UNIVERSIDAD DE GRANADA

PRÁCTICA 2: Preprocesado de documentos
Parte II. Análisis del texto

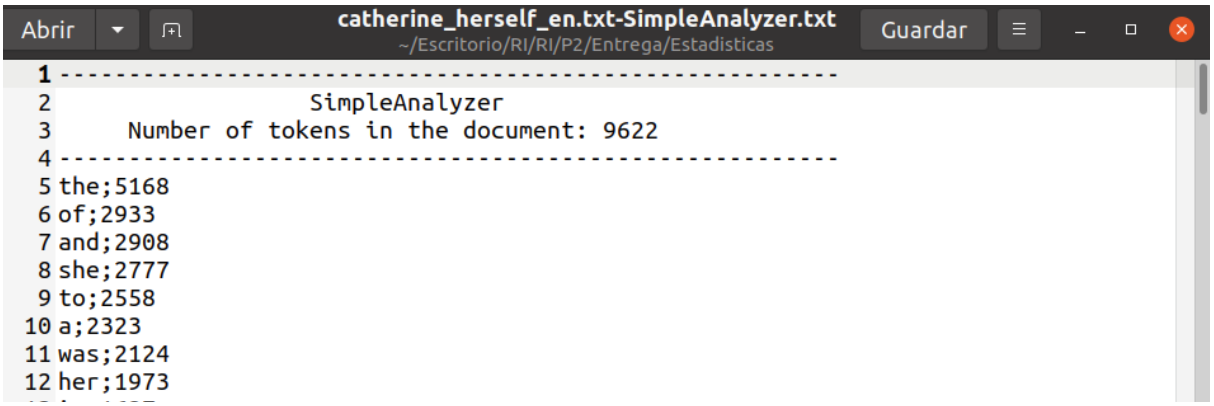
Anne Serrano Andrades
Pablo Huertas Arroyo

