



# Subreddit Classification

Group 1:  
Huimin, Ramkumar,  
Rahayu, Shi Min, Eric

# Context

## Who are we?

We are a team of Data  
Analysts in a Tech  
Magazine Firm



## Audience

Editor-in-chief and  
technology journalists



# Agenda

01

Problem Statement

04

Exploratory Data Analysis

02

Data Scraping

05

Modelling & Interpretations

03

Data Filtering

06

Conclusion



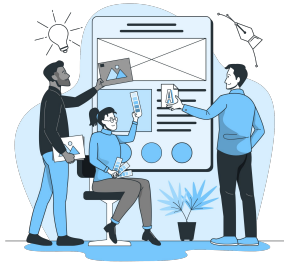


# 01. Problem Statement

# 01. Problem Statement



**Editor-in-chief** wants to come up with a big year-end story on "Brand Dominance between Samsung and Apple"

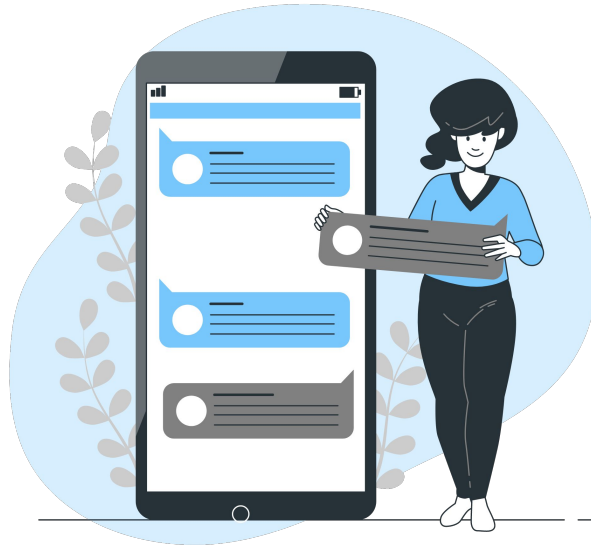


**Technology Journalists** need a quick sense of the online chatters on Samsung and Apple



**Data Analysts** will develop a text classifier that is capable of distinguishing whether an online post is about Samsung or Apple and derive meaningful insights

# How will a Successful Model look like?



## **Fairly Accurate**

At least 80% accuracy and minimally better than baseline of about 50% accuracy in classifying a post

## **Little Overfitting**

Difference between train and test accuracy is  $< 5\%$

## **Interpretable**

Able to identify common features that distinguish each class of post



## 02. Data Scrapping

## 02. Data Scraping



### Data acquisition

Data was scraped from the two subreddits ([r/Apple](#) and [r/Samsung](#)) using Pushshift's API and selecting <https://api.pushshift.io/reddit/search/submissions/> endpoint.

Example of a function created for scraping data:

```
# Multi-scraping loop for subreddit
params_samsung = {
    'subreddit': 'samsung',
    'size': 100,
    'selftext:not' : '[removed]'
}
frames_to_concat = []
frame_count = 0
while frame_count < 100:
    res_samsung = requests.get(url, params_samsung)
    samsung_json = res_samsung.json()
    frame = pd.DataFrame(samsung_json['data'])
    frames_to_concat.append(frame)
    frame_count += 1
    try:
        params_samsung['before'] = frame.tail(1).iloc[0]['created_utc']
    except:
        IndexError

# Collect frames in dataframe
df_samsung = pd.concat(frames_to_concat, ignore_index=True)
```



# Examples of Data Scrapped

From  
r/Apple :

author	created_utc	num_comments	selftext	subreddit	title
scorpionman	1561468602	0	NaN	apple	Apple Watch a DOMINAT Piata Smartwatch-urilor si in 2018, Iata Evolutia Apple
gustav83	1561467142	0	I've been looking into getting on the iPhone upgrade program, but I'm a little concerned that Ap...	apple	Has anybody had an iPhone stolen whilst being on the iPhone upgrade program?
michael44445	1561465307	0	NaN	apple	The tourist met a strange creature. the tourist was traveling in the forest with a camera, accid...
mennej	1561464236	0	I have a problem with apple. I bought a mac and they sent the wrong one. I contacted them and th...	apple	Problem returning a computer
AjPicard913	1561462399	1	NaN	apple	Introducing Splash — New charging concept for Apple Devices

From  
r/Samsung :

author	created_utc	num_comments	selftext	subreddit	title
[deleted]	1612993141	8	[deleted]	samsung	Didn't receive my Smart Tag with my credit bonus order. Gave up after talking with customer supp...
gointhrou	1612992992	7	Hey guys,\n\nSo, I have the S19 right now and I'm thinking of upgrading.\n\nSo far I've had the...	samsung	Does the S21 still show the same problems related to charging as previous entries?
[deleted]	1612992657	0	NaN	samsung	Samsung s20, charging twice a day vs charging thrice a day.
xenmaster35	1612992408	2	Hi there I just purchased this sound bar but I found out it doesn't support LPCM audio. Is this ...	samsung	Question about the HW Q950T sound bar with a PS5
Oblimix	1612992261	0	Don't go shopping until you've checked GPay still works, or bring a regular card with you. \n\nT...	samsung	To those that have Google Pay and got the February 2021 update

