Graphical user interface

Description automatically generated**Topical Segmentation of Text Documents**

Figure 1 source: [IBM](https://www.slideshare.net/MichaelBeatty/ibm-cloud-storage-cleversafe)

Before the topic of this article can be addressed, we must first start with the problem it is trying to solve. It is a cliché at this point, but a true one, that over 90% of the data ever generated, has been generated in the last 2 years as can be seen in figure 1.

However, what is rarely discussed is the fact that more than 80% of that data is unstructured. Most of the data generated today is in the form of videos, text, images, 3D art and so on.

This creates a great difficulty in making sense of this data, as unlike tabular data, there is no easy solution for categorizing, clustering, or filtering it. In many cases that is because the data itself has no labels that could be used for filtering. Due to this lack of order, most of it, is not openly accessible to the end user.

**The Problem with Text Data**

Text is by far the most widely spread form of unstructured data and it is often the one that holds the most information. The old saying is that a picture is worth a thousand words, but nobody learns algebra from pictures, and nobody studies for engineering exams using pictures. But we are not talking only about technical text documents, most of the text data found today is in the form of articles, reviews, comments, historical documents, and many other examples of information posted without labels or a way of making sense of it. The issue could be solved by implementing a labeling convention from the very beginning of any process using, storing or sorting text data, but often by the time a convention is implemented, there are already many terabytes of text data that need to be reprocessed. And even in the case of a preestablished convention, there would be no guarantee that the text data generated would fit within those constraints.

In that regard, this article, aims to prove that the problem encountered when dealing with text data is not as hopeless. In many cases Machine Learning can resolve the issue, or at least make sense of what would have been a very time-consuming problem to solve.

**Data Sources**

For this article, three sources of text data have been used, each presenting the problem of having no structure, thus making any filtering impossible. It would be difficult to understand the length of these text files just by a word count, so to make it easily understandable, a simple point of reference has been used, namely the volume of text when compared to the length of the Bible.

The data sources in question are:

* NPR News Articles, 9,266,936 words, 11.47 times larger than the Bible
* Spotify Million Song Dataset, 12,653,383 words, 15,67 times larger than the Bible
* Women’s Clothing Review, 1,221,308 words, 1.51 times larger than the Bible

As can be noticed the datasets are from 3 very different sources, each presenting the same problem, there are no labels, thus there is no method of creating a filter. This combined with the larger text volumes leads to a great difficulty in discerning any useful information from the text within good time.

The most obvious solution would be to create a simple count vectorizer or a more advanced Term Frequency-Inverse Document Frequency embedding of the documents and use a clustering algorithm on the result. However, that would create clusters, but there would be no way of understanding what those clusters are referring to. Thus, the initial problem of having to go through the whole document collection and to make sense of them, would have just been split into having to go through several clusters and do the same. The time and resources problem would not be solved but divided into smaller problems.

There are however clustering techniques that can solve this issue, the ones used for this article are, Latent Dirichlet Allocation and Non-Negative Matrix Factorization.

Both algorithms rely upon the same assumptions:

* documents would cluster around a series of words that define that cluster,
* each Topic has a series of words that have a higher probability of showing up.

**NPR**

The NPR dataset consists of 11991 unlabeled news articles, but there are no labels.

