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).

Como consecuencia del crecimiento explosivo de la WEB y del uso de la redes sociales, la tarea de agrupamiento automático de texto resulta cada vez de mayor importancia debido a la necesidad de categorizar volúmenes crecientes de datos (e.g. Tweets y noticias). En varios escenarios, estos se generan de manera descentralizada, e.g. documentos recolectados por múltiples máquinas en el contexto de un motor de búsqueda[1] incluso desde diversas zonas geográficas. La generación de grandes cantidades de documentos sobrepasa hoy en día las capacidades de computadores personales e incluso de equipos de cómputo de alto desempeño. Un enfoque para atacar el problema del gran volúmen de datos ha consistido en dividir la colección en diversas máquinas y aprovechar las capacidades de cómputo locales para obtener aquellos grupos o categorías temáticas que representan a la colección. De esta manera, primero se reduce el costo de almacenamiento al dividir la colección, y luego también se reduce el costo de cómputo al procesar una cantidad menor de datos en cada máquina, aprovechando incluso su capacidad de cómputo paralelo. Por último, el tamaño excesivo de las colecciones documentales hace impracticable su transmisión por la red a un sitio centralizado para su almacenamiento o análisis. El agrupamiento distribuido ataca este punto al identificar los grupos de la colección complete mediante la integración de los agrupamientos locales de cada máquina en un solo modelo.

Existen importantes contribuciones en la tarea de agrupamiento de datos distribuídos. Sin embargo, gran parte de estos esfuerzos ha considerado conjuntos de datos de baja dimensionalidad en su representación vectorial, alcanzando como máximo una cantidad de dimensiones del orden de las centenas. Para el caso del texto, su representación vectorial es naturalmente de alta dimensionalidad, debido al tamaño de los vocabularios (cantidad de palabras distintas en la colección). Esta característica hace difícil su procesamiento debido a que muchos métodos ampliamente usados (e.g. basados en K-Means o en densidad) escalan respecto de la cantidad de datos, pero su costo es cuadrático respecto de la cantidad de atributos del espacio de características.

En este proyecto se desarrollaran nuevas técnicas de agrupamiento de documentos de texto capaces de operar sobre colecciones documentales repartidas en varios nodos de una red de computadores, aprovechando además las capacidades de computo individuales de cada nodo. Para esto, los métodos a desarrollar deberan ser capaces de integrar en un solo modelo aquellas categorías de documentos identificadas en distintas máquinas. Por otra parte, deberan aprovechar el paralelismo existente en los procesadores de cada máquina y que además sean eficientes en términos de la cantidad de pasadas realizadas sobre cada sub-colección de documentos. Esto último, debido a que cada fragmento de la colección también puede tener un tamaño considerable, e.g. varias veces la cantidad de memoria RAM disponible en cada máquina. Para validar los métodos propuestos se utilizaran varias colecciones documentales usadas en problemas de clasificación de documentos, mejora de recuperación en motores de búsqueda y agrupamiento de texto. Para medir la calidad de los modelos generados se contrastarán aquellos obtenidos de manera distribuída con los obtenidos centralizadamente (en los casos que sea posible debido al tamaño de las colecciones), usando medidas estándar en agrupamiento tales como *Rand score*, *Mutual Information*, *Homogeneity*, *Completeness score* y la media armónica de estos últimos dos que corresponde a la *V-Measure*. Las colecciones documentales disponibles no han sido diseñadas para probar algoritmos distribuídos, pero pueden ser fácilmente distribuídas en la validación como ya lo han hecho trabajos recientes.

[1] Ver por ejemplo como opera Google para indexar contenido textual en `https://www.google.com/insidesearch/howsearchworks/crawling-indexing.html` (01.08.2016)