Trabajo 3, Notebook para procesamiento de cadenas de texto.

Pensamiento Algorítmico

Profesor: Jorge Victorino

Problema.

Se requiere la implementación de funciones específicas de procesamiento de cadenas de texto con el fin de encontrar las palabras más relevantes y la relación entre ellas de un determinado texto. El trabajo tiene 3 partes que son:

1. Preprocesamiento. El resultado de este punto debe ser una función que recibe como parámetro dos archivos de texto y devuelve una cadena de texto. Los archivos de entrada corresponden a primero, un archivo de texto cualquiera en español y segundo un archivo de palabras stop words en español. La función cumple con las siguientes tareas: elimina todos los caracteres que no son alfabéticos y numéricos (o sea los símbolos), deja todas las palabras en minúsculas, y elimina del texto todas las palabras que son stop words. Una vez se cumplan estas tareas la función retorna o devuelve el texto resultante.

2. Ranking de palabras. El resultado de este punto debe ser una función que recibe como parámetro el texto limpio de símbolos, mayúsculas y stopwords que retorna la función del punto 1, y devuelve dos listas, la primera es una lista con las 15 palabras que más aparecen en el texto ordenadas descendentemente por ocurrencia, y la segunda lista de valores que indican la cantidad de veces que aparece en el texto las palabras de la primera lista.

3. Matriz de distancia entre palabras. El resultado de este punto es una función que recibe un archivo de texto y una lista de palabras, y devuelve una matriz de distancia mínima entre palabras. La distancia entre palabras es un valor entero que indica en un determinado texto cuántas palabras hay entre la primera y la segunda palabra. Por tanto, si las palabras están seguidas en el texto, la distancia es cero, si hay una palabra entre ellas la distancia es 1 y así sucesivamente. La distancia mínima entre ellas se hace encontrando, en todo el texto, lo más cerca que estén estas dos palabras teniendo en cuenta todas las ocurrencias de ambas palabras en el texto completo. Tenga en cuenta que las palabras se buscan en el texto original sin tener en cuenta la diferencia entre mayúsculas y minúsculas. Para la matriz la distancia de una palabra consigo misma es cero.

Pautas del trabajo

1. Organizar cada punto en secciones diferentes del notebook. Indicar cual es la función que corresponde a la solución de cada punto. La función debe tener los parámetros requeridos y el valor de retorno especificado. Esto es importante porque de parte del profesor se probaran las funciones con diferentes archivos de texto y la evaluación del trabajo también depende de los resultados obtenidos.

2. La entrega consiste en un archivo tipo notebook de python con el primer apellido en orden alfabético de los integrantes del grupo separado por guión y con la extensión ipynb, el cual se debe descargar y adjuntar al correo (no se acepta compartir el archivo desde drive). Por ejemplo: "***arias-duque-puerta.ipynb***". La entrega se hace al correo jvictorinog@ucentral.edu.co con el asunto: **"Trabajo 3 pensamiento algorítmico**", la fecha límite es: Lunes 24 de Mayo a las 11:59 a.m. Los trabajos entregados después de esta fecha y hora no serán calificados. El trabajo es susceptible de sustentación individual por cualquiera de los integrantes del grupo. No tener en cuenta estas indicaciones se puede bajar hasta 5 décimas en la nota final.
3. El notebook lleva un encabezado o portada con los integrantes, el curso, el grupo, profesor, logo de la central. Debe estar dividido en secciones para cada punto y se debe indicar cual es la función principal de cada punto. Los integrantes de cada grupo son máximo de 3 personas y los pueden escoger como gusten.
4. Para este trabajo no se permite el uso de ninguna librería de python. En caso de usar alguna librería el trabajo se califica sobre 3.0. El plagio de alguna función con otro trabajo o algo en internet dará lugar a una calificación de cero en todo el trabajo para ambas partes

Calificación:

Se calificará cada punto por separado sobre 5.
Cada punto se someterá a pruebas por parte del profesor y dependiendo del funcionamiento y las pautas se calificará cada punto.

1. Si la función solicitada no tiene el propósito especificado, la calificación es menor a 1.0
2. Si al ejecutar la función se producen errores de compilación la calificación es menor a 2.0
3. Si la función funciona incorrectamente, o no tiene los parámetros de entrada, o de salida que se especifican la calificación es menor a 3.
4. Si la función presenta errores de precisión en los resultados la calificación es menor a 4.
5. Para obtener 5 la función debe superar todas las pruebas del profesor con éxito y cumplir con todos los requisitos, como documentación, ortografía y formato.

