PROPOSAL ABSTRACT:

Name of Principal Investigator:	Juan Francisco Zamora Osorio
Proposal Title:	Clustering Distributed and High Dimensional Document Collections

Describe the main issues to be addressed: goals, methodology and expected results. **The maximum length for this section is 1 page** (Verdana font size 10, letter size is suggested).

As a consequence of the explosive growth of the WEB and the used of social networks the task of automatic text clustering is becoming of greater importance due to the necessity to categorize increasing volumes of data (e.g. Tweets and News). In several scenarios, these are generated in a descentralized manner, e.g. collected documents by multiple machines in the context of search engines [1], even in different geographical zones. The generation of large amounts of documents surpases nowadays the capacity of personal computers and even of hiperformance computers. One way to deal with the problem of large volumes of data consists in dividing the collection in different machines and then to seize the local computation capabilities to obtain those groups or categories that represent the document collection. By means of the latter, first the the cost of storage is decreased, and later also the computation cost is reduced, because each machine processes less amount of data, and even parallel approaches can be used. Finally, the large size of the document collection makes the transmission over the network of these data to a centralized node for its storage and analysis, impractical. The Distributed Clustering attacks this point by identifying the groups of the complete collection by means of the integration of the local clusters created in each machine into a single model.

Previously, important contributions to the Distributed Data Clustering task have been made. Nonetheless, a major portion of these efforts have been pointed to low dimensional datasets in terms of the sizes of their feature spaces (i.e. order of hundreds). In the text data scenario the vector representations built onto the Tf-Idf model and its variations, has a very high dimensional nature due to the size of the vocabulary, i.e. the number of different words employed in the contained documents. For instance, a small text collection (e.g. order of thousands) can easily have a vocabulary of 10k words, which means that the feature space onto which the document vectors are spanned has that dimension. In addition to the high dimensionality, the sparsity involved in the computational representation of texts (only few words of the overall vocabulary appear in each document) puts a difficult challenge to traditional centralized clustering algorithms due to the main memory space required to process the data and the detection of groups under distance measures that lose their effectiveness in high dimensional and sparse data.

In this project we propose to develop new Clustering techniques capable of dealing with text collections distributed into several computational nodes in a network and also capable of exploiting the computational power of each one. In order to do this the proposed methods should be able to combine in a central master node the partial models built in the other nodes. Furthermore, the algorithms must exploit the existing parallelism in each node either at CPU or GPU level and also must be efficient in terms of the number of times each document is read and copied to main memory. This last trait is also very important since the subcollections processed in each node can also have a size several times larger than the main memory available.

In order to validate the proposed methods several real text collections used for Document Classification, Search Engine Evaluation and Document Clustering will be employed. To measure the performance attained by the proposed models several standard clustering measures will be computed over the obtained clusterings (e.g. *Silhouette*, *Adjusted Rand Index*, *Adjusted Mutual Information score* and the *V measure*). For collections with group-labels available external measures will be computed and when no group labeling is available only internal measures will be computed. Finally, in order to provide a fair comparison among other methods and an objective evaluation random partitions across network nodes will be performed.

[1] See for example how does Google operate to index textual content in`https://www.google.com/insidesearch/howsearchworks/crawling-indexing.html` (01.08.2016)