# Project Nokia Competitor Intelligence Text Mining

## Using Latent Dirichlet Allocation (LDA) Method to Analyze Customer Reviews

*Final Report*

Wesley Clark, Candice Murphree, Evan Elg, Molly (Boland) Van Gieson

*Authors Note:*

Individual correspondence of participating members can be addressed at

Email:Wesley Clark, iamwesleyclark@gmail.com,

Candice Murphree, cmurphree@student.umuc.edu,

Evan Elg, elgevan@gmail.com,

Molly (Boland) Van Gieson, 62boland@cua.edu

## Introduction

In today's ever evolving financial world, Nokia is looking to increase effectiveness in positioning their product line, organizing the company as a whole, and obtaining a competitive advantage. This knowledge is especially valuable as pertaining to business intelligence and marketing. As more customer reviews and text documents have become available to Nokia, the information contained within has proven to be a valuable asset. In this assignment, Nokia has entrusted the Data Action Response Team (DART) to employ the latest web mining techniques related to reviews of the Samsung Galaxy S5. Unstructured data makes up 80-90% of all information and consists mostly of text according to research by Seth Grimes, a leading industry analyst who specializes in uncovering natural language processing (Grimes, 2008). Nokia's business intelligence division requires consumer insight from unstructured data that is mined from the text of product reviews. The DART analytics team is tasked with the Competitor Intelligence Text Mining project to mine a previously extracted data set and transform the data into actionable insight for Nokia. Insight into the features and sentiment of the new release of the Samsung Galaxy S5 is required to compare the Nokia features against leading competitors' products with the goal of extracting a competitive edge to entice consumers to choose Nokia devices. The business objective of the project is to identify four major topics reviewed for the Samsung Galaxy S5 phone. Additionally, this information will be used by Nokia to decide the trajectory of their products, as well as how to allocate their resources.

In order to accomplish this task, the DART analytics team chose to utilize the natural language processing algorithm of Latent Dirichlet Allocation (LDA). LDA is often used for object categorization, segmentation and topic modeling, which can be used to conduct the text analysis

of the Samsung Galaxy S5 reviews (Hu, n.d.). LDA breaks down each text through two drill downs: the first is to look at the topics that compile the text under analysis and the second is to consider the words that build into each topic within the text under analysis (Robinson, 2017). Under this algorithm, the data scientist can set the number of topics for which the algorithm should search. In this case, we chose to analyze the text under the premise of four major topics to avoid over-fitting of the model but to provide for precision that would be useful as we move forward in analysis of the reviews.

## Analysis and Demonstration of Model Development

The original data set consisted of 220 individual text files originating from Wikipedia news articles (Gialampoukidis, Vrochidis, & Kompatsiaris, 2016). The files consisted of fourteen different topics and originally gathered using clustering data mining. Our team collected 35 files relevant to the Samsung Galaxy S5 by searching for files in the working directory that began with the word Samsung. Using the file collected in this search, the actual text from each file was read into a list with each item of that list containing the text from a file. Before transforming the data into a term frequency matrix, our team conducted several preprocessing steps.

The purpose of these preprocessing steps is to remove irrelevant words and characters, consolidate the text, and facilitate analysis. All words were converted to lowercase so that identical words with different cases were not identified as separate words. Our team removed all punctuation as well as extraneous white spaces. Using a list of *stop words* available online, we removed common *stop words*, or words that occur frequently in the English language but are not of analytic interest, such as “the”, “it”, or “and”. Our team also conducted the practice of *word*

*stemming* to ensure that verbs and other types of words that can have different endings were not identified as separate words. *Word stemming* is the practice of shortening words to their root. For example, if the words “fishing”, “fisher”, and “fished” all appeared in the corpus of documents, those words would all be converted to the root word, “fish” (Manning, Raghavan, & Schutze, 2009).

In our analysis, we tested models with and without word stemming. Ultimately, we chose against word stemming our data because the stemmed matrix did not result in topics that were more cohesive and can cause some loss of information and interpretability.

After these preprocessing steps, our team converted the text files to a matrix where every file is represented by a row and every word in the corpus is represented by a column. The number of times that a word is in a document is represented by the count at the intersection of the row (which represents the file) and the column (which represents the particular word). This matrix displays the text data in a numerical format that is compatible with the analysis method.

