DSPM Final Report

This is the final report of the NLP Task of Consumer Post on Samsung Galaxy s8, iPhone 8, and iPhone x from Twitter and non-Twitter review social media posts.
Collaborators: Yuwei Zhu, Zirui Zheng, Chenyu Shen
Date: 10/16/2019

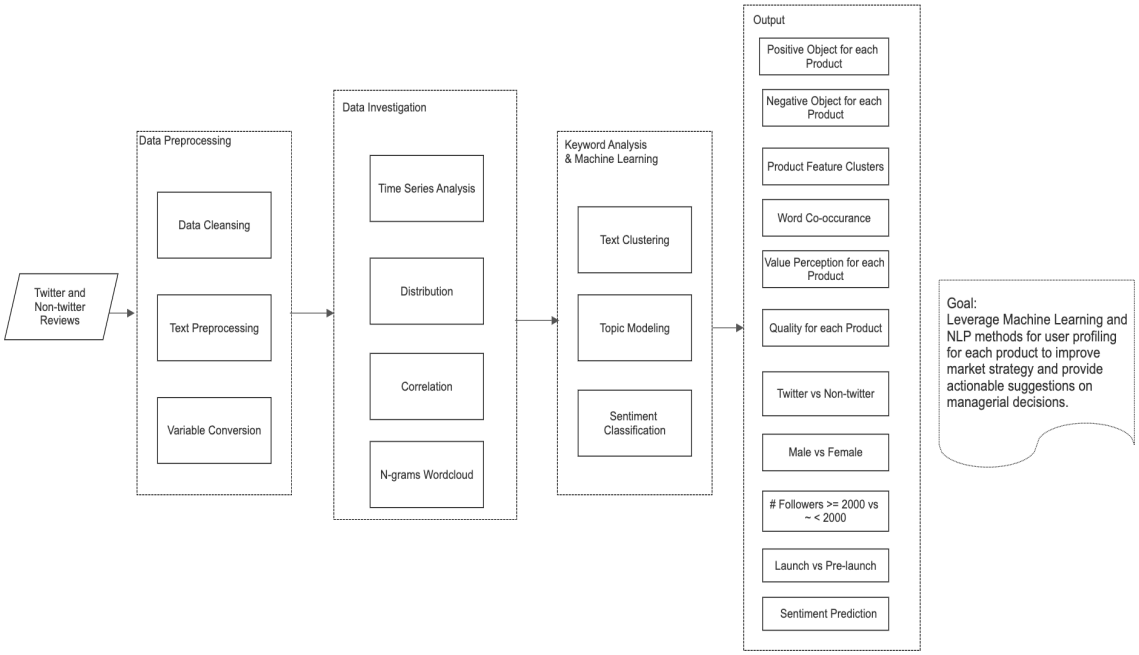
1. Introduction

1.1 Problem Statement and Task Goal

The task is to analyze the social media dataset regarding Samsung Galaxy s8, iPhone 8, and iPhone x from Twitter and non-Twitter source, snippet of those proceed and after the launch of these products. With the goal of user profiling to improve the product marketing strategy in mind, we adopt various analytical methods, machine learning and NLP techniques to conduct an in-depth research on the social media dataset.

1.2 Analytics Pipeline

The methodologies and analytics pipeline are defined as follows. With the social media dataset input, the first task is to conduct data investigation with comprehensive time series analysis, user profile analysis on their demography frequency distribution, correlations of features, and n-grams word cloud count on various levels. The second task is to conduct keyword analysis and deploy machine learning and NLP techniques, including text clustering on review text level, LDA topic modeling, and sentiment classification and prediction from various perspectives.



