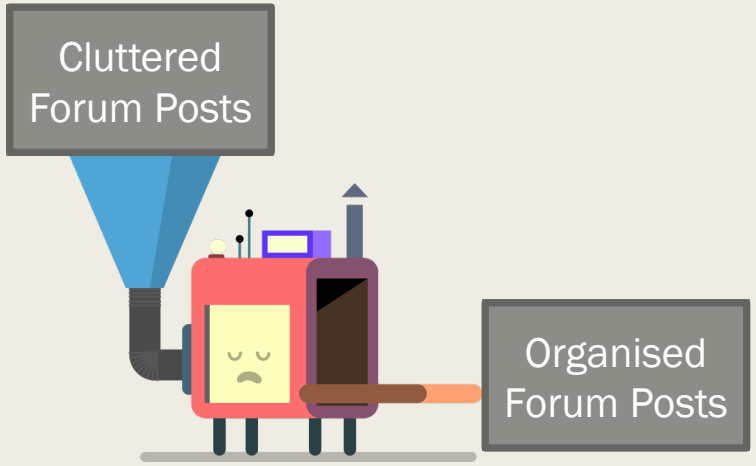
DSIF9 Project 3 – Natural Language Processing

By: Ho Kit Fai

18th March 2023

Problem Statement

- Let's Play Game Forums Pte. Ltd. hosts the largest discussion forum for RagnarokOnline and Maplestory – the hottest game at the moment
- Majority of the forum users are primary/secondary school kids that does not organize their discussions and some users are frustrated at the confusing game discussion, and unfortunately Let's Play Game Forums did not separate the game discussions into subforums at the beginnning
- Let's Play Game Forums would like to leverage machine learning to classify the forum discussions in order to help the users understand which game they are discussing on