To analyze our processed data, we chose a form of topic modeling called Latent Dirichlet Allocation (LDA). LDA is an unsupervised learning technique that will examine a group of documents and discover latent topics within those documents. This type of analysis rests on two primary assumptions. First, it assumes that each document is comprised of a finite number of topics. Second, it assumes that each topic is composed of a particular distribution of words. Among other things, the result of the analysis will describe each document as a product of various topics. Each topic can be thought of as a distribution of word frequencies. Documents are matched to topics by comparing the word distributions of a document to the available topics. By comparing the distribution of words in a document to the relative word distributions in the

topics, each document is explained as a mixture of each of the available topics. After understanding the documents in terms of the topics, we can investigate the topics to develop an intuitive understanding of what each topic represents.

The assignment of word distributions to topics and topics to documents is a complex process but essentially boils down to the following steps:

1. Randomly assign each word in document to one of the *k* topics (where *k* is the number of topics chosen beforehand)
2. Go through each word in each document and
    a. Find the proportion of words in the current document that pertain to the topic currently assign to the word being examined;
    b. Find the proportion of assignments from the current word in all documents to the topic currently assigned to it;
    c. Use the Bayes theorem to reassign the current word to a new topic using the information calculated in the preceding two steps.
3. Repeat this process many times.
4. Acknowledge that the algorithm does not guarantee arrival to a global optimum nor will it necessarily stabilize; therefore, multiple iterations are advised to verify results.

The following parameters were initialized: burn-in, iterations, thin, seed, number of starts, sampling method, and most importantly the number of topics. The sampling method used in the LDA process is known as Gibbs sampling. The parameters burn-in, iterations, thin, and number of starts tweak the sampling process. Given that the LDA process begins by randomly

assigning words to topics and improving the model by using other randomly assigned words, it is possible to get a bad sample or start the analysis with a poor word distribution. Burn-in, iterations, number of random starts, and thin will help us avoid a poor sample and performance. The random seeds will make our analysis reproducible by helping us find the same set of random numbers used in our final analysis.

The most important parameter is the number of topics (see Appendix 1, Figure 1). LDA assumes that each document is composed of topics, but it does not tell us how many topics are in a corpus of documents. Furthermore, there is no precise method for determining how many topics are in a corpus. In order to select this parameter correctly, our team needed to understand the data and perform some initial testing. It is important to remember the purpose of our analysis, which is to analyze the reviews of the Samsung Galaxy S5. It is reasonable to assume that the number of topics in this type of article would be limited; our initial hypothesis was that there would probably be approximately five topics in this corpus. Several models were built with identical parameters except for the number of topics. After creating each model, the top words contributing to the makeup of each topic were examined to see if they represented a cohesive, sensible topic. This manual examination led us to believe that four topics was a reasonable choice for the corpus. This number of topics results in topics that are interpretable and diverse. In addition to the steps taken above, we created a visualization to better understand our choice in number of topics. The area of the circles is proportional to the relative prevalence of each topic in the corpus and the x-axis and y-axis are the two first principal components of the topics (Sievert & Shirley, 2014).

Examining this visualization (see Appendix 1, Figure 1) gives us deeper understanding

about the number of topics we chose. The circles have a similar area, showing a stable spread of topics throughout the corpus. Furthermore, the location of the circles plotted on the principal components of the corpus's distribution shows a good spread between the topics. The topics are largely distinct, with the exception of some overlap between topics 1 and 2. This spread indicates that the topics are diverse. If the topics overlapped significantly or frequently, we would worry about whether the content of our topics were truly unique.

## Results and Model Evaluation

After conducting the LDA method on the data, we had three main outputs. The first output is a table that details which links each text document to a particular topic. This output allows Nokia to determine which areas to dedicate research and development. The second output is a table detailing the top ten words for each topic. This output provides Nokia with more context for determining areas of improvement and success for the four identified topics. The third output is a table of the probabilities for each topic by each text document. This output informs Nokia of connections between the topics as well as what product functionalities customers associate together. Additionally, we developed an interactive visualization that provides a breakdown of the word relevancy by topic.

