INSTITUTO TECNOLÓGICO DE COSTA RICA

I SEMESTRE 2020

INGENIERÍA EN COMPUTACIÓN
IC-8020 ELECTIVA TEXTO
TAREA PROGRAMADA I

ESTUDIANTE
HEYLER JOHEL MORA CALDERÓN

PROFESOR
JOSE ENRIQUE ARAYA MONGE

Tabla de contenido

INTRODUCCIÓN

Esta 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 de todo el 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, 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

Posteriormente 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.

SOFTWARE

GENERE_LM

Para generar un LM es necesario comprender que existen dos tipos de LM; el general y el especifico, el primero se genera a partir de un directorio y el segundo a partir de un archivo HTML a continuación se demostrará con un ejemplo como generar los LM, que también generan archivos txt:

```
                        '''*** LM Especifico ***'''
>>> Cuba_LM = genere_LM('Geografia\América\Estados_soberanos\Cuba.htm')
>>> Isla_LM = genere_LM('Geografia\Asia\Dependencias\Isla_de_Navidad.htm)
>>> Fiyi_LM = genere_LM('Geografia\Oceanía\Estados_soberanos\Fiyi.htm')

                         '''*** LM General ***'''
>>> Europe_Dep = genere_LM('Geografia\Europa\Dependencias')
>>> African_LM = genere_LM('Geografia\Africa')
>>> General_LM = genere_LM('Geografia')
```

OPR

Este comando usa la técnica vista en clase para obtener las palabras más semánticamente relacionadas con algún concepto.

EXTRAER_REFS

Extrae todas las referencias mencionadas en el archivo de la colección o directorio especificado en *Ruta*, también genera archivos txt en la carpeta ArchivoRefs. Algunos ejemplos claros para usar esta función en Python son:

```
>>> extraer_refs ('Geografia\América\Estados_soberanos\Costa_Rica.htm')
>>> extraer_refs ('Geografia\América\Estados_soberanos\Cuba.htm')
>>> extraer_refs ('Geografia')
```

LÍNEAS DE ARCHIVOS

Debido a la gran cantidad de elementos no pude ejecutar todos los dir, sin embargo, dejo muestra del dir de los países asiáticos con reconocimiento limitado y de **República de Nagorno Karabaj.**

ASIA/CON RECONOCIMIENTO LIMITADO

```
1. abandonar  :  1 : 0.0002048760499897562
2. abasies  :  1 : 0.0002048760499897562
3. abastecimiento  :  1 : 0.0002048760499897562
4. abbas  :  2 : 0.0004097520999795124
5. about  :  1 : 0.0002048760499897562
6. abril  :  2 : 0.0004097520999795124
7. abrumadoramente  :  1 : 0.0002048760499897562
8. abu  :  2 : 0.0004097520999795124
9. abundante  :  1 : 0.0002048760499897562
10. aca  :  2 : 0.0004097520999795124
11. academia  :  2 : 0.0004097520999795124
12. academica  :  1 : 0.0002048760499897562
13. acceso  :  3 : 0.0006146281499692685
14. accidentado  :  1 : 0.0002048760499897562
15. acciones  :  1 : 0.0002048760499897562
16. aceptando  :  1 : 0.0002048760499897562
17. aceptar  :  1 : 0.0002048760499897562
18. aceptaron  :  2 : 0.0004097520999795124
19. acerca  :  4 : 0.0008195041999590248
20. acordado  :  1 : 0.0002048760499897562
```

ASIA/CON RECONOCIMIENTO LIMITADO/REPÚBLICA DE NAGORNO KARABAJ

```
1. abandonar  :  1 : 0.0005561735261401557
2. about  :  1 : 0.0005561735261401557
3. abrumadoramente  :  1 : 0.0005561735261401557
4. aca  :  1 : 0.0005561735261401557
5. academia  :  2 : 0.0011123470522803114
6. acceso  :  3 : 0.0016685205784204673
7. acciones  :  1 : 0.0005561735261401557
8. aceptando  :  1 : 0.0005561735261401557
9. acerca  :  3 : 0.0016685205784204673
```

```
10.  actual  :  1 : 0.0005561735261401557
11.  actualizacion  :  1 : 0.0005561735261401557
12.  actualizada  :  1 : 0.0005561735261401557
13.  actualmente  :  4 : 0.002224694104560623
14.  acuerdo  :  3 : 0.0016685205784204673
15.  adicionales  :  1 : 0.0005561735261401557
16.  adjetivo  :  3 : 0.0016685205784204673
17.  adjetivos  :  1 : 0.0005561735261401557
18.  admitir  :  1 : 0.0005561735261401557
19.  affairs  :  1 : 0.0005561735261401557
20.  after  :  1 : 0.0005561735261401557
```

COMENTARIOS

Uno de los grandes problemas encontrados en esta tarea es la gran cantidad de elementos a analizar por lo que el tiempo de ejecución puede durar largas cantidades de tiempo, haciendo imposible la ejecución del LM_General de "Geografia", además de que muchas de las palabras que se mantenían en el LM no tenían un verdadero valor, es decir eran letras sin un significado claro, las cuales fueron difíciles de filtrar a la hora de analizar el texto.