Documents needs to represented as vectors.

Ideally they should be normalised in respect to the whole collection, for better analysis and easier clustering.

At the moment there is not possibility to just add a new document to the analysis process while analysing and/or clustering. The process needs to be rerun from step 0 performing the whole pipeline. Adding this functionality would involve caching the process at the current stage and then adding the new document(s) to perform the analysis in regard to this document (very difficult to achieve with analysis involving algorithms such IDF where the vectors are weighted strictly in respect to the others). Perhaps it is achievable with a different method of clustering, especially flat. The weighted document could be clustered simply on respect already saved values coming from the already existing clusters. When it comes to a similarity hierarchy, it may involve the computation of an entirely new similarity matrix.

Only a few types of documents are available at the moment: Microsoft Word format (97-2003 and docx), standard txt and other related. There is some experiment

## Performance

There are still high limitations when it comes to performances. Largest set testes so far comes from every documents currently present in the file system of a user in Windows (top tree: C:/users/USER) for <***insert number of documents***> documents total of varying measures (1 word up to very large sets <***insert largest document’s words count***>.

This was when parallelisation was implemented only on certain instances of the code such the normalisation of the vectors (with relative weighting). Speaking of such, that seems to prove being the slowest task