Analytical Approach to Content Creation

Introduction

With the rise of digital marketing and social media, content creation is key to grabbing the attention of your target market and adding value to your existing customer base. Without the proper tools and processes in place to capture the attention of your audience, companies fall prey to their competitors. Realizing the importance and difficulty of creating content, I pursued this project to develop an analytical approach to the content creation process.

Problem Statement

This project will primarily be concerned with developing insights about content for the mobile device industry. Specifically, I will focus on the following problems.

- *Identify the trending topics that news outlets and technology bloggers are talking about.* The trending topics will help narrow content creation to a select group of topics that readers are interested in.
- *Identify what top technology related problems customers are currently searching online.* By identifying what problems customers are facing, better insights can be gained on which content will be more helpful.

Data Collection

The analytical approach for this project starts with identifying and collecting the appropriate data. To ensure that we have the proper data to derive quality insights, the project focused on two types of free form text: technology blogs and google queries.

Technology Blogs: Because the insights from this project will be used to drive customer facing content, it was important to identify blogs that were customer centric and focused on topics such as mobile trends and product reviews. The blogs selected for this project include: CNet, Gadget, Verge, ZDNet, Android Police, and Apple Insider. To collect data from each site, python's Beautiful Soup package was used to scrape articles over a one month period. Below is a summary of the selected blogs.

Name	URL	Blog Type	Number of Articles
CNET	www.cnet.com	Product Reviews	41
Gadget	gadgets.ndtv.com	Product Reviews	160
Verge	www.theverge.com	Tech News	77
ZDNet	www.zdnet.com	Tech News	15
Android Police	www.androidpolice.com	Niche (Android)	81
Apple Insider	appleinsider.com	Niche (Apple)	146

Google Queries: Google Trends is a platform by Google that allows individuals to search and analyze popular google queries based on attributes such as keywords, geographical location, and timeframe. To access the data, the python package pyTrends was used. Query data was collected for the keywords "iPhone", "Android", "Samsung", "HTC", and "LG" over a one month period for all US users. To collect the query data, a two-step process was used. The steps included:

1. Search for queries that are directly related to the keywords of interest – specifically, the keywords "iPhone", "Android", "Samsung", "HTC", and "LG"
2. Identify other keywords similar to the keywords of interest, and use the other keywords to find queries indirectly related to the keywords of interest.

A total of 726 unique queries were collected. Below is a summary of the number of queries for each primary keyword.

Keyword Searched	Number of Related Queries
iPhone	342
Android	88
Samsung	143
HTC	79
LG	74

Data Wrangling

After data collection, the next step in the analytical process is wrangling the data into a format that is suitable for analysis. The extent and type of data wrangling that is needed is dictated by the analysis that will be performed on each dataset. To address issues around content creation, NLP will be used to analyze blog and query data. Therefore an extensive amount of pre-processing is necessary. First the data is normalized by performing the following steps: punctuation removal, lowercase conversion, tokenization, stemming, and stop word removal. Then the data is converted into a token count matrix or a term frequency-inverse document frequency (TF-IDF) matrix as needed by the analysis being performed. Below is a visual representation of that process.

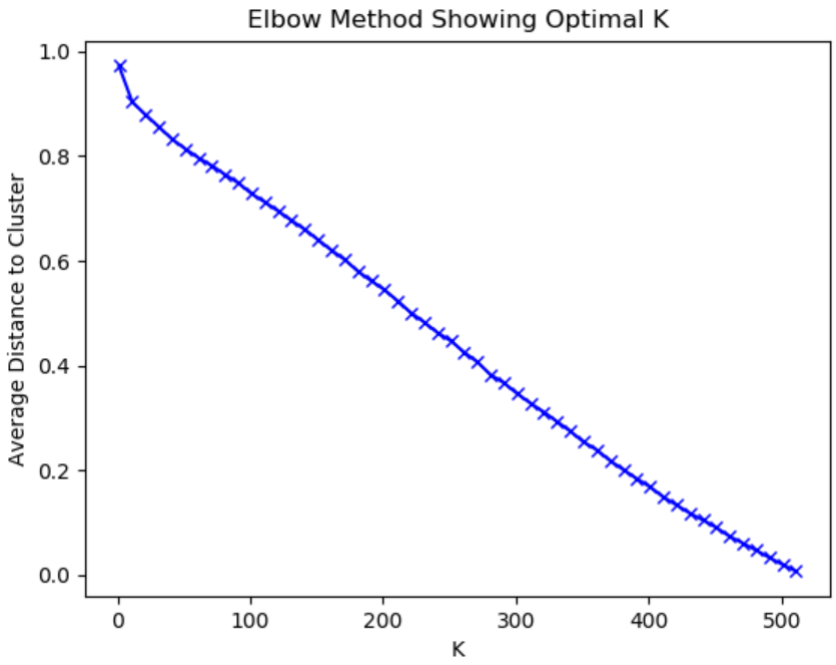
Analytical Approach

Once the data is wrangled and pre-processed it is ready for analysis. This section will further discuss the analytical strategies used to solve each issue, and the insights gained from each approach.

