Project Smart Nation

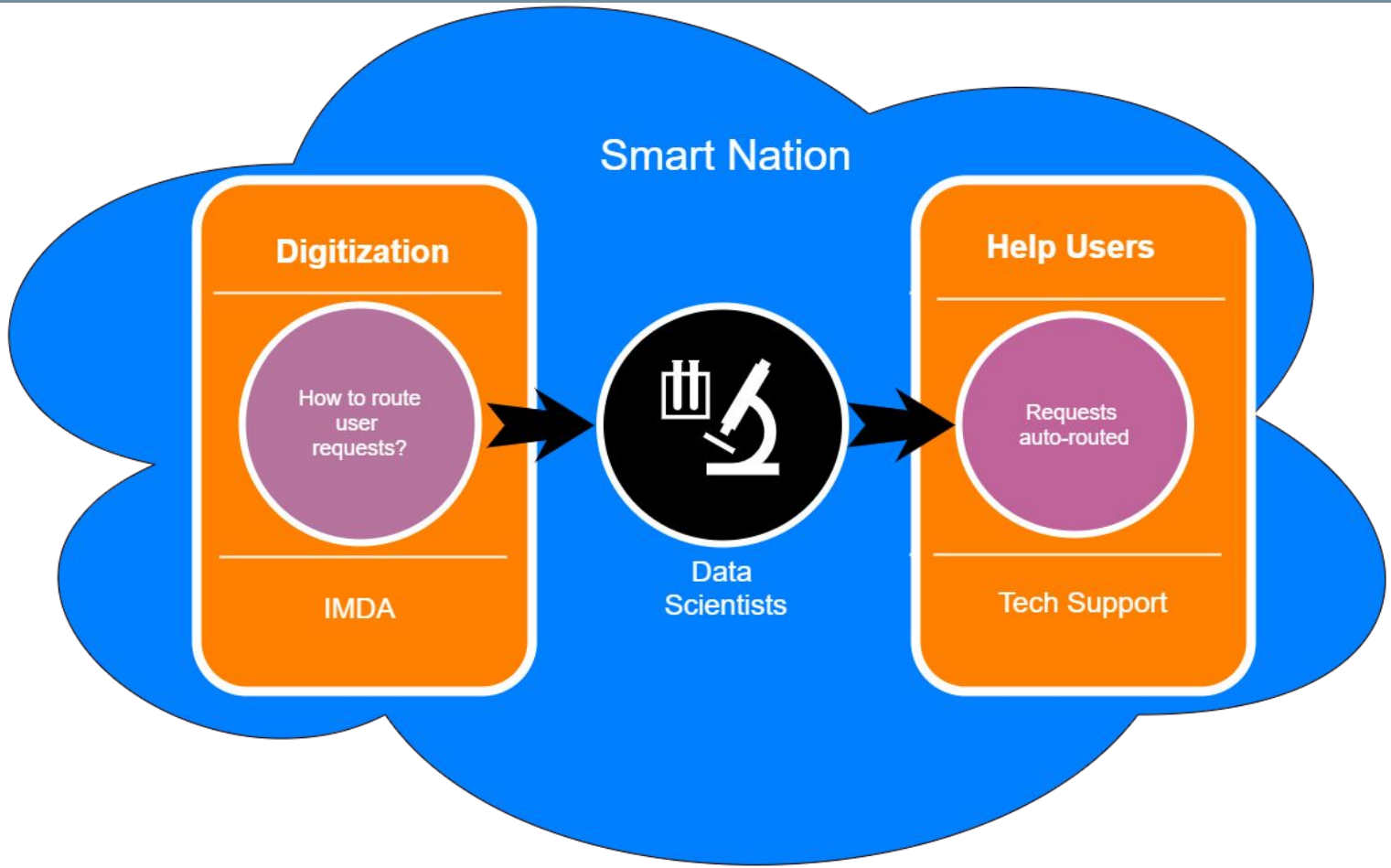
Group 1:

- Ng Heok Joo
- Kua Kee Chuah (KC)
- Leonard Ng

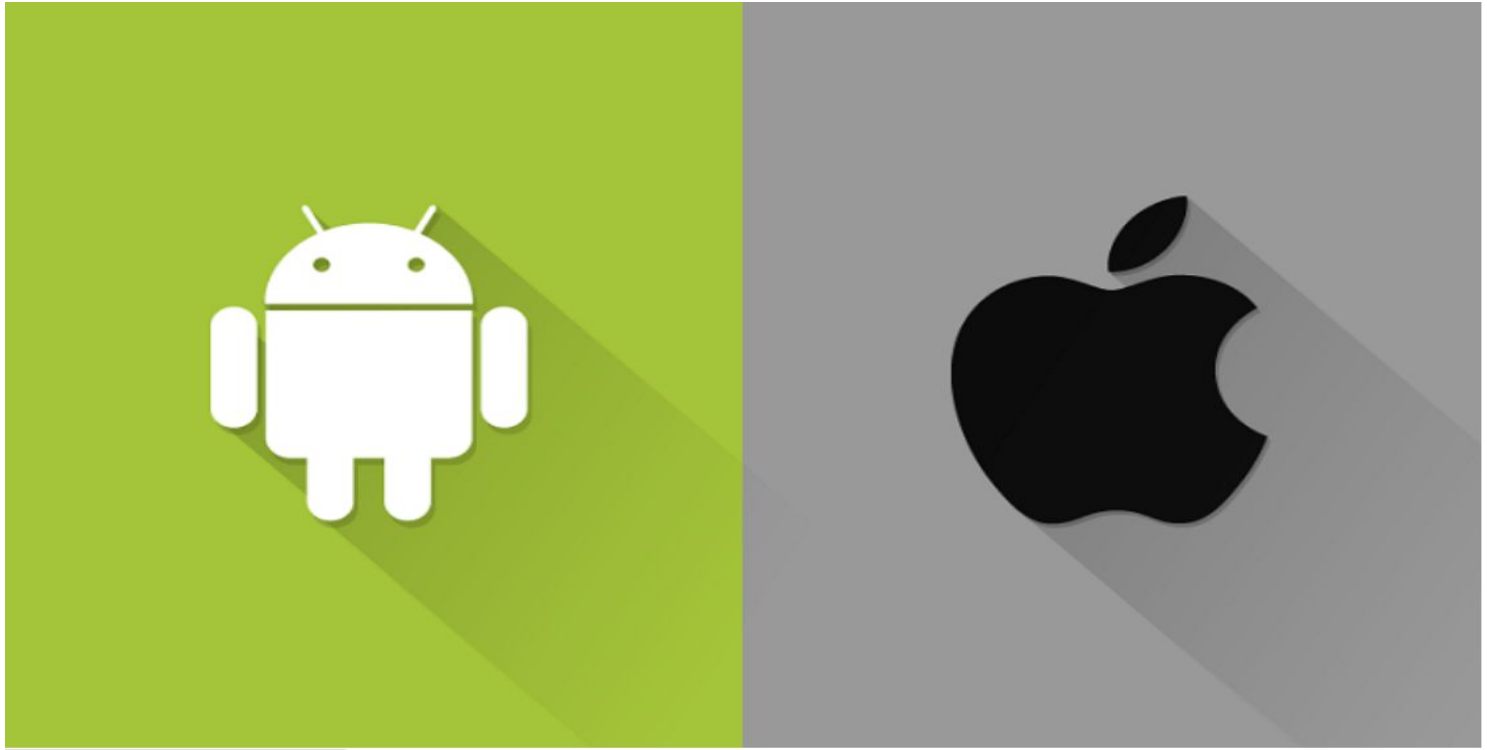
Agenda

1. Background
2. Problem Statement
3. Workflow for Reddit Data Analysis
4. Analysis of Reddit Data
5. Model Accuracy & Findings
6. Conclusion & Recommendations

