Práctica 2. Tika

Santiago Álvarez Valdivia

Ioannis Efthymiou

17 de octubre de 2023

Recuperación de Información

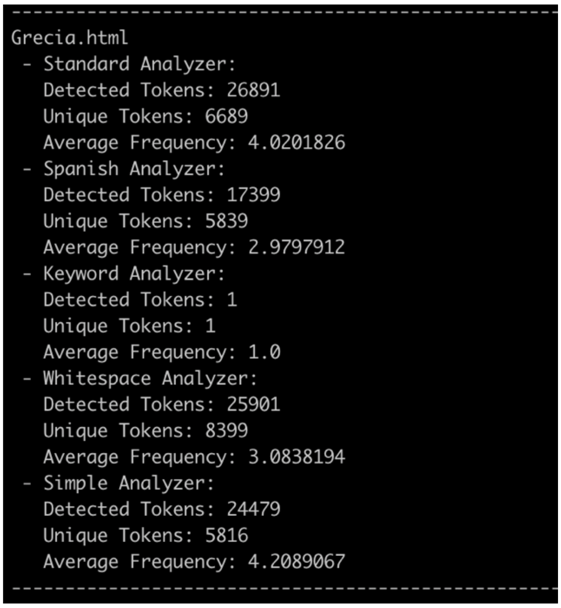
E.T.S. de Ingenierías Informática y de Telecomunicaciones

Universidad de Granada

Program Functionality

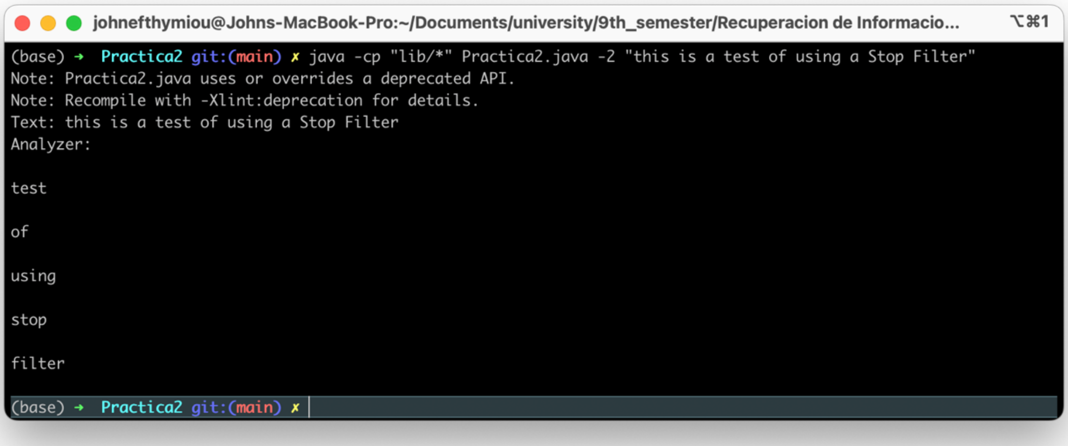
1) Analysis of documents

java -cp "lib/*" Practica2.java -1 testFiles/



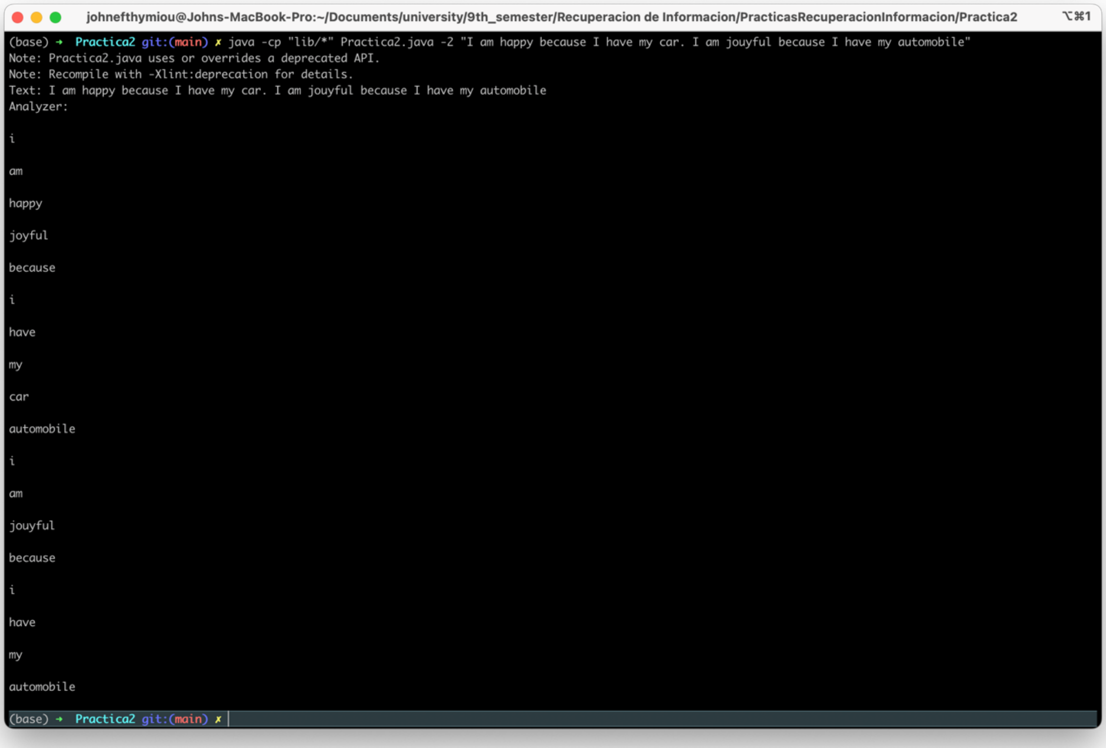
```
--------------------------------------------------
Grecia.html
 - Standard Analyzer:
   Detected Tokens: 26891
   Unique Tokens: 6689
   Average Frequency: 4.0201826
 - Spanish Analyzer:
   Detected Tokens: 17399
   Unique Tokens: 5839
   Average Frequency: 2.9797912
 - Keyword Analyzer:
   Detected Tokens: 1
   Unique Tokens: 1
   Average Frequency: 1.0
 - Whitespace Analyzer:
   Detected Tokens: 25901
   Unique Tokens: 8399
   Average Frequency: 3.0838194
 - Simple Analyzer:
   Detected Tokens: 24479
   Unique Tokens: 5816
   Average Frequency: 4.2089067
--------------------------------------------------
```

