Behavior of Symptoms on Twitter

Dennis Salcedo
El Bosque University/Bogotá-Colombia
dsalcedop@unbosque.edu.co

Alejandro León
El Bosque University/Bogotá-Colombia
alejandroleon@unbosque.edu.co

Abstract

With the amount of data available on social networks, new methodologies for the analysis of information are needed. Some methods allow the users to combine different types of data in order to extract relevant information.

In this context, the present paper shows the application of a model via a platform in order to group together information generated by Twitter users, thus facilitating the detection of trends and data related to particular symptoms. In order to implement the model, an analyzing tool that uses the Levenshtein distance was developed, to determine exactly what is required to convert a text into the following texts: 'gripa'-"flu", "dolor de cabeza"-"headache", 'dolor de estomago'-"stomachache", 'fiebre'-"fever" and 'tos'-"cough" in the area of Bogotá. Among the information collected, identifiable patterns emerged for each one of the texts.

1 Introduction

Social networks are important because of their user's opinions on diverse topics (Martos E 2010 and Soumen C 2003). Gathering, processing and analyzing those opinions is an important factor for making decisions, therefore, the study or analysis of mass opinion through social networks is an issue that has emerged as a key methodology in modern sciences (Linto C 2006) – such as psychology (Daniel T. Gilbert, Susan T. Fizke, Gardner L 1998) and economy (Ana S 2003), among many others – because it has an impact on the content generated by users (Robin B, Jonathan G, Andreas H, Robert J 2001).

Therefore, a characteristic Levenshtein distance analyzer, linking with diagrams of relationship and feeling to see how the information is behaving was needed. The result provided a close approach to the people who tweeted with a negative attitude to the symptoms.

The importance of conducting an analysis of information and structure for symptoms is important for the study of data mining and big data. This means, it can be determined how many users posting a tweet are actually sick.

2 Information and Levenshtein Distance

The information is collected using a python script in which the Twitter and json libraries are used. In order to gather tweets associated to a city, they need to be linked to a city code by using the platform of coordinates, GeoPlanet (Willi S 2010).

This way, it is possible to find all the tweets in the city of Bogotá that contains the associated symptoms. A basic algorithm used removes information such as special characters and blank spaces in the thread. Additionally, it gets a portion of the thread to perform this analysis on the desired patterns; the resulting thread is built with a maximum of 4 words.

The Levenshtein Distance shows the number of operations that you need in a thread to finish another one (Vladimir I 1965). It was used because of the simplicity of the algorithm but not for his efficiency. It is noted that there are similarities in the number of operations performed to obtain the desired pattern (Fig. 1).

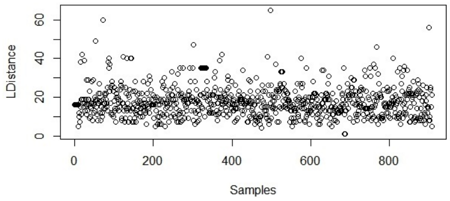

Figure 1: Levensthein Distance applied to the corpus that contains flu.

Therefore relevant information showing the pattern associated with the symptom to improve the filter using the analyzer is taken.

3 Experimentation

Clustering techniques are applied to the gathered information to determine how the information is behaving in two main aspects relationship of concepts and analysis of feeling. The relationship of concepts is established in order to show the relationship of the symptom in the information, by totaling the number of repetitions of the concepts (Fig. 2). The analysis of feeling is applied with dictionaries of positive and negative concepts, giving it a score in the remaining of the same one. Note that there is a pattern in the graph clustering associated with this symptom.

The symptom is regarded as high priority in the corpus (Jurgita M 2002), Therefore, the tree diagram or a relation of concepts are performed for observing the most relevant key points within the information.

During the experiment, a part of the information was used to primarily watch the junction of concepts you can see if the concepts have a positive or negative order in relation to the symptom. The diagram shows the relationship of the concepts:

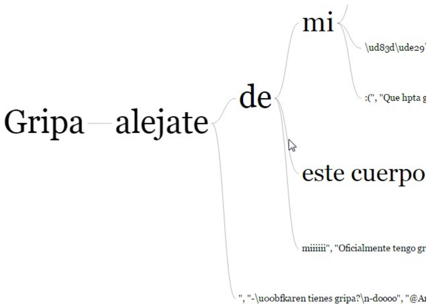
Gripa
alejate
de
mi
\ud83d\ude29'
:(", "Que hpta g
este cuerpo
miiiiii", "Oficialmente tengo gr
", "-\uoobfkaren tienes gripa?\n-doooo", "@A

Figure 2: Combination of concepts ’Gripa’-’Alejate’-“Move away” to describe the relationship between concepts. (Martin W and Fernanda B. Viégas 2008)

The relationship of concepts clearly shows that the symptom is associated with negative concepts; this is done in order to perform an analysis of feeling to the information.

The analysis of feeling performs a score to more traits of the corpus and notes that there are more negative comments than positive ones; this means that there is a high possibility that the person who posts a tweet is sick.

4 Conclusions

Performing an analysis of mass information requires a large amount of data processing and cleaning. The content and the intensity of the information are key factors in determining the maximum percentage of consistency.

The Levenshtein distance and the sentiment analysis can be considered an approach to extract relevant information about symptoms.

Future work will concentrate on obtaining more details on how to improve extraction and cleaning of the corpus to obtain accurate and clear results. The analysis of information with techniques such as artificial intelligence, machine learning and data mining helps to study the information in detail as well as in the decision making for that data set.

We need more work to indicate through the time how these symptoms spread in the city and what are the areas with more reported cases.

The common terms for the symptoms work but we must find a way to relate medical classifications to them.

References

Martos E Carrión. 2010. *Analyzes of new forms of communication via virtual communities or social networks*,Valencia, Spain,(1):2

Soumen C. 2003. *Mining the Web*. San Francisco, California.

Linto C Freeman. 2006. *The Development of Social Network Analysis*. Vancouver BC, Canada.

Daniel T. Gilbert, Susan T. Fizke, Gardner L. 1998. *The handbook of social psychology*, (2):226-227.

Ana S Garcia. 2003. *Social networks as a tool for structural analysis input-output*. University of Oviedo, Department of Economics.

Robin B, Jonathan G, Andreas H, Robert J. 2001. *Recommendation in the Social Web.*

Willi S 2010. *Free geo data solutions compared: GeoNames.org vs. Yahoo! GeoPlanet.*

Vladimir I Levenshtein 1965. *Binary codes capable of correcting deletions, incursions, and reversals.* Doklady Soviet Physics 10 (8): 707-710

Jurgita M 2002. *STUDIES ABOUT LANGUAGES No:3.*

Martin W and Fernanda B. 2008. *The Word Tree, an Interactive Visual Concordance*. Visualization and Computer Graphics, IEEE Transactions on (Volume:14 , Issue: 6)