Study of n-grams in illness related tweets

Juan Ignacio Gil Gómez

July 15, 2014

1 Description

An study of illness related n-grams in Twitter

We will study the distribution of n-grams (contiguous sequences of n-words) in the illness related tweets.

A broad outline of the study will be:

1. Study and visualization of the distribution of the n-grams[1] of the whole distribution of tweets.
2. Study and visualization of the distribution of n-grams of each group of diseases. Search for the n-grams characteristics of each group (n-grams with a high frecuency relative to the distribution of the whole tweets.
3. Study and visualization of the distribution of n-grams of each individual disease. Search for the n-grams characteristics of each group (n-grams with a high frecuency relative to the distribution of the group of diseases tweets).

By isolating these n-grams with a higher frecuency I expect to identify issues unique to each of the diseases (as symptoms, or social consecuences.)

2 Tools

We will use Python to analyze the tweets, probably with the NLTK libraries.

[1] It's still to decide if we will study the n-grams to a given n as a unique set, or as a different n sets. We will take this desicion after reading more bibliography and some experimenting.