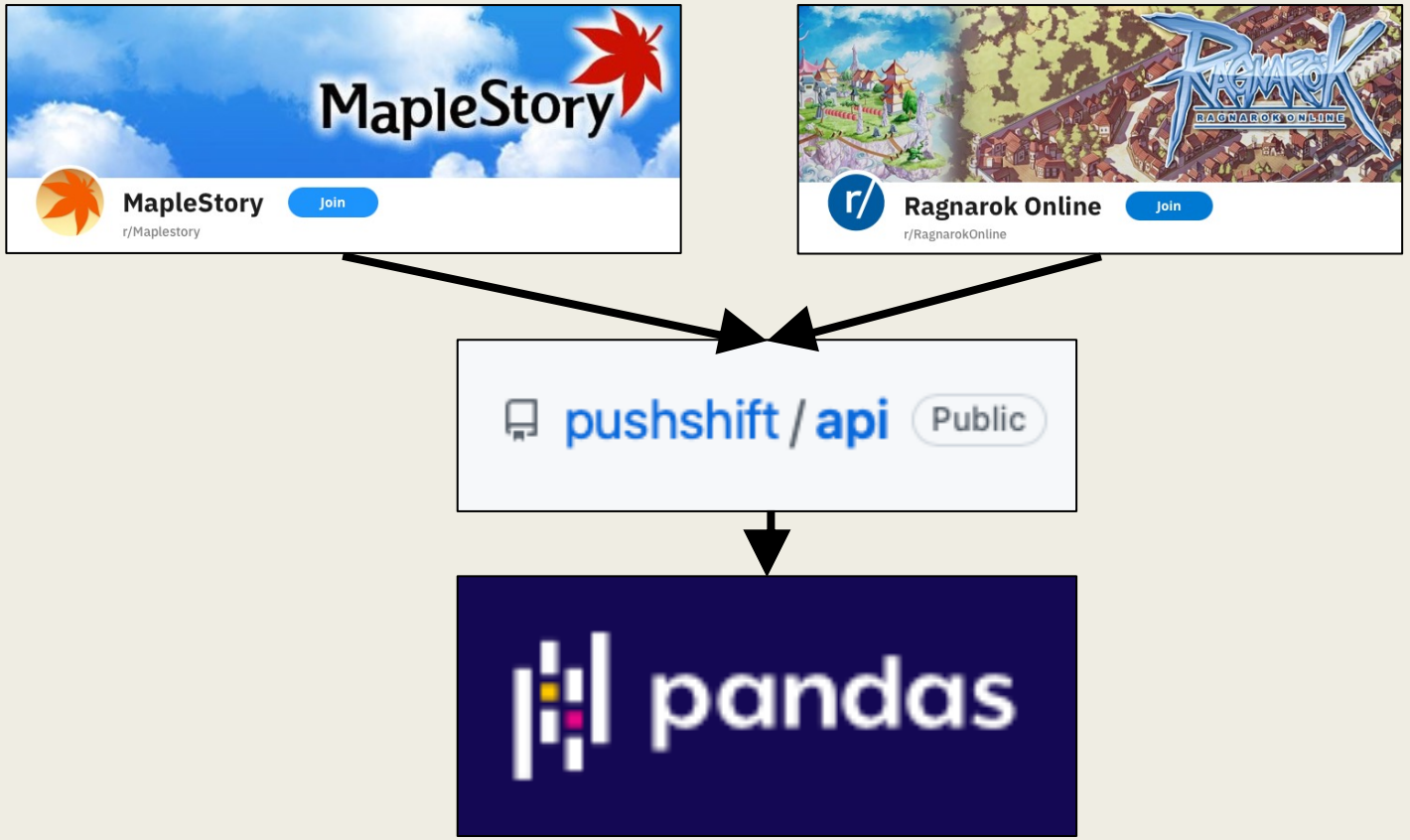


Data Science Process

1) Data Collection – Pushshift API on Reddit's r/RagnarokOnline and r/Maplestory
2) Data Cleaning
3) Exploratory Data Analysis
4) Preprocessing
5) Modelling – combination of transformer (CountVectoriser & TFIDF) and estimator (Naïve Bayes & Logistic Regression)
6) Evaluation
7) Conclusion
8) Future Work

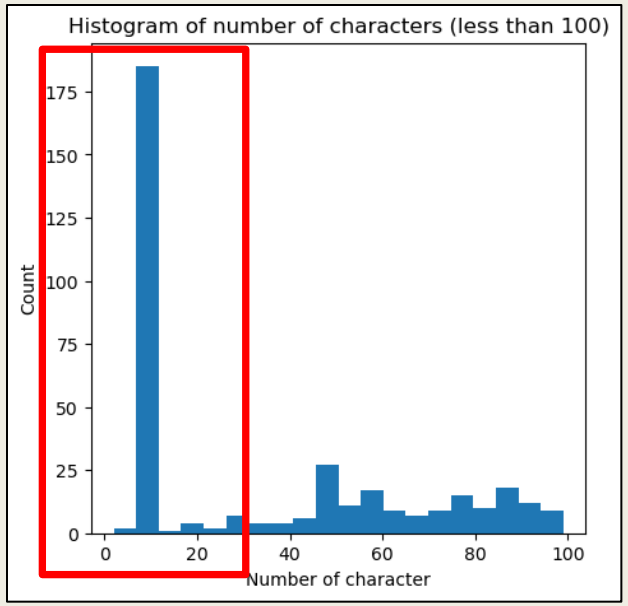
Data Science Process – Data Collection

- Leverage existing Pushshift API to pull posts from r/RagnarokOnline and r/Maplestory, the loaded into a Pandas Dataframe to be further processed

