PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

Atribución de autoría mediante grafos

ANDRÉS MAURICIO VEGA YÁÑEZ
IVÁN ANTONIO ZAMORANO CATALDO

Profesor Guía: **Rodrigo Alfaro Arancibia**

INFORME FINAL DE PROYECTO
PARA OPTAR AL TÍTULO PROFESIONAL DE
INGENIERO CIVIL EN INFORMÁTICA

DICIEMBRE, 2016

Índice

Resumen iii
Lista de Figuras iv
Lista de Tablas v
1 Introducción 1
2 Objetivos 2
2.1 Objetivo General 2
2.2 Objetivos específicos 2
3 Problema 3
3.1 Definición 3
3.2 Avances 3
3.3 Aplicaciones 3
4 Marco teórico o estado del arte 4
4.1 Planteamiento inicial 4
4.1.1 Modelos Tradicionales de Representación de Textos 4
4.2 Representación de Texto Mediante Grafos 5
4.2.1 Grafo 5
4.2.2 Grafo Sintáctico 6
4.2.3 Grafos de POS Tagging 7
4.2.4 Grafos de Coocurrencia 8
4.2.5 Grafos Semánticos 8
4.2.6 Características Estilométricas 8
4.2.7 Atribución de autoría 9
4.2.8 The Small World of Human Language 10
4.3 Métodos de comparación entre grafos 10
4.3.1 Similitud de la frecuencia de las palabras entre grafos 10
4.3.2 Simrank 10
4.3.3 Simrank Adaptado 11
4.3.4 Similitud de Vectores Propios 12
4.3.5 Ponderación entre métodos 12
5 Experimentación 13
5.1 Set de Datos 13
5.1.1 Primera Parte 13

5.1.2 Segunda Parte 13
5.2 Implementación de los Grafos 14
5.2.1 Implementación Grafo de Coocurrencia 14
5.2.2 Comparación de Grafos 14
5.3 Resultados 15
5.3.1 Frecuencia de Palabras 15
5.3.2 Resultados Obtenidos del Grafo de Co-Ocurrencia 16
5.3.3 Crecimiento de los nodos y aristas de un grafo de Co Ocurrencia 17
5.3.4 Análisis de Resultado del Crecimiento del Grafo de Co Ocurrencia 29
5.3.5 Atribución de Autoría Grafo de Co Ocurrencia utilizando similitud de la frecuencia de las palabras entre grafos 29
5.3.6 Atribución de Autoría Grafo de Coocurrencia utilizando Simrank Adaptado 32
6 Trabajo futuro 33
7 Conclusiones 34
8 Referencias 35
Anexos 36
A: The Ghostly Village 36
B: Stopwords 37
C: Frecuencias graficadas en la Figura 5.2 38
D: Porcentaje de stopwords en un texto 41

Resumen

El objetivo de este proyecto es buscar distintas representaciones de textos mediante la utilización de grafos, con el fin de encontrar una que logre una alta tasa de atribución de autoría en textos que se encuentran en disputa o que su autor siempre fue anónimo.

Se muestran distintas representaciones de textos mediante grafos, y como se deben implementar estos y se apela a las características para escribir intrínsecas de cada persona.

Palabras Claves: *Grafo, Grafo de Co-Ocurrencia, Simrank, Frecuencia de palabras, Estilometría.*

Abstract

The goal of this project is to search different representation of text using graphs with the purpose of find one with the highest rate of authorship attribution in texts that are in dispute or that the author was anonymous.

It shows different representation of texts using graphs, how this are implemented and are appeal to the essential characteristics of how every person writes.

Keywords: *grap, co occurrence graph, Simrank, words frequency, stylometry.*

Lista de Figuras

Figura 4.1 Representación de Espacio Vectorial 5
Figura 4.2 Representación de un Grafo 5
Figura 4.3 Representación Grafo Sintáctico 6
Figura 4.4 Frecuencia 10 primeros millones de palabras en 30 Wikipedias.....9
Figura 5.1 Histograma de Frecuencia de Palabras set 1.....15
Figura 5.2 Histograma frecuencia palabras set 2 16
Figura 5.3 Grafo de Co-Ocurrencia del cuento “The Ghostly Village”.....16
Figura 5.4 Nodos v/s Número de Textos 17
Figura 5.5 Aristas v/s Número de Textos.....18
Figura 5.6 Nodos v/s Número Textos 18
Figura 5.7 Aristas v/s Número Textos 19
Figura 5.8 Nodos v/s Número de Textos 19
Figura 5.9 Aristas v/s Número de Textos.....20
Figura 5.10 Nodos v/s Número de Textos 20
Figura 5.11 Aristas v/s Número de Textos.....21
Figura 5.12 Nodos v/s Número de Textos 21
Figura 5.13 Aristas v/s Número de Textos.....22
Figura 5.14 Nodos v/s Número de Textos 22
Figura 5.15 Aristas v/s Número de Textos.....23
Figura 5.16 Nodos v/s Número de Textos 23
Figura 5.17 Aristas v/s Número de Textos.....24
Figura 5.18 Nodos v/s Número de Textos 24
Figura 5.19 Aristas v/s Número de Textos.....25
Figura 5.20 Nodos v/s Número de Textos 25
Figura 5.21 Aristas v/s Número de Textos.....26
Figura 5.22 Nodos v/s Número de Textos 26
Figura 5.23 Aristas v/s Número de Textos.....27
Figura 5.24 Nodos v/s Número de Textos 27
Figura 5.25 Aristas v/s Número de Textos.....28
Figura 5.26 Nodos v/s Número de Textos 28
Figura 5.27 Aristas v/s Número de Textos.....29

