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

¿Qué debo hacer?

Escribe un programa en python que analice un fichero de texto y muestre estadísticas sobre él.

- Para realizar el análisis se eliminaran todos los símbolos no alfanuméricos.
- El programa recibirá 1, 2 ó 3 argumentos:
  1. el primer argumento será el nombre del fichero de texto que se analizará.
  2. el segundo argumento (opcional) será si se pasa todo el contenido a minúsculas.
     - valores posibles: (‘1’, ‘0’, ‘yes’, ‘no’, ‘True’, ‘False’).
     - opción por defecto: ‘no’.
  3. el tercer argumento (opcional si hay un segundo) será si se eliminan las **stopwords** (en inglés) del análisis. Se proporciona un fichero con las stopwords que se deben considerar.
     - valores posibles: (‘1’, ‘0’, ‘yes’, ‘no’, ‘True’, ‘False’).
     - opción por defecto: ‘no’.

¿Qué debo hacer?

Escribe un programa en python que analice un fichero de texto y muestre estadísticas sobre él.

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.

¿Qué debo hacer?

Ejemplo de funcionamiento

```
python SAR_p2_cuenta_palabras.py spam.txt yes no
```

Salida

```
Lines: 11
Number words (with stopwords): 77
Vocabulary size: 22
Number of symbols: 324
Number of different symbols: 23
Words (alphabetical order):
    a   2
    and 12
    aux 1
    bacon   6
...
Words (by frequency):
    spam    27
    and 12
    egg 9
    bacon   6
```

Salida (continuación)

```
Symbols (alphabetical order):
    a   63
    b   10
    c   8
    d   17
    e   25
    f   3
    g   23
    h   4
...
Symbols (by frequency):
    a   63
    s   40
    m   29
    p   29
    e   25
    g   23
    n   23
...
    k   1
    v   1
    x   1
```

Ampliación

Ampliación

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 un parámetro adicional con el valor “**extra**”.