An Example of Using Text Analytics

Context

One of the uses of high machine power and analytics tools is to analyze large volumes of texts and finding patterns in them – "Text Analytics". There are three major areas of analysis that can be done –

1. Frequency and pattern analysis – the words in the text can be identified and then counted to identify frequently appearing words, their patterns of appearance, and their connections with other words.
2. Sentiment analysis – a text not only contains words, but it also contains sentiments and emotions. Each word contributes to these sentiments and emotions in different ways. This analysis helps understand the tone of the text data, whether it is positive or negative and much more
3. Topic Modeling – if the text consists information on many different topics, this analysis helps identify the major topics and the terms in each of these topics. So, without reading, the analyst can guess the topics that are appearing more often in the text.

For the example in this document, we will be using a dataset containing 500 positive and negative comments about the workplace in Amazon, by its employees. The data looks like the following:

Review ID	Positive	Negative
1	You're surrounded by smart people and the projects are interesting, if a little daunting.	Internal tools proliferation has created a mess for trying to get to basic information. Most people are required to learn/understand SQL Database queries to get any actionable data.
2	Brand name is great. Have yet to meet somebody who is unfamiliar with Amazon. Hours weren't as bad as I had previously heard. But I guess can be long for corporate finance	Not the most stimulating work. Good brand name to work for but the work itself is mundane as it can get. As a financial analyst you do very little finance.
3	Good money. Interaction with some great minds in the world during internal conferences and sessions. Of course, the pride of being a part of earth's most customer centric experience	No proper growth plan for employees. Difficult promotion process requiring a lot more documentation than your actual deliverable.

In the rest of the document, we will look at the insights that can be generated from this document of texts.

Frequency of terms and pattern analysis

In this section we will look at which terms are frequent and other patterns. First, we can look at the top ten words that appear in the positive comments

word	n
pay	92
people	88
benefits	86
company	83
Amazon	78
time	71
lot	61
hours	47
job	44
environment	40

From this table, we can conclude that people like Amazon for the payment and the benefits, the brand of the company, and people who work there. Now let's look at the top ten words in the negative comments:

word	n
hours	78
people	70
management	61
time	60
life	57
balance	46
company	46
employees	39
job	39
lot	39

Interestingly, people have found their peers to be also a problem, besides working long hours and a tough management

Its always said that a picture is better than numbers and words. For visualizing tables like the word frequencies, we have something called the word clouds. So, let's look at the positive comments word cloud:

Figure 1 - Positive Comments Word Cloud

And now let's look at the negative comments word cloud:

Figure 2 - Negative comments Word Cloud

Both the word clouds to some extent reflect the pattern we found in the frequency table. Now instead of analyzing positive and negative comments separately, we can create a word cloud showing which are the common terms in both types of comments. These clouds are called commonality clouds and are very handy when we have many different texts and want to find out what are the common patterns in these texts. The commonality cloud looks like this:

Figure 3 - Commonality Cloud

This shows terms that people both are happy about and like to complain about. Mostly hours of work, work environment and the people. Another interesting cloud is the Comparison Cloud, which highlights the words that are frequent but different in positive and negative comments.

Comparison cloud

Figure 4 - Comparison Cloud

Based on the comparison cloud, we can see that for positive comments, pay, learning and Amazon company distinctly stand out, and for negative comments, management, working hours and employees stand out. One advantage of the comparison cloud analysis is that now we know that people/employees in Amazon is more distinctive in negative comments.

Besides word clouds, another useful visualization is a pyramid plot of the terms are common in both the positive and negative comments but vary in their frequency. One such plot is as follows:

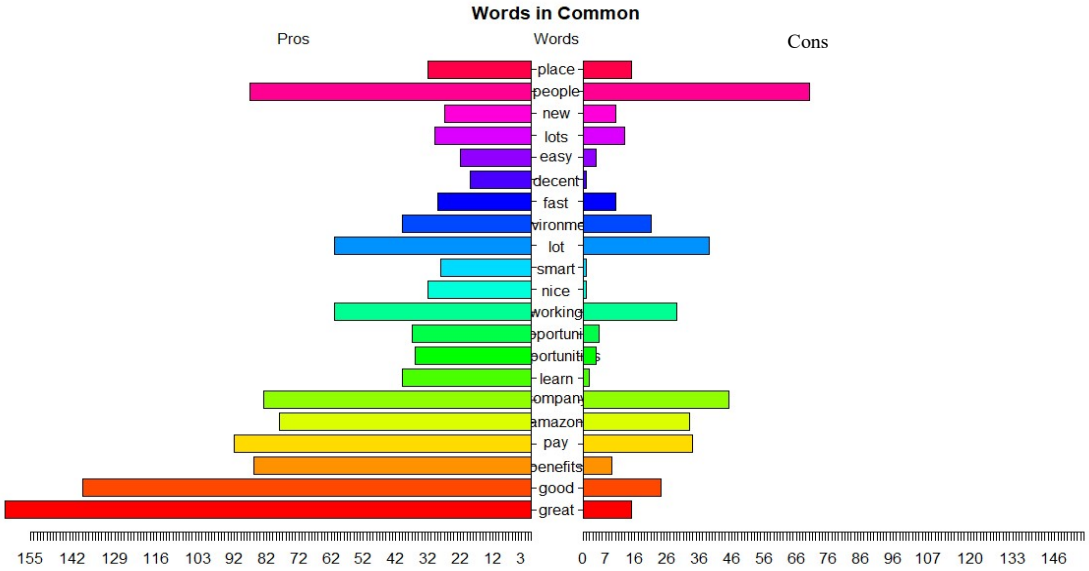


Figure 5- Pyramid plot of Pros and Cons for working in Amazon

Now that we have investigated term frequencies, another useful analysis is looking at how the terms in relate to each other in a document. These are called network plot of term associations. We select a certain term, and see how it is connected to other terms, and how those other terms relate to others. For our Amazon reviews scenario, we can first look at how the word benefit is connected in such a network:

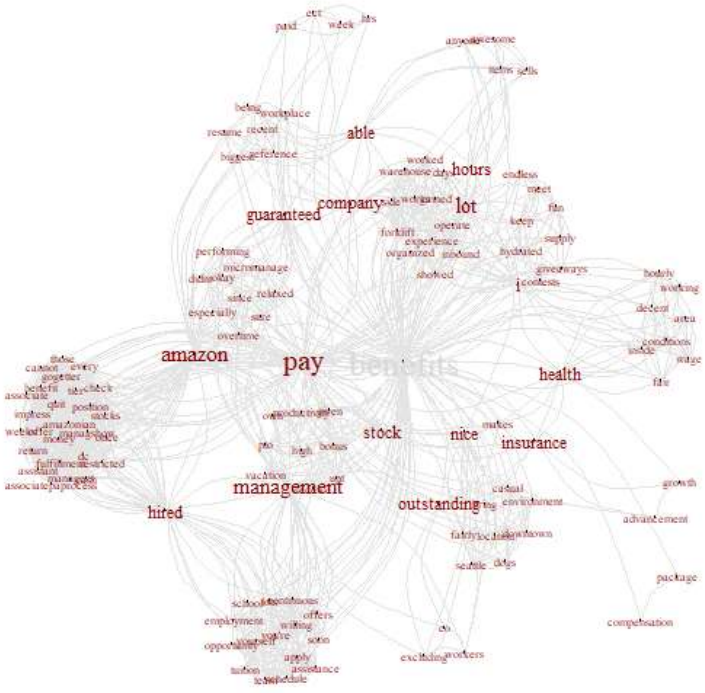


Figure 6 - Network diagram on the word "Benefit"

This diagram shows that words connected (related) to benefits are payment, health insurance, stock, management, company name. These are from the positive comments. Now we can also look at the negative comments and see how the word "stressful" is connected:

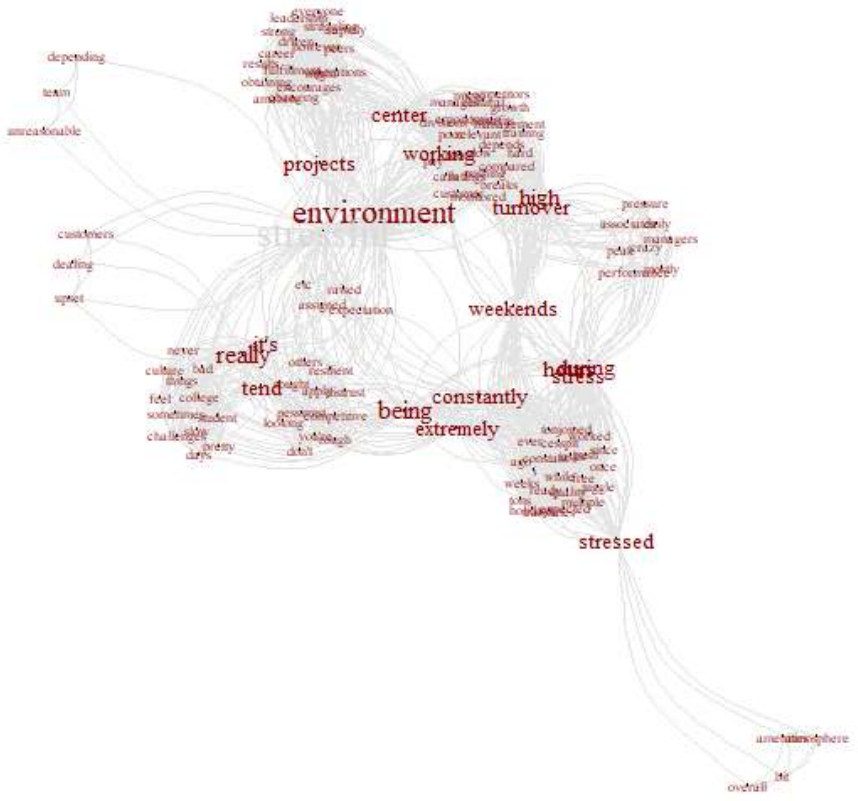


Figure 7 - Network Diagram on the word Stressful

This diagram shows that Amazon environment is stressful, and stress is due to projects and that it is constant. These network diagrams give us a sense how certain words relate to other themes, giving a clearer idea of the document.

The next set of analysis looks in to a visualization known as word association maps, which show to what extent two terms appear together. The idea is simple – if two words appear beside each other, a score of 1 is given, and if they don't appear side by side, a score is zero is given. The total appearances over the document is calculated and then an average is calculated, which shows the percentage of times words have been together. For our case, lets explore a word association map for the work "benefit", for words that appear with "benefit" more than 30% times:

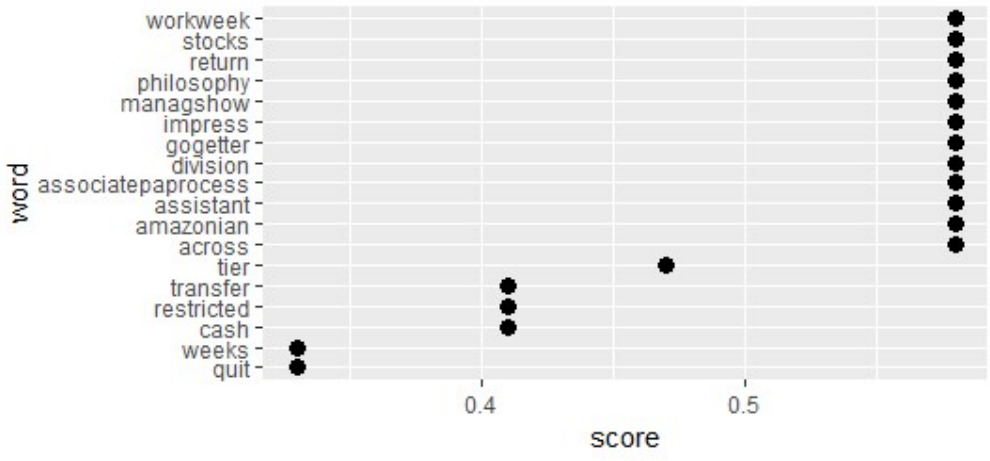


Figure 8 - Word association map for "Benefits"

We again get very similar insights to that of the word network diagram. Looking at the word association map of the word “stress” in the negative comments:

