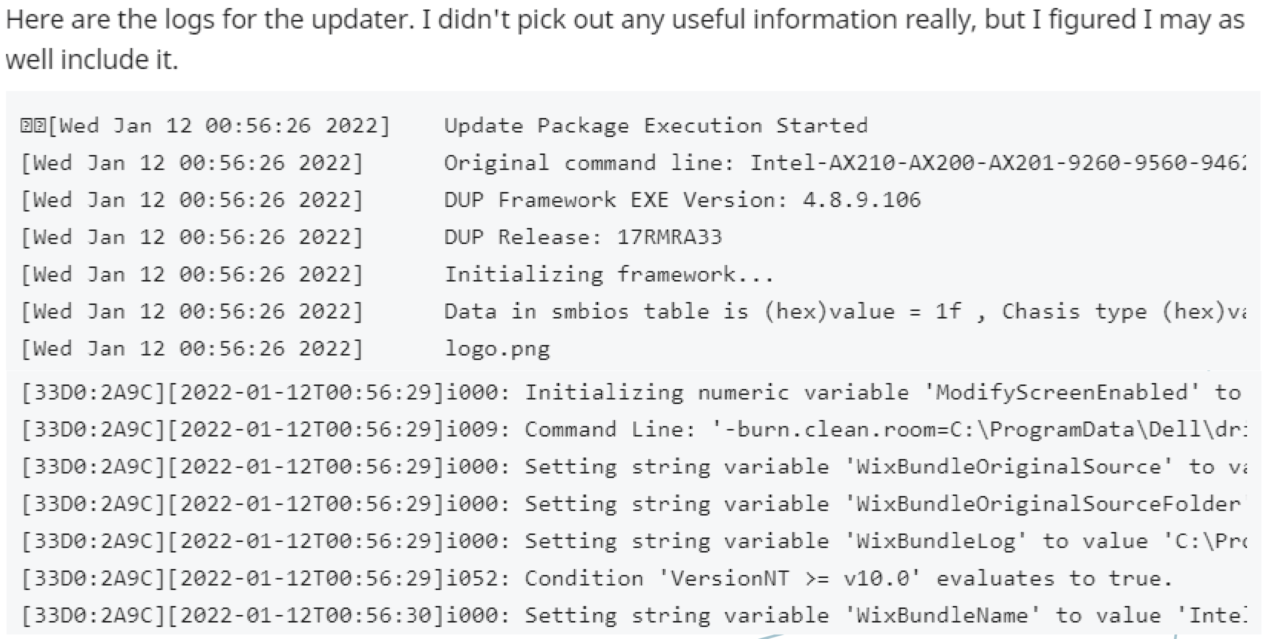
Intel wireless adapter stopped working with code 10 and I am completely stumped

Open | Windows

I have a Dell Inspiron 7586 2-in-1 laptop running on Windows 10 21H1 and a while back the wireless adapter (Intel-AC-9560) just totally stopped working. In the device manager the status says that it cannot be started, code 10. Things that I have tried so far include

- Doing a clean re-install of the driver
- Attempt to update the driver (I'll get to that)
- Resetting the BIOS options.
- Doing a fresh install of Windows 10

Updating the driver is a whole new issue. I have read that some users were able to resolve the issue by updating the driver to a later version, implying it is an issue with Windows 10, and that versions 22.70.x.x would resolve the issue. On Dell's website there is this driver, which according to them is supported by my model.
Intel AX210/AX200/AX201/9260/9560/9462 Wi-Fi UWD Driver | Driver Details | Dell US



Here are the logs for the updater. I didn't pick out any useful information really, but I figured I may as well include it.

```
⊠⊠[Wed Jan 12 00:56:26 2022]    Update Package Execution Started
[Wed Jan 12 00:56:26 2022]      Original command line: Intel-AX210-AX200-AX201-9260-9560-946
[Wed Jan 12 00:56:26 2022]      DUP Framework EXE Version: 4.8.9.106
[Wed Jan 12 00:56:26 2022]      DUP Release: 17RMRA33
[Wed Jan 12 00:56:26 2022]      Initializing framework...
[Wed Jan 12 00:56:26 2022]      Data in smbios table is (hex)value = 1f , Chasis type (hex)v
[Wed Jan 12 00:56:26 2022]      logo.png
[33D0:2A9C][2022-01-12T00:56:29]i000: Initializing numeric variable 'ModifyScreenEnabled' to
[33D0:2A9C][2022-01-12T00:56:29]i009: Command Line: '-burn.clean.room=C:\ProgramData\Dell\dr
[33D0:2A9C][2022-01-12T00:56:29]i000: Setting string variable 'WixBundleOriginalSource' to v
[33D0:2A9C][2022-01-12T00:56:29]i000: Setting string variable 'WixBundleOriginalSourceFolder
[33D0:2A9C][2022-01-12T00:56:29]i000: Setting string variable 'WixBundleLog' to value 'C:\Pr
[33D0:2A9C][2022-01-12T00:56:29]i052: Condition 'VersionNT >= v10.0' evaluates to true.
[33D0:2A9C][2022-01-12T00:56:30]i000: Setting string variable 'WixBundleName' to value 'Inte
```