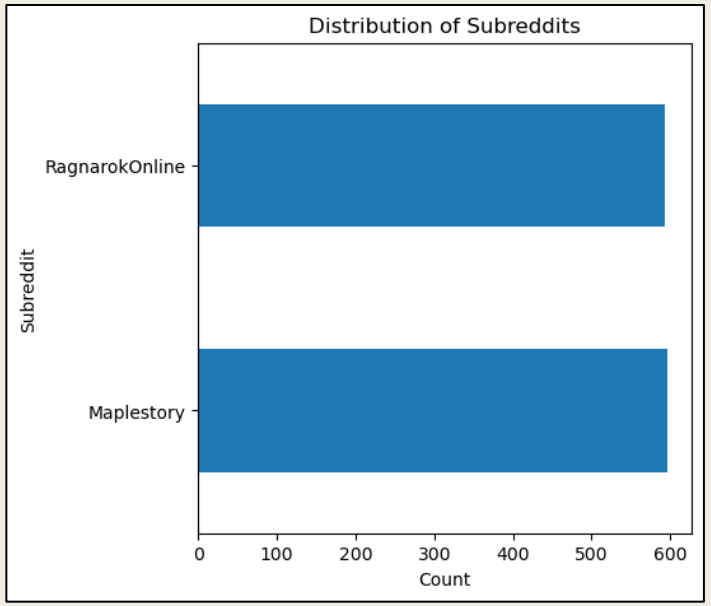


Data Science Process – Data Cleaning and Exploratory Data Analysis

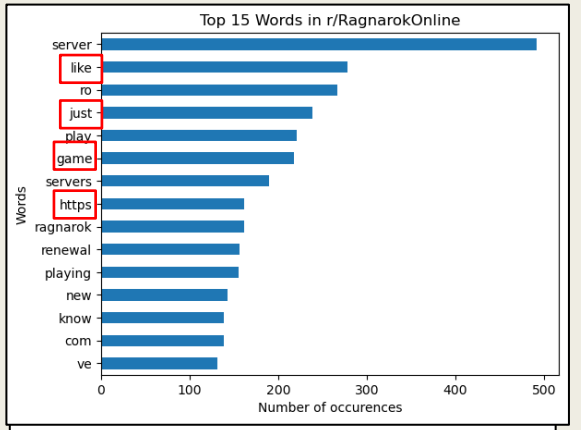
- Converted strings for main content (“selftext” column) into number of characters, then look out for anomalies

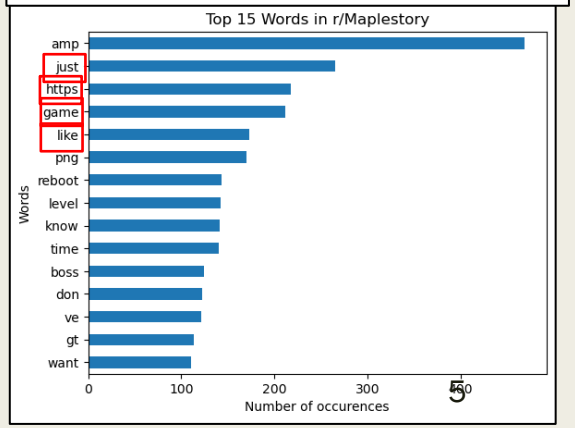


- Checked distribution of subreddits post cleaning



- Tokenised and lemmatized ”selftext” column and explored the top words





Data Science Process – Preprocessing, Modelling, and Evaluation

- Mapped the subreddits into 0 and 1
- Split data into training and testing

- Used GridSearchCV (5 folds) to find best parameters for the below models
- Accuracy is the main metric for evaluation. We would recommend Model 6.1 or 6.4 as they have a good accuracy score, and both models are computationally inexpensive

Model	Transformer	Estimator	Train Score	Test Score a.k.a. Acccuracy	Precision	Recall a.k.a. Sensitivity
6.1	Count Vectoriser	Naive Bayes	0.9900	0.9008	0.9115	0.9133
6.2	TFIDF	Naive Bayes	0.9787	0.8931	0.9144	0.9184
6.3	Count Vectoriser	Logistic Regression	0.9987	0.9059	0.8922	0.8878
6.4	TFIDF	Logistic Regression	1.000	0.916	0.8905	0.8827

Data Science Process – Some Observations

1) All models adopted max_df of 0.85, which means the transformers ignore terms that appear in more than 85% of the documents

2) All models adopted min_df of 2, which means the transformers includes terms that appears in at least 2 documents

3) Naive Bayes prefers transformers that adopts n_gram of (1, 2), which means 1~2 words are considered in the classification

4) Logistic Regression prefers n_gram of (1, 1)

5) All models performs best with "English" stop words

Data Science Process – Conclusion and Future Work

- The Model 6.1 and Model 6.4 achived an accuracy of 90% and 92%, which is significantly better than the baseline model (around 50%, it's as good as flipping a coin).
- The client is able to use the models developed to clean up their forums and classify the posts in the forum to help the forum achieve a better user experience.

- The dataset used is only 1000 posts per subreddit, so the gridsearch work is manageable. We will consider using more subreddits of similar contents to create a better classification model at the expense of some increase in computational requirements
- The client can consider creating subforums in their system for new and upcoming games in order to keep the forums organised from Day 1
- The NLP model can be further extended to create forum tags to further subdivide and improve the topic search for the users