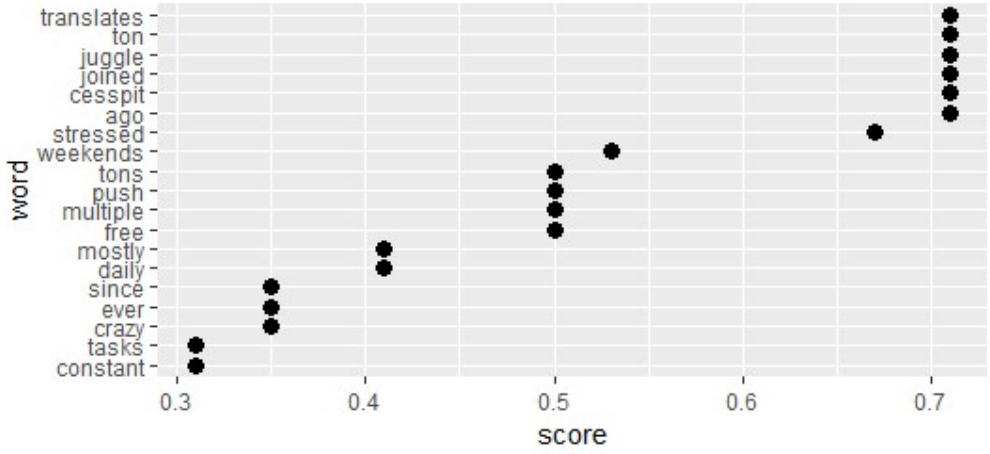


Figure 9 - Word association map for "Stress"

This shows that stress is associated with the word ton, meaning a lot of stress, juggle would mean that the workers are asked to juggle a lot of projects together.

So far, our analyses have been focused on the connections and the frequencies of one word/term. But at times, a single word has less meaning is comparison with a word pair. Such as the word “good” tells us much less than “good people”. So, for the next set of analysis, we will look at the frequency of pairs of terms appearing in the document:

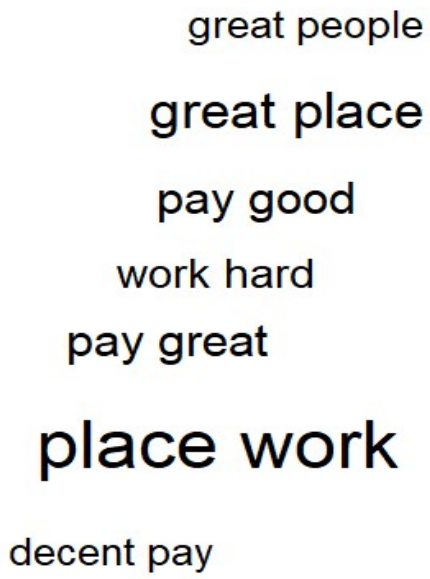


Figure 10 - Word cloud of positive term pairs

This word cloud clearly highlights the good aspects of working in Amazon, Now for the negative comments word cloud

Figure 11 - Word cloud for negative term pairs

These two paired term word clouds give clearer insights about the positive and negative comments of related to working in Amazon.

Sentiment Analysis

So far, we have looked into term frequency analysis. And it showed us what are the common positive aspects of working in Amazon and what are the common negative aspects. But words not only offer meanings, they also offer information about emotions and sentiments. So, in this section we will see what sentiment analysis tells us about working in Amazon. For the purpose of looking into overall sentiment, the positive and negative reviews are mixed together, and we observe them in aggregate. The first and the simplest thing we can observe is the overall document sentiment.

Total Sentences	Total Words	Average Polarity	Std Dev Polarity
993	19697	0.288	0.567

This information tells us the total number of sentences and the total number of words. Plus, it tells us the average polarity. The average polarity score is between -3 to +3. If its -3 then the overall document is extremely negative – most of the words used in the document has negative meaning, and vice versa if it is +3. So, from our analysis we can see that the document is on average slightly positive. This is interesting, because we used both positive and negative reviews, but the polarity is slightly positive. This means that the overall, there is many positive words compared to negative words, and that Amazon's employees, overall, consider Amazon to be quite good place to work.

We can see the distribution of polarity per document as follows:

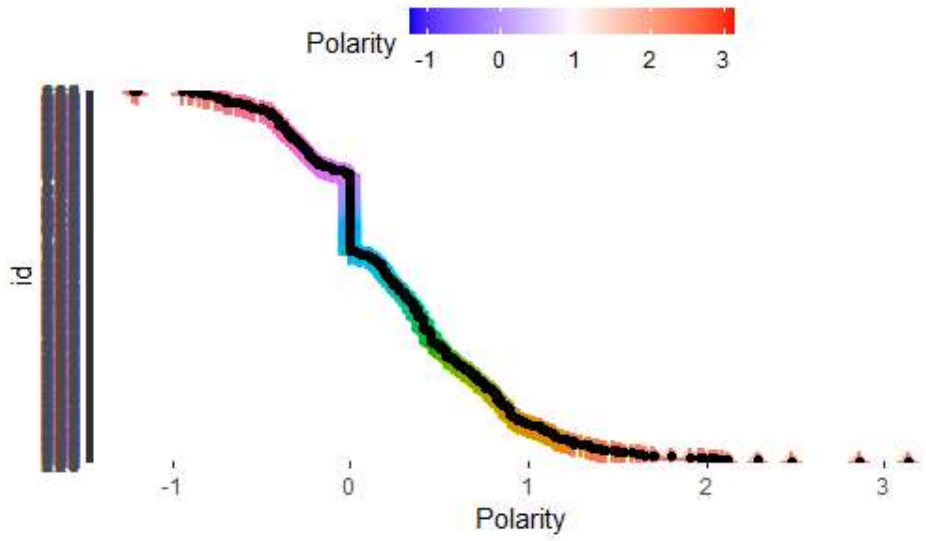


Figure 12 - Polarity score distribution per document

From this chart we can see that at max the negative polarity is around -1 and yet there is some extremely positive polarity at +3. Most of the observations (the fatness and the continuity of the line) is around 0 to +2 polarity, which means that overall there was a neutral to positive

sentiment about the work in Amazon. To further illustrate this finding, we can find out the total number of distinctly positive and distinctly negative terms in the document:

Terms	Count
Negative	561
Positive	1584

The next thing we can observe is how the sentiment of the text (consisting of multiple reviews) have changed over each review. For our case, this is not that insightful because we are looking at a cross-section of reviews (reviews collected at a fixed date and time). But if say these were comments on a product's performance, or news about a stock, where each document/review is collected per day/week, then this visualization can tell us at what time there is a negative shift and at what time there is a positive shift. Very useful in time series analysis.

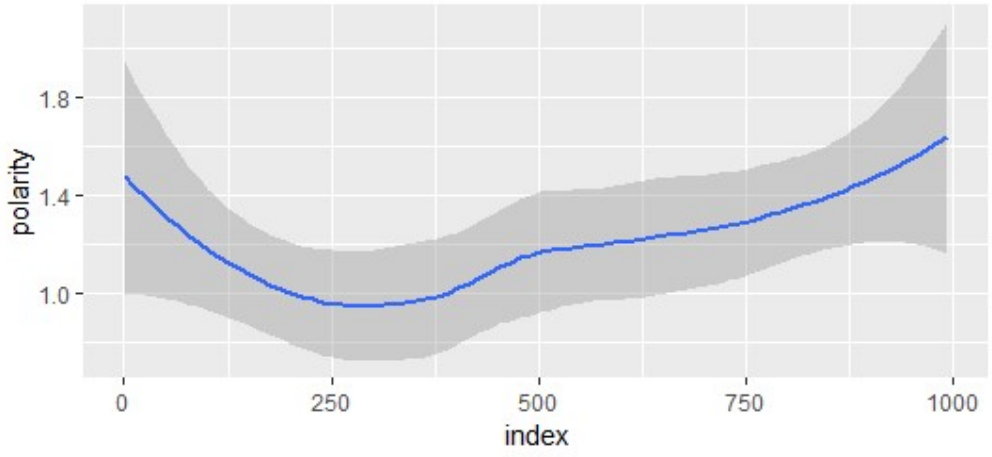


Figure 13 - Sentiment trend over the 1000 reviews

This trend again confirms out initial finding that there is not much negative sentiment in the reviews, because we can see that the polarity did not go below 0.

We are not only limited to positive and negative sentiments. We can also look at emotions conveyed by these texts. An established theory is that you can convey any form of emotion with a mixture of the basic six emotions – anger, anticipation, disgust, fear, joy and sadness. So, we can check what is the overall share of each of these emotions in our text of reviews:

Sentiment	total count
anger	181
anticipation	952
disgust	85
fear	195
joy	642
sadness	173
surprise	328
trust	927

A better way to represent the table is through a bar chart:

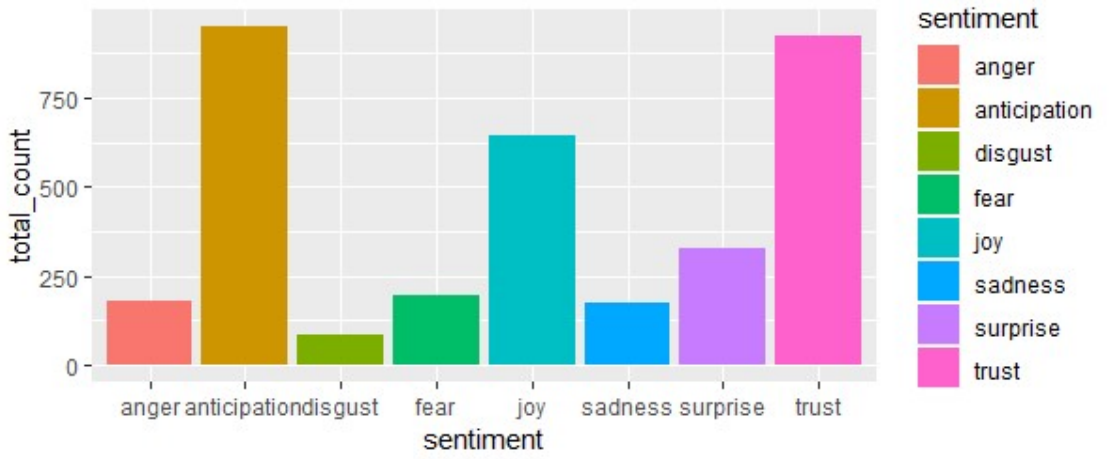


Figure 14 - Distribution of Six Emotions in the Text

This shows that most of the reviews by Amazon employees were having an anticipatory emotion or showing trust in Amazon, which is very good for the company. Positive emotion such as Joy and Surprise are quite high in frequency, and disgust, fear and sadness are low, though a significant proportion of reviews showed fear – something that the management need to work with.

Besides looking at only the sentiment scores and counts, we can also connect the last sections word frequency analysis with the sentiment analysis in this section. One simple thing to observe is that which words are contributing to most positive sentiments and which words are contributing to most negative sentiments. In this way we can identify what people mostly like and dislike about Amazon. The sentiment contribution graph looks like:

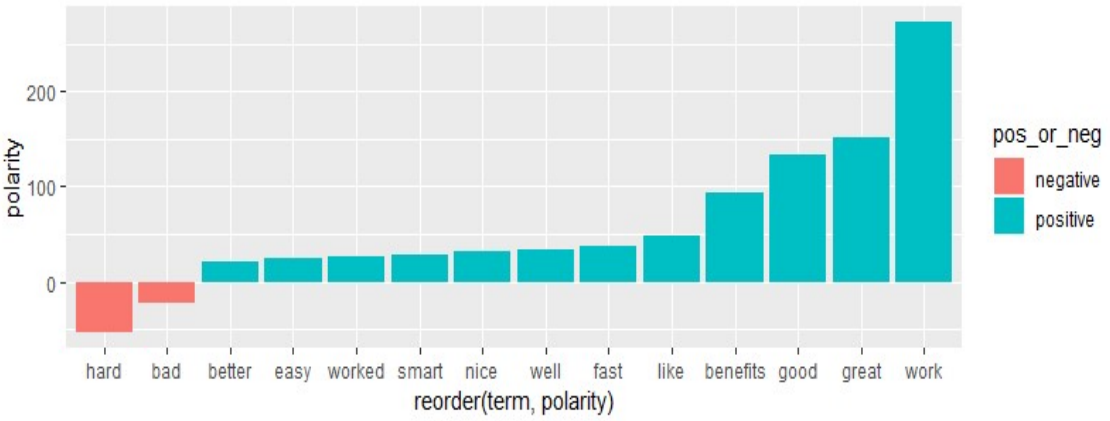


Figure 15 - Sentiment contribution per term

This shows that most of the negativities came from the word hard – which means it's the difficult nature of the work that leas to most negativities. Interestingly it's the work, good, great and benefits contributing to positive sentiment, indicating that it's the work-related benefits that are good, and this leads to the most positives in the reviews.

Another simple insight comes from word clouds, now divided between positive and negative sentiments:

Figure 16 -Positive and Negative sentiment word cloud

This clearly highlights which words are contributing to positive sentiment and which words are contributing to negative sentiments. These visualizations help pin point the source of positive and negative reviews, and can come in very handy for news analysis, customer review analysis and even electoral campaign chatter analysis.