Chart

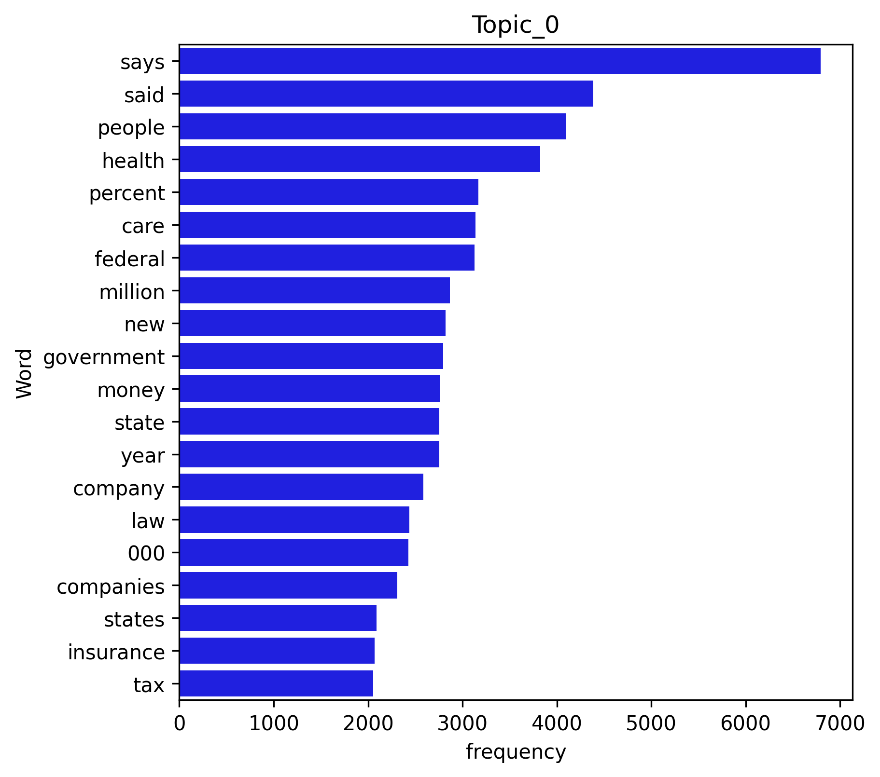
Description automatically generatedThe most frequent words found in the articles are known, those being frequently encountered stop words that can’t be used to determine a category as can be seen in figure.2, and even when eliminating these stop words, it is difficult to determine a subject as can be seen in figure.3. It can be concluded that a simple word filter could not be used to categorize the articles. As a result, we are left with the tools mentioned above, namely Latent Dirichlet Allocation and Non-Negative Matrix Factorization.

Figure 2

We start out by assuming a clusters number of seven. There is no correct number, but the higher the number the more specific the clusters become, the lower the number the more general they become.

Starting with this assumption, we end up with 7 clusters, each having a different frequency for different words. For example, for the first topic (Topic\_0), we can observe in figure.4, that the words with the highest frequency tend to be financial in nature. Notice that the clustering algorithm does not give a topical name. It is up to us to determine the appropriate name from the word frequencies obtained for each cluster.

Figure 3

**

We obtained the results as follows:

* Topic\_0: [*tax insurance states companies 000 law company year state money government new million federal care percent health people said says*]. We could classify this topic as related to **Finance**,
* Topic\_1: [*attack state military war news department country according reported president russia security npr reports told government people says police said*]. We could classify this topic as related to **International News** (from the point of view of the US)

Figure 4

* Topic\_2: [*local little land small way year make world home time day city new years just water people food like says*]. We could classify this topic as **National or Local News**,
* Topic\_3: [*brain time years research new don percent just care drug children like disease medical patients study women health people says*]. This topic is clearly dealing with **Medical Research**,
* Topic\_4: [*presidential just voters political vote donald party new people republican election white house obama state campaign clinton president said trump*]. We can notice that this topic is dealing with the **Presidential Election**,
* Topic\_5: [*black says world ve said going story years don life music way really new think know time people just like*]. This can be labeled as **Cultural News**
* Topic\_6: [*children work science kids make really way schools don university education time new think just like people students school says*], This topic can be labeled as **Education**.

We can notice that the Topics given are not labeled intuitively, but we are the ones that deduce the topics from the most frequent words used. Taking into consideration the labels determined by us we can take a few samples from the corpus to verify our assumptions, shown below.

Cultural News

Order in the court — but maybe not in movie theaters. With all the talk lately about politics and the judiciary — fights over Supreme Court vacancies, the President complaining about”” judges — I’ve been thinking about the judges I’ve seen on screen, and how their depiction might have intersected with public opinion through the years. Ce

Finance

Updated at 2:30 p. m. ET, for years, Puerto Rico has grappled with an debt crisis, watching as its bills have grown to more than $70 billion. Including what the U. S. territory owes to pension funds, that debt exceeds $120 billion. Now, Puerto Rico’s struggle with its creditors has stepped into U. S. federal court, where an unprecedented case

**Spotify**

A picture containing chart

Description automatically generatedChart

Description automatically generatedIt can be noticed that this clustering method seems to work for news articles, however different circumstance can be dealt with by the algorithm, for example on the lyrics of songs. Though expectations should be limited in regard to what the algorithm can do. It should not be expected for it to detect the genre of music. As the name suggests it can segment text on its topic but often the topic of a song is not the same as the genre. The Spotify dataset has a total of 57650 song lyrics. It shall be explored if they can be segmented via a topic.