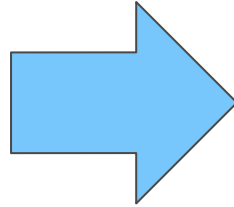


# 03. Data Filtering

## 03. Data Filtering

`df_ap.shape`

`(39535, 96)`

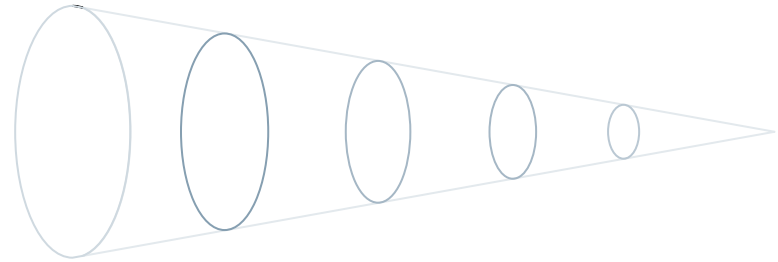


`df_ap.shape`

`(9305, 97)`

**Rows**

**Columns**





# Filtering the pulled data



## Links

### Removed

Regex was utilized to remove links

## Newline characters

### Removed

Found using python  
string methods



\n

## Moderator posts

### Removed

Author value counts/  
Looking at authors' flairs



PSA: How submissions will work tomorrow

submitted 1 month ago by **aaronp613** [M] to r/apple  
119 comments source share save hide give award

# Filtering the pulled data

## Null values in 'selftext'

*Replaced*

Singled out using boolean  
filtering/ built-in methods

## ['deleted'] as 'selftext'

*Replaced*

Conditional filtering

title	selftext
'phone worth it in the long term android devices when it com...	Is the price of an iPhone worth it in the long term when compared to Android devices when it com...
ne know the adobe crack link suitable for the m1 system?	hello, does anyone know the adobe crack link suitable for the m1 system?
last test	last test
A	A
neur-Legend Success Secrets	Steve Jobs-Entrepreneur-Legend Success Secrets
Hi	Hi
atter health be after having an Phone X for nearly 3 months?	What should my batter health be after having an iPhone X for nearly 3 months?
2019 13' able to drive the Pro Display XDR?	Is the MacBook Pro 2019 13' able to drive the Pro Display XDR?

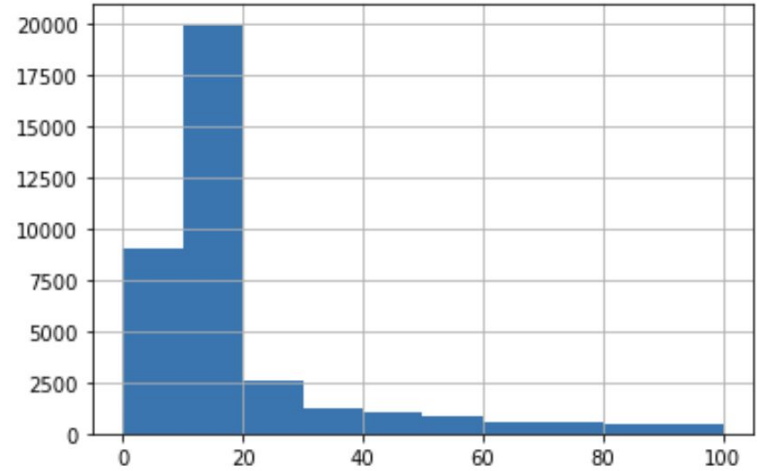
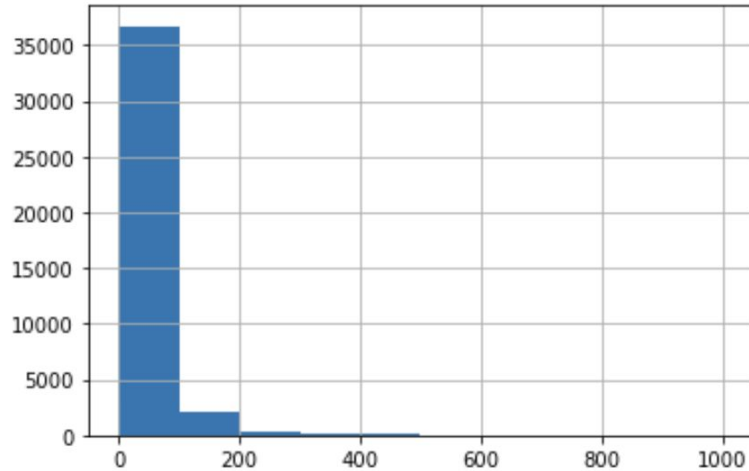
**'selftext' = 'title'**

*Replaced* and later **Removed**

# Filtering the pulled data

Number of Words per post in pulled data

(r/Apple)



# Filtering and Feature engineering

- **Combined the title and selftext**
- Number of words in the combined text
- Average sentence length in general is 15-20
- **Dropped rows with title + selftext having < 20 words**