Flowchart of the analytics pipeline

2. Data Preprocessing

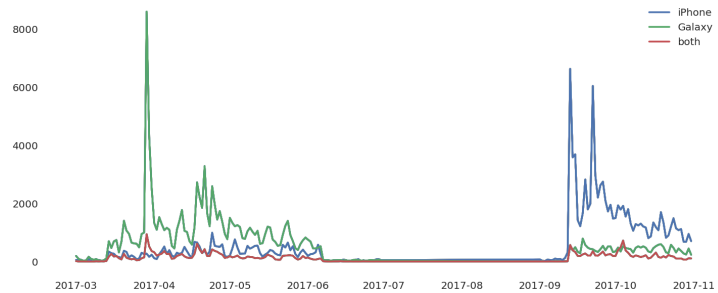
- Step 1: General Data cleaning
  - Drop “Sound Bite Text” NaN rows
  - Drop duplicate rows
  - Drop “Sound Bite Text” = “Post deleted by the author”
- Step 2: NLP Data Preprocessing
  - Drop special characters (e.g. punctuations, tags, url)
  - Stop words removal for specific business problems
  - Tokenization
  - Lemmanization
- Step 3: Labelling
  - Product type
    - Using keyword extraction
    - Labeling each post as ‘iPhone’, ‘Galaxy’, or ‘both’
  - After/Before product launch
    - Using comparison between post time and launch time
    - For “both” type, labeling post between “Mar-29-2017” and “June-6-2010” as “After” and post after “Sep-22-2017” as “After”. Otherwise, labeling as “Before”

3. Exploratory Data Analysis

3.1 User Profile[1]

In this dataset, we have 47319 male users, 14058 female users and 265851 users with unknown gender. Users come from 192 countries with 196 professions and 546 interests. The maximum number of followers is 127,774,881 while the minimum number is 11.

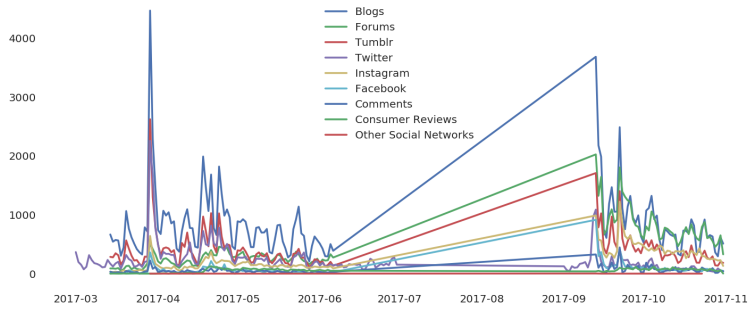
3.2 Time Series Analysis[2]



The number of posts all had two peaks for iPhone and Galaxy. Interesting thing is, Galaxy S8 reached its second peak after the product release nearly 1 month. And this peak was far less

[1] This part of code contains in “clustering-text-document-kmeans” notebook
[2] This part of code contains in “topic_modeling_wordcloud.ipynb” notebook

than the first peak. However, iPhone's two peak reached similar number of posts. More interesting thing is the first peak happened before the product release. Usually, information leakage happened before the new iPhone release. This might be the reason of the first peak.



For different platforms, the fluctuation of number of posts looks similar, which means there is no significant different between platforms when it comes discussion trend. (It doesn't not mean customers on different platforms present similar attitude towards these two products. For details, we need further analysis)

3.3 Word Cloud[3]

We use word cloud to see if the keyword has been changed after product launch.

iPhone

There is no significant difference between before and after the product release in terms of the keywords for iPhone 8 & X. Because one week before the launch, information about the new iPhones already released.

Before	After
	

Galaxy

After Galaxy S8 released, infinity display, pixel, Bixby became keywords of posts, which means Samsung marketing strategy successfully delivered these features to customers.

[3] This part of code contains in "topic_modeling_wordcloud.ipynb" notebook

Before	After
	

4. Text Analytics and NLP Tasks

4.1 Topic Modeling[4]

1) Methodology

Topic modeling is a type of unsupervised modeling for discovering the abstract “topics” that occur in a collection of documents. Latent Dirichlet Allocation (LDA) is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

Here we use the LDA topic model to figure out what are the key topics that our customers were talking about on social media. Since our interests focus on customers feedback after the product launch, we only use data that has been labeled as “After” product launch to train the LDA model. Also, we use the “product” column to separate post on products and train three models, respectively.

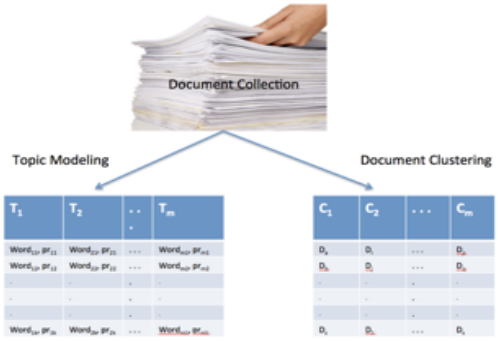
To find the best model, we used the grid search to identify the number of topics. The range we have been searched for each model 2-10[5]. All three models get the optimal number of topics at 5.