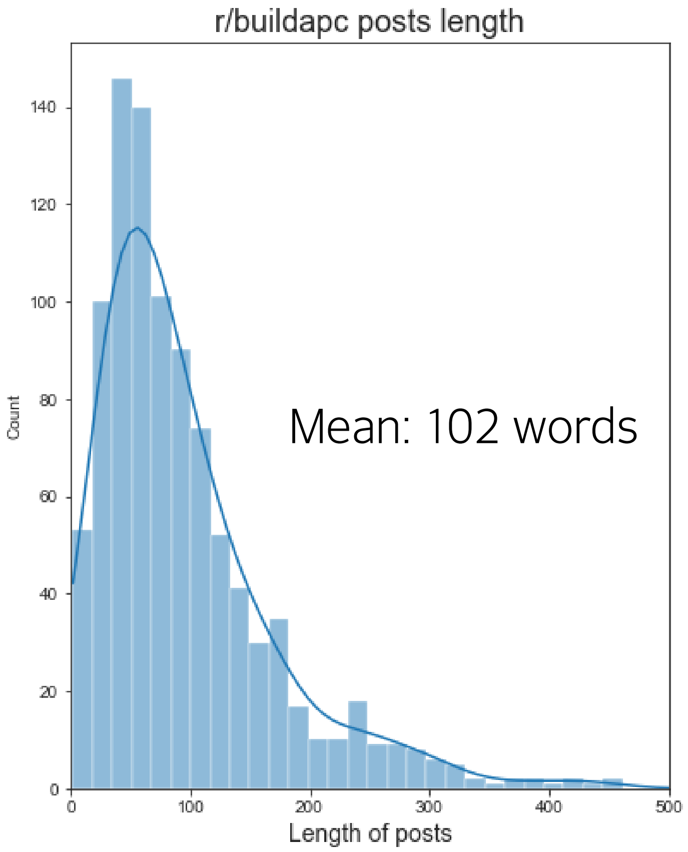
Lemmatization & Stop Words

- Will be used to try to reduce the words into their base form to eliminate these words from the modeling process.
- These words will be then included in the stop words list if needed

- Additional words like *'build', 'building', 'pc', ' computer', 'computers', 'tech', 'technology', 'support', 'supporting', 'thanks', 'thank’, 'appreciated', 'appreciate', 'appreciates’, 'help’* will be added to default English stop words list.

