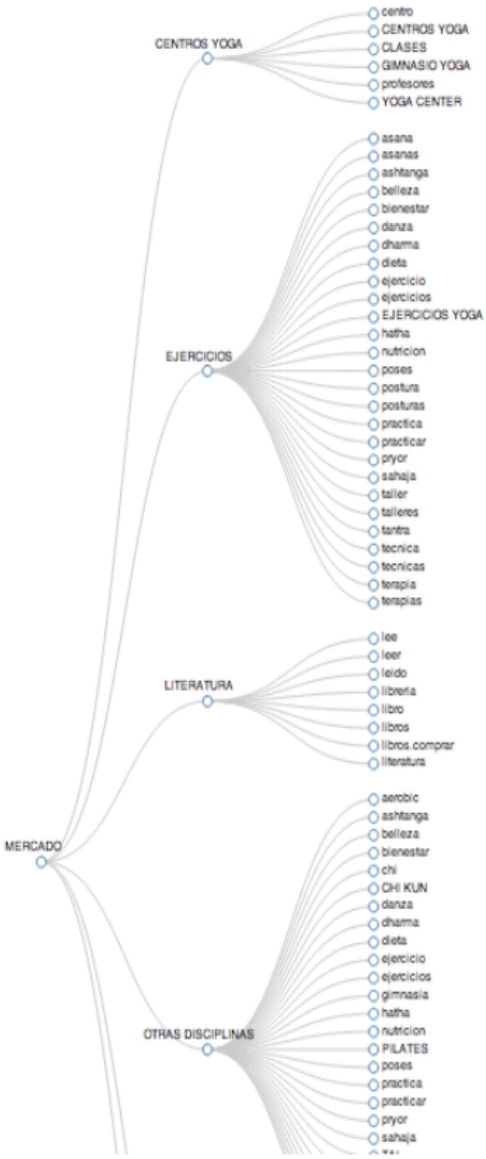
Lexical Distance-based Tagger

francisco.rangel@autoritas.es

We are interested on identifying topics people is talking about in social media. For this task, our consultants define a strategical dimensioning consisting on categories, tags and words (see figure on the left).



Whenever a word is identified in the contents, the corresponding tag is assigned. If different words from different tags are present in the contents, the contents are tagged in the different tags. This fact allows us to cross the information in order to better identify what is people talking about.

But we want to improve the tagging process in order to take into account the lexical distance between words. For example, if we would have the following tags with their corresponding words:

* TAG: Rita Barbera WORDS: Rita Barbera, Rita, Barbera, Alcaldesa Valencia...
* TAG: Fallas WORDS: Falla, Fallas...

Both of the following contents that talk about Rita and Fallas would be tagged in both tags:

“Rita Barberá está en la Falla de Nou Campanar”

“Rita Barberá, el día de la cridá, dio un discurso muy entretenido desde el balcón del Ayuntamiento. Con el tema del caloret, quedaron inauguradas oficialmente las Fallas del 2015”

Although both contents should be tagged with both “Rita Barberá” and “Fallas” tags, we are interested on include a measure of quality that indicates us that the first contents relates Fallas to Rita in a strongest lexical relationship than the second one.

We propose the measure of lexical distance between words, but other machine learning techniques are not discarded. We would provide a set of documents and a complete strategical dimensioning. Our experts on the field would help on the evaluation. The aim is at improving the current tag-independent process.