### Topics by Document

The first output is a table that aligns each document to a particular topic based on the highest probability of the topics for that document (Topics by Document output). Topic 1 (fingerprint function) is the primary topic for eight documents. Topic 2 (camera) is the primary topic for two documents. Topic 3 (comparing the Galaxy Samsung S5 to other phones) is the primary

topic for sixteen documents. Topic 4 (battery life) is the primary topic for nine documents. Nokia can use this information to focus research and development energies on these four topics, starting with the most prominent topic, Topic 3: how the Samsung Galaxy S5 compares to other phones, followed by topics 4 and 1: the battery life and the fingerprint function. Although this function is less of a priority than the other topics, some research can be dedicated to the phone camera.

## Top Words by Topic

The second output is a table with the most used words for each of the four topics (Top Words by Topic output). For all four topics, the word *Samsung* was one of the top ten words. For three of the fourth topics, the word *Galaxy* was in the top ten words. Another trend across the four topics was the use of comparative words, such as *compare, well, new,* and *like.* The four topics reflect that many of the reviewers seemed to have knowledge of other smartphones or devices to which they could compare the Samsung Galaxy S5 and were not first-time smartphone owners.

The first topic focuses on the fingerprint function of the phone. The top ten words for the first topic are *phone, like, one, can, fingerprint, back, Samsung, even, use,* and *time*. Reviewers did provide some positive feedback with regards to the fingerprint function, indicating that they liked it, used it often, and that it saved them time. The positive feedback from these reviews indicates that this feature is something that Nokia should continue to incorporate into its products. Moreover, the frequency (namely, that the fingerprint function was one of the top four topics for all the reviews) indicates that Nokia should also look at why this particular function came up. Nokia should also consider any constructive criticism from the reviews and incorporate necessary changes to ensure that Nokia can continue to market this feature successfully.

The second topic focuses on the camera function of the phone. The top ten words for the second topic are *camera, image, Galaxy, can, Samsung, good, low, light, well,* and *app*. The reviewers noted the image quality of pictures and videos taken by the Samsung Galaxy S5. They also reflected on the lighting associated with photography when using the Samsung Galaxy S5. This topic can help Nokia to continue to refine and improve its camera function on the Samsung Galaxy. Perhaps even more noteworthy to Nokia is the emphasis that reviewers place on the camera. The Samsung Galaxy S5 has better resolution on both the forward and rear facing cameras than Apple's iPhone 5s (Andronico, 2014). Is Nokia reaching photography-focused consumers with its marketing? Nokia should also use the review analysis to verify or determine if marketing strategies are reaching the intended audiences, and if they are not, determine why they are not. Apple markets the camera function of the iPhone lavishly, but Nokia does not seem to make quite the same splash in camera marketing, despite having a superior product.

The third topic focuses on comparing the Galaxy Samsung S5 to other phones. The top ten words for the third topic are *Samsung, Galaxy, new, Android, active, year, Google, will, phones,* and *sales*. The Samsung Galaxy S5 uses Android software as well as relies on Google for other functions, such as Play Store (for application downloads and updates) and other services. In contrast, the biggest competitor, Apple, uses IOS software for the competing product, the iPhone. Using these reviews to assess customer satisfaction with the software can assist Nokia in making software changes that reflect consumer desires. The other words for this topic include *new, active*, and *year*, which reflect the reviewers' regard for newer and upgraded technology. Nokia can gain a lot from this topic because this topic can directly influence what software updates can be pushed out to current users of the Samsung Galaxy S5, as well as incorpo-

rate consumer feedback into the design of newer models of the Samsung Galaxy. Pleasantly for Nokia, there is a demand for newer models and new products, which hopefully will result in sales for newer versions of Nokia products, including the Samsung Galaxy. Timing the release of new products, especially the Samsung Galaxy, is also an important take-away for Nokia, to target the consumer audience and to ensure better success against competing products.

