Tarea 1

ArXiv.org es una colección de archivos que contienen preprints de publicaciones científicas. En esta tarea procesaremos un conjunto de abstracts de papers enviados durante el año 2011 a ArXiv.org en matemáticas.

Descargue de moodle la colección arxiv.tar.gz. Encontrará una carpeta con archivos, cada uno de los cuales contiene el abstract. Recupere el título y el abstract, existen otros campos que no son considerados en esta tarea, como por ejemplo autores.

Debe procesar cada archivo para obtener una nueva versión del abstract y el título, usando NLTK. Además debe construir el vocabulario de la colección completa.
La tarea, paso a paso, consiste en lo siguiente:

1. Para la colección: Detecte collocations y regístrelas en el vocabulario.
2. Para cada documento: Separe abstract y título. Detecte sentencias.
3. Para cada sentencia: Pase todo a minúsculas, elimine signos de puntuación, tokenize y elimine stopwords. Use lematización Wordnet. Cada nuevo unigrama debe agregarlo al vocabulario.
4. Para cada palabra del vocabulario: Registre las ocurrencias de cada palabra en la colección. Para cada palabra, cree dos listas invertidas, una para ocurrencias en el título y otra para ocurrencias en el abstract. La lista debe contener el ID del archivo.

Importante: Puede hacer la tarea en grupos de hasta 3 personas. Empiece a hacerla de inmediato, puede encontrarse con dificultades en algunos pasos.

Entregables:

1. Código generado: subir scripts NLTK python.
2. Datos: subir los archivos con las listas invertidas. Las palabras deben estar ordenadas alfabéticamente. Comprima los archivos a un .tar.gz
3. README: integrantes del grupo e instrucciones básicas de su código.

Evaluación: Las tareas serán evaluadas de acuerdo a eficacia.

Entrega: Lunes 4 de Mayo, en moodle. A las 24:00 Hrs. el link de *upload* se cerrará.