Práctica 1. Tika

Santiago Álvarez Valdivia

Ioannis Efthymiou

6 de octubre de 2023

Recuperación de Información

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

Universidad de Granada

Program functionality

-d option -> Table with file name, type, coding and language

```
File Summary:
------------------------------------------------------------------------------------------------------------------------------------------
Filename                       Content-Type                                                               Encoding            Language
------------------------------------------------------------------------------------------------------------------------------------------
file_example_PNG_2500kB.jpg    image/jpeg                                                                 Unknown             Unknown
file-sample_1MB.doc            application/msword                                                         Unknown             ca
Grecia.html                    text/html                                                                  UTF-8               es
Pedro_Paramo-Rulfo_Juan.epub   application/epub+zip                                                       Unknown             es
WikipediaArticle.html          text/html                                                                  UTF-8               en
IvV2y.png                      image/png                                                                  Unknown             Unknown
file-example_PDF_1MB.pdf       application/pdf                                                            Unknown             ca
test.txt                       text/plain                                                                 ISO-8859-1          es
file_example_ODS_5000.ods      application/vnd.oasis.opendocument.spreadsheet                             Unknown             en
file_example_PPT_1MB.pptx      application/vnd.openxmlformats-officedocument.presentationml.presentation  Unknown             ca
random.txt                     text/plain                                                                 ISO-8859-1          en
CienAñosDeSoledad.pdf          application/pdf                                                            Unknown             es
```

-l option -> All links that extracted from each document

```
File: file_example_PNG_2500kB.jpg
--------------------------------------------------------------------------------------------------------------------------
No links found
--------------------------------------------------------------------------------------------------------------------------
File: file-sample_1MB.doc
--------------------------------------------------------------------------------------------------------------------------
https://products.office.com/en-us/word
embedded:image1.emf
embedded:image2.jpg
--------------------------------------------------------------------------------------------------------------------------
File: Grecia.html
--------------------------------------------------------------------------------------------------------------------------
/w/load.php?lang=es&modules=codex-search-styles%7Cext.cite.styles%7Cext.tmh.player.styles%7Cext.uls.interlanguage%7Cext.visualE
/w/load.php?lang=es&modules=ext.gadget.imagenesinfobox&only=styles&skin=vector-2022
/w/load.php?lang=es&modules=site.styles&only=styles&skin=vector-2022
/w/load.php?lang=es&modules=noscript&only=styles&skin=vector-2022
//upload.wikimedia.org
//es.m.wikipedia.org/wiki/Grecia
/w/index.php?title=Grecia&action=edit
/static/apple-touch/wikipedia.png
/static/favicon/wikipedia.ico
```

```
/static/images/footer/poweredby_mediawiki_88x31.png
https://www.mediawiki.org/
------------------------------------------------------------------------------------------------------------------------
File: IvV2y.png
------------------------------------------------------------------------------------------------------------------------
No links found
------------------------------------------------------------------------------------------------------------------------
File: file-example_PDF_1MB.pdf
------------------------------------------------------------------------------------------------------------------------
https://products.office.com/en-us/word
https://products.office.com/en-us/word
https://products.office.com/en-us/word
https://products.office.com/en-us/word
https://products.office.com/en-us/word
https://products.office.com/en-us/word
------------------------------------------------------------------------------------------------------------------------
File: file_example_ODS_5000.ods
------------------------------------------------------------------------------------------------------------------------
No links found
------------------------------------------------------------------------------------------------------------------------
File: file_example_PPT_1MB.pptx
------------------------------------------------------------------------------------------------------------------------
No links found
------------------------------------------------------------------------------------------------------------------------
File: CienAñosDeSoledad.pdf
------------------------------------------------------------------------------------------------------------------------
No links found
------------------------------------------------------------------------------------------------------------------------
```

Note: We used the AutoDetectParser to extract the links for each document, but this does not handle the plain text files, so we did not extract the links from .txt files.

