Ejercicio 1:

Para esta comparativa, hemos utilizado la clase creada para la práctica 1. Y vamos a realizar dos análisis. Uno con SimpleAnalyzer() y otro con WhitespaceAnalyzer(). Y lo compararemos con la extracción que realizamos en la primera práctica.

Para ello, vamos a utilizar el archivo del "quijote.txt". En donde sabemos, que las tres palabras que más se encuentran[1] son "de", "que" e "y".

Palabras	AnalizadorPractica1	SimpleAnalyzer	WhiteAnalyzer	%DiferenciaSimple	%DiferenciaWhite
que	20337	20414	19236	-0,378620249	5,413777843
de	17756	17953	17745	-1,109484118	0,06195089
y	16973	17982	15706	-5,944735757	7,464797031
la	10075	10227	10073	-1,508684864	0,019851117
a	9497	9776	9483	-2,937769822	0,147414973
el	7840	8076	7835	-3,010204082	0,06377551
en	7782	8110	7773	-4,214854793	0,115651503

Tomando de referencia los datos de la primera práctica, vemos como usando el analizador "Simple", obtenemos una mejora. Sin embargo, vemos como la "y", tienes mas apariencias que la "de", algo que es erróneo. Con el analizador White, obtenemos unos peores resultados, pero no en todos los datos. Ya que en la mayoría la diferencia es tan solo de un 0.05%. Podemos concluir, que nuestro analizador es, por tanto, "mejor" que los ofrecidos por la librería Lucene.

Ejercicio 2:

Probar sobre un texto relativamente pequeño el efecto que tienen los siguientes tokenFilters:

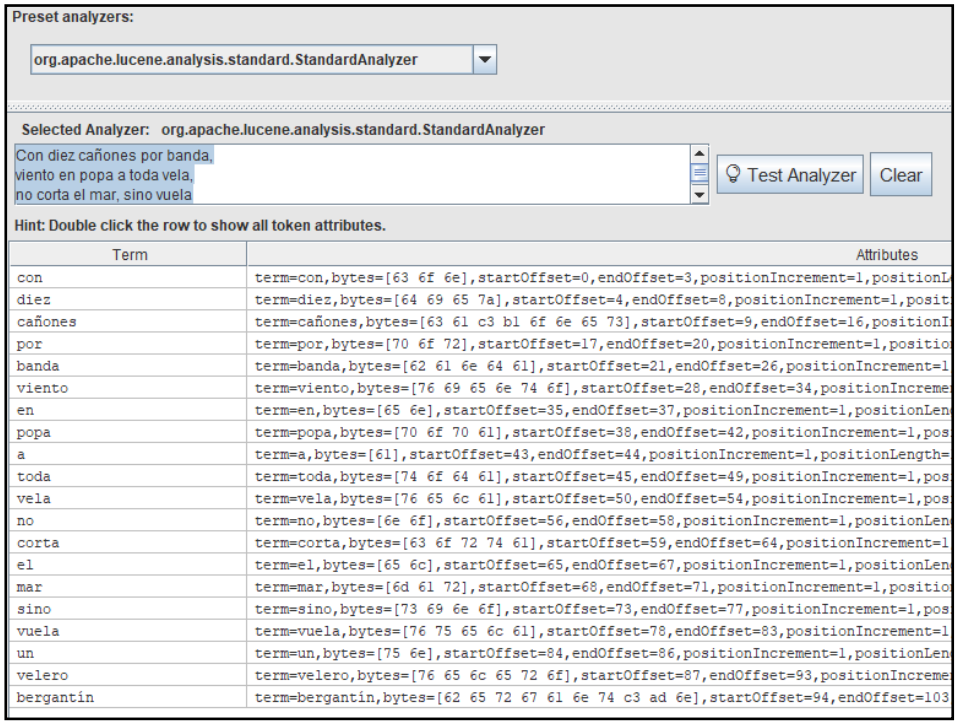
Para este ejercicio y los sucesivos, hemos usado el texto de José Espronceda: "Canción del pirata".

- StandardFilter: Este es el analizador más sofisticado y es capaz de manejar nombres, direcciones de correo electrónico, etc. Minúscula cada token y elimina las palabras comunes y los signos de puntuación, si los hubiera. Este analizador, nos ha dado problemas al intentar realizarlo en java, así que es el único que hemos probado con la interfaz gráfica.