Goal: *Identify the main trending topics that technology bloggers are talking about*
Analytical Strategy:

1) *Given text from technology blogs use text clustering to group similar articles together.*
2) *Given the clusters of articles use topic modeling to identify the main topic for each cluster*

To identify the main trending topics that news outlets and technology bloggers are talking about, a two-step approach was used to analyze the technology blog dataset. First, K-Means clustering was used to group similar articles together. To determine the appropriate number of clusters (K), an elbow graph was initially plotted over a wide range of clusters. Because of the sparsity of the term-document matrix, the elbow graph proved to be ineffective. See graph below.



As an alternative, the following formula was used to identify the optimal K for text clustering [1]:

$$(\text{number of documents} \times \text{number of terms}) / \text{number of non-zero entries}$$

Many of the clusters that were returned from the K-Means algorithm only contained one document, and therefore was removed. Also a few of the clusters that were returned contained articles from a single source (i.e. A cluster with 5 articles from CDNet). Because the goal of the project is to identify topics being discussed across various blogs, these clusters were removed as well. After identifying and filter for the appropriate clusters, the second step in the process was to use Latent Dirichlet Allocation (LDA) to perform topic modeling on each cluster. This gave a better idea of the topic being discussed across each article within each cluster.

Given the technology blog dataset of 520 articles, the analytical approach described above was able to successfully identify 54 trending topics across various blog sites. The clustering algorithm was able to group articles around one of three central themes: current events, product features, and general product news. Due to the specificity of their articles, clusters centered on current events and product features were more likely to be better suited for content creation. Each cluster was able to discuss a single topic from different perspectives. Unlike its counterparts, clusters centered on general product news were less likely to be a great source of content. Many of those clusters touched on various topics, with the only thing in common between articles is the device being discussed. These clusters also tended to be much larger, than other clusters. It was not uncommon for these clusters to have 10 to 15 articles, whereas other clusters would have 3 to 7 articles. Below is an excerpt of some of the clusters that were identified.

Cluster 1

Article Title 1: Apple Watch Series 4 reportedly had some trouble with daylight saving time
Article Title 2: Apple latest watch is crashing and rebooting due to daylight saving time bug
Article Title 3: Apple Watch Series 4 suffers from daylight savings bug rebooting loop in Australia
Topic: apple, watch, time, daylight, issue, series, savings, Australia, hour, reboot

Cluster 2

Article Title 1: Google Assistant now helps you compare ride-hailing prices and summon a car
Article Title 2: Google Assistant redesign brings bigger visuals, new controls, and more
Article Title 3: Redesigned Google Assistant settings page is rolling out
Topic: assistant, google, new, voice, developers, settings, smart, user, services, design