1. Background

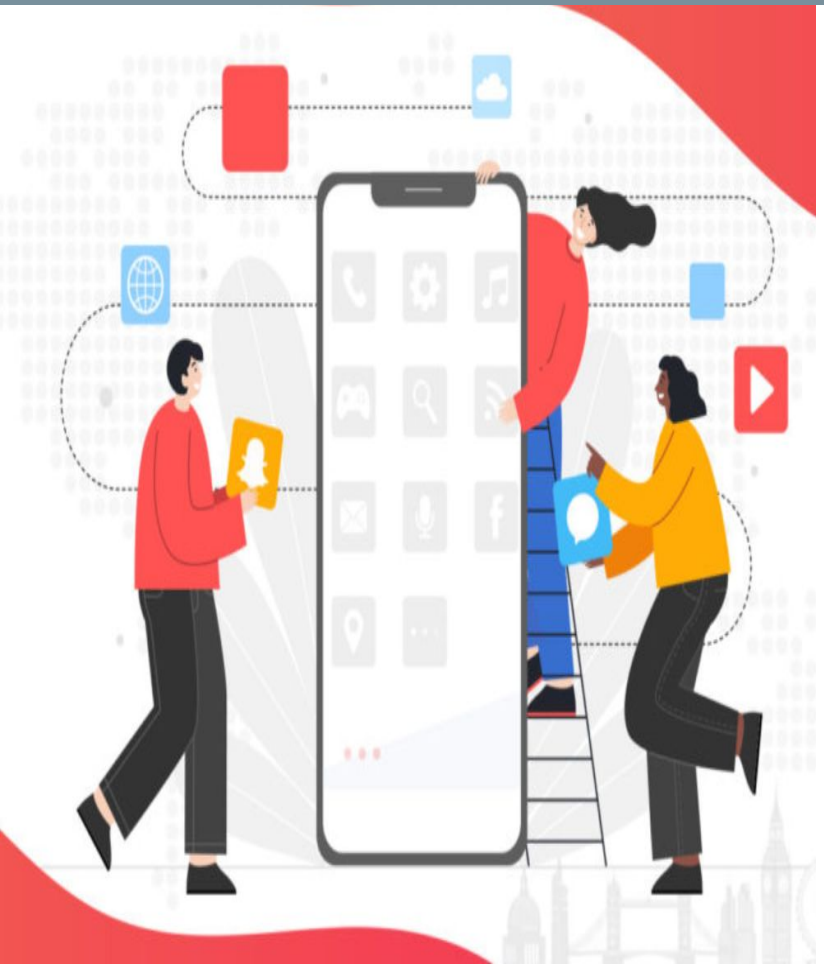


2. Problem Statement



Can we classify a query to be Iphone or Android related based on subreddits?

3. Workflow for Reddit Data Analysis



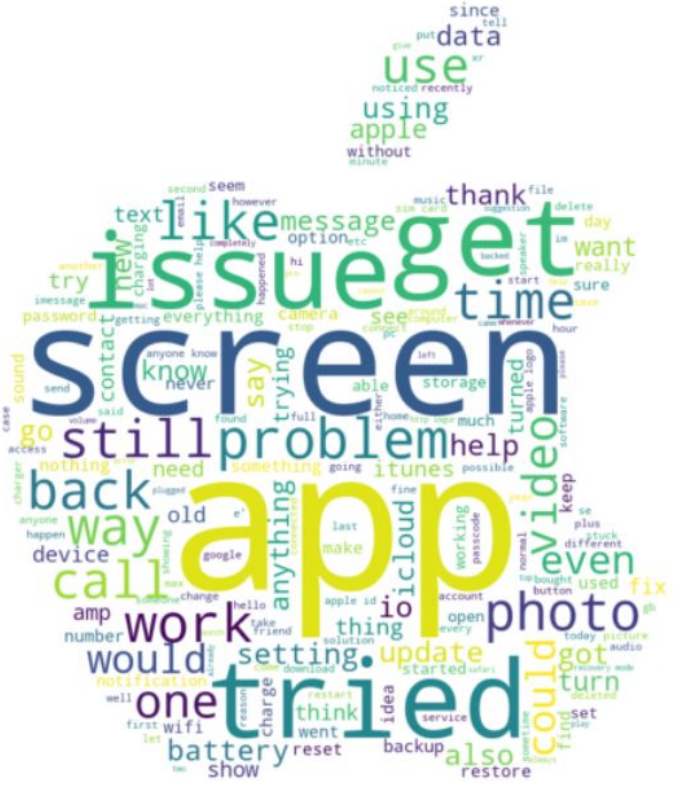
- Data Scraping
- Data Cleansing
- Data Preprocessing
- Exploratory Data Analysis