# Cleaning the text

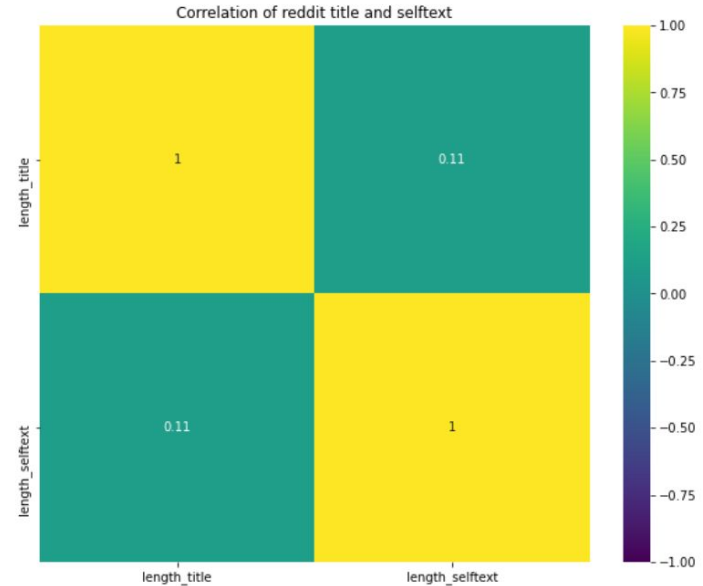
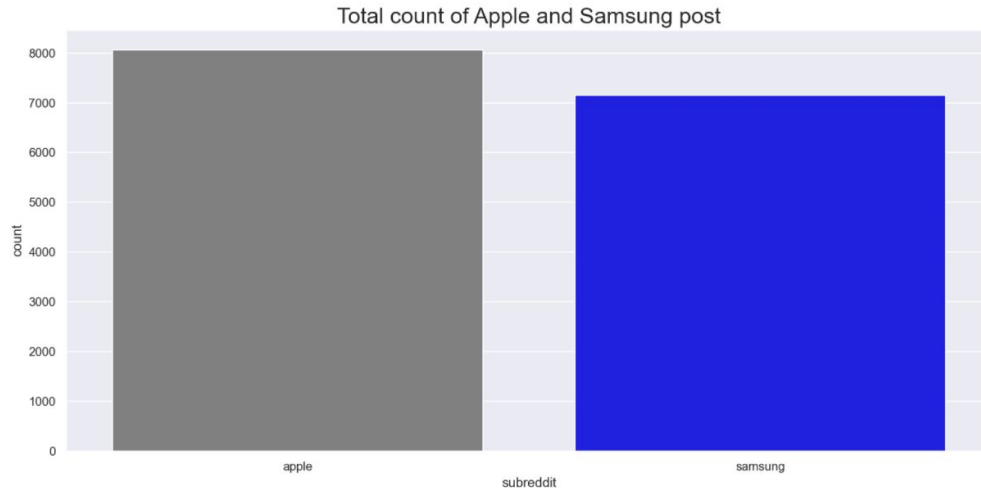
- Tokenization - Splitting text into a list of words
- Punctuation removal
- Removing words with non-ASCII printable characters
- Removing all empty strings
- Lowercasing
- Stopwords removal - Excluding words that occur frequently in all text
- Lemmatization - Reducing all words to their base form





# 04. EDA

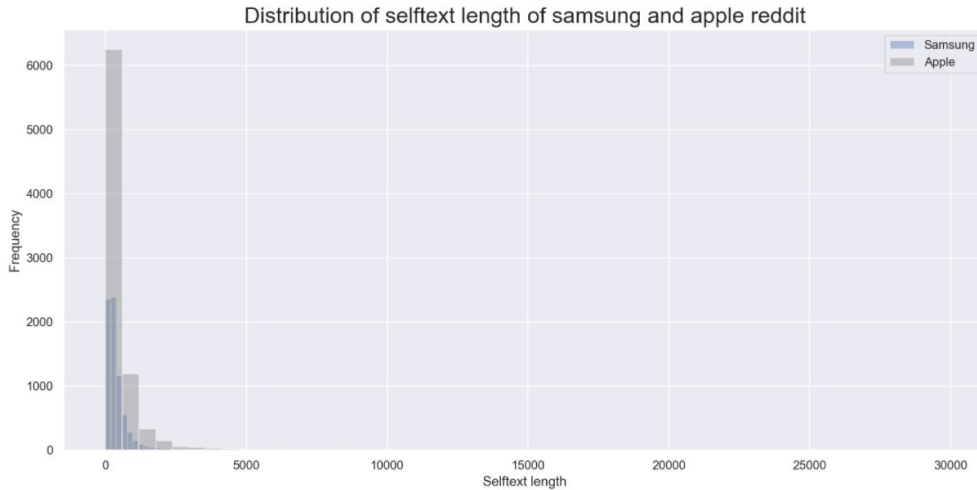
## 04. Exploratory Data Analysis



- Post counts after the data filtering steps
- Apple has a higher amount of posts at 8065
- Selftext and title length are not highly correlated

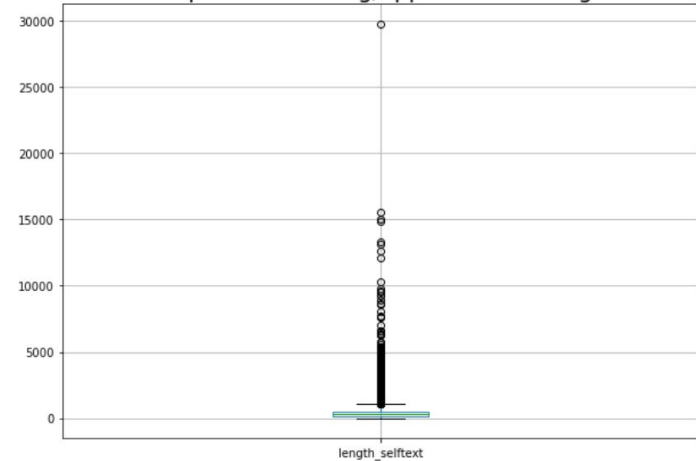
# 04. Exploratory Data Analysis

## Distribution of self text length



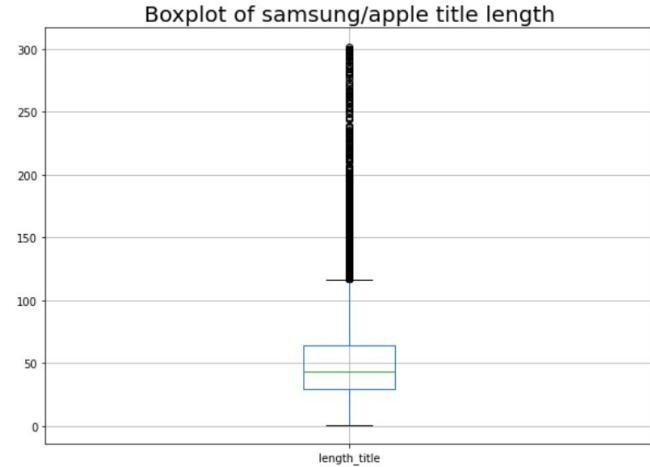
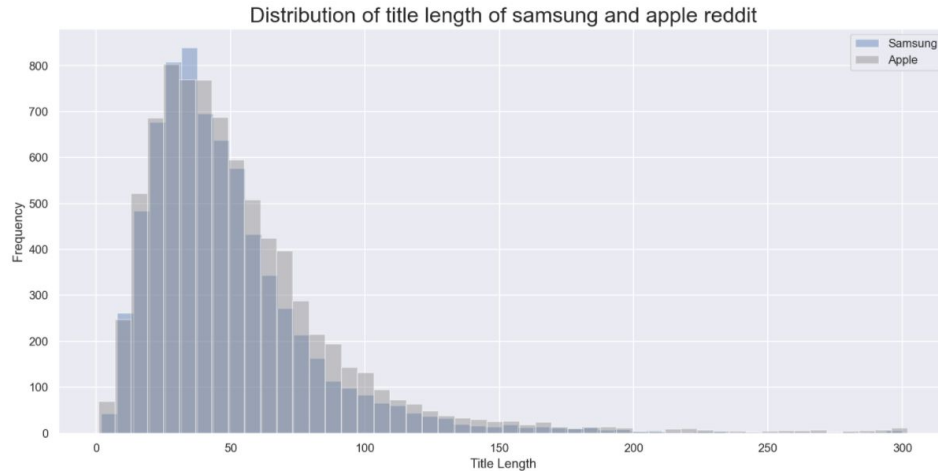
- Right-skewed for both
- Range of characters between 1 - 29,770
- Median at 286 char

