**Topical Segmentation of Text Documents**

**Graphical user interface

Description automatically generated**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 year as can be seen in **fig 1.**

However what is rarely discused is the fact that more than 80% of that date is unstructured. The majority of the data generated today is in the form of videos, text, images, 3D art and so on. This creates a great dificulty in making sense of this data, as unlike tabluar date, there is no easy solution for categorizing, clustering or filtering it. In many cases that is simply due to the fact that the data itself has no labels that could be used for filtering. As a consequence of this lack of order in the data we generate, most of it is not openly accesible to the end user, and it is a very resource and time consuming process to make it accesible.

**The Problem with Text Data**

Text is by far the most widly spread form of unstructured data, but it is not the one that ocupies the most memory. However it often the the one that holds the most information. The old saying is that a picture can hold a 1000 words, but nobody learns algebra from pictures, and nobody studies for engineering examns using just pictures. But we are not talking about jus technical text documents, most of the text data found today is in the form of articles (news or otherwise), reviews, comments, historical documents and many other examples of information posted without labels or a way of making sense of this data. The issue could be easily solved by implementing a labeling convention from the very beginning of any platform, but often by the time anyone thinks of a convention, there are already many Terbites of text data that need to be sorted.

In that regard we aim to prove that the problem encountered when dealing with text data is not so hopless, in many casses 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**

In this article we have used three sources of text data, each presenting the problem of having no structure, thus making any filtering imposible.

The data sources in question are:

* [NPR News Articles](https://www.kaggle.com/datasets/gauravduttakiit/npr-data)
* [Spotify Million Song Dataset](•%09https:/www.kaggle.com/datasets/notshrirang/spotify-million-song-dataset)
* [Womens Clothing Review](https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews)

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

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

Both algorithms run upon the same assumption:

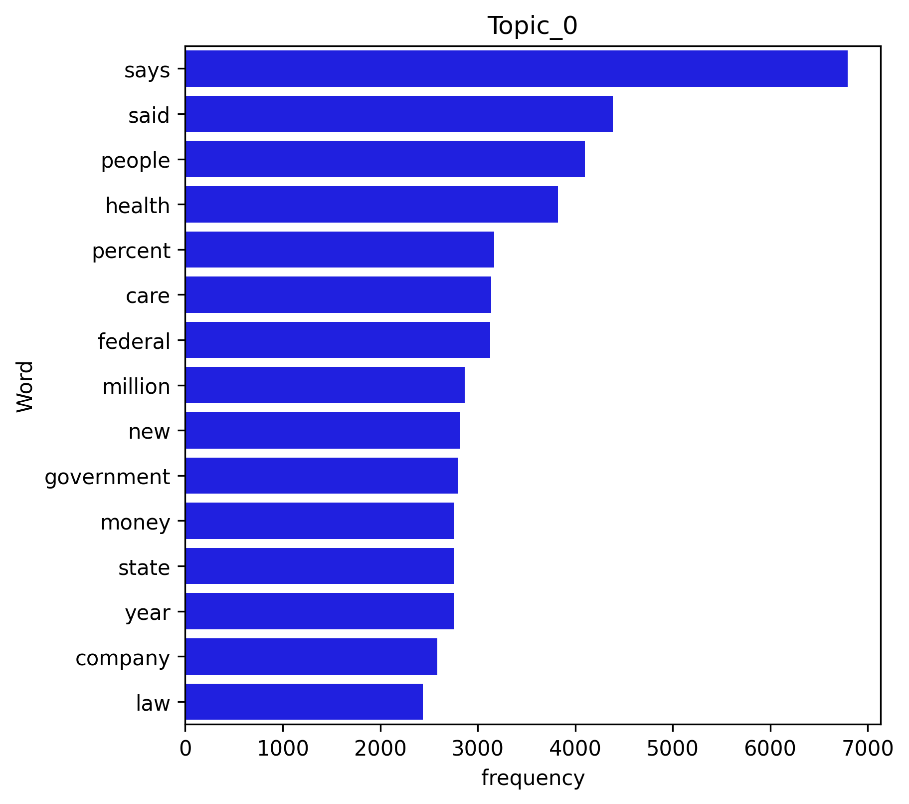
* Documents would cluster around a series of words that define that cluster
* Topics have a higher probability that words related to their topic would show up

**Text

Description automatically generated with medium confidence**Chart

Description automatically generated**NPR**

The NPR dataset consists of 11991 unlabled news articles. We know that these are news articles but we do not have a label, we do not know if it is about politics or weather.

We do know the most frequent words found in the articles, those beeing frequently encountered stop words that cant be used to determine a category as can be seen in **fig 2,**  and even when eliminating these stop words is is very hard to determine a subject as can be seen in **fig 3.** We can conclude that a simple word filter could not be used in order to categorize our articles. As a result of this conclusion we are left with the tools mentioned above, nameley Latent Dirichlet Allocation and Non-Negative Matrix Factorization.