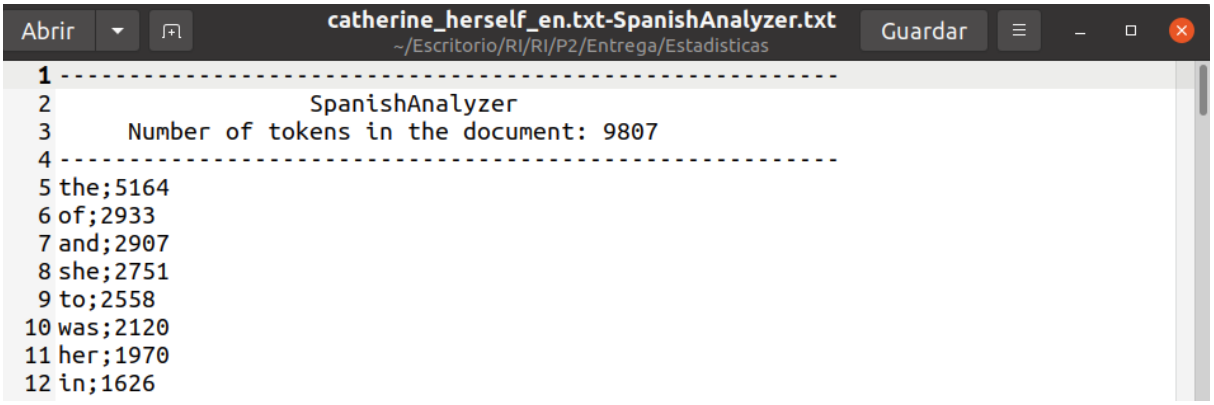
1. EJERCICIO 1

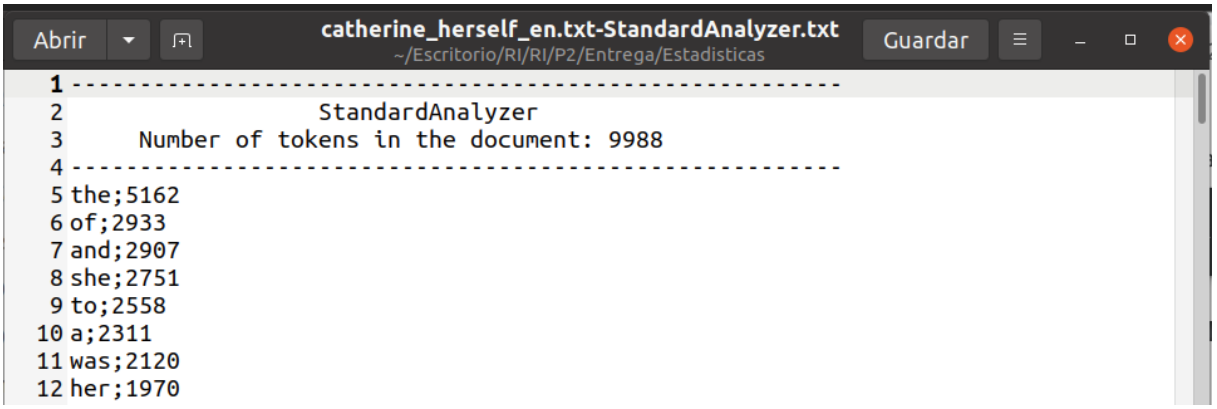
Se nos pide que sobre los documentos de la práctica anterior hagamos un estudio estadístico sobre la diferencia entre los tokens obtenidos usando diferentes tipos de análisis predefinidos, además de un análisis comparativo entre los resultados

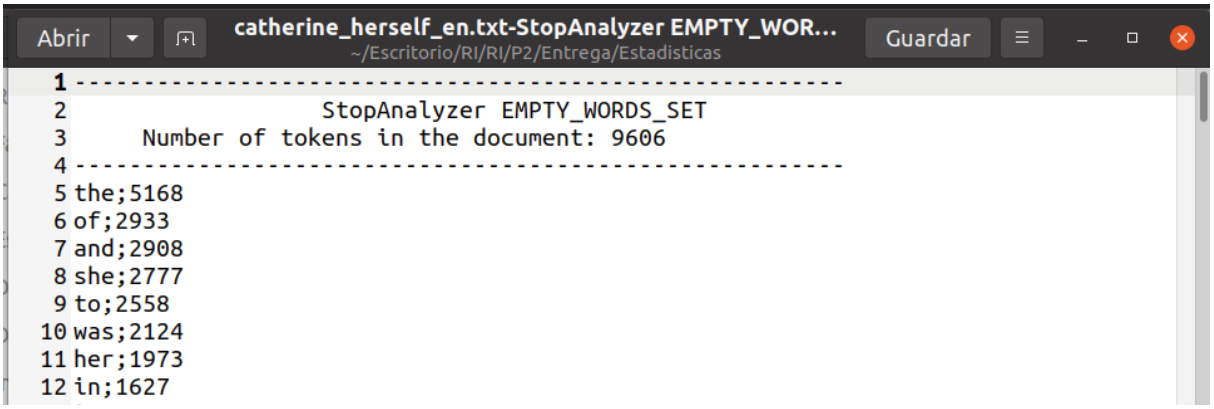
Con la práctica anterior hemos mejorado su eficacia almacenando los tokens extraídos con los analizadores en un map, de esta manera se escribirá en un fichero la clave (token) y su valor almacenado (número de veces que aparece).

