**INSTITUTO TECNOLÓGICO DE COSTA RICA I SEMESTRE 2020**

**INGENIERÍA EN COMPUTACIÓN**

**IC-8020 RECUPERACIÓN TEXTUAL**

**TAREA PROGRAMADA 1**

La tarea consiste en programar una herramienta que crea un modelo de lenguaje unigrama (LM unigram) a partir de la colección de páginas web en español con información sobre países y dependencias del mundo.

Para cada archivo de la colección se debe extraer su texto eliminando las etiquetas HTML. El texto debe ser dividido en palabras, las cuáles consisten en secuencias de uno o más caracteres tomados del siguiente conjunto: [a-zñ]. Antes de separar el texto en palabras, se deben convertir todas las letras a minúsculas y se deben eliminar todos los acentos, con excepción del de la eñe.

Para cada término de la colección se debe contar el número de veces que aparece en todos los documentos; esto es, para cada término se debe contar el número de veces que aparece en cada documento y luego sumar dichas cuentas. Luego se procede a calcular las probabilidades del LM tal como fue visto en clase:  
 p(término w) = conteo(término w) / M,  
dónde M es la suma del conteo de todos los términos de la colección.

Adicionalmente se debe programar una herramienta que extraiga todas las referencias usadas en los atributos href de HTML. Se debe producir una salida que para cada archivo de la colección escogida liste todas las referencias extraídas, eliminando casos repetidos.

**Funcionalidad**

A continuación se describe con más detalle la funcionalidad que debe ser implementada. Se puede cambiar la interfaz, por ejemplo integrar todo en una sola herramienta, siempre y cuando no se restrinja la funcionalidad descrita a continuación.

**genere\_LM *Ruta* *ArchivoLM***  
Crear un LM a partir de la colección especificada en ***Ruta***. Si ***Ruta*** es un directorio, entonces procesar todos los archivos con sufijo "htm" o "html" en ese directorio y subdirectorios. Si ***Ruta*** es un archivo, entonces procesar ese archivo si es tipo "htm" o "html". El LM resultante es escrito en el archivo especificado por la ruta ***ArchivoLM***. Dicho archivo debe estar ordenado alfabéticamente.

**opr *LM\_específico* *LM\_general* *Escalafón***  
Este comando usa la técnica vista en clase para obtener las palabras más semánticamente relacionadas (***opr***: **o**btener **p**alabras **r**elacionadas) con algún concepto. El ***LM\_específico*** da la ruta de un archivo que contiene el LM de un caso específico (por ejemplo, la página de Costa\_Rica). El ***LM\_general*** da la ruta de un archivo que contiene el LM de una colección mucho más grande (en este caso toda la colección geográfica).  
Para cada término en ***LM\_específico*** debe calcularse su probabilidad condicional:  
 P(termino en ***LM\_específico***) / P(término en ***LM\_general***).  
El resultado de dicho cálculo debe guardarse en la ruta especificada por ***Escalafón*** ordenado descendentemente por el valor obtenido.

**extraer\_refs *RutaArchivo* *ArchivoRefs***  
Extraer todas las referencias mencionadas en el archivo de la colección especificado en ***Ruta***. Este debe ser un archivo con tipo "htm" o "html". La lista resultante debe ser escrita en el archivo especificado por la ruta ***ArchivoRefs***.

**Ejemplos**

* La siguiente secuencia obtiene LMs para la colección total y para Costa\_Rica; luego calcula las palabras más relacionadas semánticamente con Costa Rica:

**genere\_LM "Geografía" LM\_geo**

**genere\_LM "Geografía\América\Estados\_soberanos\Costa\_Rica.htm" LM\_CR**

**opr LM\_CR LM\_geo opr\_CR**

Para este ejemplo en particular, se espera que las siguientes palabras tengan un valor destacado: metates, jade, esferas, café.

* El siguiente comando extrae todas las referencias para cada uno de los países soberanos de América:

**extraer\_refs "Geografia\\Asia\\Dependencias\\Sinaí.htm" refs\_sinaí**

La salida tendría una forma como:

**/wiki/Suez\_(gobernaci%C3%B3n)**

**/wiki/Mar\_Mediterr%C3%A1neo**

**/wiki/Puerto\_Sa%C3%ADd**

**/wiki/Monte\_Sina%C3%AD**

**...**

**Observaciones finales**

Para esta tarea los archivos pedidos pueden ser archivos con números expresados en forma textual.

Para la extracción del texto a partir del documento HTML aproveche la biblioteca ***Beatiful Soup*** de Python, la cual permite manipular fácilmente documentos HTML y XML. Hay bibliotecas similares disponibles en otros lenguajes de programación.

Esta tarea vale 6 de los 40 puntos de tareas programadas. La tarea es individual y la fecha de entrega es jueves 12 de octubre a las 10:00am.