**Analysis of Latent Dirichlet Allocation Methods for Categorization of Cell Phone Reviews**

By Pravin Raja

**LDA Explained**

Latent Dirichlet allocation is a great and popular method for fitting a topic model in the subject of Natural Language Processing. It treats each document as a mixture of topics, and each topic as a mixture of words. This allows documents to overlap each other in terms of content, rather than being separated into different groups in a way compared to a **natural language**.

So, let us say you have a collection of documents, like articles in a magazine. Now we want to find classify those articles into topics, but we do not know the topics. So, ether we decide on a fixed number of topics and clump all the articles that are similar enough together into one topic, or we decide how similar document need to be for them to have their own topic.

So, to break it down:

1. Words make a document those words also make a topic, so that topic information is hidden inside the documents. that information is known as Latent.
2. We can assume that a document will contain words from more than one topic, but we also assume that one document is mainly about one topic, so many of the words in a document will be about one topic and there won't be very many topics in one document. that is called a Dirichlet allocation.
3. Finally, we guess, what words go in what topics.

* We can assume that words in a document are usually about one topic and we assume that different words from a topic are usually one document. Then we put all the words into random topics and check if our assumptions hold. We then check if the distribution is a Dirichlet distribution.
* We can use the words in the topics to check the words in the documents and we use the words in the documents to check the words in the topics. If the words do not fit in the topic distribution than we change the topic the word is in.
* This is repeated until we notice that we do not change that many words.

**Our Goal**

* Use Latent Dirichlet Allocation methods for categorization of Social and News Feeds of users to make it easier for a user to traverse through the various sites of historical cellphone reviews.
* This will be done in a multi hierarchical model method for traversing as and will include a graphical representation of general cellphone reviews of various brands and their sentiments and see what common likes and dislikes people have for various phone brands from Apple, Nokia, and Samsung. Our specific python library integrated for this is LDAvis: <https://cran.r-project.org/web/packages/LDAvis/index.html>
* Prediction: The one that has the higher duration of battery life and overall speed snappiness of the smartphone choices will probably come out as the most common occurrence.

**Part 1: Building a Data Set**

**Python Scrapy to Scrape for Review Data**

For our project we will first create a clean data set for cellphone reviews. We will use a python library called Scrapy to do so and do sentiment analysis with it.

Sentiment analysis is an important part of Natural Language Processing and it can be performed over the reviews scraped from products on Amazon. This helps in identifying the user’s emotion towards a particular product. This can even help sellers or even other prospective buyers in understanding the public sentiment related to the product. Which we will later integrate into LDA model for our project.

**What is Scrapy?**

Scrapy is a web crawling framework for a developer to write code to create, which defines how a group of websites will be scrapped. This will be done in a pattern matching way to complete with the review URL passed in the code. The most significant and very important feature is on Scrapy that it is built on is called Twisted. Twisted is an asynchronous networking library, which makes the spider performance is very significant. Keep in mind this method can be used across all sites including news feeds, twitter feeds, RSS feeds, and Facebook posts.

**Importance of Data Collection and Parsing**

Once the sites are scraped, the parser can dump out the results in any format you wish such as CSV or JSON. This is the final output while in which your scraped data resides. For this specific project we will use CSV because it will be easier to later transform it into a corpus.

* Now go to the Scrapy Reviews.py file and follow the directions to build your own data, once ran the python scrapy python program will output CSV file with data. Techniques like installing libraries, reading, and implementing a panda’s data frame, plotting, and visualizing data, and working with your computer directory will be discussed. Make sure you have Python and an IDE installed on your computer.
* For this project I have already compiled a data set with 1000’s of user’s data of cellphone reviews ranging from phone brands across various stores, sites, and countries. The file is name Finaldata.csv. Feel free to explore and build a data set with features you find the most intuitive to problems you are solving.
* As we can see with the FinalData.csv we have scraped over 5000 historical data reviews into separate categories (doing this will be important later for breakdowns and file i/o)
* The data is across various stores and brands like Apple, Samsung, Blackberry, Nokia, and many others and across 4 countries of USA, India, Canada, and Great Britain.

**Part 2: Writing the CODE**

* After building a clean data set, go over to the Project.py file. Make sure your csv and python notebook are in the same folder location to avoid runtime errors down the line
* In our detailed python notebook file we will go through how to install and import the correct packages, executing file operations for our data set, NLP techniques like stop words, regular expressions for cleansing, splitting and tokenizing words, and frequency and word distributions will also be used. After completion of all this we will assemble a corpus with our collected data that will be translated into a pandas data frame so we can easily sort our topics for LDA visualization.
* For our LDA model we will use the library LDAvis- this will help create an interactive web-based visualization of a topic model that has been fit to a corpus of text data using Latent Dirichlet Allocation. Given the estimated parameters of the topic model, it computes various summary statistics as input to an interactive visualization built with D3.js that is accessed via a browser. The goal is to help users interpret the topics in their LDA topic model. <https://cran.r-project.org/web/packages/LDAvis/index.html>

**Part 3: Analysis of Phone Reviews**

* Originally when implementing the NLP techniques, I wanted to use the porter stemmer algorithm to try to remove potential prefixes and suffixes from words but later decided against it. This was mainly because a word like impossible or possibly could have been broken into possible and completely changed the meaning of the entire topic. This could be more useful in a field like Information retrieval.
* <https://tartarus.org/martin/PorterStemmer/>
* After visualizing a few models, I had to manually extend the stop words and take out a lot of words manually that didn’t add much value such as phone, mobile, smartphone, Apple as this kept repeating and didn’t really describe anything. The word case was also removed as that does not actually physically describe the phone itself.
* The numbers 1-9 kept getting mixed into topics and the range did not seem to fit and make sense with some of the topics, for example, the number #2 was paired with the topic good at one point, which didn’t make sense as the word “good” should get a paired number higher than average like 6-9. So, all numbers 1-9 were also manually removed using the extend stop words.
* Increasing the number of topics was also useful as it gave more insights.
* Insights:
* For topic of money. It was surprising to see that most were pleased with the cost and overall seemed happy for the value provided. There were not much mixed values.
* My prediction from earlier was correct as battery performance a huge impact on sentiment. As we saw from our model people were generally pleased or very unhappy with the overall battery life of these various phone selections. The overall reviews were varied as many seemed to either have issues or seemed to have everything just work fine.
* For topic of use: It seems most people were easily able to navigate the interface and the most common uses were for apps, texting and sending things. Most people seem to love their phone.
* For topic of love it seems most people love their phone for the camera, lots of features, apps, and overall usability.
* For topic Camera: It seems almost all phones had good camera quality despite some dated reviews. This may not be too surprising as phone camera has improved a lot over time and even beat out some stand-alone camera and not to mention you do not have to carry a separate device to maintain.
* Overall, from our model we can safely assume that most people that bought phone in new condition overall seemed pleased with the condition and everything worked as expected and advertised. The main concerns had to do more with battery issues and end user experience at times. To end off people seemed satisfied with this and got they wanted.