The fourth topic focuses on the battery life of the phone. The top ten words of the fourth topic are *battery, power, new, Galaxy, display, life, compared, Samsung, mode*, and *video*. Many of the reviewers commented on how the video mode affected the battery life, as well as other modes, such as the display modes. Another important trend to note about this topic is how this new Galaxy model compared to older models in terms of the longevity of battery life. Battery life is an important issue on which Nokia can focus research and development. Some smartphones have a great battery life, and there are many accessories on the market that can increase the battery life of a smartphone (i.e. the charging phone case). In order to keep up with competing manufacturers (and potentially move ahead of the pack), Nokia should carefully consider their customer reviews of the Samsung Galaxy S5 battery life, consider what apps drain the battery life, and look for better ways to preserve and maximize battery life in later editions of the Samsung Galaxy.

## Topic Probabilities by Document

The third output is a table that details the probabilities of each topic for every single document (Topic Probabilities by Document output). The probabilities for each document total 1 (or 100%). In a sense, the probabilities represent the percentage that the document talks about a certain topic based on the words associated with that topic that are found within the document.

In some cases, a single topic dominated a particular document. For example, Topic 1 (fingerprint function) dominated document 9 (.731) and document 13 (.732), Topic 2 (camera) dominated document 7 (.839), Topic 3 (comparison to other phones) dominated document 28 (.726), and Topic 4 (battery life) dominated document 26 (.846). In other cases, the most prominent topic is not as clear as the probabilities are much closer among the topics. For example, document 11 had a near tie among the topics (Topic 1 (.299), Topic 2 (.235), Topic 3 (.214), and Topic 4 (.252)), document 15 had two topics that were close to the same probability (Topic 3 at .396 and Topic 4 at .356), and document 24 had two topics that were even closer to the same probability (Topic 3 at .311 and Topic 4 at .334). Other cases involve a higher probability of a particular topic over the other topics but none of the topics had a particular high probability. For example, document 18 was most likely to be categorized as Topic 1 (.397), followed by Topic 4 (.286), Topic 2 (.184), and Topic 1 (.166). Topic 1's probability of .397 in document 18 is not a strong of a probability as Topic 1's probability of .731 in document 9, but, at the same time, it was not as close to the other topics' probabilities for this document as the topic probabilities are for document 11.

This third output provides insight into the first output (Topics by Document). While a particular topic may technically have the greatest probability for a particular document, that topic may only have a slightly greater probability than another topic, or upon further review by the data analyst, another topic may be pronounced or prominent in the document despite the model's assessment. In order to account accurately for a topic's prominence, it is important to compare the Document by Topic output to the Topic Probabilities by Document output to ensure that the Document by Topic output has captured the probabilities accurately. Nokia can use this third

output to look at the reviews that are relevant to each topic, even if the documents did not initially show up in the Document by Topic output for a particular topic and can thereby conduct a more thorough analysis for each topic.

This output is also important because it can show links between particular topics. For example, document 15 and document 24 both had very close probabilities for Topic 3 (comparison to other phones) and Topic 4 (battery life). Nokia should take into consideration that these two topics are intertwined and also look at customer reviews for competing products with regards to battery life. In order to continue to improve the longevity of battery life as well as tend to customer needs, Nokia can look at what customers think about Apple's iPhone 5s battery life. By considering Apple customers' compliments and complaints, Nokia can use these reviews to increase the battery life of the Samsung Galaxy by reducing the apps or functions on the phone that cause the battery to drain faster. This process can be applied to any topics that are closely linked by the reviews in order to analyze why the topics overlap and what information Nokia can draw from these connections.

## Visualizations

In order to increase understanding of the relevance of the top words, we have an interactive visualization tool that shows the overall word relevancy for the corpus with a sliding weight parameter that allows the user to adjust between rarely used words that are exclusive to a particular topic or most used words that are common to several topics (Sievert & Shirley, 2014). Each of the topics can be selected for comparison of the top words for that topic against the frequency of the words in the whole entity as shown below. This interactive visualization can be accessed through the R script (Appendix 3).

In figures 2 through 5 (see Appendix 1), the top 30 most relevant words for a particular topic are displayed with their frequency (the red bar) in relation to their use in the overall corpus (the blue bar). In Figure 2, the top 30 words for Topic 1 (fingerprint) shows that the most frequently used word for Topic 1 is *phone*. Of these 30 words, *Samsung* is the most frequently used word overall. In Figure 3, the top 30 words for Topic 2 (camera) shows that the most frequently used word for Topic 2 is *camera*. Of these 30 words, *Samsung* is the most frequently used word overall. In Figure 4, the top 30 words for Topic 3 (comparison to other phones) shows that the most frequently used word for Topic 3 is *Samsung*, which is also the most frequently used word overall within this subset. In Figure 5, the top 30 words for Topic 4 (battery life) shows that the most frequently used word for Topic 4 is *battery*. Of these 30 words, *Samsung* is the most frequently used word overall.