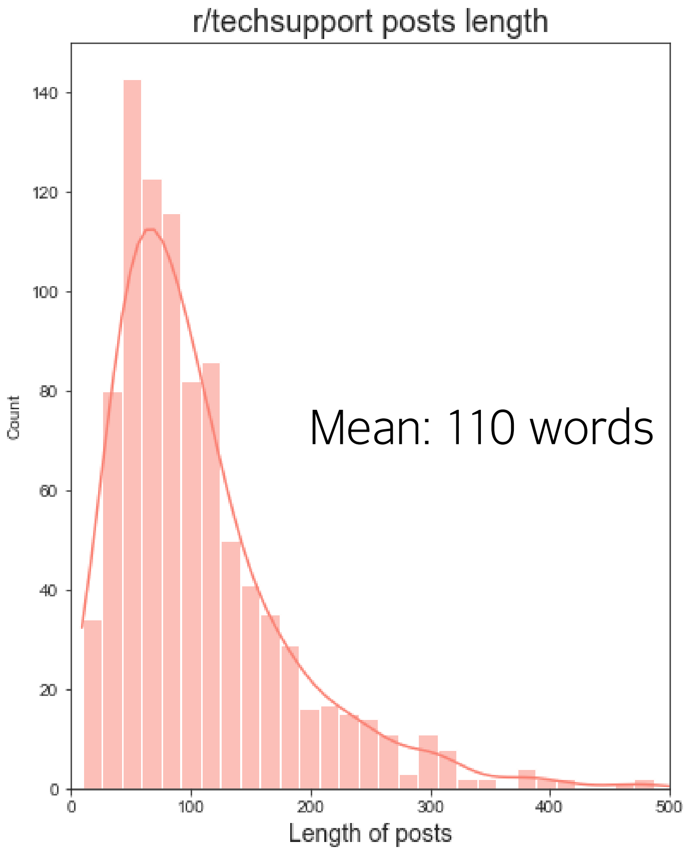
Feature Engineering

- Create a new feature of length of posts for better understanding



r/buildapc posts length

Mean: 102 words

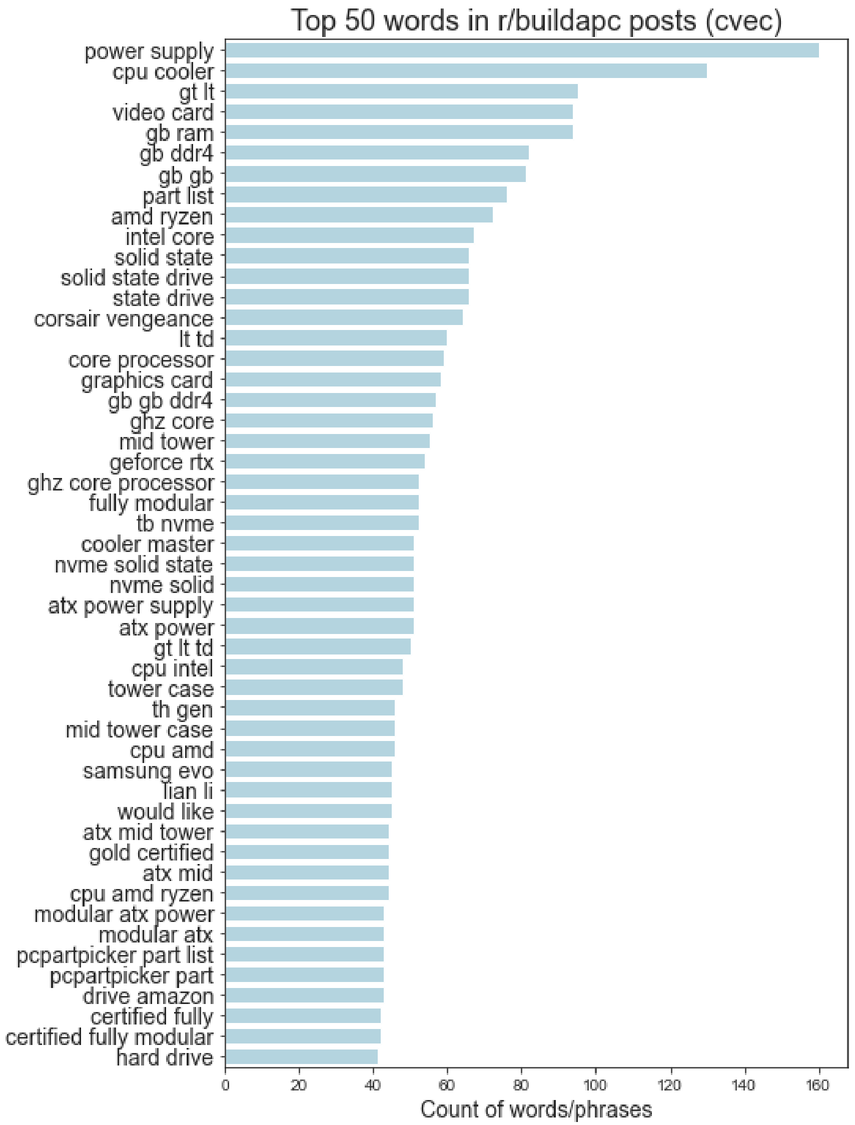


r/techsupport posts length

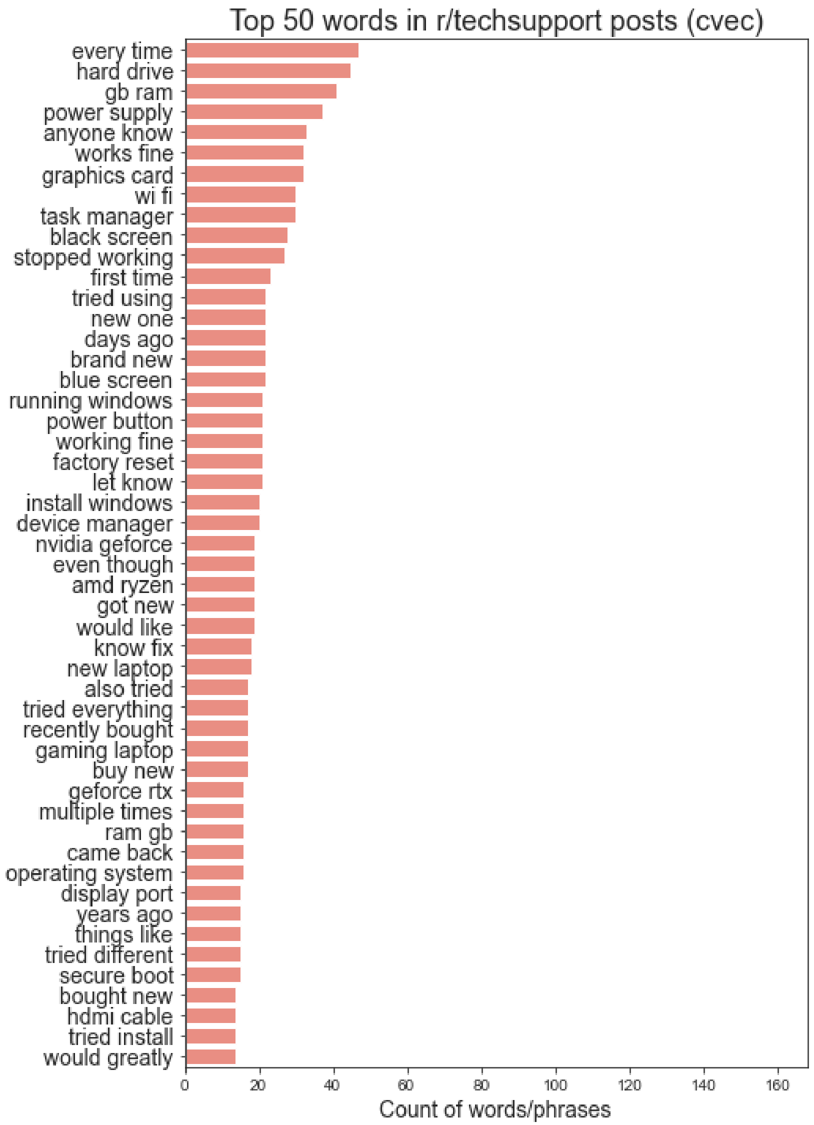
Mean: 110 words

EDA

Using CountVectorizer



Top 50 words in r/buildapc posts (cvec)

