Practica 10

En este práctica, ilustramos un método para ayudar y mejorar los esfuerzos de los investigadores al permitir un análisis semiautomático de grandes volúmenes de datos no estructurados (en forma de artículos de revistas publicadas) a través de la aplicación de minería de textos

Utilizando bibliotecas digitales estándar y motores de búsqueda de publicaciones en línea (por ejemplo Google Scholar[1], descargue y recopile artículos disponibles. Extraiga esta información al menos 30 publicaciones de los autores listados abajo y de al menos tres revistas diferentes. Para cada artículo, extraiga el título, resumen, lista de autores, palabras clave publicadas, volumen, número de edición y año de publicación. Si es posible incluya además el tipo de artículo para futuros análisis de patrones.

Se recomienda usar información de estos autores:

- Mauricio Espinoza Mejía
- Victor Saquicela Galarza
- Guadalupe Aguado de Cea
- Jim Hendler
- Frank van Harmelen
- Natasha Noy
- Priscila Cedillo Orellana
- Miguel Angel Zuniga
- Dario Rodriguez
- Yaskelly Yedra
- Jeanette Riley
- Kenneth Palacios Baus
- Otto Parra González
- Don Batory
- Alexander Egyed
- Marten van Sinderen
- Lizandro Damian Solano-Quinde
- Alex Nicolau
- Bill Punch
- Zhenjiang Hu
- Juan Pablo Carvallo
- Luis I. Minchala
- Liping Guo
- Angel de Castro
- Javier Garrido
- Amit Sheth

[1] https://scholar.google.com/schhp?hl=es

Para esta práctica se usará únicamente el resumen de un artículo como única fuente de información. No se debe incluir las palabras clave del artículo por dos principales razones: en circunstancias normales, el resumen incluye las palabras clave enumeradas, y, por lo tanto, la inclusión de las palabras clave enumeradas para el análisis significaría repetir la misma información y potencialmente otorgarles un peso inmerecido; y las palabras clave enumeradas pueden ser términos con los que los autores desearían que se asociara su artículo (a diferencia de lo que realmente figura en el artículo), lo que podría introducir un sesgo no cuantificable en el análisis del contenido.

Cada uno de los artículos deben ser grabados en archivos de Excel y entonces se propone ejecutar procesos que permitan

1. Representar la relación entre las palabras/términos y los documentos
2. Ejecutar pre-procesamiento que incluya al menos Transform Cases, Tokenization (usando la opción all nonletter characters), Filter Stopwords, y Filter Tokens (by Length) (por ejemplo menos de dos caracteres de longitud). Use algunos procesos adicionales en caso de considerarlo necesario.
3. Determinar el número de grupos para clasificar los artículos. En todo momento se debe incluir el autor, año y revista

Tareas

1. Verifique los documentos que pertenecen a cada clúster y explique el valor sobre el número de cluster
2. Usando algún software de visualización o análisis (e.j. tabla dinámica de Excel) ejecute los siguientes reportes
   - Listado de autores que trabajan en temas similares (pertenecen al mismo cluster)
   - Cluster x Año x Autor
   - NombreArticulos x Cluster x Autor
   - Un histograma que muestre en el tiempo los tópicos sobre los que ha investigado un autor
3. Explique el significado de los reportes