Boxplot of samsung/apple selftext length



# 04. Exploratory Data Analysis

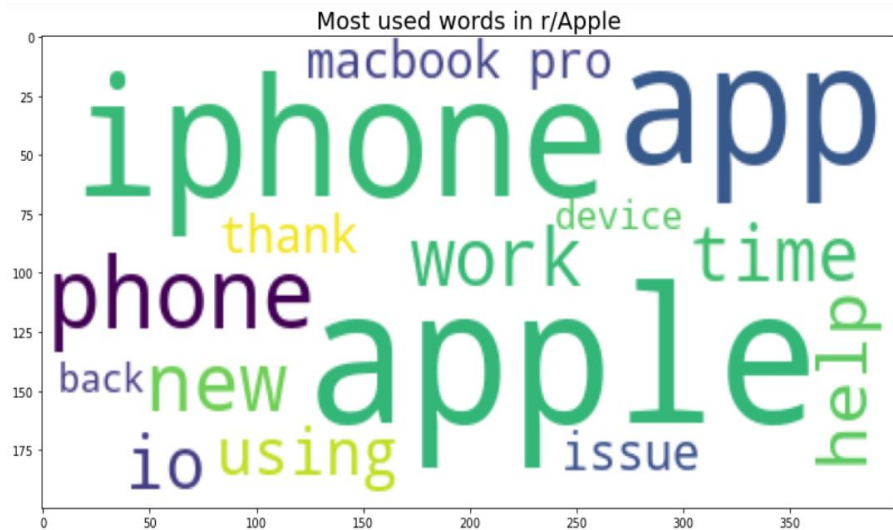
## Distribution of title length



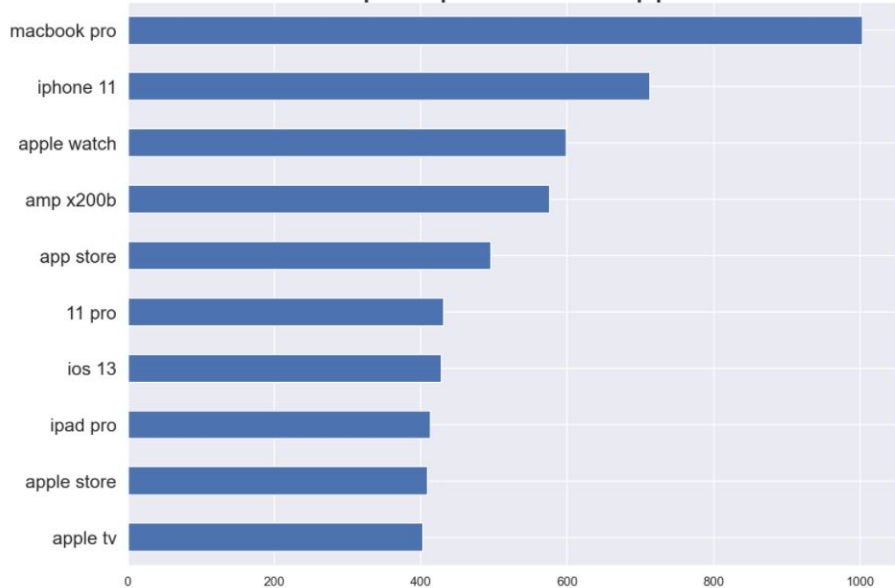
- Right-skewed for both
- Range of characters between 1 - 302
- Median at 43 char