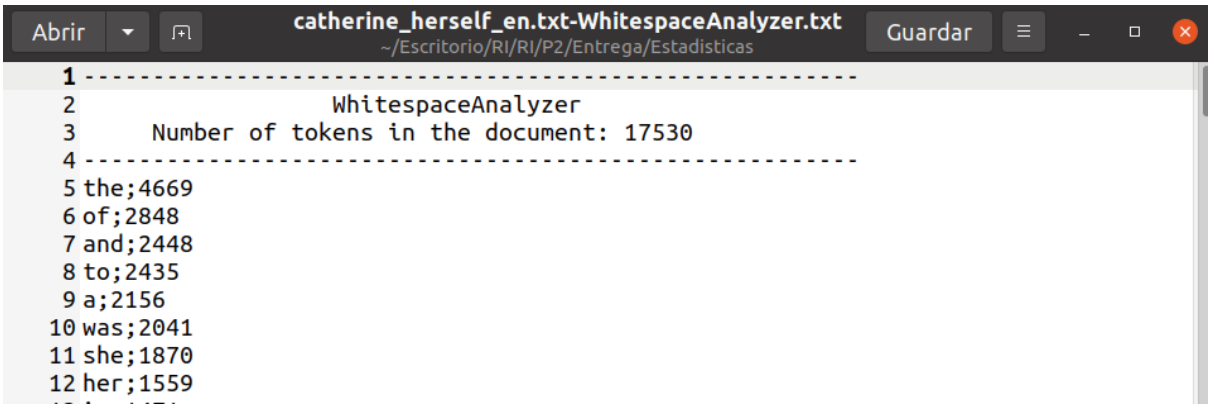
- ESTUDIO ESTADÍSTICO



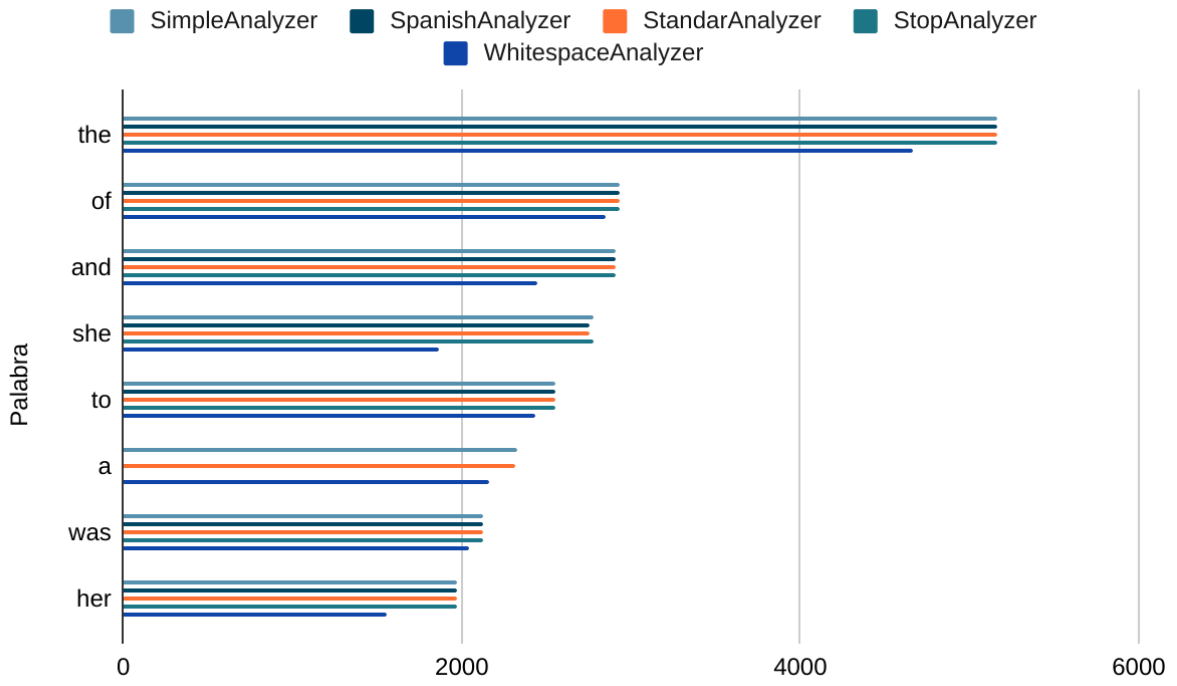




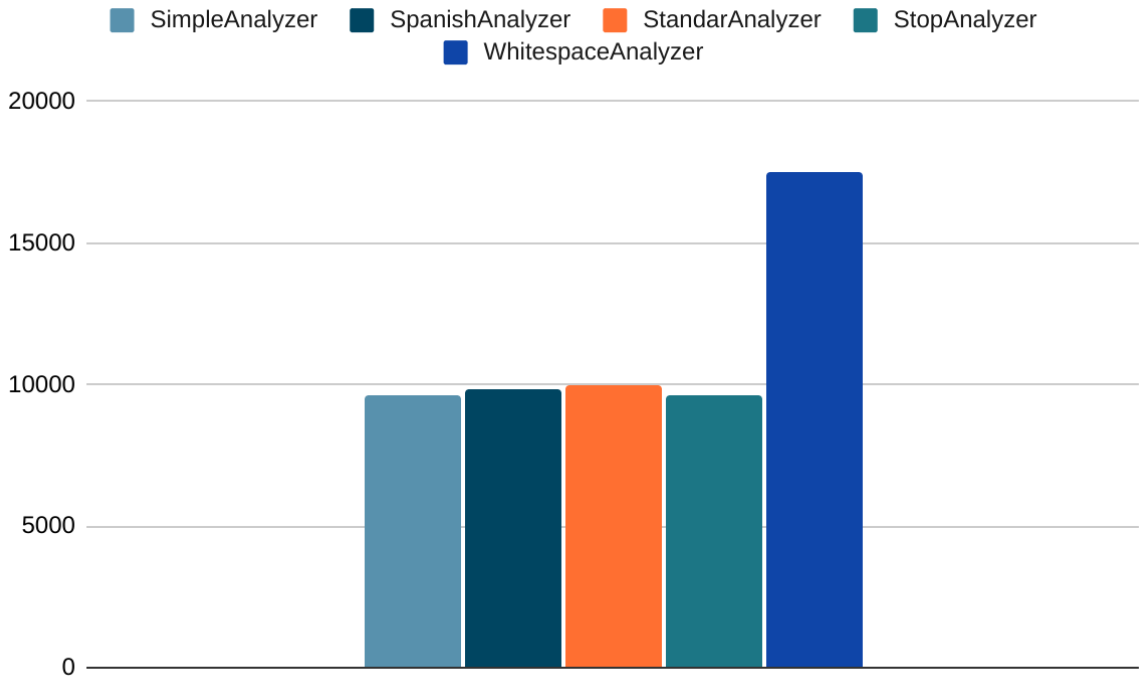




Palabra	SimpleAnalyzer	SpanishAnalyzer	StandarAnalyzer	StopAnalyzer	WhitespaceAnalyzer
the	5168	5164	5162	5168	4669
of	2933	2933	2933	2933	2848
and	2908	2907	2907	2908	2448
she	2777	2751	2751	2777	1870
to	2558	2558	2558	2558	2435
a	2323	0	2311	0	2156
was	2124	2120	2120	2124	2041
her	1973	1970	1970	1973	1559



SimpleAnalyzer	SpanishAnalyzer	StandarAnalyzer	StopAnalyzer	WhitespaceAnalyzer
9622	9807	9988	9606	17530



2. EJERCICIO 2

Probar sobre un texto relativamente pequeño el efecto que tienen los siguientes tokenFilters: StandardFilter, LowerCaseFilter, StopFilter, SnowballFilter, ShingleFilter, EdgeNGramCommonFilter, NGramTokenFilter, CommonGramsFilter, SynonymFilter.

Para algunos tokenFilters necesitaremos añadir algunas modificaciones por ejemplo:

- SynonymFilter: necesitamos un diccionario con sinónimos.
- CommonGramsFilter: necesitamos un array con las common words.
- StopFilter: necesitamos un contenedor para las palabras vacías.