After finding the best model, we apply the model to each post and then use the maximum score for each post to assign them into dominant topics. Based on the assignment, we are able to observe post contents under each topic and generate corresponding word cloud.
Please note the difference between topic modeling and text clustering (illustrate later). Below image clearly shows the difference between these two models. In general, the result of topic

[4] This part of code please refer to notebook: “topic_modeling_wordcloud.jpynb”

[5] You may find in the notebook the search range are 2-7. We first run the grid search with range of [5,10] and get all the best model with number of topics = 5, then we re run the search with range of [2,7] to validate our result. However, running such grid search takes long time, so for the second time we use range [2,7] instead of [2,10].

modeling are groups of words co-occurred frequently; clustering gives us group of users that their posts are similar.



Difference between topic modeling and document clustering. Pic from: https://iksinc.online/2016/05/16/topic-modeling-and-document-clustering-whats-the-difference/

2) Result

A. iPhone

When people talk about iphone X & 8, the most common topics are cameras, ios system, battery (there was a swollen battery scandel for iPhone 8) and wireless charging. Post in topic 4 are most meaningless posts, more like ads for selling Apple products.

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Counts	16651	9980	10579	8393	10291
Focus	camera	system	battery	charing	Meaningless posts
Word cloud					
Sample text	DxOMark gives Pixel 2's camera a record score of 98, four points above iPhone 8 Plus and Note8...[6]	For those of you who are using iOS 11 on your iPhone or iPad, it 's time to check for an update again.	A fresh case of Apple Inc's new iPhone popping open due to a swollen battery has been reported in state media in China	If you're wondering whether your car's wireless phone charger will work with this year's iPhones, and Apple support document has the answer.	Apple iPhone 8 64gb por 3.750,00

B. Galaxy

For Samsung Galaxy s8, people like to discuss it's infinity display screen, system and the new Bixby virtual assistant. Similar to the result we got from iPhone, there are two topics talking about meaningless stuff. Topic 2 contains posts that took pictures using Samsung Galaxy. And

[6] For more sample text, please check the notebook.

the topic 4 are advertisement for screen protector (might because people are concerned about the infinity display screen).

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Counts	32797	27825	16843	10279	15516
Focus	Infinity display	system	picture	Bixby	Screen protect
Word cloud					
Sample text	Rumors suggest the Pixel 2 XL will feature a QHD+ display with an 18:9 (2:1) aspect ratio, 4GB of RAM, a 12-megapixel rear camera	you may have noticed how the Galaxy S8 automatically sets your default apps the first time you choose them from the Android intents list.	Captured by Samsung Galaxy S8+	With the latest Samsung smartphones and DeX Station, Amazon WorkSpaces and other services from the Ingram Micro Cloud Marketplace	The secret to Defense Clear's ultra protective design is two layers of shock absorbing rubber.

C. Both

Posts mentioned both products gave us some interesting results. Under this category, people comparing two products. From the topics they were talking about we can understand the key features people care about when comparing two smartphones.

First, post belongs to topic 1 looks like posted by iPhone lovers. They simply love the product because of the brand. Post in topic 0, 2 and 3 all compared two products. Posts under topic 0 just generally compare. They seldom mention a specific feature. Similarly, posts under topic 2 expressed their intention to switch but still didn't mention features they cared. Posts under topic 3 focus on compare cameras in both products. For topic 4, it's a group of posts that comparing a lot of features in both products.

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Counts	4292	6068	6905	1849	4158
Focus	General compare	Love iPhone	Consider switch	Compare camera	Features
Word cloud					
Sample text	Probably gonna get a new phone Samsung Galaxy S8 or Iphone 7	That would be an extreme downgrade switching from the iPhone 1 to the Galaxy S8. Any iPhone, even a dead one, is better than an S8.	Should I switch to a galaxy s8 from the iPhone 7+??	Galaxy Note 8 ties iPhone 8 Plus in camera tests	5 Features iPhone 8 Should Steal from Samsung Galaxy S8

Keyword Network Graph

Using the word co-occurrence matrix, we also build a keyword network graph to visualize connections between keyword. However, the result was not satisfying because connections weight between top keywords are quite similar. If you are interested in the keyword network graph, please check the last part of “topic_modeling_wordcloud.ipynb” notebook.