"*Con diez cañones por banda, viento en popa a toda vela, no corta el mar, sino vuela un velero bergantín;*"

Con este fragmento, el resultado será:

[1] http://www.cervantesvirtual.com/descargaPdf/las-palabras-del-quijote--notas-introductorias-0/



Term	Attributes
con	term=con,bytes=[63 6f 6e],startOffset=0,endOffset=3,positionIncrement=1,positionL
diez	term=diez,bytes=[64 69 65 7a],startOffset=4,endOffset=8,positionIncrement=1,posit
cañones	term=cañones,bytes=[63 61 c3 b1 6f 6e 65 73],startOffset=9,endOffset=16,positionI
por	term=por,bytes=[70 6f 72],startOffset=17,endOffset=20,positionIncrement=1,positio
banda	term=banda,bytes=[62 61 6e 64 61],startOffset=21,endOffset=26,positionIncrement=1
viento	term=viento,bytes=[76 69 65 6e 74 6f],startOffset=28,endOffset=34,positionIncreme
en	term=en,bytes=[65 6e],startOffset=35,endOffset=37,positionIncrement=1,positionLen
popa	term=popa,bytes=[70 6f 70 61],startOffset=38,endOffset=42,positionIncrement=1,pos
a	term=a,bytes=[61],startOffset=43,endOffset=44,positionIncrement=1,positionLength=
toda	term=toda,bytes=[74 6f 64 61],startOffset=45,endOffset=49,positionIncrement=1,pos
vela	term=vela,bytes=[76 65 6c 61],startOffset=50,endOffset=54,positionIncrement=1,pos
no	term=no,bytes=[6e 6f],startOffset=56,endOffset=58,positionIncrement=1,positionLen
corta	term=corta,bytes=[63 6f 72 74 61],startOffset=59,endOffset=64,positionIncrement=1
el	term=el,bytes=[65 6c],startOffset=65,endOffset=67,positionIncrement=1,positionLen
mar	term=mar,bytes=[6d 61 72],startOffset=68,endOffset=71,positionIncrement=1,positio
sino	term=sino,bytes=[73 69 6e 6f],startOffset=73,endOffset=77,positionIncrement=1,pos
vuela	term=vuela,bytes=[76 75 65 6c 61],startOffset=78,endOffset=83,positionIncrement=1
un	term=un,bytes=[75 6e],startOffset=84,endOffset=86,positionIncrement=1,positionLen
velero	term=velero,bytes=[76 65 6c 65 72 6f],startOffset=87,endOffset=93,positionIncreme
bergantín	term=bergantín,bytes=[62 65 72 67 61 6e 74 c3 ad 6e],startOffset=94,endOffset=103

A partir de ahora, se usará el texto del poema de Espronceda[2].

- LowerCaseFilter: Convierte a minúscula todas las letras en mayúscula.

"*con diez cañones por banda viento en popa a toda vela no corta el mar sino vuela un velero bergantín bajel pirata que llaman por su bravura el temido en todo mar conocido del uno al otro confín la luna en el mar riela en la lona gime el viento y alza en blando movimiento olas de plata y azul y ve el capitán pirata cantando alegre en la popa asia a un lado al otro europa y allá a su frente stambul navega velero mío sin temor que ni enemigo navío ni tormenta ni bonanza tu rumbo a torcer alcanza ni a sujetar tu valor veinte presas hemos hecho a despecho del inglés y han rendido sus pendones cien naciones a mis pies qué es mi barco mi tesoro qué es mi dios la libertad mi ley la fuerza y el viento mi única patria la mar*"

- StopFilter: Elimina las palabras vacías.

"*Con diez cañones por banda viento en popa toda vela corta el mar sino vuela un velero bergantín bajel pirata que llaman por su bravura el Temido en todo mar conocido del uno al otro confín La luna en el mar riela en la lona gime el viento y alza en blando movimiento olas de plata y azul y ve el capitán pirata cantando alegre en la popa Asia un lado al otro Europa y allá su frente Stambul Navega velero mío sin temor que ni enemigo navío ni tormenta ni bonanza tu rumbo torcer alcanza ni sujetar tu valor Veinte presas hemos hecho despecho del inglés y han rendido sus pendones cien naciones mis pies Qué es mi barco Mi tesoro Qué es mi Dios La libertad Mi ley La fuerza y el viento Mi única patria La mar*"

[2] http://recursosdidacticos.es/textos/texto.php?id=270

- Snowball-Filter: reduce las palabras a su raíz. En este caso utilizaremos el SpanishStemmer.