We start out by assuming a clusters number of 7. 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 the this assumtion we end up with 7 clusters, each having a different frequencies for different words. For example for the first topic (Topic\_0), we can observe in **fig 4**, that the words with the highest frequency ten to be financial in nature. Notice that the clustering algorithm does not give a topical name. It is up to us to determine the appropiate name from the word frequencies obtained for each cluster.

We get the final result 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 **Finace**,
* Topic\_1: [attack state military war news department country according reported president russia security npr reports told government people says police said]. We could clasiffy this topic as related to **International News (from the point of view of the US)**
* 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 intuitivly but rather 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 in order to verify our assumptions.

*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 — Ive been thinking about the judges Ive seen on screen, and how their depiction might have intersected with public opinion through the years. Ce*

*Finance*

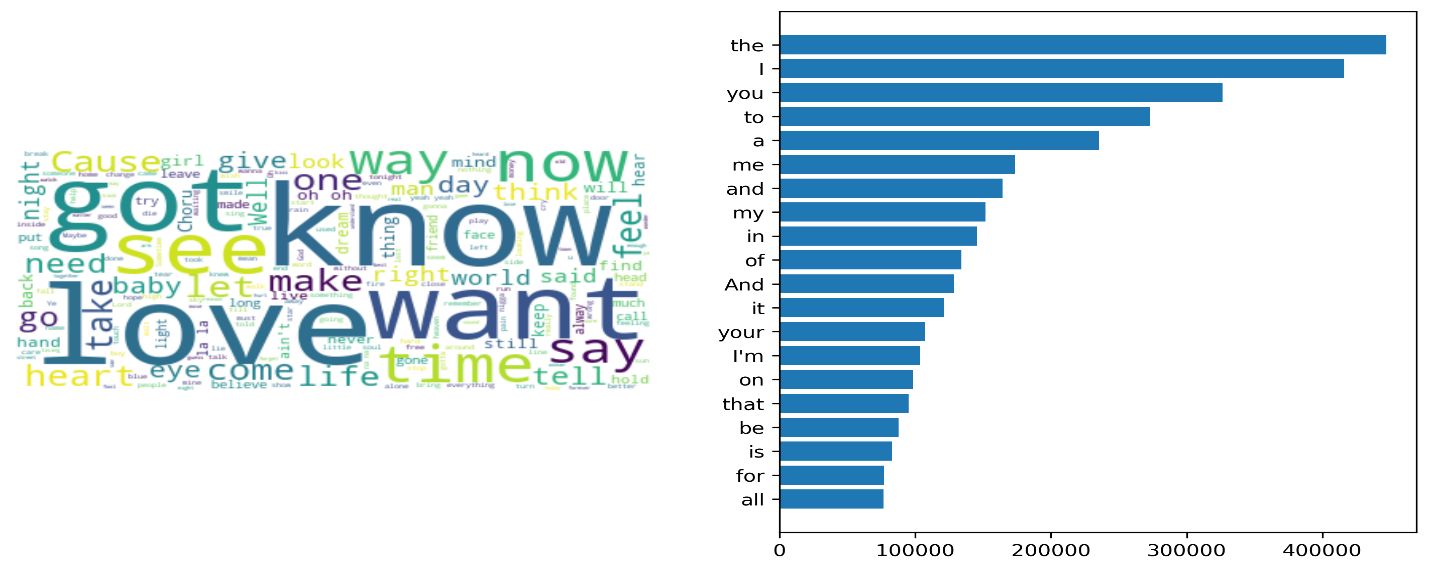
*Updated at 2:30 p. m. ET, For years, Puerto Rico has grappled with an debt crisis, watching as its bills have grown to more than $70 billion. Including what the U. S. territory owes to pension funds, that debt exceeds $120 billion. Now, Puerto Ricos struggle with its creditors has stepped into U. S.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

The sample above appears to confirm our assumptions.

