29/10/2017

Tarea Corta 2

Aplicación de Stemming

Andrey Mendoza Fernández - 2015082908
INSTITUTO TECNOLÓGICO DE COSTA RICA

¿Cómo ejecutar el programa?

1. Abrir una consola de comandos
2. Ubicarse a la ruta donde se encuentra ubicado el directorio de código
3. Utilizar el siguiente comando:
    - python main.py n_documentos nombre_resultados ruta_coleccion

    Donde cada parámetro significa:

    - n_documentos: cantidad de documentos a procesar de la colección
    - nombre_resultados: nombre a otorgar al archivo final con los resultados
    - ruta_colección: ruta relativa donde se encuentra la colección

Stemmer utilizado

Natural Language Tool Kit (NTLK) es una herramienta que provee muchas herramientas para trabajar con lenguaje natural. Entre ellas se ofrecen dos métodos para realizar *stemming:* usando Porter y Snowball. Para esta tarea corta se solicitó explícitamente el uso de Snowball, el cual es muy simple de utilizar con esta biblioteca.

Casos de prueba y resultados obtenidos

# directorios procesados	**#archivos tomados**	**# bytes procesados**	**# términos encontrados**	**# palabras distintas**	**tiempo**
1	1	734.346	4898	11526	0.56 seg
1	3	1.610.733	8259	21236	0.9 seg
1	6	4.516.877	20048	45763	1.69 seg
1	16	11.305.210	42619	83171	3.66 seg
1	40	27.786.527	72711	131863	7 seg
1	101	70.269.147	129269	219571	15.43 seg
3	303	216.713.045	291588	459534	48.79 seg
6	588	414.603.183	392705	619528	1.73 min
16	1534	1.212.641.742	556167	909811	5.05 min
40	4002	3.632.227.195	950018	1470854	16.49 min
100	9436	8.810.101.297	2058994	2977696	50.42 min

Comentarios

Los requerimientos solicitados fueron completados al 100%. La duración para realizar toda la ejecución del programa sobre la colección entera es bastante buena. Para lograr esto, fue necesaria la utilización de la clase *Counter* que ofrece la biblioteca *collections*. Esta clase se encarga de tomar una lista con cualquier tipo de objeto y retornar un diccionario con la llave de cada elemento distinto dentro de la lista y la cantidad de apariciones como el término. Esta clase sobrecarga el operador de suma, por lo que es muy simple unir dos diccionarios.