3) Takeaways and Recommendations

From the topic modeling result, we had three interesting findings:

- Clearly, people tend to discuss certain features for each product. For iPhone, people are interested in discussing camera, system and the wireless charging. For Galaxy, infinite display, system and Bixby are the top topics among discussions.
- There are some people are iPhone lovers. Their posts showed a different pattern from other posts. Therefore, we might want to dig deeper to check if there are any other groups of people worth noticing.
- iPhone lovers posts are more positive than others. However, we cannot get people’s sentiment simply from keyword extraction. Thus, our next step of exploring this dataset is sentiment analysis.

4.2 Sentiment Analysis and Classification[7]

1) Goal

We have two main goals for Sentiment Analysis. First, we want to look at the sentiment differences from pre-launch to post-launch of different products. Second, by looking at sentiment differences under different groups of people, such as genders, different social media platform users, etc., we want to find out the product adoption patterns among these groups.

2) Methodology

In order to prepare the data, we need to do two things: 1) Sentiment Score labeling 2) Data clustering to separate data for different products.

A. Sentiment Score[8]

For sentiment analysis, since it will be almost impossible for us to manually label the data, we decide to use an existing NLP tool package in Python. As we cared more about efficiency and accuracy of the sentiment, we picked TextBlob. This tools used a rule-based learning to identify the sentiment of a sentence and returns a polarity score from -1 (worst sentiment) to 1.

[7] see ‘Sentiment.ipynb’

[8] *Reference: https://www.iflexion.com/blog/sentiment-analysis-python*

B. Clusters

In order to separate data into apple_data and samsung_data, we used CountVectorization and Latent dirichlet allocation (LDA). Besides, we also used regex matching to extract iPhone X and iPhone 8 data for further analysis. Details are shown in the file 'Sentiment.ipynb'.

3) Results

A. Pre-Launch vs Post-Launch[9]

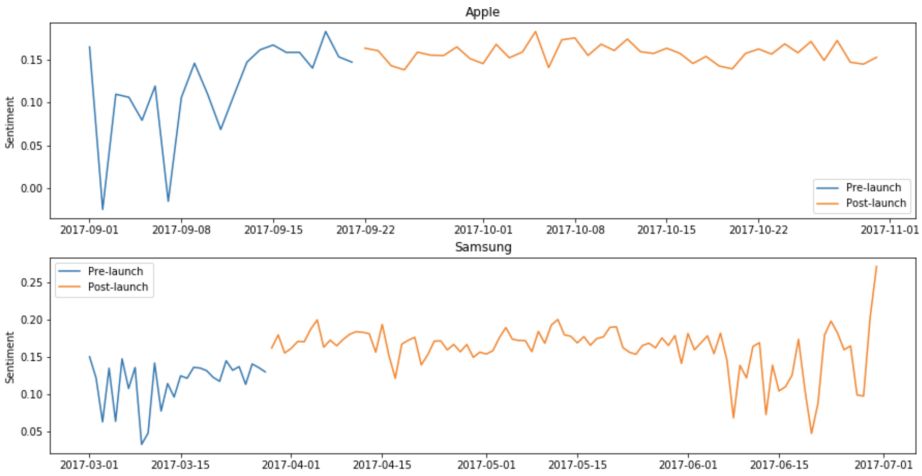


Figure: Sentiment score trends for Apple and Samsung around its launch date.

The Mean Sentiment score of each product before and after launch. See 'Sentiment.ipynb' for plots of iPhone X and iPhone 8.

Sentiment / Product	Samsung Galaxy S8	Apple	iPhone X	iPhone 8
pre-launch	0.1265	0.1556	0.1522	0.1558
post-launch	**0.1675**	0.1568	0.1499	0.1570

Samsung enjoys a 32% increase in sentiment after launch. We've looked into the press release and public reactions during that time. As S8 was the first release since Samsung's global recall of the 'exploding' Note 7, the sentiment before launch was as a result pretty low. However, with innovative design and display, Samsung earned a lot of positive feedback among the public.