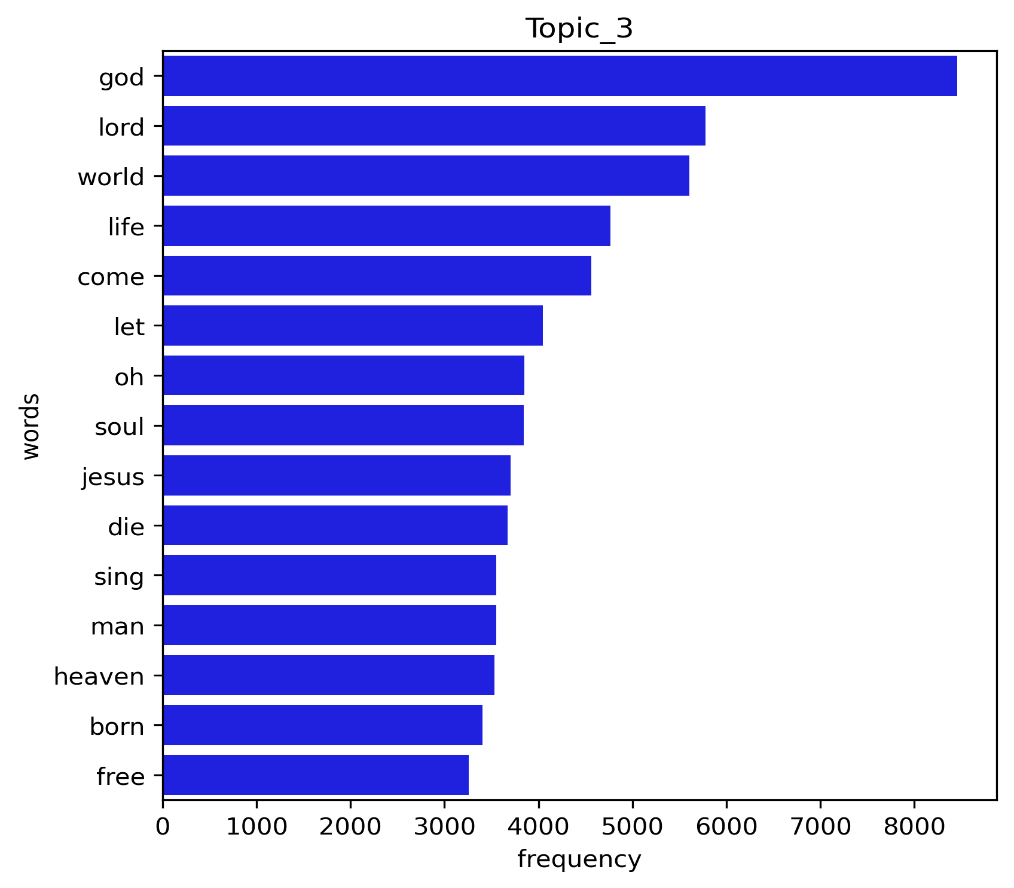
**Spotify**

We have noticed that this clustering method seems to work just fine for news articles, however we can try to use it in a different circumstance. We shall try to use it in on the lyrics of songs. Though we should limit our expectetion of what the algorithm can do. We should not expect the algorith to detect the genra of song. As the name suggests it can segment text on its topic but often the topic of a song is not the same as the genra.

The Spotify dataset has a total of 57650 song lyrics. We shall see if they can be segmented via a topic.



We can notice in **fig 5** that the most comon words in the songs are normal English stop words, as could be expected. However if we ignore these we can also notice that the most come words have to do with sentinmet. Thus we could assume that the Topical Segmetation shall lead to different topics as relating to sentiments.

The same algorithm as for the NPR data was run over the Spotify text data, with the same assumed 7 topics. We can observe in **fig 6,** the most frequent words encountered in the 3rd topic. As can be observed most of the words have a religious connotation, thus it can be assumed that the topic of the songs in the 3rd category are religious in nature.

But again the topic name is only implied by the frequency of the words, the algorithm does not provide a topic name but a number.

Thus we have to decided the topic name 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 overlapyin with **Rap Music.**

As before we need to test our assumptions with a few random examples:

Locational

Theyre 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

==================================================

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 theres no place for me

In the future you see

I dont understand you

Ive done all I c

Song : All My Love by Cliff Richard

==================================================

Our assumptions appear to be correct.

**Womens Clothing**

A picture containing text

Description automatically generatedNo 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 beying that if you sell womens clothing you migh asume that it is rather straight forward to understand what your customers are interested in when it comes to your product. And of course you can mesure their satisfaction with your product via the rating but you cant realy understand what your customer is interested in unless you go into the ratings, which is of course ustructured text data.

In such an enviroment we can use the exact same tools that have been used so far. In **fig 7**, we can noticed the most often used words but this does not give us the topics of interset. As wit the other casses we need to apply the Latent Dirichlet Allocation, as for the previous casses, with the assumption of 7 topics of interest. The result being the following topics with the adjescent 20 most frequent words:

* 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 focuse 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 focuse seems to be no the **Size,**
* Topic\_2 : *[long, colors, cute, nice, black, look, bought, fall, like, looks, jacket, comfortable, jeans, perfect, soft, wear, color, sweater, love, great],* the focuse of this topic seems to be the **Confort,**
* 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 diferentiate between Sizes 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 appear 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** seems to be the only topic that overlaps with a category of clothing, similar to rap in the music segment.

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

**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, not eno

==================================================

**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. imho i think this top might be better suited fo

==================================================

Our assumptions appear to be close to reality.

**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 greate chalanges in making sense of this data. However it appears that at least in the case of text data there are tools available that can help in achievine an understanding of unstructured text data.