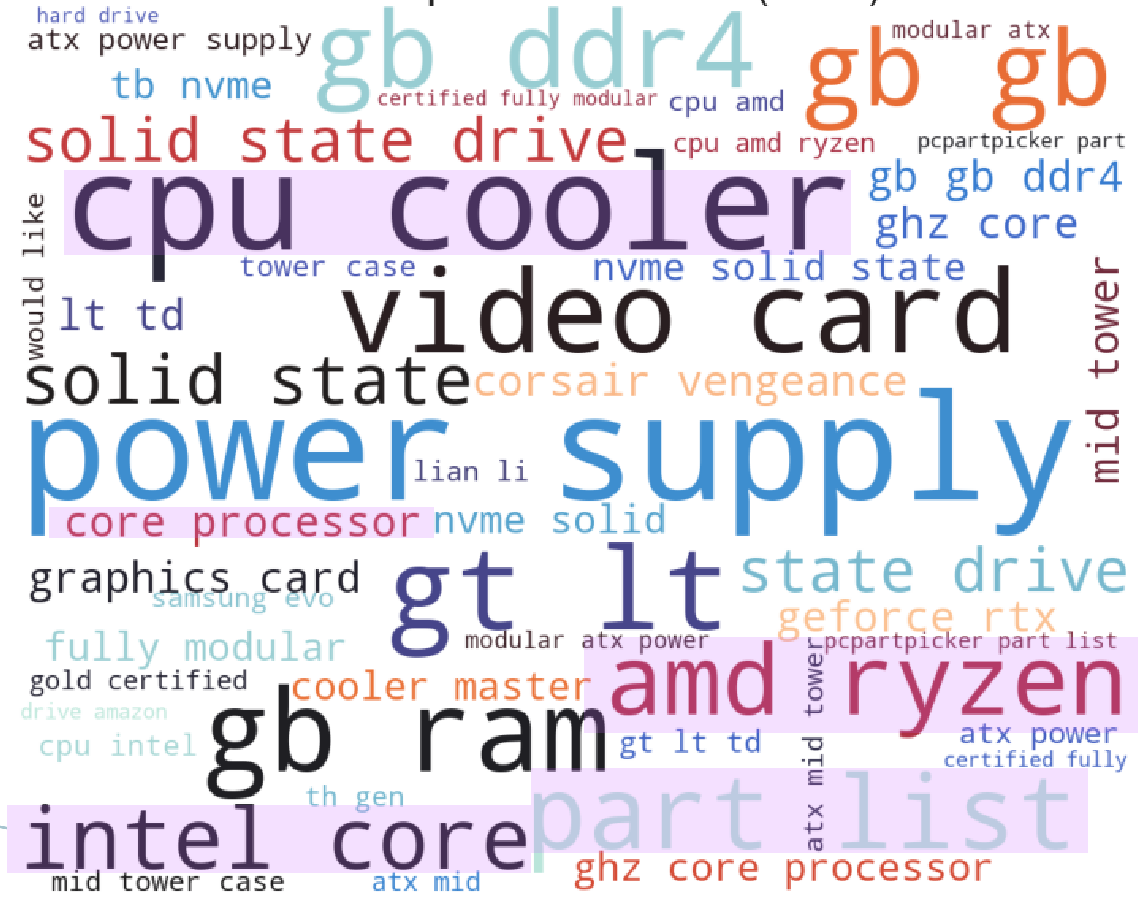


Top 50 words in r/techsupport posts (cvec)

