Práctica 1

Recuperación de Información

Amin Kasrou Aouam

UNIVERSIDAD
DE GRANADA

2020-10-25

Índice

Práctica 1

En esta práctica, vamos a obtener información de una serie de documentos usando *Apache Tika*.

Ejecución

Implementamos la práctica usando *Java* como lenguaje de programación, y *Maven* como herramienta de gestión del proyecto.

En el caso que deseemos utilizar *Maven*, debemos ejecutar los siguientes comandos:

1. Compilar el proyecto

```
mvn compile
```

1. Ejecutar el proyecto

```
mvn exec:java -Dexec.mainClass="org.RI.P1.AnalyzeDirectory" -Dexec.args
    ="data metadata"
```

Debemos modificar el argumento **metadata** según la salida que deseemos:

- metadata: obtenemos la información de los archivos (nombre, codificación, tipo)
- links: obtenemos la lista de enlaces de cada archivo
- frequency: se guarda la frecuencia de las palabras de cada documento en un archivo (se encuentran en la carpeta output).

Implementación

Lamentablemente, no hemos podido implementar la funcionalidad del gráfico para comprobar si se cumple la ley de Zipf. Además de ello, nuestro *tokenizer* no funciona demasiado bien debido al uso de una expresión regular que no toma en cuenta todos los casos.