ADVANCED MACHINE LEARNING

Assignment 1

OPINION MINING

By , Arindam Adak
Ramakrishna Mission Vivekananda University
Roll No. : B16755
Program Name : BDA
Semester : Third

Problem Release date:- 12/11/2017
Date of Submission :- 30/11/2017

1. Introduction: What is Opinion Mining ?

Opinion mining is a type of natural language processing for tracking the mood of the public about a particular product or subject.Opinion mining can be useful in several ways. It can help marketers evaluate the success of an ad campaign or new product launch, determine which versions of a product or service are popular and identify which demographics like or dislike particular product features. Or as we will discuss in our dataset the kind of person our country wants as their Prime Minster.

1.1 Dataset Description

The dataset consists of opinion of 38 persons when they were asked what kind of person they want as their Prime Minister or what should be the ideal qualities of there Prime Minister. The data set is supplied to us in an excel file. Each row from second row onward in the excel file indicates the comments of a particular person.

1.2 Problem Statement

The task is to find the significant comments from this data.Some experts had already reviewed the data set and found significant qualities manually from the data, which results are already given to us. The objective is to analyze the performance of your method using the human coding.

2. Methodologies

2.1 Abstract

• Firstly I changed all the uppercase words to lower case.
• Next removed the comma, fullstops and punctuations from the dataset.
• Next we find out all the words that are nouns and adjectives in the dataset (using Wordnet), this also takes care of removing the stopwords from the dataset.
• Then we create a vocabulary of words using the above mentioned words and all there synonyms and derivationally related words.
• Then using a distance measure between words, in this case I have used Wu-Palmer Similarity measure[1]
• The next step is to cluster the words that are in our vocabulary. The Clustering Algorithm is discussed in the next subsection.

[1] Wu-Palmer Similarity

2.2 Details

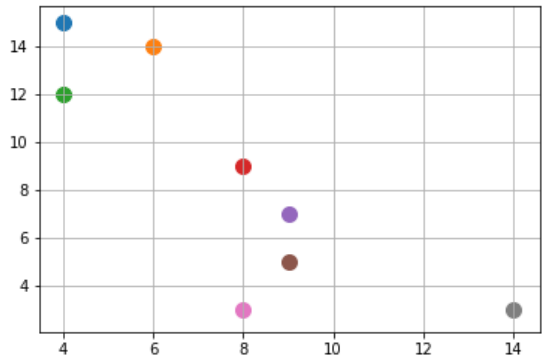
• We picked out only nouns and adjective from the document because they describe the personality of the Prime Minister best, following are the table of them and there number of occurrence in the document :
'leadership': 9, 'honesty': 7, 'intelligence': 5, 'good': 4, 'president': 4, 'experience': 4, 'integrity': 3, 'determination': 3, 'country': 3, 'smart': 3, 'people': 3, 'others': 3, 'ability': 3, 'decisiveness': 2, 'teamwork': 2, 'understanding': 2, 'i': 2, 'dedication': 2, 'someone': 2, 'commitment': 2, 'calm': 2, 'politics': 2, 'morality': 2, 'compassion': 2, 'skills': 2, 'different': 2, 'groups': 2, 'strong': 2, 'passion': 1, 'u': 1, 'qualified': 1, 'work': 1, 'ethic': 1, 'systems': 1, 'progressive': 1, 'common': 1, 'sense': 1, 'sure': 1, 'lot': 1, 'qualities': 1, 'confident': 1, 'eloquent': 1, 'logical': 1, 'wise': 1, 'humble': 1, 'negotiator': 1, 'appeals': 1, 'whole': 1, 'county': 1, 'character': 1, 'course': 1, 'leader': 1, 'capable': 1, 'crucial': 1, 'decisions': 1, 'time': 1, 'empathy': 1, 'management': 1, 'goals': 1, 'achievement': 1, 'long': 1, 'term': 1, 'levelheadedness': 1, 'inclusivity': 1, 'making': 1, 'promises': 1, 'idiot': 1, 'peaceful': 1, 'deliberate': 1, 'proud': 1, 'background': 1, 'knowledgeable': 1, 'agreeable': 1, 'able': 1, 'selfless': 1, 'level-headed': 1, 'sophisticated': 1, 'mature': 1, 'honest': 1, 'ideas': 1, 'polite': 1, 'nice': 1, 'unity': 1, 'prejudices': 1, 'willingness': 1, 'kindness': 1, 'umm': 1, 'equality': 1, 'trustworthiness': 1, 'problem': 1, 'genuine': 1, 'care': 1, 'bravery': 1, 'hard': 1, 'trustworthy': 1, 'political': 1, 'office': 1, 'ulterior': 1, 'motives': 1, 'stressful': 1, 'situations': 1, 'international': 1, 'relations': 1, 'constitution': 1, 'underprivileged': 1, 'join': 1, 'approachable': 1, 'basic': 1, 'sociology': 1, 'diplomacy': 1, 'humility': 1, 'vast': 1, 'knowledge': 1
• At the time of vocabulary creation the problem is that Wordnet Wu-palmer similarity measure only allows us to get a similarity measure between two nouns but not between a noun and a adjective. And when I try to get the similarity the results are not satisfactory. That is why i took all the similar nouns of the adjectives and add them to the vocabulary. Then we can easily get the Wu Palmer similarity between any two such nouns.Then we created a vocabulary
• The next step is our clustering process.