## 04. Exploratory Data Analysis

Most used words and phrases (without stopwords) in r/Apple

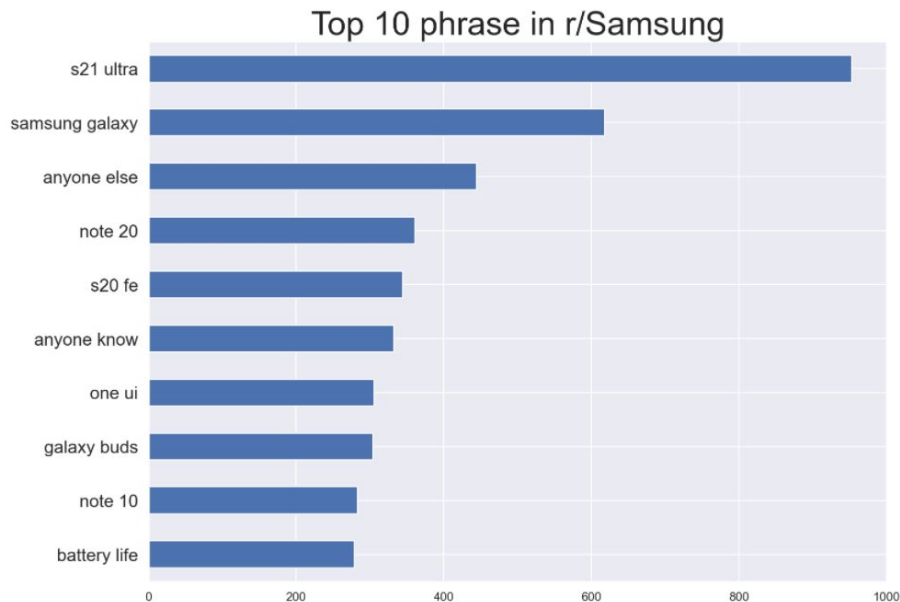
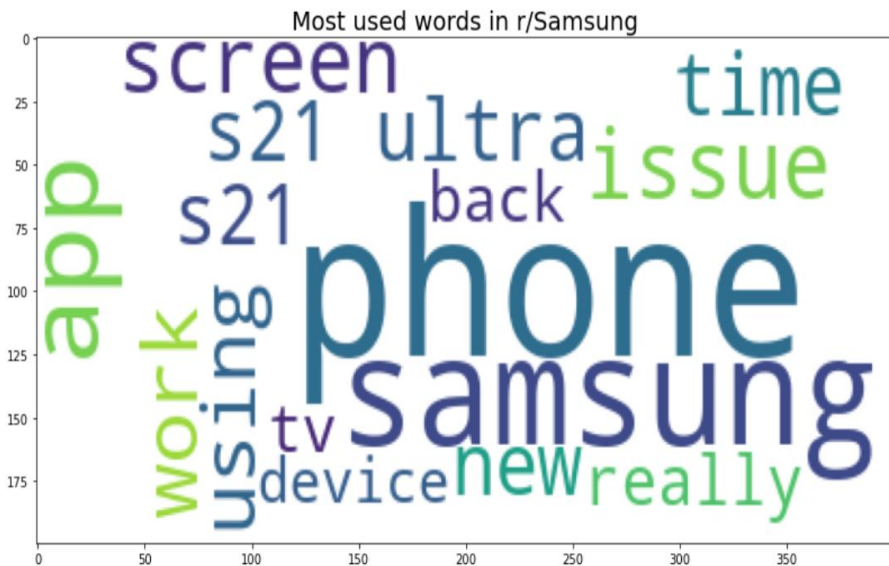


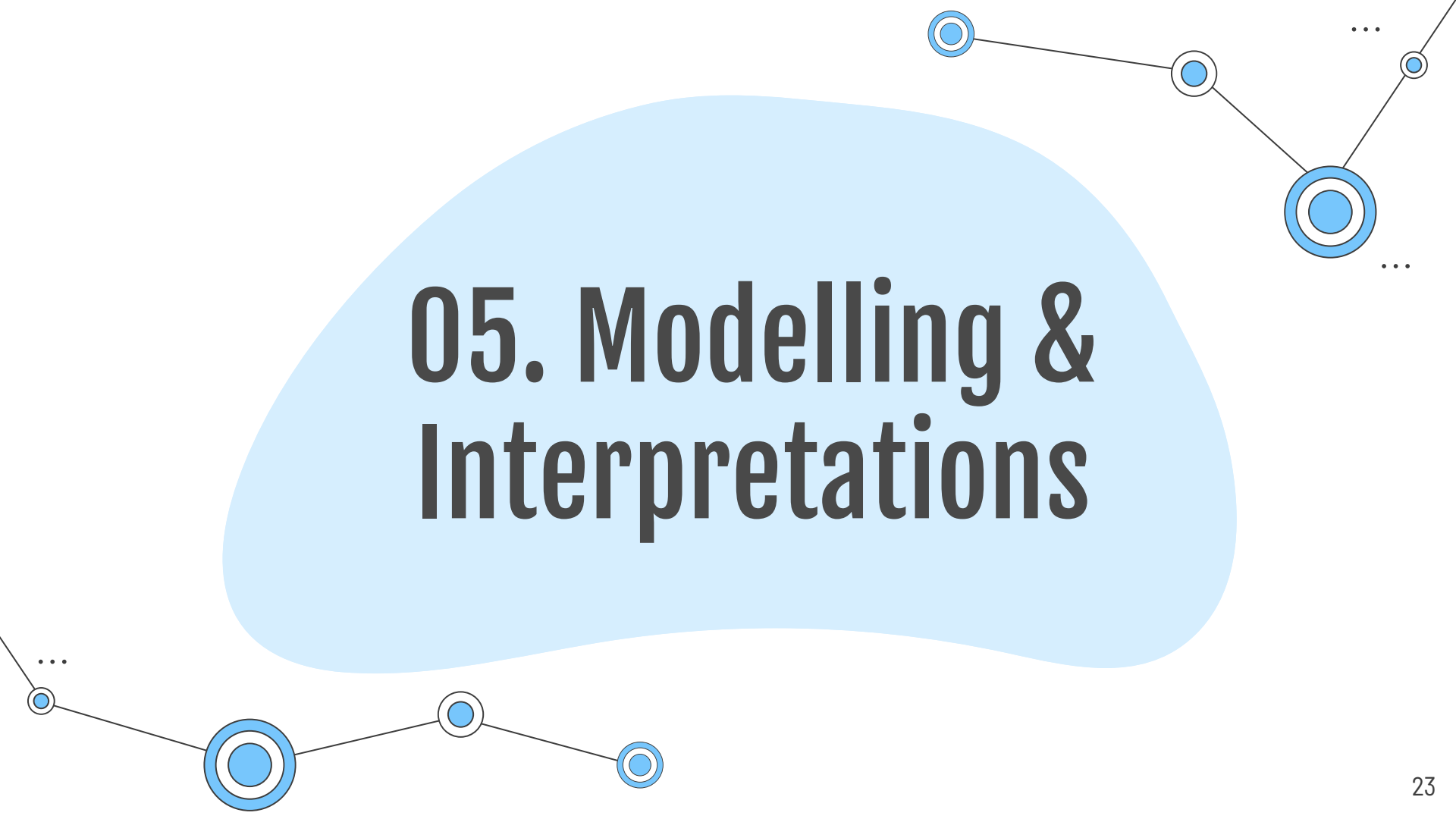
Top 10 phrase in r/Apple



## 04. Exploratory Data Analysis

Most used words and phrases (without stopwords) in r/Samsung





# 05. Modelling & Interpretations



# Defining the Modelling Problem



## Recap of the Problem:

- Tech journalists need to assess the brand dominance of Samsung vs Apple
- Need a quick way to sieve through numerous online word posts and classify them as Samsung or Apple, so as to assess which brand is more talked about
- Bonus: To get insights on discussions that are specific to each brand, to know where the brand's edge is

