Description of the tutorial's pipeline

[Salta al contenido principal](https://egela.ehu.eus/mod/page/view.php?id=4325219&inpopup=1#maincontent)

1. R basics. Link to resources of courses
2. Download dataset. Set path. Load raw text-dataset
3. Create a corpus from a set of archives organized in a subdirectory. Inspect corpus
4. Basic text transformations: remove punctuation, stopwords, stemming...
5. Create a "DocumentTerm matrix": remove sparse terms, correlations among terms
6. Convert the "DocumentTerm matrix" to WEKA's \*.arff data format (just a curiosity)
7. Build a wordcloud
8. Retrieve text and create a corpus from html pages
9. Retrieve text and create a corpus from files (stored in disk)
10. Retrieve text and create an annotated corpus from Twitter: using rtweet or twitteR packages. The annotated corpus will be the starting point to learn further supervised classification models.

Última modificación: martes, 21 de diciembre de 2021, 10:03