"*Con diez cañon por band vient en pop a tod vel no cort el mar sin vuel un veler bergantin bajel pirat que llam por su bravur el Tem en tod mar conoc del uno al otro confin La lun en el mar riel en la lon gim el vient y alza en bland movimient olas de plat y azul y ve el capitan pirat cant alegr en la pop Asi a un lad al otro Europ y alla a su frent Stambul Naveg veler mio sin temor que ni enemig navi ni torment ni bonanz tu rumb a torc alcanz ni a sujet tu valor Veint pres hem hech a despech del ingles y han rend sus pendon cien nacion a mis pies Que es mi barc Mi tesor Que es mi Dios La libert Mi ley La fuerz y el vient Mi unic patri La mar*"

- ShingleFilter: construye una nueva frase, que se realizará con cada término. El proceso será, él mismo y a éste en pareja con el siguiente.

"Con Con diez diez diez cañones cañones cañones por por por banda banda banda viento viento viento en en en popa popa popa a a a toda toda toda vela vela vela no no no corta corta corta el el el mar mar mar sino sino sino vuela vuela vuela un un un velero velero velero bergantín bergantín bergantín bajel bajel bajel pirata pirata pirata que que que llaman llaman llaman por por por su su su bravura bravura bravura el el el Temido Temido Temido en en en todo todo todo mar mar mar conocido conocido conocido del del del uno uno uno al al al otro otro otro confín confín confín La La La luna luna luna en en en el el el mar mar mar riela riela riela en en en la la la lona lona lona gime gime gime el el el viento viento viento y y y alza alza alza en en en blando blando blando movimiento movimiento movimiento olas olas olas de de de plata plata plata y y y azul azul azul y y y ve ve ve el el el capitán capitán capitán pirata pirata pirata cantando cantando cantando alegre alegre alegre en en en la la la popa popa popa Asia Asia Asia a a a un un un lado lado lado al al al otro otro otro Europa Europa Europa y y y allá allá allá a a a su su su frente frente frente Stambul Stambul Stambul Navega Navega Navega velero velero velero mío mío mío sin sin sin temor temor temor que que que ni ni ni enemigo enemigo enemigo navío navío navío ni ni ni tormenta tormenta tormenta ni ni ni bonanza bonanza bonanza tu tu tu rumbo rumbo rumbo a a a torcer torcer torcer alcanza alcanza alcanza ni ni ni a a a sujetar sujetar sujetar tu tu tu valor valor valor Veinte Veinte Veinte presas presas presas hemos hemos hemos hecho hecho hecho a a a despecho despecho despecho del del del inglés inglés inglés y y y han han han rendido rendido rendido sus sus sus pendones pendones pendones cien cien cien naciones naciones naciones a a a mis mis mis pies pies pies Qué Qué Qué es es es mi mi mi barco barco barco Mi Mi Mi tesoro tesoro tesoro Qué Qué Qué es es es mi mi mi Dios Dios Dios La La La libertad libertad libertad Mi Mi Mi ley ley ley La La La fuerza fuerza fuerza y y y el el el viento viento viento Mi Mi Mi única única única patria patria patria La La La mar mar"

- EdgeNGramCommonFilter: Con el número entero que recibe X, elimina las palabras que tengan un tamaño L < X, deja las palabras con L == X y de las palabras cuyo tamaño L > X sólo deja los X primeros caracteres.

"Con die cañ por ban vie pop tod vel cor mar sin vue vel ber baj pir que lla por bra Tem tod mar con del uno otr con lun mar rie lon gim vie alz bla mov ola pla azu cap pir can ale pop Asi lad otr Eur all fre Sta Nav vel mío sin tem que ene nav tor bon rum tor alc suj val Vei pre hem hec des del ing han ren sus pen cie nac mis pie Qué bar tes Qué Dio lib ley fue vie úni pat mar"