Binary Classification Problem

Able to identify keywords driving  
the classification model

Natural Language Processing  
needed



# Modelling Pre-processing



Posted by u/damian\_ 18 hours ago

481

Posted by u/damian\_ 18 hours ago

481

Posted by u/damian\_ 18 hours ago

481

Posted by u/damian\_ 18 hours ago

481

**Mac** The new MBP M1 Max 14" battery life is incredible after turning on "Low Power" mode

I got a new MBP M1 Max 14" on launch day, and have found the battery life OK-but-not-great for my usage. I spend a lot of time in zoom calls, Slack, coding (node.js) and general web apps and have been getting approx 5 hours on battery at normal workload.

Well, today I turned on "Low Power" mode and the difference has been amazing! The system performance feels absolutely fine - most apps it's not noticeable, but things like building code take an extra few seconds... however battery life has been amazing!



Posted by u/josh89rea 10 hours ago

5

Posted by u/josh89rea 10 hours ago

5

Posted by u/josh89rea 10 hours ago

5

Posted by u/josh89rea 10 hours ago

5

**Discussion** Trying to recover old photos off of a Samsung Galaxy S7

A few years ago I dropped my phone, and while the phone system is still working, the screen completely broke and is black. The photos and files on it are not backed up. I tried using Droidkit to recover the data, but you need to allow USB debugging for this to work, and I can click on anything since the screen is black. Is there anything I can do to

Tokenize,  
Remove  
Stopwords,  
Lemmatize,  
Vectorize



Individual words from all posts

	M1	Galaxy	...	Subreddit
Post 1	1	0		Apple
Post 2	0	1		Samsung

Frequency of M1  
appearing in post 1

# Model Train-Test

Individual words from all posts

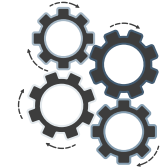
	M1	Galaxy	...	Subreddit
Post 1	1	0		Apple
Post 2	0	1		Samsung

Frequency of M1  
appearing in post 1

Train the model

75% of posts


25% of posts



Apply on unseen  
posts and predict a  
classification



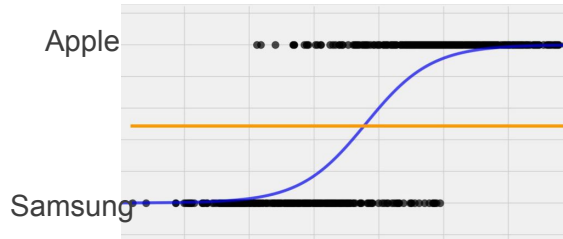
Prediction:  The unofficial Apple community  
r/apple

Prediction:  Samsung  
r/samsung

# Shortlist of Model Choices

## Multinomial Naive Bayes

Calculates likelihood that a post belongs to Apple/Samsung, given the words in the post (conditional probability)

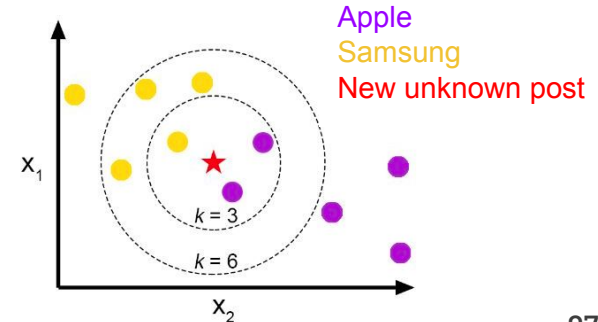


## Logistic Regression

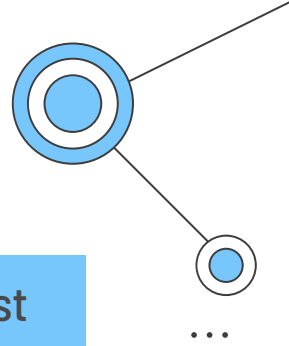
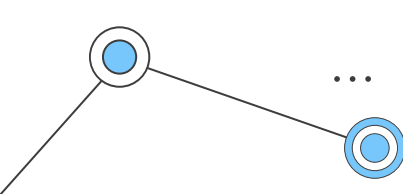
Predicts odds of a post belonging to Apple, based on its relationship with the words that appear in the posts

## K-nearest neighbors

Predicts classification of each new post, by observing how similar the words in the post are to the training observations, and labelling the post similarly to their "neighbors"



# Model Results



Test (Train)	Dummy Classifier	Multinomial Bayes	Logistic Regression	K-nearest neighbors
Accuracy	0.50 (0.50)	0.94 (0.96)	0.95 (0.99)	0.85 (0.86)
F1-score For Apple	0.67 (0.67)	0.94 (0.96)	0.95 (0.99)	0.83 (0.85)
F1-score For Samsung	0.00 (0.00)	0.94 (0.96)	0.95 (0.99)	0.87 (0.88)