Apple has a smooth sentiment with no significant differences before or after launch. It looks quite surprising given the revolutionary design of iPhone X. However, many newspapers and reports were actually showing negative attitude towards Apple during that time, basically saying that all features were the same as what was leaked, iPhone 8's position was awkward, the price was too high, etc. which made the sentiment look reasonable.

---

[9] *Reference:* *https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b*

In addition to visualization, we ran a traditional Time Series model using ARIMA with grid search for the best parameters. However, the results didn't show close matches for the prediction for the two products.

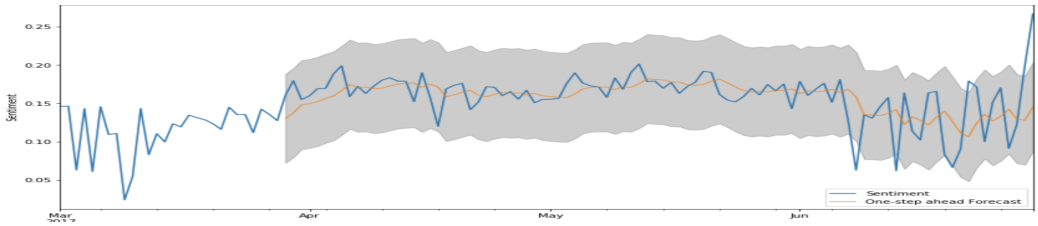


Figure: The One-step ahead Forecast for Samsung post-launch sentiment. (blue line is the initial data, and orange line is the prediction. 90% Confidence Interval is shown with grey zone.

We looked into the data and concluded that the traditional Time Series forecast might not work well since our dataset lacks data before launch. In the future, our next steps will include Time Series forecasting with more advanced algorithms and more complex dataset.

B. Professionals/Celebrities vs Normal Users

We did not detect significant differences between Professionals/Celebrities vs Normal Users. However, Professionals/Celebrities has a high standard deviations compared to the other group, which is a result of their deliberate extreme / biased opinion in public. More details and graphs can be found in 'Sentiment.ipynb'.

C. Female vs Male

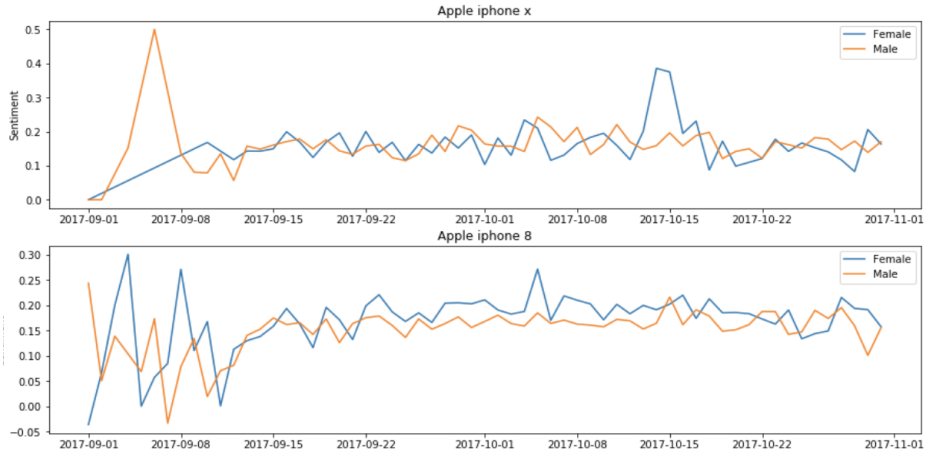


Figure: Difference of sentiment across different genders for iPhone X and iPhone 8.

We didn't notice a significant difference between genders in Samsung, but female users prefers iPhone more than males do, especially iPhone 8. We looked into the word frequencies for different genders when they talk about iPhone.