2) Java -cp “lib/*” Practica2.java -2 “this is a test of using a StopFilter”


johnefthymiou@Johns-MacBook-Pro:~/Documents/university/9th_semester/Recuperacion de Informacio...
⌥⌘1

```
(base) ➜  Practica2 git:(main) ✗ java -cp "lib/*" Practica2.java -2 "this is a test of using a Stop Filter"
Note: Practica2.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
Text: this is a test of using a Stop Filter
Analyzer:

test

of

using

stop

filter

(base) ➜  Practica2 git:(main) ✗ 
```

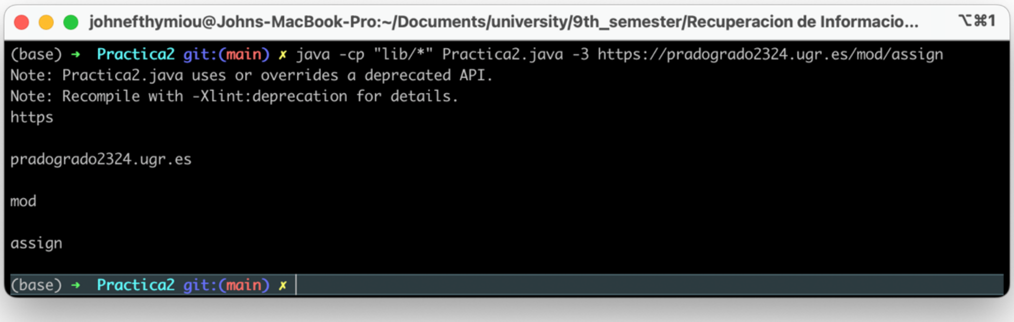
We pass the sentense, which will be analyzed. In this example we use the Stop Filter to remove words like “this” and “a”.


johnefthymiou@Johns-MacBook-Pro:~/Documents/university/9th_semester/Recuperacion de Informacion/PracticasRecuperacionInformacion/Practica2
⌥⌘1

```
(base) ➜  Practica2 git:(main) ✗ java -cp "lib/*" Practica2.java -2 "I am happy because I have my car. I am jouyful because I have my automobile"
Note: Practica2.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
Text: I am happy because I have my car. I am jouyful because I have my automobile
Analyzer:

i

am

happy

joyful

because

i

have

my

car

automobile

i

am

jouyful

because

i

have

my

automobile
(base) ➜  Practica2 git:(main) ✗ 
```

Another example using of a Synonym Filter, which for an array set of common words, create a same token for synonym words. For example here the sentence is "I am happy because I have my car. I am joyful because I have my automobile". And for the words "car" and "automobile" we have the same token "automobile".

3) Java -cp "lib/*" Practica2.java -3 https://pradogrado2324.ugr.es/mod/assign/


johnefthymiou@Johns-MacBook-Pro:~/Documents/university/9th_semester/Recuperacion de Informacio...
⌥⌘1

```
(base) ➜  Practica2 git:(main) ✗ java -cp "lib/*" Practica2.java -3 https://pradogrado2324.ugr.es/mod/assign
Note: Practica2.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
https

pradogrado2324.ugr.es

mod

assign

(base) ➜  Practica2 git:(main) ✗ 
```

We pass a URL, which will be analyzed with our custom Analyzer. We used a Tokenizer which splits the URL whenever finds "/". Also, we used a LowerCaseFilter which converts letters to lowercase. At the end we print the tokens.