## Conclusion, Limitations, and Suggestions

Team DART set out to analyze Samsung reviews for primary topics along with sentiment in which Nokia could focus on competitive intelligence. The four major topics are fingerprint function (Topic 1), camera (Topic 2), comparing the Galaxy Samsung S5 to other phones (Topic 3), and battery life (Topic 4). The four topics were analyzed to determine sentiment by extracting the top ten words for each topic. The analysis sentiment for Topic 1 is positive for the desire for and use of the fingerprint function. The analysis for Topic 2, which was the least significant of the topics, expressed negative sentiments regarding lighting and positive sentiment regarding resolution. The analysis sentiment for Topic 3 focuses on upgraded software. The analysis sentiment for Topic 4 is negative towards the battery drain while playing videos.

The use of the LDA algorithm to identify topics by document proved extremely effective

with natural language processing. Sentiment extraction required additional research and was guided by the ranking of words per topic. Improvement of the text analysis utilizing LDA includes tailoring the *stop words* eliminated from the initial search, modifying the initial parameters regarding burn-in, iterations, number of starts, thin, and number of topics. Running the algorithm on additional articles could increase the accuracy of the parameter settings. Caution should be placed on the cantankerous nature of text and topics. Rarely will two data sets result in the same conclusion in text analysis. Additional algorithms should be applied to create an ensemble of approaches to handle the complexities of text analysis.

## Continued Analysis

The LDA Analysis has provided significant insight into the Nokia reviews. In order to provide additional understanding, we recommend a three-pronged approach. First, we recommend continued analysis of the more recent Nokia reviews, utilizing the same methodology. These results are time-sensitive, as technology is quickly evolving. To ensure that the results are pertinent, we recommend applying the LDA model every six months, using the latest articles. As new phones evolve along with new technology, customer demands will continue to change. It is imperative that Nokia stay one step ahead of these demands to allow sufficient time to develop coinciding technology. Second, we recommend performing analysis on additional reviews external to Nokia. While our analysis has inspected articles pertaining directly to Nokia, it has not taken into account similar products that are made by other companies. In order to position Nokia strategically, we recommend a broader analysis of all similar products on the market. This analysis would include a replicated study of Apple iPhone reviews. Such an analysis would yield suggested improvements for the Apple iPhone as well as a continued understanding of customer de-

mand. A comparison of the Apple and Samsung studies would garner a broader consumer view, enabling Nokia to make a more educated decision of how to prioritize its resources. The most important information should come from within the Nokia Corporation. Nevertheless, reviews of Apple iPhones are important – it is critical that Nokia consider the direction of the market as the whole. Finally, we recommend an analysis of current Nokia sales and how existing products sell based on battery life, fingerprint sensor, and camera. Nokia must know if one of these topics, or a combination of them, drives sales. This additional analysis will allow Nokia to optimize their phones in terms of meeting customer demand while appropriately allocating money to research, resulting in a more efficient sales stream. The ability to analyze text is critical to the progress of Nokia. Our Team DART has succeeded in laying the initial groundwork of mining text, providing valuable information and enabling a more clairvoyant understanding of how to move forward in a constantly changing marketplace.

