(Informal) Description of the tutorial's pipeline

[Salta al contenido principal](https://egela.ehu.eus/mod/page/view.php?id=4325225&inpopup=1#maincontent)

* Starting from the "DocumentTerm matrix" built in previous tutorial
* Clustering words-terms with similart patterns of occurrences: hierarchical clustering
* Classification of documents: predict the "type" of a document
  + "Concatenate" the annotation column: type of training documents
  + Classification by R-package "caret", covering a classic data-mining analysis pipeline:
    - create a "Train-Test" partition for classifier validation
    - general machine learning preprocess filters
    - internal performance estimation in the training partition
    - selection of supervised classification algorithms (at least 2)
    - tuning of the parameters of the selected classifiers (only in the training partition)
    - performance estimation: confusion matrix and derived scores
    - predict the class-type of future unseen-unlabeled texts
    - statistical comparison between two classifiers by means of t-test

Última modificación: martes, 19 de enero de 2021, 16:18