Figure 5

We can notice in figure.5 that the most common words in the songs are normal English stop words, as could be expected. However, if these are ignored, it can also be noticed that the most common words have to do with sentiment. Thus, it could be assumed that the Topical Segmentation will lead to different topics as relating to sentiments.

The same algorithm as for the NPR data was applied to the Spotify text data, with the same assumed 7 topics. In figure.6, the most frequent words encountered in the 3rd topic can be observed. Most of them have a religious connotation, thus it can be assumed that the topic of the songs in the 3rd category is religious in nature. Again, the topic name is only implied by the frequency of the words, the algorithm does not provide a topic name but a number.

Figure 6

Thus, the topic names were decided by looking into the frequency of the words for each topic. They are as follows:

* Topic\_0: *[wanna like hey know girl come let love want got don gonna yeah baby oh]*, We can observe that the topic of these songs seems to be the **Subject of a Courtship,**
* Topic\_1: [*wind blue time dream rain day come sky eyes light sun away ll night like*] this topic seems to be focused on **Nature,**
* Topic\_2: [*feel away life want like way heart say time ve ll just don know love*] this topic seems to be focused on **Love**, which the algorithm seems to consider a distinct topic from that of courtship.
* Topic\_3: [*free born heaven man sing die jesus soul oh let come life world lord god*], as we have noticed from this graph these songs seem to have the topic of the **Divine**,
* Topic\_4: [*way day ll long town good little ve got just old said home man la*], this topic appears to be **Locational** in nature,
* Topic\_5: [*la santa gimme music di ba happy ha doo roll dance rock da christmas na*], this topic seems to be focused on ***Holidays***,
* Topic\_6: [*nigga chorus fuck man shit money em just cause yaain know don got like*] and the last topic seems to be indeed overlapping with **Rap Music**.

As before a random test is needed to test the assumptions made in the topics:

***Love***

*All my love*

*Came to nothin' at all my love*

*When I woke up to find*

*You were no longer mine*

*All my love*

*Throw away after all this time*

*Now there's no place for me*

*In the future you see*

*I don't understand you*

*I've done all I c*

*Song: All My Love by Cliff Richard*

***Locational***

*They're really rockin' Boston*

*In Pittsburgh, P. A.*

*Deep in the heart of Texas*

*And 'round the Frisco Bay*

*All over St. Louis*

*And down in New Orleans*

*All the cats want to dance with*

*Sweet little sixteen*

*Sweet little sixteen*

*Song: Sweet Little Sixteen by Chuck Berry*

The assumptions appear to be correct. It can now be verified which topics are most popular in the Spotify dataset, as can be seen in figure.7:

Chart, bar chart

Description automatically generated

Figure 7

**Women’s Clothing**

No that is not the beginning of a joke, from a technical point of view it is more the beginning of a tragedy. The reason behind it being that if one sells women’s clothing the assumption might be made that it is Chart, bar chart

Description automatically generatedstraight forward to understand what your customers are interested in when it comes to the product. And of course, their satisfaction can be measured via the rating, as can be observed, in figure.8.

Figure 8

Overall, the customers are satisfied with the products. But we can’t really understand what the customers are interested in unless we go into the reviews, which is of course unstructured text data.

Text

Description automatically generatedIn such an environment we can use the exact same tools that we have used so far. In figure.9, we can notice the most often used words, but these do not give us the topics of interest. As with the other cases we need to apply the Latent Dirichlet Allocation, as for the previous cases, with the assumption of 7 topics of interest. The result being the following topics with the adjacent twenty most frequent words:

Figure 9

* Topic\_0: *[don, ordered, went, got, jeans, did, pants, try, love, bought, just, online, price, fit, retailer, tried, saw, size, sale, store*] it appears that here the focus is on the **Sale** of the item,
* Topic\_1: [*petite, bit, right, nice, little, great, short, just, love, hips, look, fabric, flattering, long, like, length, skirt, size, fit, waist*], in this topic the focus seems to be on the **Fit** of the item,
* Topic\_2: [*long, colors, cute, nice, black, look, bought, fall, like, looks, jacket, comfortable, jeans, perfect, soft, wear, color, sweater, love, great*], the focus of this topic seems to be the **Comfort**,
* Topic\_3: [retailer, run, like, love, bit, fits, lbs, big, little, wear, usually, runs, petite, medium, fit, xs, ordered, large, small, size], the focus of this topic appears to be **Small Sizes**, the algorithm seems to think women differentiate between **Fit** and **Small Sizes**,
* Topic\_4: [*looks, colors, material, blouse, pretty, bit, sheer, soft, really, underneath, little, bra, nice, love, wear, color, white, shirt, like, fabric*], this topic appears to be focused on **Material and Color**,
* Topic\_5: [*cut, cute, loved, beautiful, thought, wanted, work, model, material, looks, fit, way, didn, looked, really, fabric, look, just, dress, like*], of course **Appearance** appears to be a topic,
* Topic\_6: [*work, gorgeous, fabric, true, quality, recommend, summer, dresses, fit, compliments, size, fits, comfortable, beautiful, flattering, great, perfect, wear, love, dress*], **Dresses** are the only topic that overlaps with a category of clothing, similar to rap in the music segment.

As before the assumptions need to be tested with a few random examples:

Appearance

I love natural colors and nature themes but this top was a disappointment. i ordered both the small and the xs because retailer can run very large. the small was too big and the xs did not hang correctly, the arms hit at a weird spot and the neckline wouldn't sit right. overall the fit was boxy and unflattering. the fabric is pretty but the style is well....boring. there was just no wow factor for me. i think the fabric deserved a much better design.

Material and Color

I'm not usually a fan of simple crewneck sweaters but this one has some nice added details that make it a keeper. the dark red oxblood color is really nice (it manages to be a red sweater without feeling overly holiday), the detailing at shoulder/chest is interesting yet subtle enough to not be distracting and doesn't feel bulky at all, the dropped shoulder seams are done correctly and without the body becoming boxy. it does feel like it might run a tad bit large, but only very slightly,

The assumptions are close to reality; however, this can be taken even further. We have the ratings of the products; thus, we can dig deeper and see the topics for the liked products and what are the topics for the disliked products, thus getting valuable insight about the customers.

Figure 10

The convention is assumed that the liked products are those with a rating of 4 stars or above and the rest are considered disliked products. Separating the data into two datasets, one with positive ratings and one with a negative ratings. Then, the clustering algorithm is applied on each dataset individually. We assume just 4 topics for each of them. We can see in figure.10 what the topics of discussion tend to be when it comes to products that are positively rated. In much the same way we can notice what the topics of the reviews are when it comes to the products that are negatively rated in figure.11.

We can observe that positively reviewed products tend to get reviews regarding their:

* Well-fitting small sizes
* Appearance
* Casual Style
* Price to Quality Ratio

Whilst negatively reviewed products tend to be focused on:

* Tops that do not fit or look well
* Dresses made of cheap fabric

Figure 11

* Tops with a bad Color or Material
* Items of clothing that are labeled as a small size but are larger than expected

Thus, through a simple clustering algorithm we have managed to gain valuable insight into the customers’ interests and complaints, even though this data is stored as text reviews.

**Conclusion**

When taking into consideration that most of the data generated today is unstructured, and furthermore a large part of this data is in the form of text, companies and individuals face great challenges in making sense of this data. However, it appears that at least in the case of text, there are tools available that can help in the analysis and understanding of unstructured text data, and despite the difficulties valuable insight can be obtained from such data sources.

**References:**

* [cod and images](https://github.com/alex-parvu/Articles/tree/main/02.Text%20Document%20Clustering)
* [NPR Data](https://www.kaggle.com/datasets/gauravduttakiit/npr-data)
* [Spotify Data](https://www.kaggle.com/datasets/notshrirang/spotify-million-song-dataset)
* [Women’s Clothing Data](https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews)
* [IBM estimations](https://www.slideshare.net/MichaelBeatty/ibm-cloud-storage-cleversafe)