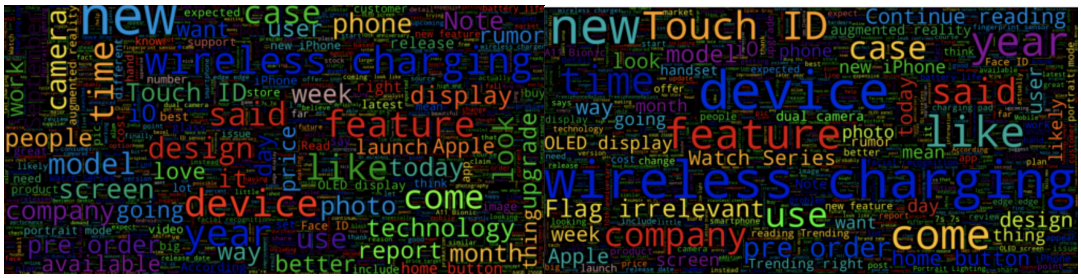


Figure: Female word cloud for Apple (left) and male word cloud for Apple (right)

Unique keywords Female is talking about: display, photo, camera...
Unique keywords Male is talking about: Touch ID, OLED display, wireless charging...

Apparently, females and males have different focus on iPhone, and it would be worth for PM to invest in this segment and to direct target different genders with different selling point.

D. Twitter vs Other Sources

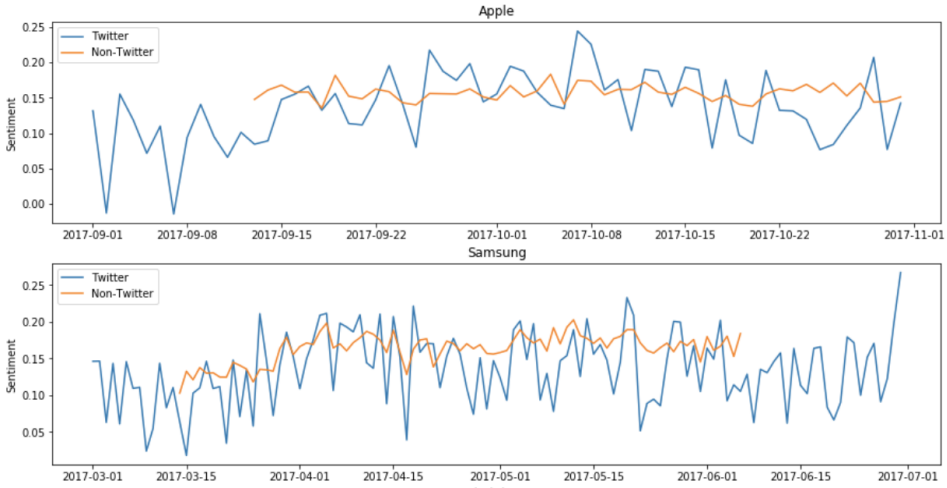


Figure: Difference of sentiment across users from different platforms.

It's interesting to notice that Twitter users have significantly lower sentiment mean and higher standard deviation. It alerts us that when using social media data to analyze product, we want to consider different types and profiles of the platforms.

4) Takeaways and Recommendations

The Sentiment Analysis shows the importance of customer segmentation during product design and product marketing, as all kinds of groups of people tend to have different adoption patterns for a new smartphone. It is necessary for PMs and the whole team to clearly portrait their customer profiling, to thoroughly understand the user needs for different groups, and to target them with specific selling points. This can be an essential step in user acquisition and profit optimization.

4.3 Text Clustering[10]

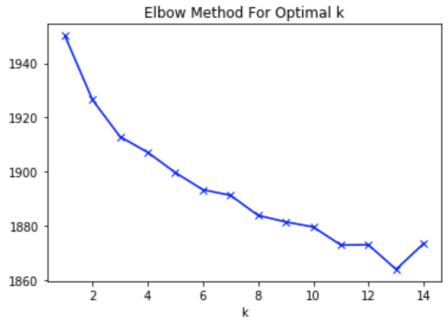
1) Goal

The goal of text clustering is to identify the quality, price, and value perceived by users for each product. The clustering result provides insights into the market segmentation and sheds lights on the user profiling.

2) Methodology

Clustering is applied on the text review level to refine the user profiling findings for each user cluster. The reason behind choosing the review text as feature is that the review post reflects users' perception of the quality, price, and most importantly, the value of each product.