Funcionamiento

A continuación se muestran algunas pruebas de ejemplo, para que ustedes puedan verificar el correcto funcionamiento de sus funciones utilizando los archivos fragmento.txt y stopwordsSPA.txt

Punto 1.

Ejemplo 1: Utilizando la función con los archivos dados:
`nombre_funcion('fragmento.txt', 'stopWordsSPA.txt')`
Da como resultado un texto en minúscula, sin símbolos o signos de puntuación que contiene 4020 palabras.
Ejemplo 2: Si el archivo de prueba tiene este contenido:

RESUMEN: La novela está estructurada en capítulos sin nombrar. Sin embargo, para facilitar la comprensión del argumento, hemos ordenado y separado el relato en cuatro etapas que identifican, a grandes rasgos, los pasajes más emblemáticos. Etapa "fundación y primeros años de Macondo". Desde que Úrsula Iguarán se casó con su primo José Arcadio Buendía, teme engendrar un niño con cola de cerdo como consecuencia del parentesco. Por ello, se niega temporalmente a consumar el matrimonio. Esto es causa de que Prudencio Aguilar se burle de José Arcadio Buendía quien, ofendido, lo mata en duelo para salvar su honor. Desde entonces, el fantasma de Aguilar lo persigue y José Arcadio decide irse del pueblo.

El resultado de **nombre_funcion('prueba.txt', 'stopWordsSPA.txt')** es el siguiente texto:

resumen novela estructurada capítulos nombrar facilitar comprensión argumento ordenado separado relato etapas identifican rasgos pasajes emblemáticos etapa fundación años macondo úrsula iguarán casó primo josé arcadio buendía teme engendrar niño cola cerdo consecuencia parentesco niega temporalmente consumar matrimonio causa prudencio aguilar burle josé arcadio buendía ofendido mata duelo salvar honor fantasma aguilar persigue josé arcadio decide irse pueblo

Haciendo la cuenta de las palabras debería dar lo siguiente:

```
Cantidad de palabras en prueba.txt:  113
Cantidad de palabras tipo stopwords  55
Cantidad de palabras no stopwords    58
```

Punto 2:

El resultado de llamar a la función del punto 2 con el texto de fragmento.txt limpio de símbolos, mayúsculas y stopwords, son dos listas como se muestra a continuación:

Palabra	**Frecuencia**
arcadio	62
buendía	60
josé	54
úrsula	42
casa	23
melquíades	22
años	21
mujer	19
niños	17
noche	17
macondo	16
aldea	16
aquella	16
mundo	14

```
hombres       13
```

Punto 3.

Utilizando la lista de palabras obtenidas en el punto anterior y usando el texto original de fragmento.txt en minúscula. Se obtiene la siguiente matriz de distancia mínima entre palabras:

PALABRA	arcadio	buendía	josé	úrsula	casa	melquía	años	mujer	noche	niños	macondo	aquella	aldea	mundo	hombres
arcadio	0	0	0	3	8	6	12	7	3	5	7	19	5	26	4
buendía	0	0	1	2	17	5	4	6	2	6	8	3	6	25	3
josé	0	1	0	4	7	7	11	6	4	4	6	7	4	27	5
úrsula	3	2	4	0	3	24	8	13	0	2	8	12	20	36	17
casa	8	17	7	3	0	87	44	1	58	9	13	31	16	7	209
melquía	6	5	7	24	87	0	11	238	36	27	61	10	13	11	41
años	12	4	11	8	44	11	0	48	37	22	26	13	7	64	44
mujer	7	6	6	13	1	238	48	0	13	122	144	0	147	87	309
noche	3	2	4	0	58	36	37	13	0	111	76	23	190	37	14
niños	5	6	4	2	9	27	22	122	111	0	29	27	67	22	11
macondo	7	8	6	8	13	61	26	144	76	29	0	12	2	37	94
aquella	19	3	7	12	31	10	13	0	23	27	12	0	0	7	80
aldea	5	6	4	20	16	13	7	147	190	67	2	0	0	8	27
mundo	26	25	27	36	7	11	64	87	37	22	37	7	8	0	6
hombres	4	3	5	17	209	41	44	309	14	11	94	80	27	6	0