Prácticas de SAR

Sistemas de Almacenamiento y Recuperación de información

Práctica 2: Cuenta Palabras

Descripción del problema

Para hacer estudios sobre la autoría de unos documentos, se desea obtener estadísticas del estilo literario del autor (centrándonos en el uso del vocabulario).

Ejercicio Mínimo

¿Qué debo hacer?

Escribe un programa en python que analice ficheros de texto, calcule estadísticas sobre cada fichero y las escriba en disco.

- Recibirá uno o más nombre de fichero obligatoriamente.
- Por cada fichero de texto generará otro con las estadísticas.
- Aceptará los siguiente argumentos opcionales:
  - -h, - -help: mostrará el mensaje de ayuda.
  - -l, - -lower: pasa todo el contenido a minúsculas.
  - -s, - -stop fichero: fichero con las **stopwords** que se deben eliminar.
  - -f, - -full: se deben mostrar las estadísticas completas. En caso contrarío solo se muestras 20 entradas de cada categoría.
- Para realizar el análisis se eliminaran todos los símbolos no alfanuméricos.

Ejercicio Mínimo

Ejercicio Mínimo, ¿Qué debo hacer?

```
python SAR_p2_cuenta_palabras.py --help
usage: SAR_p2_cuenta_palabras.py [-h] [-s STOPWORDS] [-l] [-b] [-f]
                                 file [file ...]

Compute some statistics from text files.

positional arguments:
  file                  text file.

optional arguments:
  -h, --help            show this help message and exit
  -s STOPWORDS, --stop STOPWORDS
                        filename with the stopwords.
  -l, --lower           lowercase all words before computing stats.
  -b, --bigram          compute bigram stats.
  -f, --full            show full stats.
```

Ejercicio Mínimo, ¿Qué debo hacer?

Escribe un programa en python que analice ficheros de texto, calcule estadísticas sobre cada fichero y las escriba en disco.

El programa en python mostrará la siguiente información:

- Número de líneas.
- Número de palabras.
- Número de palabras sin stopwords (en el caso de elegir eliminarlas).
- Vocabulario: número de palabras distintas que aparecen en el texto.
- Símbolos: número de letras que aparecen en el texto.
- Símbolos distintos: número de letras distintas que aparecen en el texto.
- Número de veces que aparece cada palabra: ordenado alfabéticamente y por el número de veces que aparecen.
- Número de veces que aparece cada letra: ordenado alfabéticamente y por el número de veces que aparecen.

Ejercicio Mínimo, ¿Qué debo hacer?

Ejemplo de funcionamiento

```
python SAR_p2_cuenta_palabras.py spam.txt --lower
```

Salida

```
Lines: 11
Number words (including stopwords): 77
Vocabulary size: 22
Number of symbols: 324
Number of different symbols: 23
Words (alphabetical order):
    a: 2
    and: 12
    aux: 1
...
    spam: 27
    thermidor: 1
    top: 1
Words (by frequency):
    spam: 27
    and: 12
    egg: 9
...
    thermidor: 1
```

Salida (continuación)

```
Symbols (alphabetical order):
    a: 63
    b: 10
    c: 8
    d: 17
...
    s: 40
    t: 9
    u: 7
    v: 1
Symbols (by frequency):
    a: 63
    s: 40
    m: 29
    p: 29
...
    f: 3
    l: 2
    w: 2
    y: 2
```

Ampliación

Ampliación, ¿Qué debo hacer?

Se proponen como ampliación:

- Realizar un análisis de los pares de palabras consecutivas (bigramas) que aparecen en las frases.
  - se mostrarán los resultados ordenados por orden alfabético y por frecuencia.
  - se considerará cada linea del fichero como una frase.
  - se deberá añadir un símbolo (‘$’) como primera y última palabra de cada frase.
- Realizar un análisis de los pares de letras que aparecen en cada palabra.
  - se mostrarán los resultados ordenados por orden alfabético y por frecuencia.

El análisis adicional se activará mediante el argumento -b, - -bigram.

Nombre de los ficheros de estadísticas

Nombre de los ficheros de estadísticas

Por cada fichero que analice, el programa debe generar un fichero con las estadísticas. El nombre del fichero de estadísticas debe cumplir las siguientes normas:

- Debe tener la misma extensión que el fichero original.
- Al nombre del fichero original se le debe añadir:
  - un código con las opciones elegidas, **lsbf**.
  - '_stats' al final del nombre original.

Ejemplo de nombres de fichero:

- python SAR_p2_cuenta_palabras.py **spam.txt -l -b >> spam_lb_stats.txt**
- python SAR_p2_cuenta_palabras.py **tirant -f -s** english.txt **-l >> tirant_lsf_stats**
- python SAR_p2_cuenta_palabras.py **spam.txt >> spam_stats.txt**