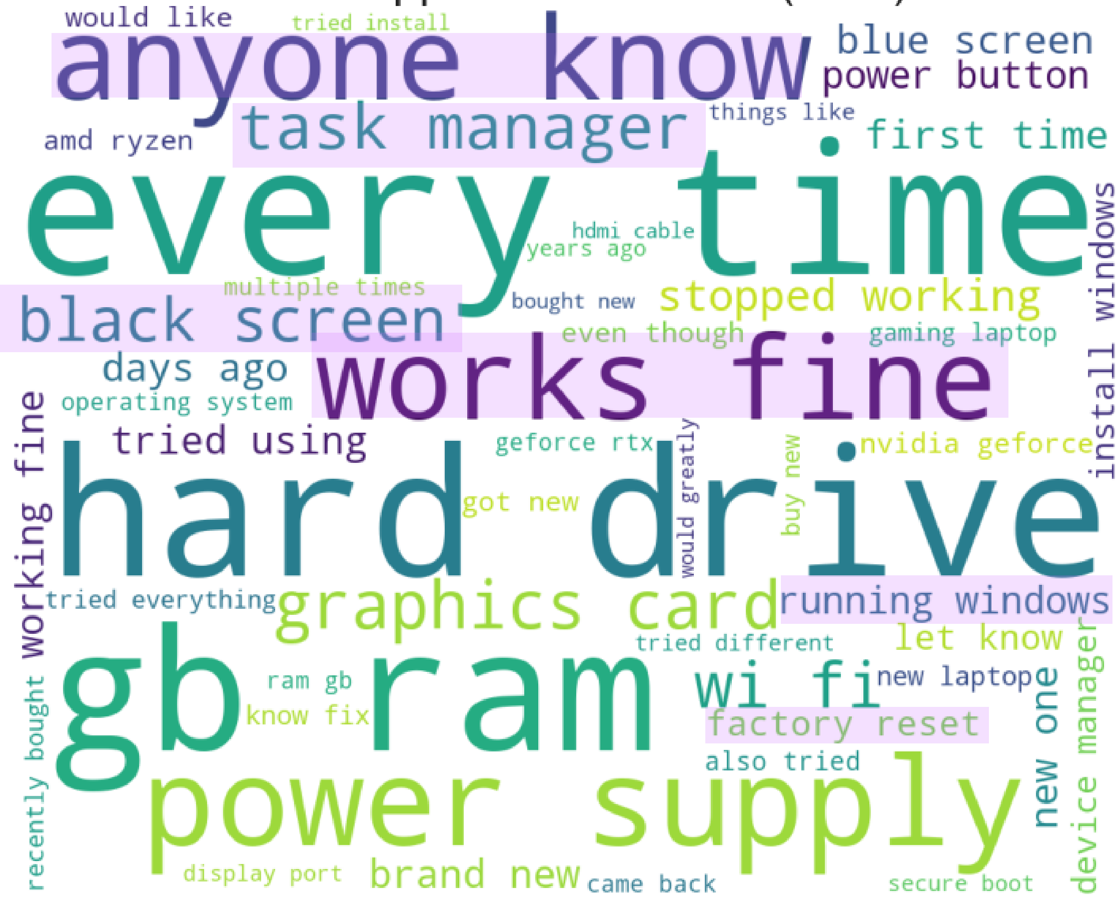
EDA

Using CountVectorizer

r/buildapc Word Cloud (cvec)

