UNSUPERVISED OPINION MINING FROM TEXT OPINIONS USING WORDNET

Nabarup Maity
Roll No : B17705
Ramakrishna Mission Vivekananda Educational and Research Institute
MSc Big Data Analytics

Problem Release date: 09/04/2018
Date of Submission: 04/21/2018

1.Introduction :

Sentiment mining and opinion mining from text have recently drawn a lot of attention because of their many potential applications for companies and organizations such as extracting customer sentiments, automated recommender systems, or deducing public opinion about a certain topic[1] . Two main goals of sentiment mining are: (1) to determine whether a given text contains an opinion as opposed to being factual or objective, and (2) to extract the sentiment of a given text by classifying it as positive, negative, or neutral with respect to the given target [2]. Furthermore, sentiment mining can be performed at the sentence level or document level. In general, sentence-level mining is made complex by the fact that the semantic orientation of words is highly context-dependent, and document-level mining is made complex by the fact that one document may contain several contradictory opinions about the same target.

2. Problem Statement :

The data set contains opinions of 38 different people on the following question:
”What qualities do you think are necessary to be the prime minister of India?”
The given data set is supplied to us is in an excel file. Each row from second row onward in the excel file indicates the comment of a particular person. The task is to find the significant comments from this data.

3. Methodology

1*. Form the vocabulary of the data set.*

We consider only the nouns and adjectives and simply ignore the other words, since for opinion mining these two are important only and since we try to find the qualities of the PM.

Steps:

a) Convert the file into lower case.

b) Remove puntuation, commas and replacing space.

c) Tokenising with nltk(tokenise each row)

d) Putting into a single list and tagged the list using nltk.

e) Take only the nouns and adjectives.

.	JJ	Adjective
.	JJR	Adjective, comparative
.	JJS	Adjective, superlative
.	NN	Noun, singular or mass
.	NNS	Noun, plural
.	NNP	Proper noun, singular
.	NNPS	Proper noun, plural

f) Find the frequency of each word and order the words in order of decreasing frequency .

2. Find the synonyms for each word (using Wordnet synsets) and include them in the vocabulary :

Steps:

a) Finding the noun and adjective synsets by extract the synonyms that are nouns or adjective.
b) Removing the empty lists.

3. Find the derivationally related forms of each word, if any and add them to the vocabulary :

Steps :
a) We lemmalize the vocabulary by using the function lemmas() and derivationally_related_forms() .
b) Remove the blank list and for each lemma take their unique name.
c) Find their noun and adj synsets of the names(as they are only string).
d) Remove the empty lists and add these names to new_voc list

4. Find similar words from the vocabulary and merge them to form a cluster :

Steps:
a) Clustering using wup/path similarity.
b) We fix a particular threshold to stop merging clusters after certain steps.
c) Remove the empty clusters

5. Recalculate the frequency of each cluster and order them in decreasing order of frequency :

Steps :
a) We consider only the top frequency clusters (here the clusters those frequency are greater than 1.)
b) Compare the words in these clusters with the
human coding .

TOOLS :

- **WORDNET :**

WordNet is a Lexical database for the English language[3]. It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. WordNet can thus be seen as a combination of dictionary and thesaurus.

- **Wu-Palmer Similarity**

TheWu-Palmer calculates the relativeness by considering the depths of the two synsets in the wordnet Taxonomies, along with the depth of the LCS(least common subsumer) .

The formula is **score = 2*depth (lcs)/(depth(s1)+depth(s2) .**

This means that 0<score<=1. This score can never be zero because the edpth of the LCS is never zero (the depth of the root of a taxonomy is one). The score is one if the two input concepts are same.

4. ANALYSIS THE RESULT :

- Using **Wu-Palmer Similarity** we first check the clusters. The threshold was between 70 to 95 . IF we increase the threshold Then the clusters become more sparse. We look a cluster ['skill', 'leadership', 'ability', 'intelligence'] .It is when threshold is 90. Now if we decrease the threshold it become ['brightness', 'memory', 'reason', 'ability', 'capability', 'leadership', 'cognition', 'skill', 'sense', 'intelligence', 'novelty', 'mind'].

- For most of the threshold values(between 80-90) the human codes honesty, Intelligence/knowledge, leadership and humble are present in the clusters.

- If we take all cluster(including frequency 1) then we get the remaining qualities confident, determination, diplomacy.

- We can not get any cluster for any threshold for the term - long term vision, political skills, relate to diverse groups. Even that are not given in the given data.

- For **Wup similarity** we get the best cluster that is similar to human coding. The threshold value is 90 and we get 42 clusters.

- Now we run this method using **Path similarity**. The result is same but for good clusters(that similar to the human code) the threshold is lies between 32 to 45.

- For **Path similarity** we got best threshold at 49. [NB: If the threshold become 50 all words become cluster with frequency 1. i.e, we got the words as clusters.] The clusters are similar as we got using **wup similarity**. We get 63 clusters.

- So after comparing the human code with our cluster we can say that the qualities - confident, determination, diplomacy, intelligence, knowledge, honesty, leadership and humble are necessary to be the prime minister of India.

LIMITATIONS of our Process:

WordNet does not include information about the etymology or the pronunciation of words and it contains only limited information about usage. WordNet aims to cover most of everyday English and does not include much domain-specific terminology. In this type of problem it is good enough but We should mine opinions in a better way.

5. CONCLUSION and FUTURE WORKS :

Opinion mining become popular research area due to the increasing number of internet users, social media etc.
Here we have work on aspect level analysis using WordNet. The method used here is very simple and
domain independent. In this assignment we present our experiment with reviews which generate great result.

In future we will work it with the blogs and other dataset. Also try to improve the method.

6. REFERENCES :

[1] B.Pang and L.Lee, “A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts” in Proceedings of the ACL, 2004

[2] A. Abbasi, H. Chen, and A. Salem, “Sentiment Analysis in Multiple Languages: Feature selection for opinion classification in Web forums,” ACM Transactions on Information Systems (TOIS), v. 26, no. 3, pp. 12, 2008.

[3]G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller. 1990. WordNet: An online lexical database. Int. J. Lexicograph. 3, 4, pp. 235–244.