The initial assumption of the optimal number of clusters is 6. Since there are three products and the features for each product that would be most actionable are their value and quality, the optimal number of clusters is assumed to be 6 (eg. cluster one: value of samsung; cluster two: quality of samsung; cluster three: value of iPhone 8; cluster four: quality of iPhone8, etc.). After applying the statistical method of the sum of squared distances, the optimal number of k is set to 6. The text input is encoded as the tf-idf vector for each document (each review text), with stop words removed, and the clustering algorithm is k-means clustering.

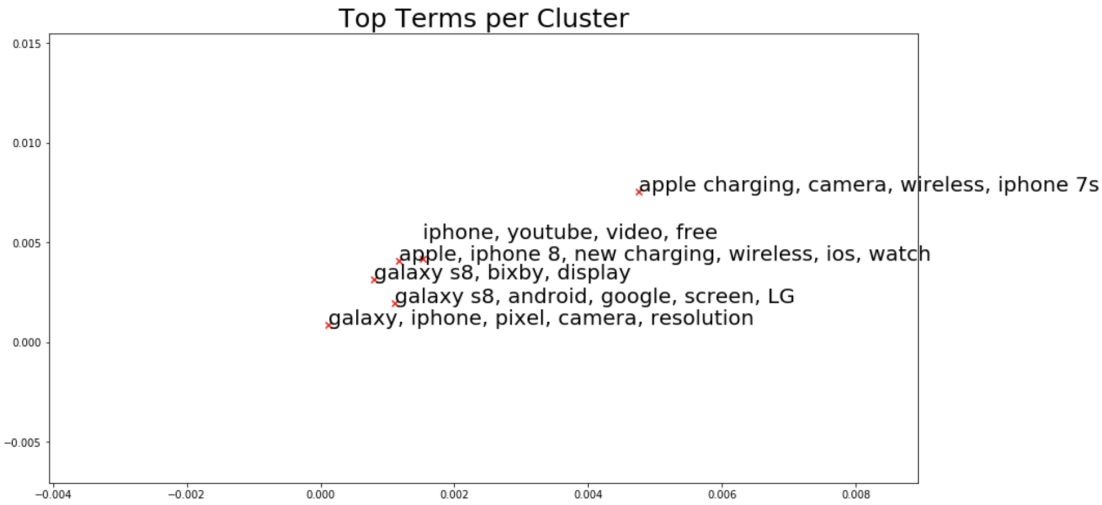


Sum of squared distances for different k on tf-idf vectors of text reviews

3) Results

The k-means clustering model result gives us the following extraction of text keywords for each cluster, the distance of each cluster from each other is marked as their centroid axis number.

[10] This part of code please check the "clustering-text-documents-kmeans" notebook



Top Terms per k-means Cluster

The cluster result gives us six clusters of which keywords of their review posts are summarized in the figure above. Each cluster has a specific focus of topic that sheds light on the quality and value of each product. Cluster one likes iPhone 8 plus's new wireless charging, iPhone x's OLED, but does not like the design of both model too much. Cluster two likes Galaxy's camera, and image resolution, but is not a big fan for Galaxy's Bixby. Cluster three compares iPhone 8 plus's portrait mode with Pixel, and does not like about Phone's high price. Cluster four has a mixed review on Bixby. Cluster five compares across different android smartphone models, Samsung Galaxy s8, LG, Google's Pixel. Cluster six focuses on the review for iPhone models, including 8 plus, x, and 7s, and they like about the new wireless charging, camera, Apple's brand, but generally don't like about the high price.

4) Takeaways and Recommendations

Clustering analysis as an unsupervised method provides the suggestions on the quality, and value feature for each product. By analyzing the user attributes of these clusters, product managers can look into the market segmentation. Another use case of clustering result is to look at the user clusters for different market country. For example, US has mostly cluster three and four, whose concern for iPhone is the high price, and the main concern for Galaxy s8 is the device issues that could link back to the massive recall. By gathering more user information, product managers are able to identify the market segmentation and user profiling for each product.

Work Distribution:

- Yuwei Zhu: Topic modeling, EDA, data preprocessing, presentation deck & final report
- Zirui Zheng: Sentiment Analysis, EDA, data preprocessing, presentation deck & final report
- Chenyu Shen: Text clustering, EDA, data preprocessing, presentation deck & final report