-t option -> count frequency of words in documents, and store them in a csv file per document

Sample output:

```
Counted words for file WikipediaArticle.html. Output in word_count_WikipediaArticle.csv
Counted words for file file-sample_1MB.doc. Output in word_count_file-sample_1MB.csv
Counted words for file IvV2y.png. Output in word_count_IvV2y.csv
Counted words for file file_example_ODS_5000.ods. Output in word_count_file_example_ODS_5000.csv
Counted words for file test.txt. Output in word_count_test.csv
Counted words for file file_example_PNG_2500kB.jpg. Output in word_count_file_example_PNG_2500kB.csv
Counted words for file file-example_PDF_1MB.pdf. Output in word_count_file-example_PDF_1MB.csv
Counted words for file Pedro_Paramo-Rulfo_Juan.epub. Output in word_count_Pedro_Paramo-Rulfo_Juan.csv
Counted words for file Grecia.html. Output in word_count_Grecia.csv
Counted words for file file_example_PPT_1MB.pptx. Output in word_count_file_example_PPT_1MB.csv
Counted words for file CienAñosDeSoledad.pdf. Output in word_count_CienAñosDeSoledad.csv
```

Word art generated from one of the csv files:

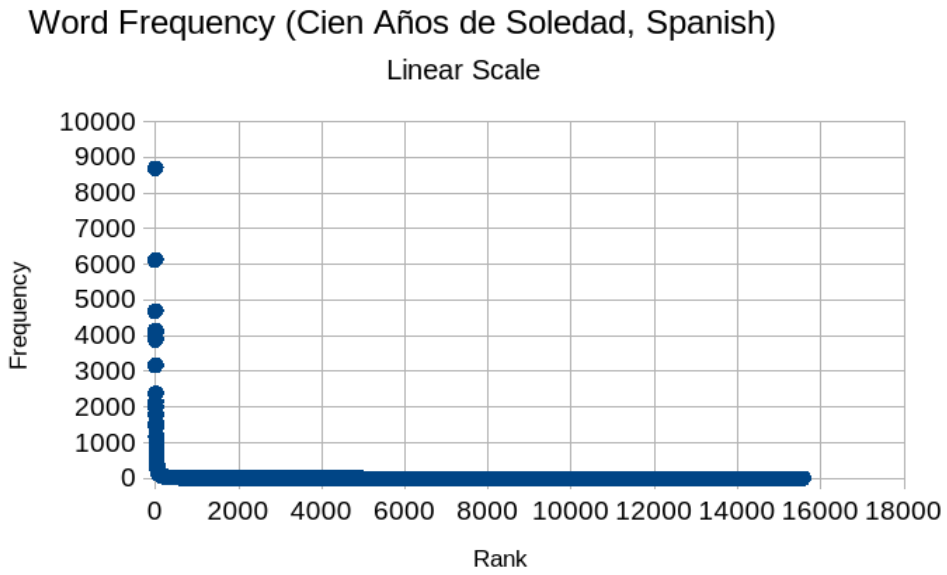
Word Frequency Analysis

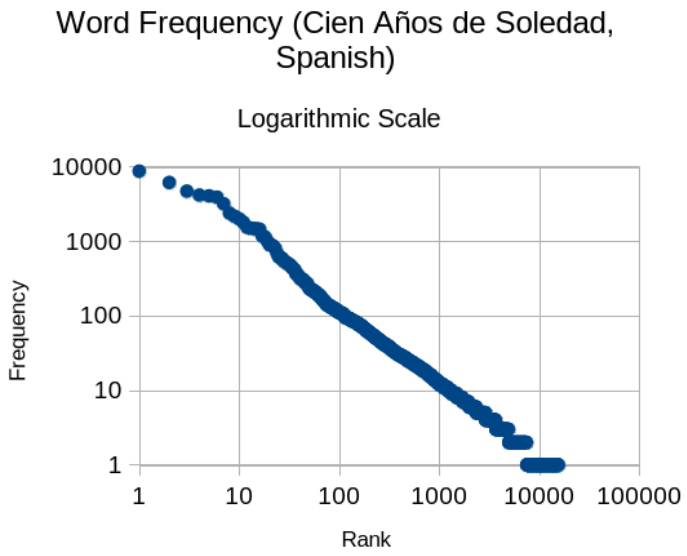
For verifying if a language follows the Zipf's Law, we measured the word frequency for three books, all of them in different languages:

- Cien Años de Soledad, by Gabriel García Márquez, in spanish
- The Collection of Hitchhiker's Guide to the Galaxy books, by Douglas Adams, in British English
- The Illiad, by Homer, in greek

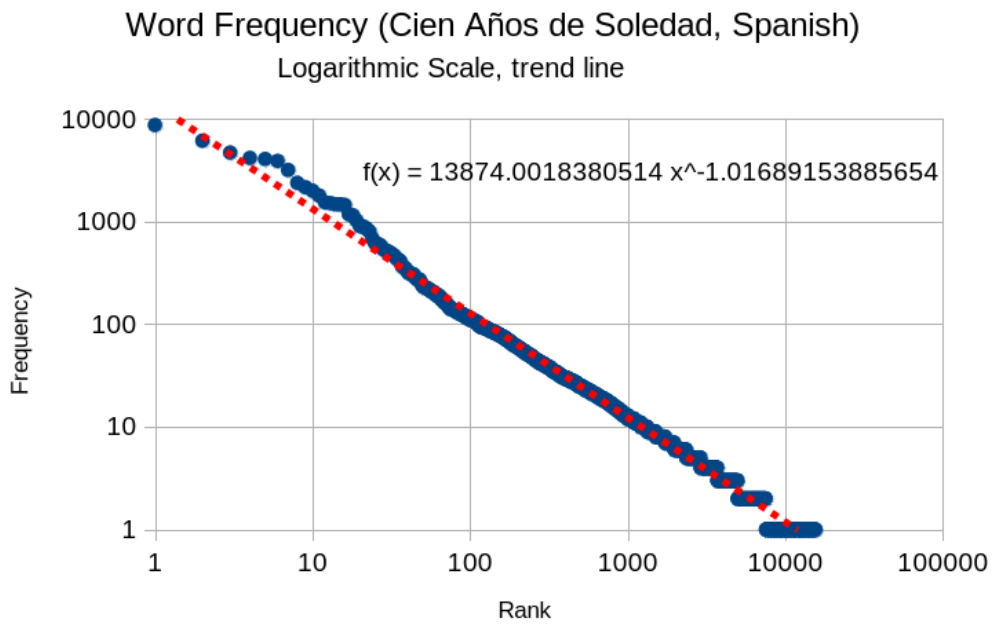
Spanish

Cien Años de Soledad has 137,865 words, which, when plotted for frequency, look like this.