A similar word cloud can be drawn for the 6 basic emotions, where the color of the words shows which words correspond to which emotion. The six emotions word cloud looks like:

Figure 17 - Six emotions Word Cloud

Another handy (and fancy) visualization is that of a radar chart, which shows the weight of each emotion in the document.

Figure 18 - Radar chart of basic 6 emotions

This radar chart shows again that the most prevalent emotions conveyed by the reviews were anticipation, trust and joy.

Topic Modeling

The last but definitely not the least tool available to text analyst is the concept of topic modeling. In simple terms, topic modeling helps you understand the main topics that are being talked about in the document. The mathematics behind identification of these topics is very complicated, and this is an emerging area of application of machine learning algorithms. Like Neural Networks, the idea originated quiet long back, but computers with high processing powers are making this tool useful in recent times. For a body of text, the algorithm tries to identify the optimal group of clusters of words which represent most of the things talked about in the document. Each of these groups are basically a topic that people are talking about in the document. Then the analyst can look at the words in the group to identify a central theme of that topic. The result – an idea of what are the topics being talked about in the document. For our case, we will look into the positive reviews for Amazon and identify what topics these positive reviews belong to. For our case, after running the topic modeling algorithm, we got 4 topics that the document was talking about. The words and the frequency of those words that belong to each topic are shown below as word clouds:

Topic 1

Figure 19 - Word cloud for terms belonging to Topic 1

From these words, we can conclude that the first topic theme is – "Benefits and payments". Mostly these reviews talk about how Amazon is a good place to grow, gives good salaries, have many benefits, etc.

Topic 2

Figure 20 - Word cloud for terms in Topic 2

These words indicate that the second topic theme is – "Nature of the Work", where people are talking about their coworkers, number of projects, overtime etc.

Topic 3

Figure 21 - Word Cloud of terms in Topic 3

Here the theme is "Work Environment", where the reviews are talking about the pace, fast mobbing nature of the work.

Topic 4

Figure 22 - Word Cloud of terms in Topic 4

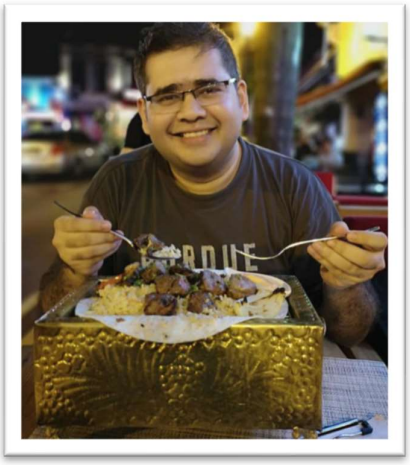
The theme for this topic is "Company" – reviews talking about working in amazon, career growth, amazing experience etc.

This sums up the concept of topic modeling. Its core advantage is that it helps us understand the themes of the text, and we can use these themes to separate the contents of the text and then do further analysis, to understand why these themes are coming up and what even how the sentiments and the emotions relate to these themes.

Concluding Remarks

Text analytics is slowly starting to get increasing attention in very diverse areas – electoral campaigns, stock news analysis, customer feedback analysis, forensic and judiciary science, just to name few. This document highlights some insights that can be generated using text analytics. If you have enjoyed going through the report, then it is time for you to start learning how to create these insights.

About the Author



Khan Muhammad Saqiful Alam is presently a PhD Student and Commonwealth Scholar in National University of Singapore (NUS). His area of research is in Innovation and Application of Machine Learning in Business Strategy. Besides this, he is also an adviser of Intelligent Machines Limited (a company in Bangladesh working with big data), and a trainer of a World Bank project for helping entrepreneurs to scale up. Previously he used to be a Senior Lecturer in North South University and part time guest lecturer in Institute of Business Administration (IBA), Dhaka University. Also, he is involved in Grameen Phone Accelerator program (the largest startup accelerator in Bangladesh), Upskill (a project by Startup Bangladesh) and Light Castle Partners (a Management Consultancy) as a trainer for Data Analytics and Marketing Analytics. Finally, he also acts as one of the community managers of Google Business Group Sonargaon as a Google Analytics expert and as a mentor for Google Business Group Singapore. Saqif pursued his MSc in Operations, Project and Supply Chain Management in Manchester Business School, University of Manchester, where he did his dissertation in risk management frameworks and the use of simulation in Risk Management. Before that, he finished his Bachelor of Business Administration from IBA, Dhaka University, Bangladesh.