Appendix 1

## Figures


*Figure 1. Inter-topic Distance Map*

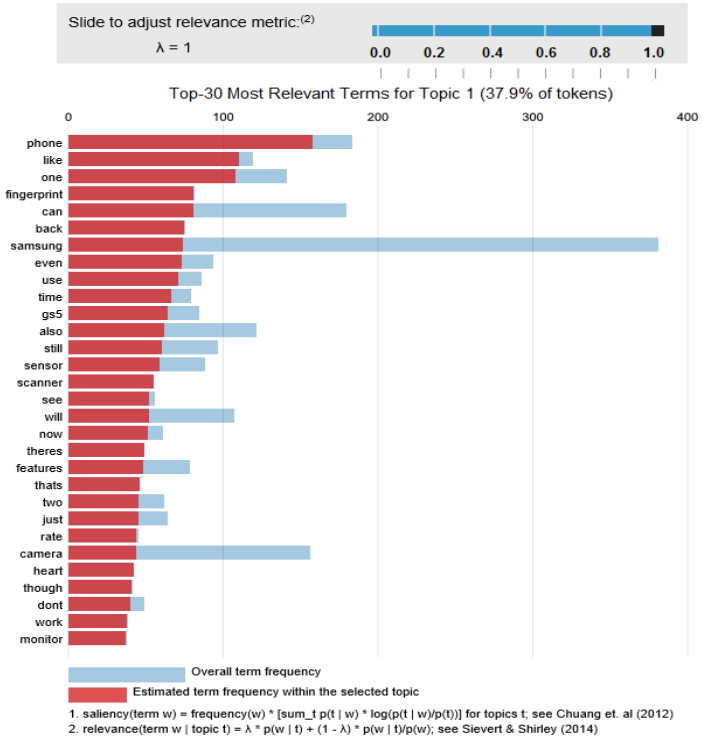


*Figure 2. Word relevancy for Topic 1 (fingerprint function) out of all overall frequency.*