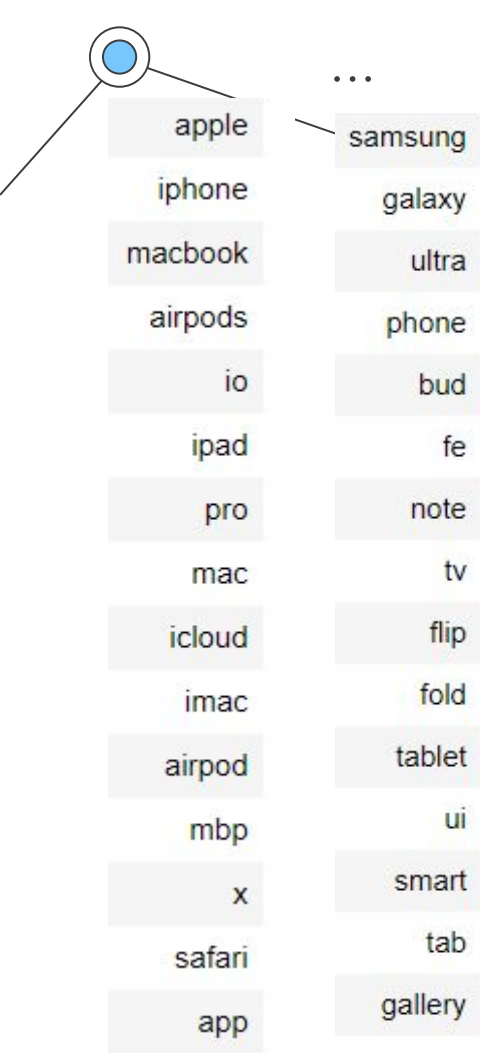
# The Chosen Model: Logistic Regression

Criteria	Multinomial Bayes	Logistic Regression	K-nearest neighbors
Accuracy	★ ★ ★	★ ★ ★	★ ★ ☆
No Overfitting	★ ★ ★	★ ★ ★	★ ★ ★
Ease of Interpretability	★ ★ ☆	★ ★ ★	★ ★ ☆

Accuracy: >0.9: 3-stars, 0.8-0.9: 2-stars, 0.5-0.8: 1-star, <0.5: No stars

No Overfitting: Difference between train and test accuracy less than 0.05: 3 stars, 0.05-0.1: 2 stars, 0.1-0.2: 1 star, >0.2: 0-star

Ease of Interpretability is more subjective; takes into account how easy it is to explain how the model works, identify the key variables and how to interpret how important those variables are



# Key Predictors for Logistic Regression

- Brand names and product names
- May indicate which products are discussed most
  - Information for the upcoming magazine article?
- Consistent with what we saw from the EDA section

What if we want to look beyond these words?






# Words filtered out



'apple', 'iphone', 'iphones', 'macbook', 'ipad', 'pro', 'pros', 'airpods', 'airpod',  
'ios', 'mac', 'catalina', 'itunes', 'imac', 'icloud', 'safari', 'macos', 'max', 'air',  
'mbp', 'ipod', 'samsung', 'galaxy', 'ultra', 'note', 'bud', 'buds', 'fold', 'ui', 'tab',  
'flip', 'z', 'fe', 'exynos', 'oneui', 'bixby', 'siri', 'mini', 'xr', 'smart', 'gallery',  
'applecare', 'lightning', 'id', 'x', 'xs', 'se', 'pencil', 'airplay', 'pen', 'pens',  
'earbuds', 'edge', 'gen', 'snapdragon', 'lite', 'phone', 'phones', 'tablet', 'tablets',  
'tv', 'tvs', 'dex', 'thunderbolt'





## Predictors for Apple (product/ brand names excluded)

Apple Christmas sales  
surge to \$111bn amid  
pandemic

© 27 January

BBC



NEWS

### Words

product

store

computer

help

buy

music

feedback

subscription

library

passcode

amp

christmas

password

photo

free

## Predictors for Samsung (product/ brand names excluded)

SAMSUNG



Fast charge your  
Galaxy phone

### Words

google

factory

fast

anyone

setting

sd

active

notification

unlocked

remote

good

camera

screen

frame

fingerprint





# 06. Conclusion



# A Successful Model



Metric	Logistic Regression
Accuracy On Test Set	95%
Difference between Train and Test	4%
Interpretability	Key predictive words can be identified

## Fairly Accurate

At least 80% accuracy and minimally better than baseline of about 50% accuracy in classifying a post

## Little Overfitting

Difference between train and test accuracy is  $< 5\%$

## Interpretable

Able to identify common features that distinguish each class of post



# Summary and Recommendations



- The model is in a good position to be used as a tool by the Editor-in-chief
- Conduct Share of Voice analysis more efficiently to measure brand dominance
- Exploratory data analysis can reveal additional insights other than modelling

## Sentiment analysis



- Positive and negative words
- Context of those words

## Types of text classified

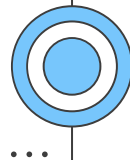
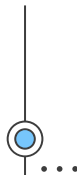
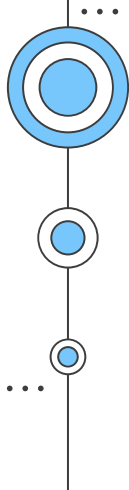


- Social media
- Product reviews

A decorative graphic consisting of blue circles of varying sizes connected by thin black lines, forming a network-like structure around the central text. Some circles have concentric rings. Ellipses (...) are placed at several points along the lines.

**Thank you!**

# Appendix



# Apple posts classified as Samsung posts

text
considering whether switch samsung apple ive samsung user year typically upgrade new flagship sa...
exsamsung user regret switch curious making switch samsung apple note galaxy bud frontier watch...
need advice would smart stupid upgrade samsung galaxy base iphone 128gb know screen quality woul...
question switching samsung iphone anyway transfer data android transfer google acc
pixel bud capability ive using pixel bud iphone way check battery life bud case feature update a...
lock screen colour inverted half second turning phone phone updated recent update phone keep in...
yeet apple big gay samsung better
android bad like post think android bad
samsung dexfor iphone know hook android phone monitor activate dex mode essentially turn phone ...
issuesi noticed getting iphone hour ago android phone press pause button bluetooth headphone you...
sort note folder alphabetically trying use note note taking app replacement cant figure one thi...