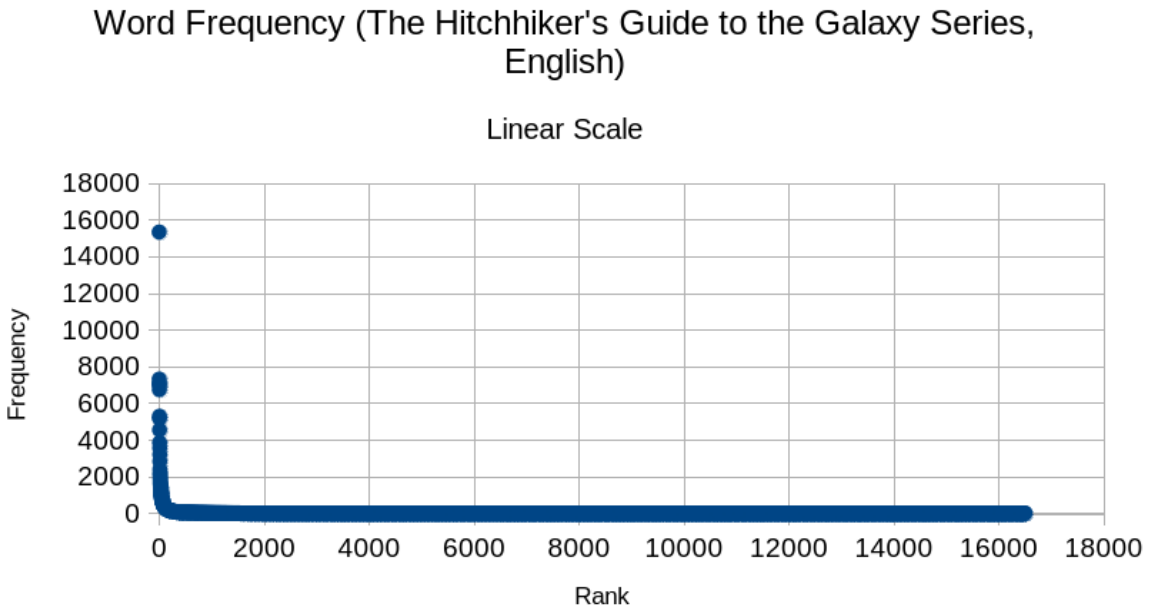


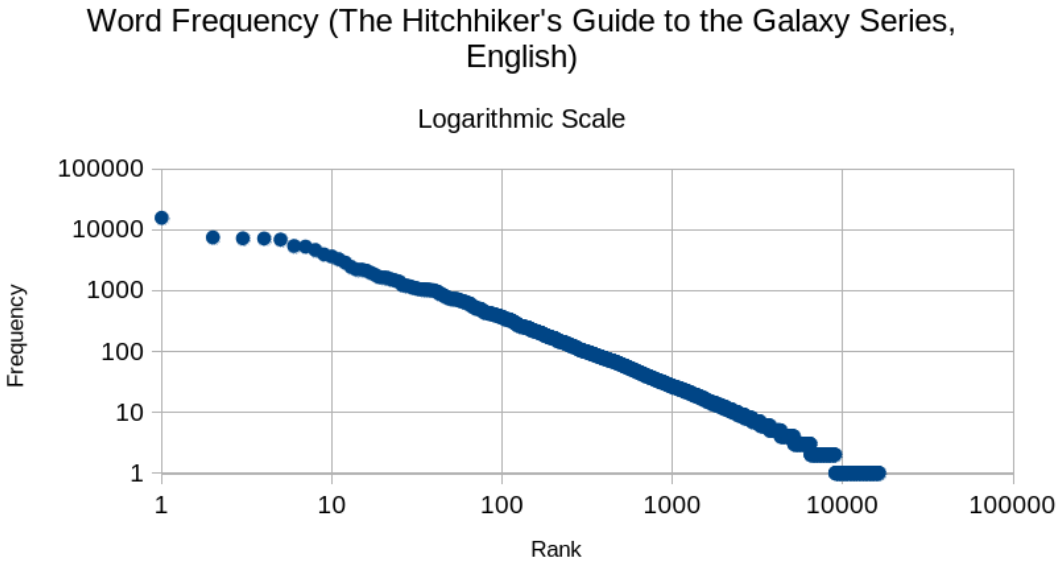


When generating a trend line with a spreadsheet tool, we get a rather close relationship with the actual data. Our language trend is $F = \frac{k}{R^m}$, **where k = 13874 and m = 1.01689**

