Práctica 1 - Análisis léxico

Módulo I – Lenguaje Estructurado

Procesamiento de Lenguaje Escrito

Grado en Ciencia e Ingeniería de Datos

2022/2023

Profesores: Carlos Dafonte – carlos.dafonte@udc.es
Daniel Garabato – daniel.garabato@udc.es

Departamento de Ciencias de la Computación y Tecnologías de la Información
Universidade da Coruña

Diseño e implementación mediante la herramienta `SLY` de un analizador léxico para el tratamiento de mensajes de *log* centralizados del servicio de autenticación de una máquina Linux.

La entrada del analizador será un fichero de texto plano como el que se adjunta junto con este enunciado (`auth_example.log`). Cada línea representa un mensaje de *log*, compuesto por los siguientes campos:

- **Fecha:** Compuesto por una secuencia de letras que representan el mes, seguidas de una secuencia de dígitos que representa el día (separados por un espacio o un tabulador).
- **Nombre de la máquina**: Secuencia de caracteres alfanuméricos que representan el nombre del equipo en que se registró el evento.
- **Servicio:** Secuencia de caracteres que identifica el proceso o servicio que registra el evento. Nos centraremos en el servicio SSH: `sshd[%PID%]`.
- **Mensaje:** Encapsula el tipo de evento que se ha registrado, junto con información adicional. Trataremos especialmente aquellos asociados a un usuario autenticado correctamente ("`Accepted password for...`"), una contraseña errónea para un usuario ("`Failed password for..`"), o un usuario no autorizado ("`Invalid user...`" y "`Failed password for invalid user`"). Cabe destacar que existen muchos otros tipos de eventos que, no obstante, no consideraremos para el desarrollo de la práctica y simplemente catalogaremos como "otros".

El objetivo de la práctica consiste en implementar los siguientes contadores:

- Número total de eventos.
- Número de eventos por mes y quincena (primera o segunda).
- Número de eventos según la franja horaria: mañana (08:00h – 15:59h), tarde (16:00 – 23:59h) y noche (00:00h – 07:59h).
- Número de eventos por máquina.
- Según el tipo de evento identificado:
  - Usuarios autenticados correctamente ("`Accepted password for...`").
  - Contraseña errónea ("`Failed password for...`").
  - Usuario no autorizado ("`Invalid user...`" y "`Failed password for invalid user...`")

  Se obtendrá:
  - Número total de ocurrencias.
  - Número de eventos por máquina.
  - Número de eventos por usuario.

    - Número de eventos por clase (A, B o C) y tipo (privada o pública) de IP.
- Número total de ocurrencias para los demás tipos de eventos (otros).

Deben definirse adecuadamente cada uno de los ***tokens*** a identificar, especialmente aquellos que se necesiten contabilizar de algún modo. Aquellos elementos que no requieran tratamiento podrán ignorarse mediante expresiones regulares, pero no pueden dar lugar a errores de procesamiento.

Junto con el enunciado, se proporciona un esqueleto en python (`p1_base.py`) del que tendréis que partir y cuya sección `__main__` **NO debéis modificar**. En ella se lee íntegramente el fichero de *log* proporcionado como entrada por el sistema estándar. Así mismo, procesa la entrada con el analizador léxico e invoca el método `print_output` incluido en el esqueleto, que debéis completar adecuadamente para mostrar por pantalla el valor de los contadores arriba indicados, siguiendo este formato:

```
#contadores_generales
total_eventos,x
total_aceptados,x
total_fallidos,x
total_no_autorizados,x
total_otros,x
#eventos_por_fecha
mes_1,primera,x
mes_1,segunda,x
mes_2,segunda,x
[...]
#eventos_por_hora
manana,x
tarde,x
noche,x
#eventos_por_maquina
maquina_1,x
maquina_2,x
[...]
#eventos_aceptados_por_maquina
maquina_1,x
maquina_2,x
[...]
#eventos_aceptados_por_usuario
usuario_1,x
usuario_2,x
[...]
#eventos_aceptados_por_ip
clase_a,privada,x
clase_a,publica,x
clase_b,privada,x
clase_b,publica,x
```

```
clase_c,privada,x
clase_c,publica,x
#eventos_fallidos_por_maquina
maquina_1,x
maquina_2,x
[...]
#eventos_fallidos_por_usuario
usuario_1,x
usuario_2,x
[...]
#eventos_fallidos_por_ip
clase_a,privada,x
clase_a,publica,x
clase_b,privada,x
clase_b,publica,x
clase_c,privada,x
clase_c,publica,x
#eventos_no_autorizados_por_maquina
maquina_1,x
maquina_2,x
[...]
#eventos_no_autorizados_por_usuario
usuario_1,x
usuario_2,x
[...]
#eventos_no_autorizados_por_ip
clase_a,privada,x
clase_a,publica,x
clase_b,privada,x
clase_b,publica,x
clase_c,privada,x
clase_c,publica,x
```

Obviamente, los valores `x` tendrán que corresponderse con el número de ocurrencias identificadas para cada caso. Los nombres de los meses, así como los nombres de máquinas y de usuarios tendrán que corresponderse con los que figuren en los registros. Las cabeceras (líneas precedidas por `#`) deben ser las que se indican, mientras que la secuencia `[...]` se ha utilizado a modo de ejemplo para indicar una sucesión de elementos (los que sean). Si para algún caso no se detectasen ocurrencias, la cabecera debe aparecer igualmente, pero sin sucesión de elementos (pasando directamente a la siguiente cabecera). Si se omite algún elemento en las listas, se asumirá que tiene 0 ocurrencias.