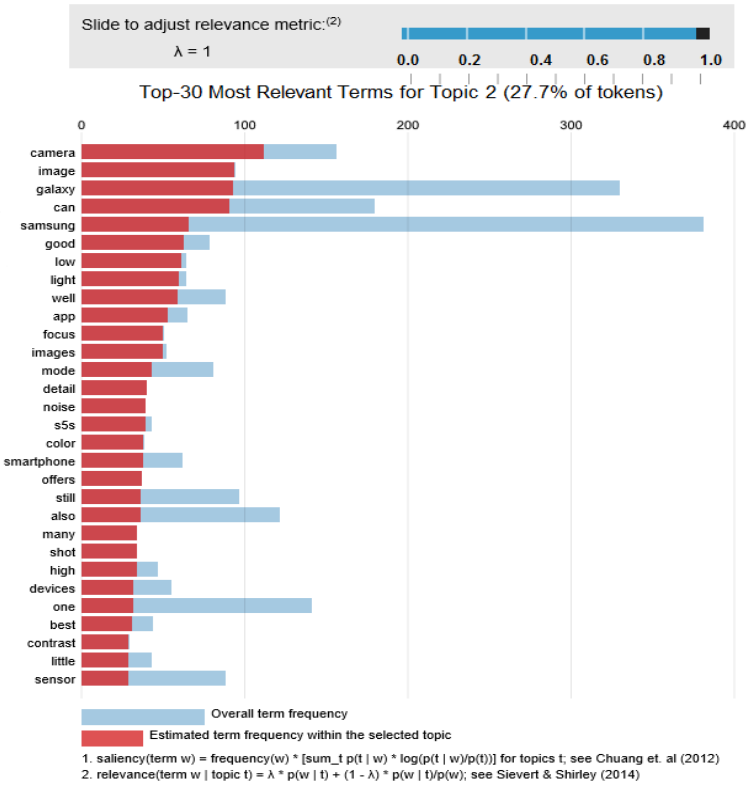

*Figure 3. Word relevancy for Topic 2 (camera) out of all overall frequency.*

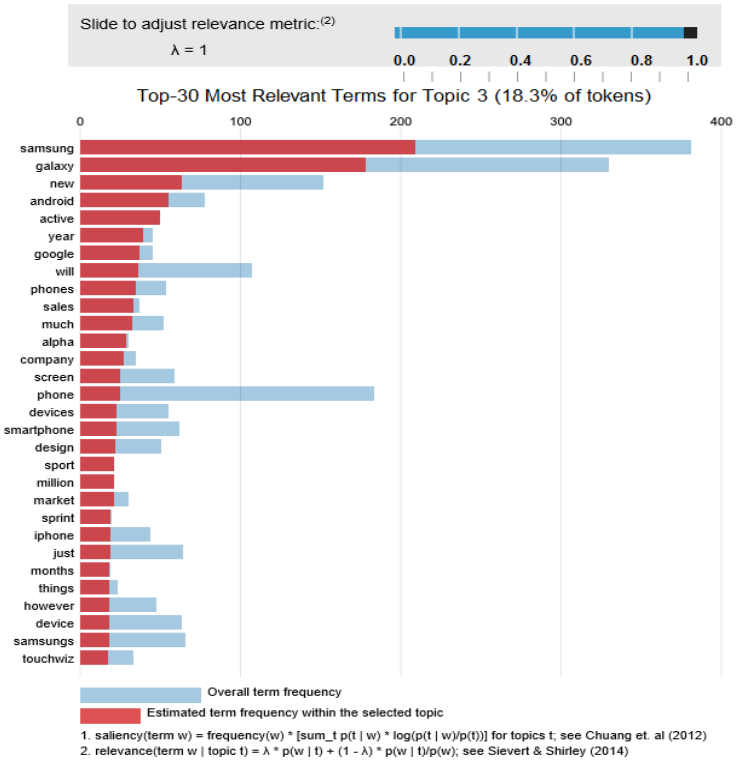


*Figure 4. Word relevancy for Topic 3 (compared to other phones) out of all overall frequency.*

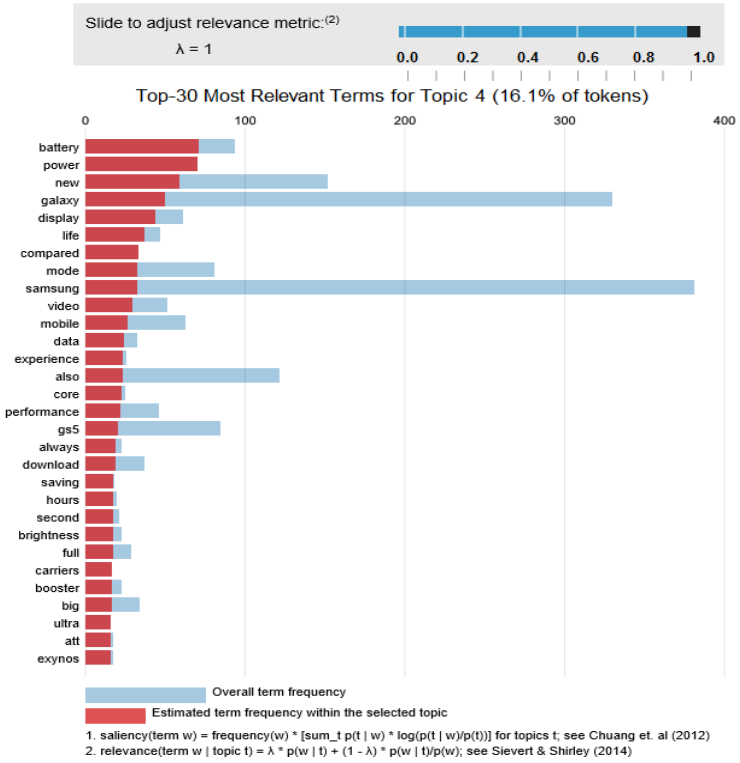

*Figure 5. Word relevancy for Topic 4 (battery life) out of all overall frequency.*

## Appendix 2

### Sources

Andronico, M. (2014, June 05). Camera Face-Off: iPhone 5s vs. Galaxy S5. Retrieved August 12, 2017, from https://www.tomsguide.com/us/iphone-5s-vs-galaxy-s5-cameras,review-2190.html

Awati, K. (2016, October 19). A Gentle Introduction to Topic Modeling Using R. Retrieved August 12, 2017, from https://eight2late.wordpress.com/2015/09/29/a-gentle-introduction-to-topic-modeling-using-r/

Gialampoukidis, I., Vrochidis, S., & Kompatsiaris, I. (2016). A Hybrid Framework for News Clustering Based on the DBSCAN-Martingale and LDA. In *Machine Learning and Data Mining in Pattern Recognition* (pp. 170-184). Springer International Publishing.