2.2.1 Clustering

We could not use any other types of clustering like - in case of K-means we need to give prior information about how many clusters are needed and in case of hierarchical clustering we need to make a tree to fix a distance above which we want to cut , but as the word similarity matrix is of dimensions 526*526 we cannot really visualize it properly,that is why I used the following way of clustering.I will discuss it using an example:
We will try to group next set of points: (4,15), (6,14), (4,12), (8,9), (9,7), (9,5), (8,3), (14,3) which looks like this:
After that we have we can move on to creating similarity matrix. We will do this by calculating similarity distance between all objects in the set. The distance measure in this case take is Euclidean Distance[2].We need to define the threshold value which will be used to separate objects into different groups if the similarity distance is greater than

[2] Euclidean Distance

Figure 1: Data Points



the provided threshold value. In this example let's define our threshold as d = 4.00.Now, generating similarity matrix should be easy. Rows and columns are representing the exact object, so for example if row is at index 1 and column is at index 2, it means we are considering similarity distance between object 1 and object 2. But we are not putting that distance into the matrix, because we need only the chain of similarity between the objects. So, comparing the distance between objects with the threshold value should decide if we are putting 1 or 0 at the chosen place in the matrix. D12=2.2360, so D34 < t which means that we are putting number 1 into that place. So our Similarity Matrix is

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Now we have similarity matrix, we find the row with the largest sum. In this example there are 5 rows with the largest sum so we pick the first one. We have selected first row in the matrix (i=0) and now all the columns which contain number 1 in the selected row should be added to a new group which will be named G1,G1 = 1,2,3

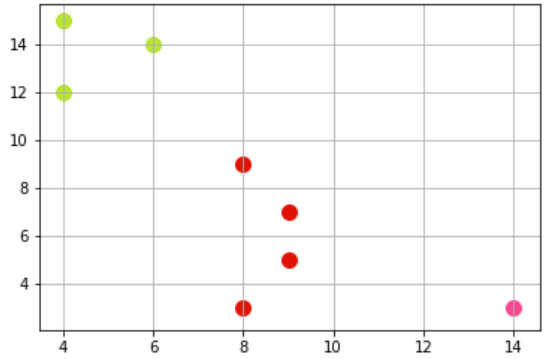
Next step in adding objects to a current group is to take those indices from G1, and repeat the previous procedure. If we look at the row with the index value of 1, there are again 0, 1 ,2 indices so we don't have to add nothing to a group. Then we take index value of 2 and look at that row but there's also nothing new to add.
Next step after we got a new group is to delete or reset values in the matrix, but only in the rows and columns with indices of newly grouped objects. So if we grouped objects with

indices 0, 1, 2 then the matrix should look like this:

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Now we are going back to the step no. 2, where we choose a new row to be processed and we repeat the same procedure until our whole matrix only contains zeroes as values. Result of the clustering procedure : (0, 1, 2), (3, 4, 5, 6), (7)

Figure 2: Data Points



In our case I used the distance between the words as Wu-Palmer similarity (which is between 0 and 1) and used threshold point as 0.8 and computed the clusters. I used a threshold point of 0.8 since ,if we took 0.85 as cut off if I give an example the words "honesty" , "morality" are in two different groups but that is not very good since both are kind of similar sensed words, and if I take 0.7 as a cut off it brings more words as garbage and these doesn't make much of a sense.

3. Results From The Clustering and Analysis

After running the clustering algorithm I implemented on the 576*576 similarity matrix I get 345. (The clusters are given in the appendix) out of which there are 267 clusters that contain only one element. As anyone can see the biggest drawback is that I have lots of clusters and we do not get the words "confident" "humble" into my clusters which are in the human coding since we cannot get the adjectives inside the total word list. But I did get all the nouns that are in the hand coded versions. But among the human coding we see a limited numbers of words very few. Since obviously the amount of data was very low. But from my coding i get a lot more options , I get words like ("background", "experience") ("fearlessness","humilty") in one group which makes quite a lot of sense in this case because people would want a good background and experience for the Prime Minister of their country . And also people would want there Prime Minister to show both the qualities fearlessness when needed as well as "humility".

4. Conclusion and Future work

Clearly, in my way of working the biggest drawback is building up the vocabulary because I couldn't include adjectives in there original forms.So clearly in that case the human coding is better but this process by using the wordnet does give a lot many options. Also the computation time of building up a similarity matrix is also time consuming because the number of words included in the vocabulary. The future work should to find out a way to measure similarity between a noun and an adjective and noun. And also a filtering of the vocabulary so that one can bring the the computation time of building up the similarity matrix.