# Samsung posts classified as Apple posts

switching to android apple ecosystem iphone pro ipad pro magic keyboard apple watch series airp...

enough enough goodbye iphone apple dont know many yall like use galaxy iphone time time convinc...

cord need connect 2019 macbook pro purchased new macbook pro thunderbolt port cord purchase get...

frame yet see apple app

use iphone work suffering year note screen died bit bullet repair screen since beast phone next ...

tried download youtube video grandma iphone aid people use io restricted title

samsung itest working iphone keep getting prompt code say need scan iphone literally using iph...

samsung apple watch hello everyone looked reddit post found everyone pretty aggressive others re...

make advertisement le annoying annoying make bought iphone spite

share calendar someone hello everyone long time ago created shared calendar sister event added w...

reason chose samsung apple people die imessage day made choose android apple feel free shsre ex...

moved iphone s21 best way move stuff iphone since 3gs till finally thought change android due p...

iphone user year switched s21 plus make right decision phone yet still shipped switching iphon...

thinking switching iphone worth dumb question probably pretty obvious answer ask pretty serious...



text



...



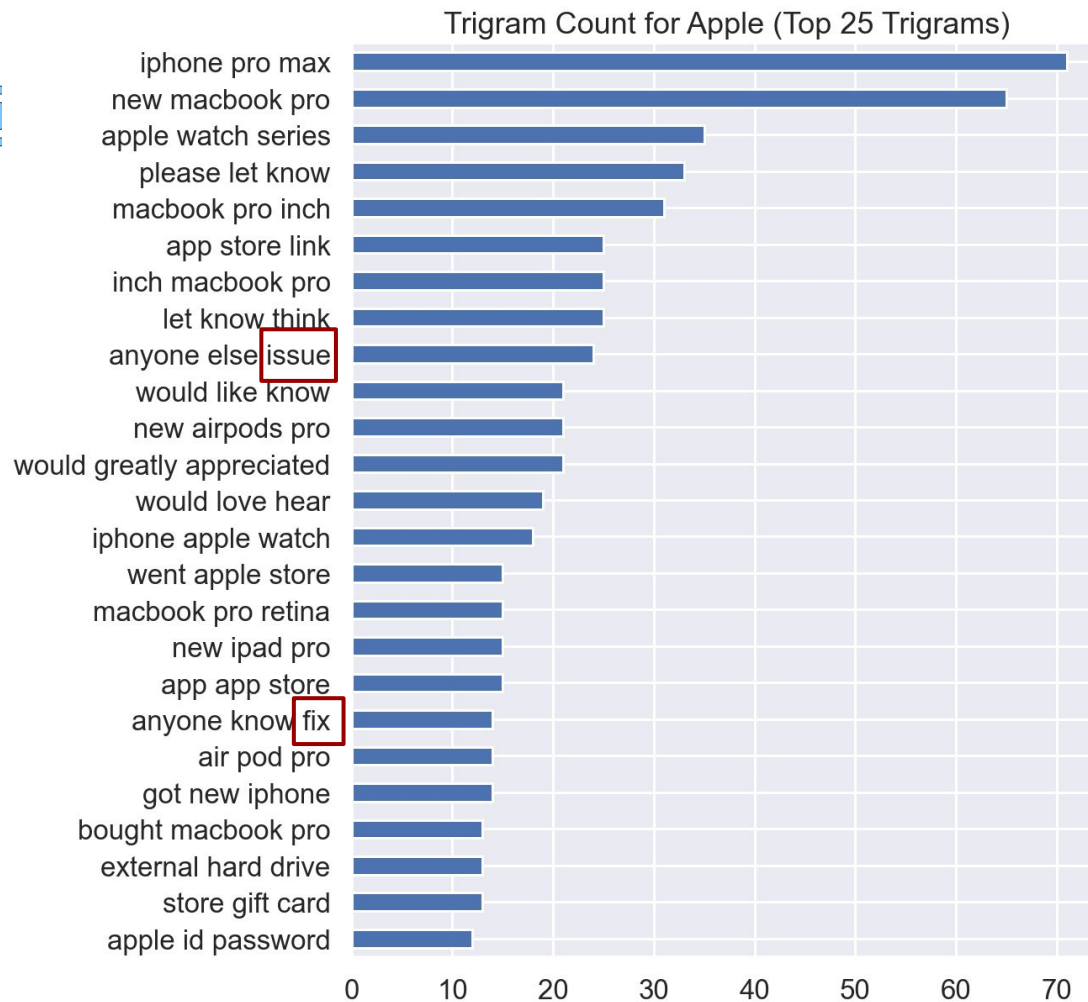
## Samsung post classified as Apple post

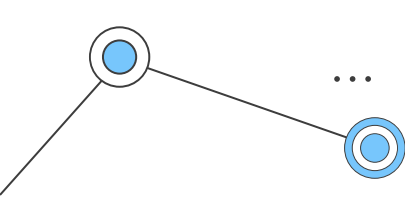
**iPhone 12 Pro Max** to S21 Ultra worth the switch? Hello everyone, I've currently got an **iPhone 12 Pro Max** which I've had for about 10 months and have had issues with (battery life, software issues, etc) I can get an S21 Ultra on contract from my carrier for around £56 a month and was wondering if it's worth the switch this late into the year.

## Apple post classified as Samsung post

LPT: On safari if you open **tab** view and put your phone into landscape mode, you can long press on any **tab** and sort them by title or sort **tabs** by website. This is quite useful to sort **tabs** if you have around 200-300 open. Also a bonus tip, while in **tab** view, if you scroll all the way to the top and search for a certain **tab** and then long press the cancel button you can close all **tab**s matching your search result.







Trigram Count for Samsung (Top 25 Trigrams)