Meaningful Words

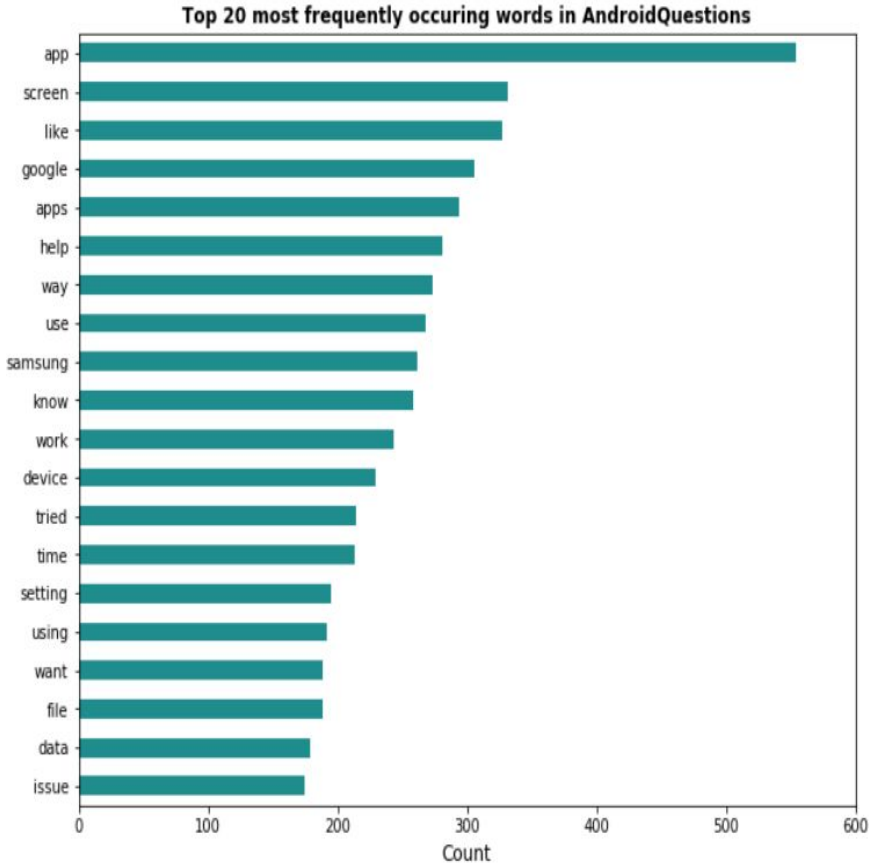
4. Analysis of Reddit Data

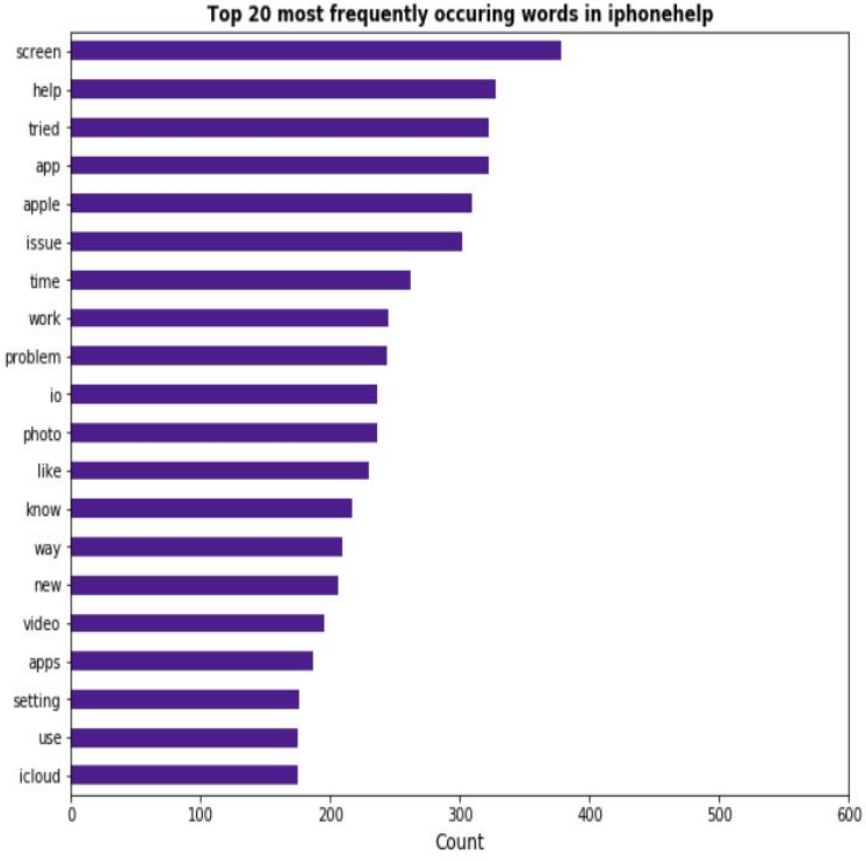
Wordcloud for Androidquestions

Wordcloud for iphonehelp

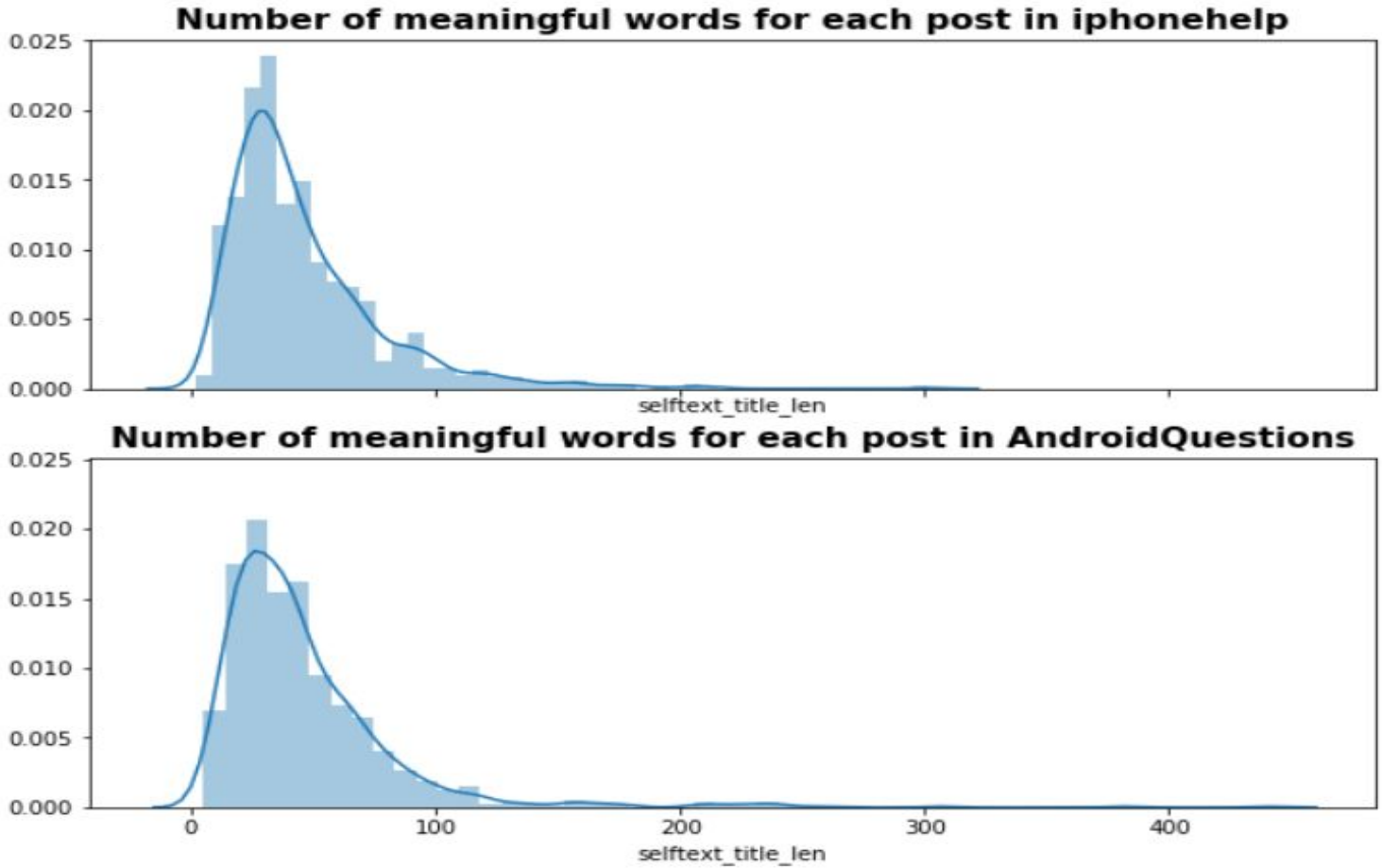


4. Analysis of Reddit Data





4. Analysis of Reddit Data



5. Model Accuracy

Our optimized model classifies

83%

of unlabelled posts accurately

- Accuracy: 0.83 (Accurately classified posts/ Total posts)
- Based on an optimized Multinomial Naive Bayes model

4. Findings

Data from ‘title’, ‘text’ or ‘comments’?

- It does not matter
- Better to feed all data into the model

More data is better

- Adding more data > optimising hyperparameters

Model may never be perfect

- Many common issues between the two systems

5. Conclusion and Recommendations

Problem Statement: Can we classify a query to be Iphone or Android related based on subreddits?

1. Consolidate and throw all ‘text’-related data at the model
2. The model will improve in accuracy over time
3. Consider using other sources of ‘text’.

Project Smart Nation

Thank you!

Any Questions?

References

https://bestandroidapps.com/android-versus-ios-by-region-infographic/

https://www.fluper.com/blog/analysing-the-crucial-differences-between-ios-and-android-app-development

https://www.fluper.com/blog/category/resources/infographic

https://www.itgsnews.com/singapore-smart-nation/

https://www.datacamp.com/community/tutorials/wordcloud-python