UNIVERSIDAD DE GRANADA
E.T.S.I INFORMÁTICA Y TELECOMUNICACIÓN

Recuperación de Información (RI)

Preprocesado de documentos

Parte II. Análisis del texto

Curso 2017-2018

Cuarto Curso del Grado en Ingeniería Informática

María Camarero Granados

Javier Gómez Luzón

Francisco Porcel Molina

• Análisis comparativo entre los distintos resultados obtenidos.

Hemos añadido los analizadores WhitespaceAnalyzer, SimpleAnalyzer, StandardAnalyzer para cualquier tipo de documento de texto y EnglishAnalyzer, FrenchAnalyzer y FinnishAnalyzer para cada uno de los lenguajes.

·WhiteSpaceAnalyzer: Nos divide el texto separado por espacios en blanco.
·SimpleAnalyzer: Nos separa el texto por todo aquello que no sean letras.
·StandardAnalyzer: Nos convierte a minúscula, elimina las palabras vacías de los textos en inglés.
·English, French y Finnish Analyzer nos analiza las palabras de cada uno de los idiomas permitiéndonos eliminar palabras vacías.

• Analizador específico para indexar código fuente.

Creamos un CustomAnalyzer con el tokenizador StandardTokenizerFactory para que nos genere los tokens. Añadimos con addTokenFilter el LowerCaseFilterFactory para pasar a minúsculas. Por último con addTokenFilter añadimos StopFilterFactory para eliminar las palabras vacías de java que previamente hemos añadido en un fichero txt.

• Analizador propio.

Para crear nuestro analizador sobreescribimos el método createComponents para usar los tokenizadores y filtros adecuados para obtener los términos que nos interesan. Para ello, utilizamos un StandardTokenizer y, además, una serie de filtros para lograr que no haya palabras vacías y que los términos estén lematizados, entre otras cosas.

• Trabajo en grupo:

María Camarero Granados: Realización de la documentación y analizador propio.

Javier Gómez Luzón: Análisis comparativo y analizador específico.

Francisco Porcel Molina: Analizador específico y analizador propio.