- NGramTokenFilter: Con el número entero que recibe X, elimina las palabras que tengan un tamaño L < X, deja las palabras con L == X y de las palabras cuyo tamaño L > X muestra cortes de palabras con el tamaño X. Ejemplo: Con la palabra caracol, mostrará caraco, aracol. (Sólo muestra de X en X, hemos supuesto X=6 y va avanzando de carácter en carácter)

"*cañone añones viento velero bergan ergant rgantí gantín pirata llaman bravur ravura Temido conoci onocid nocido confín viento blando movimi ovimie vimien imient miento capitá apitán pirata cantan antand ntando alegre Europa frente Stambu tambul Navega velero enemig nemigo tormen orment rmenta bonanz onanza torcer alcanz lcanza sujeta ujetar Veinte presas despec espech specho inglés rendid endido pendon endone ndones nacion acione ciones tesoro libert iberta bertad fuerza viento patria*"

- CommonGramsFilter: Cuando encuentra una palabra vacía, la muestra añadida a su anterior, ella misma y la posterior: Hola a todos -> Hola_a a a_todos.

"Con diez cañones por banda viento en popa popa_a a a_toda toda vela vela_no no no_corta corta el mar sino vuela un velero bergantín bajel pirata que llaman por su bravura el Temido en todo mar conocido del uno al otro confín La luna en el mar riela en la lona gime el viento y alza en blando movimiento olas de plata y azul y ve el capitán pirata cantando alegre en la popa Asia Asia_a a a_un un lado al otro Europa y allá allá_a a a_su su frente Stambul Navega velero mío sin temor que ni enemigo navío ni tormenta ni bonanza tu rumbo rumbo_a a a_torcer torcer alcanza ni ni_a a a_sujetar sujetar tu valor Veinte presas hemos hecho hecho_a a a_despecho despecho del inglés y han rendido sus pendones cien naciones naciones_a a a_mis mis pies Qué es mi barco Mi tesoro Qué es mi Dios La libertad Mi ley La fuerza y el viento Mi única patria La mar"

- SynonymFilter: Dado un mapa de sinónimos, añade al lado de la palabra el sinónimo correspondiente. Nuestro mapa es el siguiente:

```
SynonymMap.Builder builder = new SynonymMap.Builder(true);
        builder.add(new CharsRef("cañones"), new CharsRef("(disparadores)"), true);
        builder.add(new CharsRef("barco"), new CharsRef("(velero)"), true);
        builder.add(new CharsRef("diez"), new CharsRef("(10)"), true);
        builder.add(new CharsRef("mar"), new CharsRef("(playa)"), true);
```

"*Con diez (10) cañones (disparadores) por banda viento en popa a toda vela no corta el mar (playa) sino vuela un velero bergantín bajel pirata que llaman por su bravura el Temido en todo mar (playa) conocido del uno al otro confín La luna en el mar (playa) riela en la lona gime el viento y alza en blando movimiento olas de plata y azul y ve el capitán pirata cantando alegre en la popa Asia a un lado al otro Europa y allá a su frente Stambul Navega velero mío sin temor que ni enemigo navío ni tormenta ni bonanza tu rumbo a torcer alcanza ni a sujetar tu valor Veinte presas hemos hecho a despecho del inglés y han rendido sus pendones cien naciones a mis pies Qué es mi barco (velero) Mi tesoro Qué es mi Dios La libertad Mi ley La fuerza y el viento Mi única patria La mar (playa)*"

Ejercicio 3:

El filtro creado a partir de otros será el siguiente:

```
Analyzer ana = CustomAnalyzer.builder(Paths.get("."))
        .withTokenizer(StandardTokenizerFactory.class)
        .addTokenFilter(LowerCaseFilterFactory.class)
        .addTokenFilter(StopFilterFactory.class, "ignoreCase", "false", "words", "stopwords.txt", "format", "wordset")
        .build();
```

Usamos el filtro de LowerCase y el StopFilter. Por tanto, convertimos a minúscula todo el texto y luego eliminamos palabras que tenemos en el txt. Que son las siguientes:

{Cañones, viento, por, ley, tormenta, valor, naciones, velero, pirata, presas, }

El resultado del texto citado anteriormente quedaría finalmente será:

"*con diez banda en popa a toda vela no corta el mar sino vuela un bajel que llaman su bravura el temido en todo mar conocido del uno al otro confín la luna en el mar riela en la lona gime el y alza en blando movimiento olas de plata y azul y ve el capitán cantando alegre en la popa asia a un lado al otro europa y allá a su frente stambul navega mío sin temor que ni enemigo navío ni ni bonanza tu rumbo a torcer alcanza ni a sujetar tu veinte hemos hecho a despecho del inglés y han rendido sus pendones cien a mis pies qué es mi barco mi tesoro qué es mi dios la libertad mi la fuerza y el mi única patria la mar*"

Ejercicio 4:

Para crear nuestro analizador, hemos extendido dos clases Analyzer y TokemFilter:

Para la clase Analyzer, realizamos un código similar al StandardAnalyzer y le pasamos nuestro filtro específico.

```
public class AnalizadorSufijoEjer4 extends Analyzer{

    @Override
    protected TokenStreamComponents createComponents(String string) {
        final Tokenizer tokenizer = (Tokenizer)new LetterTokenizer();
        return new Analyzer.TokenStreamComponents(tokenizer, (TokenStream)new filtroSufijosEjer4((TokenStream)tokenizer));
    }
```

Para el filtro usado de plantilla la clase LowerCase. Por tanto, realizamos tres condiciones. La de si es un token para que el bucle siga y dentro de este, si la palabra tiene longitud menor que tres, eliminará la palabra y si es mayor a tres, cogerá sólo las últimas cuatro posiciones.

```
public class AnalizadorEjer4 extends TokenFilter
{
    private final CharTermAttribute termAtt;

    public AnalizadorEjer4(final TokenStream in) {
        super(in);
        this.termAtt = (CharTermAttribute)this.addAttribute((Class)CharTermAttribute.class);
    }

    @Override
    public final boolean incrementToken() throws IOException {
        if(this.input.incrementToken()){
            if(this.termAtt.length()>4){
                char aux1 = termAtt.buffer()[termAtt.length()-1];
                char aux2 = termAtt.buffer()[termAtt.length()-2];
                char aux3 = termAtt.buffer()[termAtt.length()-3];
                char aux4 = termAtt.buffer()[termAtt.length()-4];
                termAtt.setEmpty();
                termAtt.append(aux4);
                termAtt.append(aux3);
                termAtt.append(aux2);
                termAtt.append(aux1);
                return true;
            }
            if(this.termAtt.length()<4){
                termAtt.setEmpty();
            }
            return true;
        }
        return false;
    }
}
```

En ambos casos eliminamos la palabra, pero en el segundo, le añadimos los caracteres, extraídos anteriormente.

Luego en la función de filtrado, le añadimos la cláusula de que si lo que vamos a añadir es algo vacío, no lo añada.

Tras darle, el texto de prueba, el resultado del poema sería el siguiente:

"*diez ones anda ento popa toda vela orta sino uela lero ntín ajel rata aman vura mido todo cido otro nfín luna iela lona gime ento alza ando ento olas lata azul itán rata ando egre popa Asia lado otro ropa allá ente mbul vega lero emor migo avío enta anza umbo rcer anza etar alor inte esas emos echo echo glés dido ones cien ones pies arco soro Dios rtad erza ento nica tria*"

Trabajo en grupo

En esta práctica, en un principio estuvimos hablando sobre cuál era el mejor entorno posible para la realización de la práctica. Decidimos desarrollarlo desde Netbeans debido a que nos ha sido muy sencillo trabajar con distintas clases y funciones además de poder incorporar librerías de manera inmediata. Hemos trabajado cada uno desde su casa y usando Google Meet para poder trabajar y poder comunicarnos fluidamente. Hemos aprovechado la práctica 1 para empezar.

El método de trabajo ha sido sencillo: primero pensábamos en cómo podíamos implementar nuestras ideas en el proyecto. A continuación, buscábamos la información que nos fuese necesaria para poder implementarla ya fuese desde el PDF de la práctica o desde Internet. Por último, hemos creado el código para la práctica y hemos corregido los errores que nos ha producido.

Bibliografía

https://lucene.apache.org/core/4_3_0/analyzers-common/org/apache/lucene/analysis/snowball/SnowballFilter.html

https://lucene.apache.org/

https://lucene.apache.org/core/7_4_0/core/org/apache/lucene/analysis/LowerCaseFilter.html

https://lucene.apache.org/core/6_4_0/core/org/apache/lucene/analysis/Analyzer.html

https://www.baeldung.com/lucene-analyzers