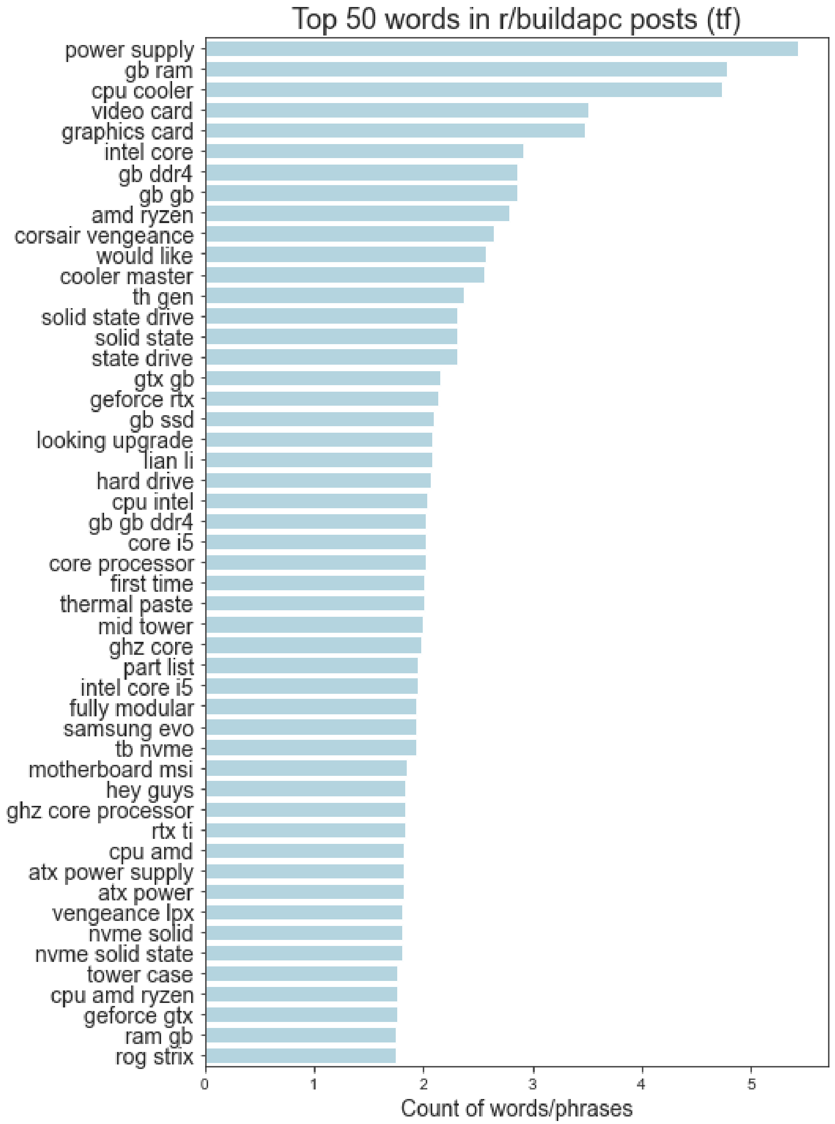


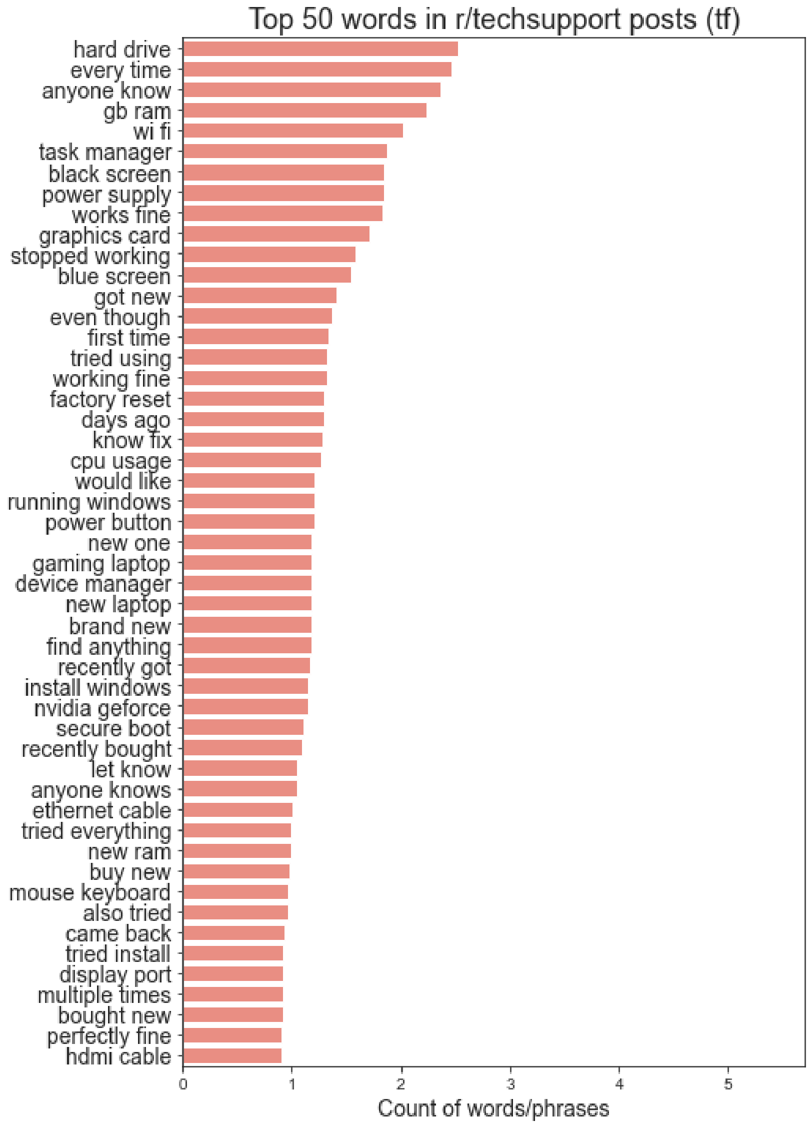
r/techsupport Word Cloud (cvec)



EDA

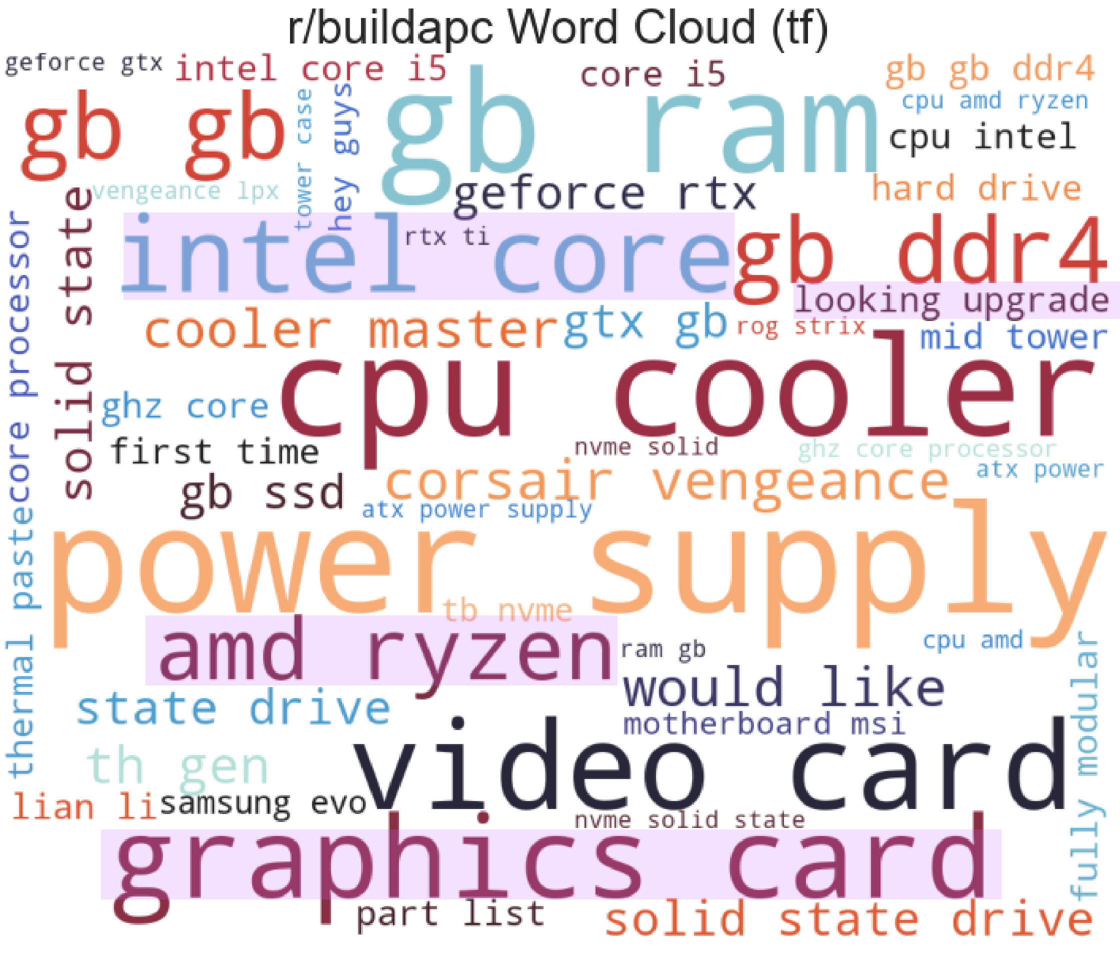
Using TfidfVectorizer


Top 50 words in r/buildapc posts (tf)