Grimes, S. (2008) Unstructured data and the 80 percent rule. Retrieved August 11, 2017, from https://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/

Hu, D. (n.d.). Latent Dirichlet Allocation. Retrieved August 04, 2017, from http://cseweb.ucsd.edu/~dhu/docs/exam09.pdf

Manning, C. D., Raghavan, P., & Schutze, H. (2009, April 9). *Stemming and Lemmatization*. Retrieved from NLP Stanford: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Multimedia Knowledge and Social Media Analytics Laboratory. (2015). Web News Article Dataset. Retrieved August 12, 2017, from http://mklab.iti.gr/project/web-news-article-dataset

Robinson, Julia Silge and David. "Topic Modeling." May 07, 2017. Accessed July 22, 2017. http://tidytextmining.com/topicmodeling.html.

Sievert, C., & Shirley, K. E. (2014, June 27). NLP Stanford. Retrieved from LDAvis: A method for visualizing and interpreting topics: https://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf

Appendix 3

**R Script**

```
#Create list of necessary packages
packages = c("tm","SnowballC","topicmodels", "bindr", "dplyr", "stringi", "LDAvis")

#check if package is on machine, load if yes, install and load if no
package.check <- lapply(packages, FUN = function(x) {
 if (!require(x, character.only = TRUE)) {
   install.packages(x, dependencies = TRUE)
   library(x, character.only = TRUE)
 }
})

setwd("C:/Text")

# Get List of all samsung files
dir = list.files(getwd(), pattern="samsung*")
files = lapply(dir, readLines)

documents = Corpus(VectorSource(files))

#View file 10
writeLines(as.character(documents[[10]]))

##### Preprocessing ####
# Convert all characters to lowercase
documents = tm_map(documents, content_transformer(tolower))

#Replace "-" with white space
toSpace = content_transformer(function(x, pattern) { return (gsub(pattern, " ", x))})
documents = tm_map(documents, toSpace, "-")

# Remove non characters
documents = tm_map(documents, removePunctuation)

# Remove stop words
documents = tm_map(documents, removeWords, stopwords("english"))

# Remove extraneous white spaces
documents = tm_map(documents, stripWhitespace)
```

```
# Word stemming
#documents = tm_map(documents, stemDocument)

# Create document term matrix
matrix = DocumentTermMatrix(documents)
rownames(matrix) = dir
freq = colSums(as.matrix(matrix))
ord = order(freq, decreasing = TRUE)
write.csv(freq[ord], "word_frequency.csv")

# Parameters
burn = 4000
iter = 2000
thin = 500
seed = list(1234, 4321, 123, 321, 1413)
nstart = 5

# Number of topics
k = 4

ldaOut = LDA(matrix, k, method = 'Gibbs', control = list(nstart=nstart, seed = seed, burnin =
burn, iter = iter, thin = thin))

ldaOut.topics = as.matrix(topics(ldaOut))
write.csv(ldaOut.topics,file=paste("LDAGibbs", k, "DocsToTOpics.csv"))

# Top 10 terms in each topic
ldaOut.terms = as.matrix(terms(ldaOut, 10))
write.csv(ldaOut.terms, file=paste("LDAGibbs", k, "TopicsToTerms.csv"))

# Probabilities associated with topic assignment
topicProbabilities = as.data.frame(ldaOut@gamma)
write.csv(topicProbabilities, file=paste("LDAGibbs", k, 'TopicProbabilities.csv'))

topicmodels_json_ldavis <- function(fitted, corpus, doc_term){
 ## Find required quantities
 phi <- posterior(fitted)$terms %>% as.matrix
 theta <- posterior(fitted)$topics %>% as.matrix
 vocab <- colnames(phi)
 doc_length <- vector()
 for (i in 1:length(corpus)) {
   temp <- paste(corpus[[i]]$content, collapse = ' ')
   doc_length <- c(doc_length, stri_count(temp, regex = '\\S+'))
```

```
 }
 temp_frequency <- inspect(doc_term)
 freq_matrix <- data.frame(ST = colnames(temp_frequency),
                           Freq = colSums(temp_frequency))
 rm(temp_frequency)

 ## Convert to json
 json_lda <- LDAvis::createJSON(phi = phi, theta = theta,
                                vocab = vocab,
                                doc.length = doc_length,
                                term.frequency = freq)

 return(json_lda)
}

#Display visualization in web browser
ldaOut.json = topicmodels_json_ldavis(ldaOut, documents, matrix)
serVis(ldaOut.json)
```