SALIDA DEL EJERCICIO 2

Texto original: En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ StandardAnalyzer ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

en un lugar de la mancha de cuyo nombre no quiero acordarme no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero adarga antigua rocín flaco y galgo corredor

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ LowerCaseFilter ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

en un lugar de la mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ StopFilter ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

lugar mancha cuyo nombre quiero acordarme tiempo vivía hidalgo lanza astillero adarga antigua rocín flaco galgo corredor

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ SnowballFilter ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

en un lug de la manch de cuy nombr no quier acord no ha much tiemp que viv un hidalg de los de lanz en astiller adarg antigu rocin flac y galg corredor

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ ShingleFilter ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

en en un un un lugar lugar lugar de de de la la la mancha mancha mancha de de de cuyo cuyo cuyo nombre nombre nombre no no no quiero quiero quiero acordarme acordarme acordarme no no no ha ha ha mucho mucho mucho tiempo tiempo tiempo que que que vivía vivía vivía un un un hidalgo hidalgo hidalgo de de de los los los de de de lanza lanza lanza en en en astillero astillero astillero adarga adarga adarga antigua antigua antigua rocín rocín rocín flaco flaco flaco y y y galgo galgo galgo corredor corredor

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ EdgeNGramTokenFilter ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

luga manc cuyo nomb quie acor much tiem viví hida lanz asti adar anti rocí flac galg corr

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ NGramTokenFilter ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

luga ugar manc anch ncha cuyo nomb ombr mbre quie uier iero acor cord orda rdar darm arme much ucho tiem iemp empo viví ivía hida idal dalg algo lanz anza asti stil till ille ller lero adar darg arga anti ntig tigu igua rocí ocín flac laco galg algo corr orre rred redo edor

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ CommonGramsFilter ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

en en_un un un_lugar lugar lugar_de de de_la la la_mancha mancha mancha_de de de_cuyo cuyo nombre no quiero acordarme no ha mucho tiempo que vivía vivía_un un un_hidalgo hidalgo hidalgo_de de de_los los los_de de de_lanza lanza lanza_en en en_astillero astillero adarga antigua rocín flaco y galgo corredor

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ SynonymFilter ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

en un lugar de la mancha de cuyo nombre apodo no quiero acordarme hacer memoria no ha mucho tiempo que vivía un hidalgo noble de los de lanza en astillero varadero adarga antigua rocín flaco y galgo lebrel corredor

3. EJERCICIO 3

Diseñar un analizador propio.

Nuestro analizador comprueba la longitud de las palabras según dicha longitud ocurren tres casos:

1. >3 caracteres <5 caracteres: la convierte a minúscula y pone la palabra al revés
2. >3 y >5 : solo se convierte a minúscula

3. <3 se queda igual

```
// Usamos CustomAnalyzer para poder usar los distintos analizadores

// Invertimos los tokens que tienen una longitud mayor que 3
// Ponemos en mayúsculas los tokens que tienen una longitud menor o igual a 5
Analyzer analyzer = CustomAnalyzer.builder()
        .withTokenizer("standard")
        .addTokenFilter("uppercase")
        .whenTerm(token -> token.length() > 3)
        .addTokenFilter("reversestring")
        .endwhen()
        .whenTerm(token -> token.length() <= 5)
        .addTokenFilter("lowercase")
        .endwhen()
        .build();

TokenStream stream = analyzer.tokenStream(null, text);

stream = new NumbersFilter(stream);

// Creamos el PrintWriter para escribir en cada fichero
PrintWriter writer = new PrintWriter("./AnalizadorPersonalizado/" + fileName + ".txt");


stream.reset();
// Obtenemos los tokens
while (stream.incrementToken())
    writer.print(stream.getAttribute(CharTermAttribute.class).toString()+"\n");

System.out.println();

stream.end(); // Se llama cuando se termina de iterar
stream.close(); // Liberamos los recursos asociados al stream
}
```

SALIDA DEL EJERCICIO 3

the
TCEJORP
GREBNETUG
koobe
of
ENIREHTAC
FLESREH
siht
koobe
is
for
the
use
of
ENOYNA
EREHWYNA
in
the
DETINU
SETATS