Cluster 3

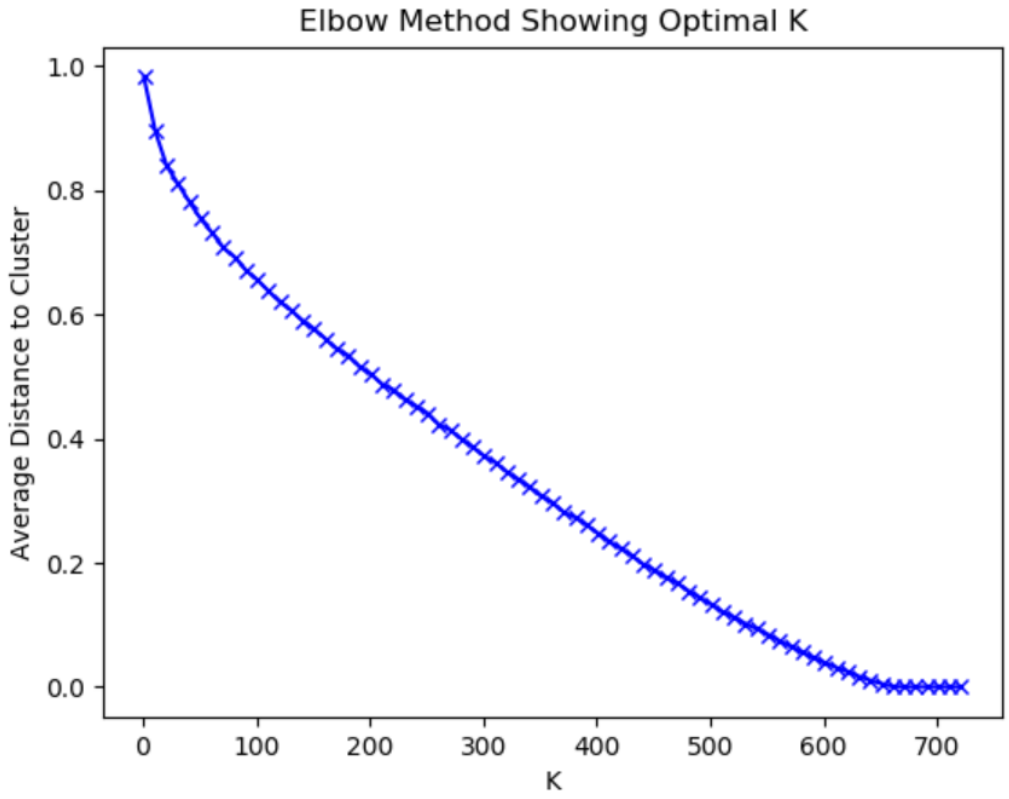
Article Title 1: How to delete your Google+ account
Article Title 2: Google is shutting down Google+ for consumers following security lapse
Article Title 3: Google hid major Google+ security flaw that exposed users personal information
Article Title 4: Google is putting Google+ out of its misery following data exposure
Article Title 5: Google+ shutting down in wake of allegations of weak user data security
Topic: google, data, users, access, company, apps, bug, profile, api, secure

Goal: Identify top technology problems users are searching online

Analytical Strategy:

1) *Given search query data for terms of interest (i.e. "iphone") use text clustering to group similar search queries together. Cluster size will be used to judge prevalence of cluster.*
2) *Given the cluster of queries use topic modeling to identify the main topic for each cluster*

To identify the top technology problems users are searching online, K-Means clustering was used to cluster similar search queries together. The queries used were the top queries entered in Google over a one month period, given a keyword of interest. Therefore cluster size will be a key indicator of the cluster topic's prevalence. To find the optimal K for the K-Means clustering algorithm, an elbow graph over a range of K values was used. Due to the sparsity of the data, the plot displayed no clear elbow. See graph below.



Because the elbow graph did not display a clear elbow, the formula below was used to determine the optimal value for K [1]. The formula is the same as the one used in the previous section.

$$(\text{number of documents} \times \text{number of terms}) / \text{number of non-zero entries}$$

The optimal number of clusters for the dataset was calculated to be 362. After the clusters were created, they were then filtered to keep those with two or more search queries. The number of queries per cluster is an indicator of how relevant that cluster's topic is to online users. Clusters with one query are least prevalent, and therefore was removed. This step reduced the number of clusters from 362 to 180 clusters. For the last step, LDA was used to model the topic for each cluster.

From the query dataset, which contained 726 queries, 180 prevalent issues were identified. The K-Means algorithm did a great job of isolating less common queries from the dataset. Many of the problems that were identified were iPhone or Apple related. This is less of a surprise than it is an observation, given that nearly half of the queries are related to Apple products. Below are some of the problems that were identified.

Cluster 1

Query 1: iPhone backup failed
Query 2: How to manually backup iphone
Query 3: How to backup contacts on iphone
Query 4: Where is backup on iphone
Topic: backup, iphone, manually, failed, contacts

Cluster 2

Query 1: How to find photos on iCloud
Query 2: Hot to retrieve photos from iCloud
Query 3: How to view iCloud photos on iphone
Query 4: How to turn off iCloud library
Query 5: Restore phots from iCloud
Topic: iCloud, photos, turn, library, view

Cluster 3

Query 1: How to screenshot on iphone xs max
Query 2: How to screenshot on xs max
Query 3: Screenshot on iPhone xs max
Query 4: Screenshot on xs max
Query 5: iPhone xs max screenshot
Query 6: How to take a screenshot on xs max
Topic: max, screenshot, iPhone

Cluster 4

Query 1: How to factory reset on lg phone
Query 2: Factory reset lg phone
Topic: reset, factory, phone

Conclusion

Data analytics can be a very powerful tool in the area of content creation. To address the problem of identifying trending topics and customer problems, this project relied heavily on data mining and text clustering. Although the project focused on the mobile device industry, the analytical approach to content creation described above can be applied to any industry.

Bibliography

[1] Can, F.; Ozkaranhan, A. "Concepts and effectiveness of the cover-coefficient-based clustering methodology for text database", 1990