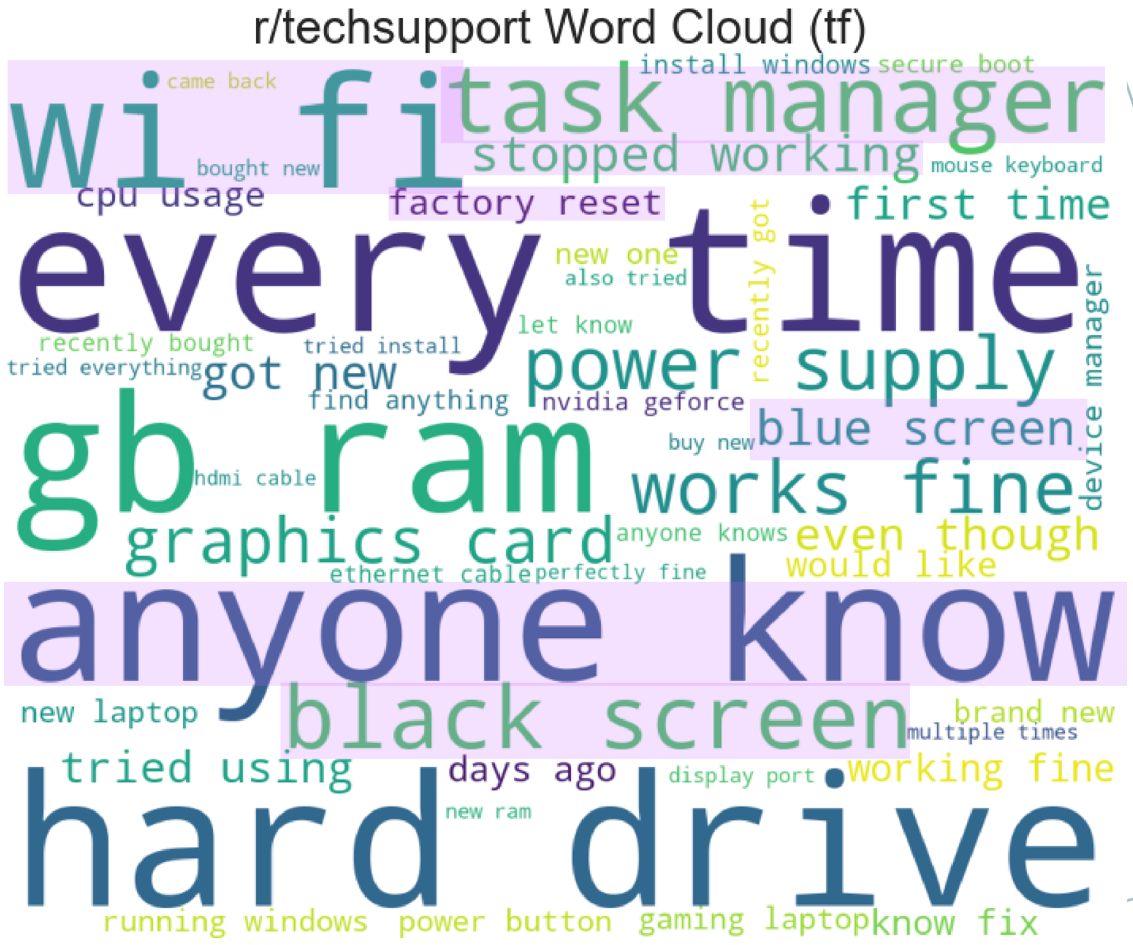

Top 50 words in r/techsupport posts (tf)

EDA

Using TfidfVectorizer



r/buildapc Word Cloud (tf)



r/techsupport Word Cloud (tf)

Model Preparation

- Check data class balance:
  - r/buildapc: 51.2%
  - r/techsupport: 48.8%

- Encoding of the subreddits to class 0 and 1:
  - r/buildapc will be represented by 0
  - r/techsupport will be represented by 1

Model Testing

- CountVectorizer with Multinomial Naive Bayes
- TfidfVectorizer with Multinomial Naive Bayes
- CountVectorizer with Random Forest
- TfidfVectorizer with Random Forest
- CountVectorizer with Logistic Regression
- TfidfVectorizer with Logistic Regression

