***From Tweet to Graph: Social Network Analysis for Semantic Information Extraction***

This paper represents a study along the cutting edge of the current analysis of online social network in relation with the contents communicated among users. Twitter data is carefully selected around a fixed hash-tag in order to study the specified content in relation with other contents that users bring to connection. A separate network of hash-tags related (in tweets) is constructed for different days; the networks are analyzed within advanced Gephi package, providing several measures - degree, betweenness centrality, communities, as well as the longest path, by which the evolution of communication around specified concepts is quantified.

The purpose of our research is to find visualizations of the information as *networks* (nodes and links between information, in this case between the hashtags) in order to understand the meaning of the use of different elements in the tweet. How hashtags are related within other concepts and words used? Is it possible to visualize the central problematic of the movement #noalospluris by using a network? How clusters and subgroups of information can help to understand the meaning of the movement?

**Methodology**

In the proposed method, we analyze tweets using three steps. In the first step, we extract tweets from Twitter containing a certain hashtag. In the second step, we parse the extracted tweets in order to clean it and to create two different tables: one with all the labels of the important words contained in the corpus and another containing the relationships between each couple of words. These two steps are automatically performed by using R which is an interpreted computer language designed for statistical data analysis. The third step concerns the creation of the network by using Gephi that is an open source software for graph and network analysis.

**Part 1**

In this way, by using R we have developed a computer program that is able to:

1) Extract tweets containing a certain tag, in our case we extracted tweets containing #noalospluris. It is noteworthy to say that the API of Twitter that is being used only allows the recovery of tweets that have between 6 and 9 days of creation. It’s not possible to retrieve tweets older than 10 days, so in some cases where the participation is very high the extraction process should be done daily.

2) Eliminate stop words and punctuation.

3) Eliminate website addresses (anything that starts with “http”).

4) Eliminate actors (all that contains the symbol ‘@’ as the first character).

5) Delete numbers and special characters (except the symbol ‘#’ and letting the numbers and special characters in tags).

6) Change the corpus to lowercase.

7) Eliminate accents.

8) Get the frequency for each word.

9) Generate the nodes for words that are above the third quintile (Q3).

10) Determinate the edges for each node.

Once we got the tweets we make a pre-processing, manipulation, cleaning and formatting in order to compose the corpus. The main structure for managing documents, by using tm package in R, is the *Corpus* that represents a collection of text documents (in our case a collection of tweets). With tm package R, we are able to create a Term-Document Matrix from our corpus. The matrix is exported as a .CSV file in order to be used in Gephi.

**Part 2**

APPLYING SOCIAL NETWORK ANALYSIS

The tables that were generated by R (table of nodes and edges) are used in Gephi in order to:

• Generate a network of the corpus.

• Calculate the degree for each node.

• Calculate the centrality for each node.

• Detect the main communities.

***Subgroups***

When the tweets of a particular topic are analyzed they might appear like if they are disorganized without a specific center where it prevails only the spontaneity and the anarchy. But, while the tweets are set in our mind like a swarm of birds or insects attacking irrationally, as the meaning of Twitter, we find that the interior of the network is organized rational and creative, a “*swarm intelligence*”. In this logic, according to Arquilla and Ronfeldt [34], the swarm reaches its maximum effectiveness and its greater attack power when the network members aren’t together to combat their forces but they concentrate on the dispersion, sharing important information so it can reach his final destination.