Tarea 2 de Minería de Textos: *Clustering*

En esta tarea se pretende que los estudiantes utilicen un software de *clustering* y se familiaricen con este tipo de herramientas, sus requerimientos y el tipo de salida que generan. En concreto, se propone utilizar **CLUTO - Family of Data Clustering Software Tools** que se encuentra en http://glaros.dtc.umn.edu/gkhome/views/cluto/.

De las tres familias de herramientas que aparecen en esa página habrá que utilizar: **CLUTO - Software for Clustering High-Dimensional Datasets** http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview. En la pestaña "Download" se puede descargar el software. Se recomienda la descarga de: Latest stable release (2.1.2a). En esta misma página se puede encontrar información de la instalación y en el apartado "Documentation", se puede encontrar el enlace al **manual de uso** del programa.

Los programas y algoritmos que incluye CLUTO tienen dos formas de ser utilizados: como un programa en sí mismo, al que se le llama desde la línea de comando junto con los correspondientes parámetros, o como parte de un programa C. En nuestro caso, se utilizará de la primera manera y en el manual de uso habrá que consultar: **Using CLUTO via its Stand-Alone Program** (Apartado 3 del manual de uso)**.**

Se van a utilizar como datos de prueba los que proporciona CLUTO que se encuentran accesibles en esa misma página como "Datasets": datasets.tar.gz. La descripción de lo que contienen los conjuntos de datos se puede encontrar en el artículo "**Criterion Functions for Document Clustering**" http://glaros.dtc.umn.edu/gkhome/node/165, en el apartado 4.1 Document Collections. En ese artículo también se definen algunos conceptos básicos de clustering, y se describen algunos de los elementos que intervienen en el clustering, como por ejemplo las funciones criterio que luego se pueden seleccionar para realizar el clustering.

La realización de la práctica consiste en lo siguiente:

- Descarga del software necesario, del manual de uso, conjunto de datos con los que se va a experimentar, su descripción y toda la información adicional que se considere necesaria.
- Seleccionar la colección "re0", que consiste en noticias de la agencia de noticias Reuters. Es un subconjunto de la colección Reuters-21578, tiene 1.504 documentos, 2886 términos o rasgos y 13 clases o clústeres. (ver tabla 1 en el artículo mencionado anteriormente http://glaros.dtc.umn.edu/gkhome/node/165). Se puede encontrar información adicional de esta colección, por ejemplo en http://www.daviddlewis.com/resources/testcollections/reuters21578/.
- Utilizar el programa *vcluster* que usa como representación de los objetos el modelo del espacio vectorial. La entrada de este programa es una matriz que corresponde a un fichero con extensión .mat de los que aparecen en datasets.tar.gz, en concreto en este caso "re0.mat". Mediante parámetros se le puede indicar que haga el clustering con diferentes algoritmos, funciones criterio, etc. (Ver apartado 3.1.1 del manual de uso).
- Realizar el clustering y hacer el análisis de los resultados del clustering de la citada colección en 13 clústeres con:
    - Un algoritmo de partición y el algoritmo aglomerativo.

    - Dos funciones de similitud.
    - Dos funciones criterio.
    - Habrá que combinar estos 6 elementos entre ellos en la medida de lo posible.
- Preparar una memoria en la que:
    - Se describa la colección seleccionada.
    - Se indique el método de representación de la colección en cuestión (aparece descrito en el apartado 2 de http://glaros.dtc.umn.edu/gkhome/node/165).
    - Se indiquen los parámetros de *vcluster* que se han utilizado indicando el tipo de algoritmo, funciones de similitud y criterio, así como cualquier otra variación de los parámetros por defecto que se haya utilizado.
    - Se añada la información de salida de las ejecuciones de *vcluster*. (En la sección 3.2 del manual de uso se explica la información que se obtiene). La calidad del clustering se mide con medidas internas y externas. Para obtener información de la calidad estadística del clustering comparando la solución del algoritmo con una solución manual (medida externa) se debe utilizar el parámetro *–rclassfile* en la ejecución de *vcluster* junto con el fichero correspondiente a la colección con extensión .rclass de datasets.tar.gz.
    - Se analicen los resultados en términos de calidad de las medidas internas y externas con los diferentes algoritmos, funciones de similitud y criterio, así como el tiempo de ejecución.

Parte opcional

En la parte obligatoria de la tarea habéis partido de una representación vectorial de una colección que ya estaba preparada. En esta parte opcional tendréis que generar una representación a partir de una colección diferente, en concreto la colección "CorpusClustering-Tarea2-parteAdicional.rar" que encontraréis en la carpeta de documentos públicos.

Se trata de que obtengáis una representación para esa colección en un formato que sirva de entrada al software de CLUTO y que realicéis el clustering. Los formatos válidos están descritos en el apartado 3.3 del manual de uso de la herramienta.

La realización de la parte opcional de la práctica consiste en lo siguiente:

- Eliminar las etiquetas XML que contienen los documentos.
- Seleccionar una función de pesado de entre las que se han visto en el curso, debe ser diferente de la utilizada en las colecciones precalculadas de CLUTO.
- Utilizando alguna de las librerías que se han presentado en el curso u otras, obtener una representación de la colección que sirva de entrada al software de CLUTO.
- Realizar el clustering de la colección utilizando CLUTO seleccionando los parámetros que se deseen.
- Hacer una valoración de la calidad del clustering teniendo en cuenta el contenido de los documentos.
- La memoria correspondiente a esta parte deberá contener:
    - Una descripción de la función de pesado seleccionada.

- Indicar la librería utilizada o en su caso si se ha desarrollado expresamente.
- Descripción de los parámetros utilizados para realizar el clustering.
- La salida de la ejecución.
- Valoración de la calidad considerando el contenido de los documentos. La colección es pequeña por lo que se espera una revisión manual.

La **fecha tope de entrega** de esta tarea para la **convocatoria de febrero** será el **2 de febrero de 2021**.