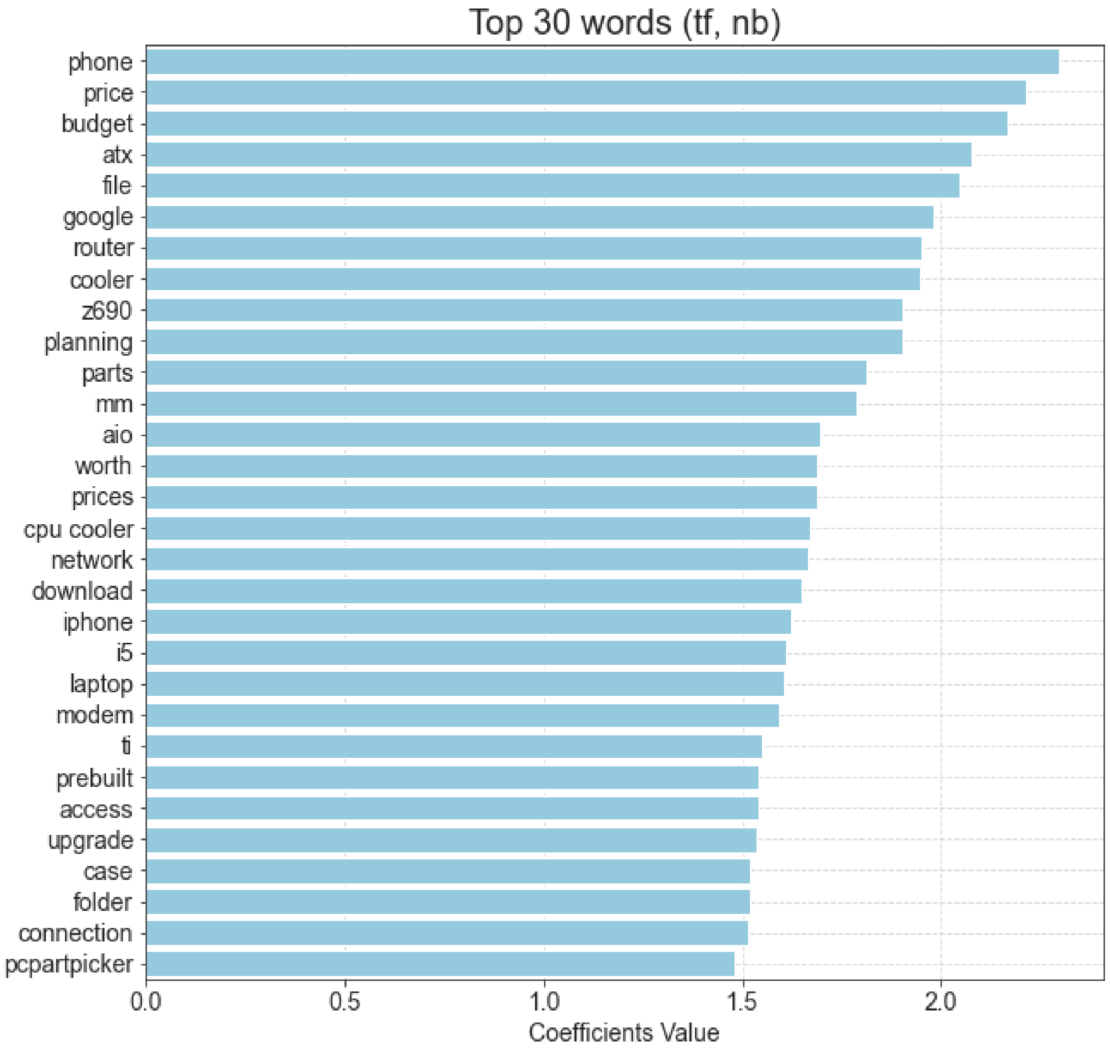
- Scoring metrics: Accuracy

Model Results

Results from the models	Train Score	Test Score	Best Score
Baseline Accuracy	0.512	0.510	--
CountVectorizer with Multinomial Naive Bayes	0.883	0.859	0.837
TfidfVectorizer with Multinomial Naive Bayes	**0.877**	**0.869**	**0.845**
CountVectorizer with Random Forest	0.872	0.838	0.827
TfidfVectorizer with Random Forest	0.883	0.832	0.822
CountVectorizer with Logistic Regression	0.962	0.827	0.815
TfidfVectorizer with Logsitic Regression	0.926	0.861	0.834

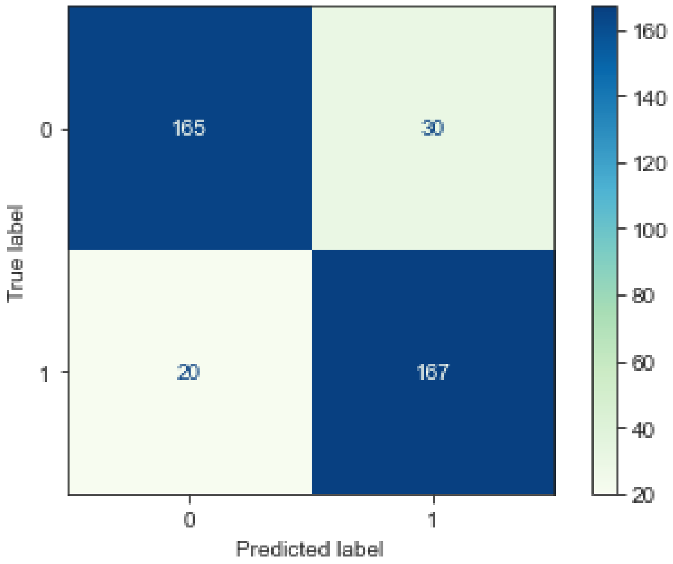
- From the 6 models tested, the best performing model is the **TfidfVectorizer with Multinomial Naive Bayes**.

Model Results



Confusion Matrix

- r/buildapc - 0
- r/techsupport - 1



Recommendations

- Implement this model as automated first step of classification for incoming queries without need for human oversight as it is able to correctly classify the incoming queries at a relatively high success rate.
  - Help to reduce the labour costs required to manually sort and classify the incoming queries.

- Further improvements:
  - As the customer service staff reviews the queries and finds any mistakes, they can transfer the queries across to the correct department while flagging the wrong query for further analysis to improve the model.

Limitations

- Reddit posts from the two subreddits (r/buildapc, r/techsupport) were used instead as the training and test dataset to train the model. The phrasing used in Reddit could be significantly different from the queries the company will receive from local customers here in Singapore.

- The list of stop words might not be extensive enough to filter out all the words that are not useful to the modelling process hence these stop words can be continually be identified and added to further improve the model.

Thank you!

ANY QUESTIONS?