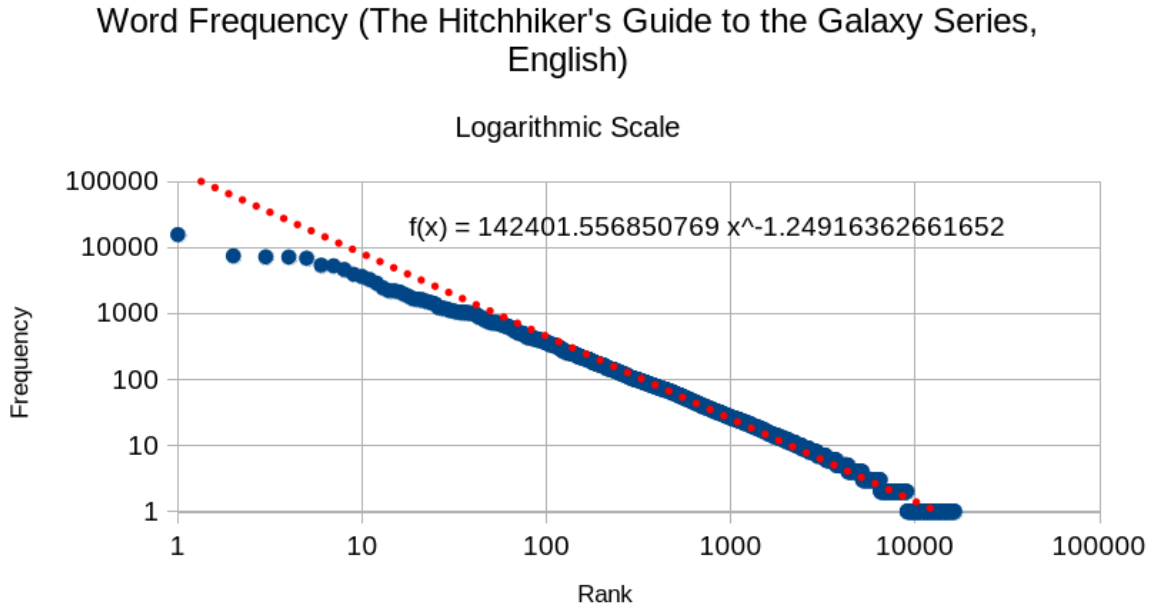


English

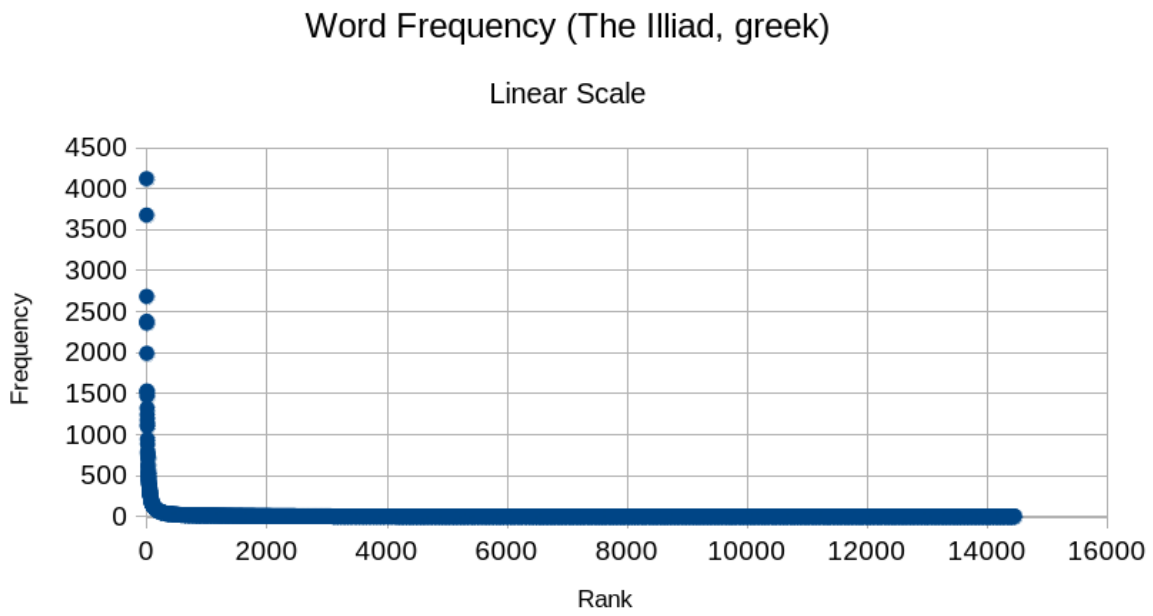


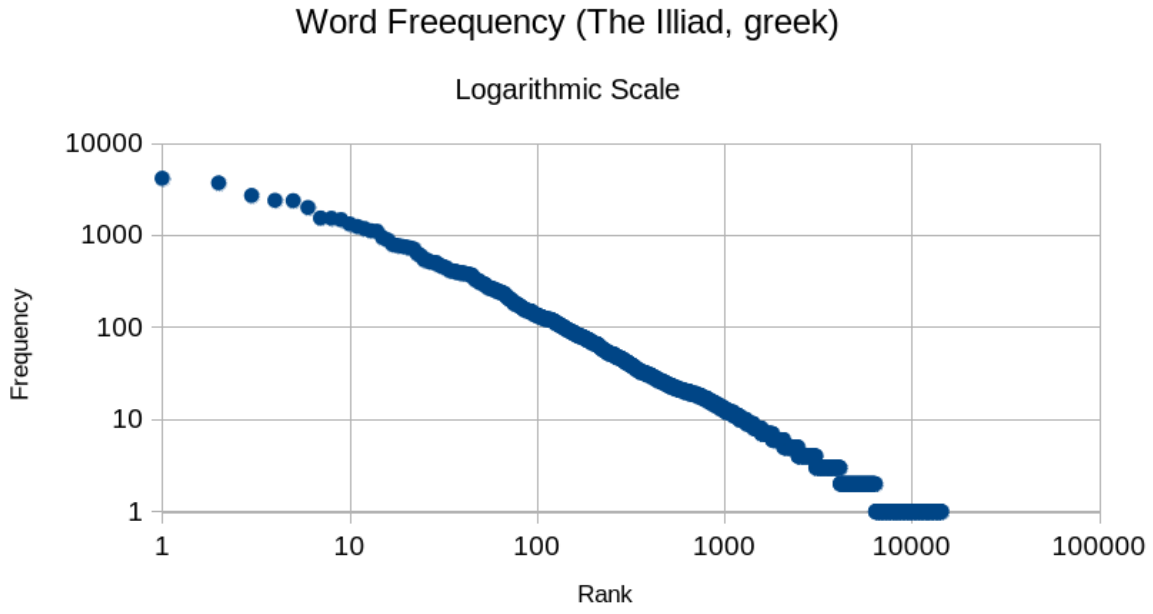


The logarithmic plot trends towards Zipf's Law, but not as closely as our spanish analysis. Our language trend is $F = \frac{k}{R^m}$, **where k = 142401 and m = 1.24916**

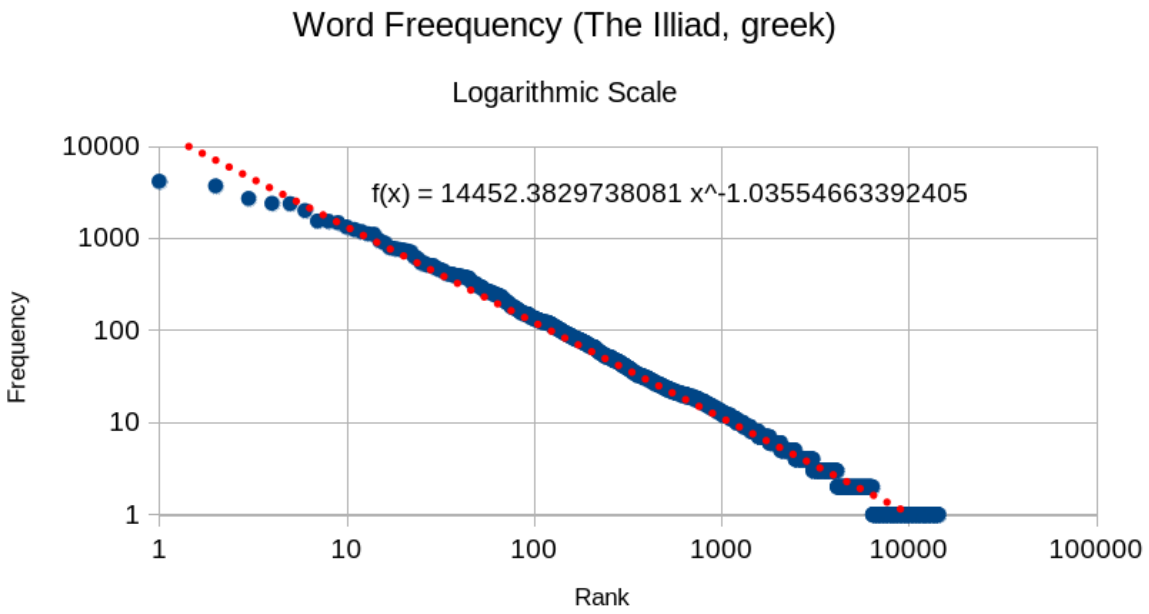


Greek





The logarithmic plot trends towards Zipf's Law rather closely. Our language trend is $F = \frac{k}{R^m}$, **where k = 14452 and m = 1.035546**


Word Freequency (The Illiad, greek)
Logarithmic Scale
f(x) = 14452.3829738081 x^-1.03554663392405
Frequency
Rank
10000
1000
100
10
1
1
10
100
1000
10000
100000