Lista de Tablas

Tabla 5.1 Número de textos por autor del set de datos 13
Tabla 5.2 Resumen de atribución mediante frecuencias de palabras. 30
Tabla 5.3 Resumen de atribución mediante frecuencias de palabras con stopwords.....30
Tabla 5.4 Resumen de atribución mediante frecuencias de palabras de stopwords.....31
Tabla 5.5 Resumen Atribución de Autoria 32

1 Introducción

Desde principios de la humanidad, el humano ha tenido la necesidad de comunicarse con otros, principalmente con el fin de preservar la especie, y para esto ha desarrollado diversas tecnologías que ha ido perfeccionando con el paso del tiempo. Una de éstas tecnologías y elemento comunicativo que mayor evolución ha tenido en el tiempo es la escritura. Se entiende por escritura al medio de comunicación que representa el lenguaje y las emociones a través de inscripción o grabado de signos y símbolos.

Se considera texto a cualquier que puede ser leído, ya sea que este objeto sea un trabajo de literatura, una señal en la calle, o un graffiti. Es un conjunto coherente de signos que transmiten algún mensaje. Éstos pueden ser realizados por un autor, o un conjunto de autores con el fin de trasmitir un mensaje.

La Estilometría es una de las primeras tecnologías que se utilizaron para el análisis de texto, y un uso recurrente es en el campo de la atribución de autoría. La atribución de autoría busca definir el autor de un texto cualquiera. Esto tiene diversos fines, entre ellos: atribución de mensajes amenazantes, mensajes de acoso, un texto literario, o documentos anónimos los cuales su autoría se encuentra en disputa.

Este documento se ha organizado en las siguientes secciones:

En la sección 3 se definen los objetivos que se desean conseguir con este trabajo.

En la sección 4 se muestra la definición de la problemática a tratar, la evolución de la técnica y las aplicaciones posibles que puede tener.

En la sección 5 se presenta el marco teórico necesario para entender el trabajo, las técnicas utilizadas y algunos resultados esperados.

En la sección 6 se muestra la experimentación con el marco teórico propuesto y los métodos realizados.

En la sección 7 se presentan propuestas de trabajos futuros.

En la sección 8 se presentan las conclusiones del trabajo.

Finalmente en la sección 9 se muestran todas las referencias que fueron necesarias para la elaboración del trabajo.

2 Objetivos

A continuación, se presenta el objetivo general y los objetivos específicos para el presente proyecto.

2.1 Objetivo General

El objetivo general del proyecto consiste en identificar representaciones de textos mediante grafos, para poder utilizar un algoritmo que logre identificar la autoría de un texto.

2.2 Objetivos específicos

- Definir el marco teórico de la atribución de autoría de textos.
- Buscar distintas formas de representación de texto
- Búsqueda de distintas métricas que representen características de los textos
- Búsqueda de softwares que permitan generar las representaciones de los textos
- Búsqueda de un método para poder comparar estas representaciones

3 Problema

El problema principal a evaluar en este proyecto es, la búsqueda de una representación de texto mediante grafos que se espera tenga un buen rendimiento en la atribución de autoría.

El problema principal a evaluar es la atribución de autoría, para lo cual, se buscarán diferentes formas de representar textos y así poder compararlos.

3.1 Definición

En el problema de la atribución de autoría, se busca determinar quién es el autor de un texto mediante el entendimiento, representación y comparación de patrones que cada autor tiene como característicos a la hora de escribir un texto.

Los patrones que se busca reconocer dentro del texto tienen relación con la forma en que el autor se expresa, ya sea de forma consciente o inconsciente. En ellos se encuentran rasgos como qué tan amplio es el vocabulario del autor, la frecuencia con que utiliza las palabras, el largo promedio de sus oraciones, cómo utiliza los signos de puntuación, y características que no son simples de ver, pero que se buscarán mediante técnicas de minería de textos.

Para efectuar una minería de textos adecuada y descubrir patrones y características del autor que ayuden a atribuir la autoría es importante tener una buena representación, por esto, el proyecto busca encontrar la mejor forma de representar éste.

3.2 Avances

Los primeros avances en la atribución de autoría se realizaron usando como único criterio la frecuencia con que aparecían un pequeño grupo de palabras dentro de los textos. Con el pasar de los años se llegó al concepto de estilometría, y los factores de medición para el texto incrementaron considerablemente, pero no fue así con los buenos resultados en la atribución de autoría.

Para aumentar la efectividad en la atribución de autoría, los estudios se encuentran en la búsqueda de adecuadas representaciones de texto y selección de características adecuadas para obtener mejores resultados.

3.3 Aplicaciones

Las aplicaciones de la atribución de autoría tienen un impacto grande en el universo académico, donde el plagio es recurrente. Además, una atribución de autoría con un alto grado de acierto resulta muy útil para identificar a las personas que se encuentran detrás de mensajes anónimos, ya sean de carácter inofensivo o no. En muchas ocasiones han sido publicadas novelas bajo seudónimos, o se publican columnas de forma anónima, pero también están los mensajes anónimos que resultan ser amenazas o mensajes terroristas. En estos casos resulta trascendental manejar los estilos y características inconscientes del autor, ya que son situaciones en que el escritor conscientemente tratará de ocultar su identidad.

4 Marco teórico o estado del arte

4.1 Planteamiento inicial

Para almacenar los documentos textuales en computadoras y permitir al usuario buscar a través de su contenido, uno de los primeros pasos es construir un modelo de representación de texto.

4.1.1 Modelos Tradicionales de Representación de Textos

El modelo de espacio vectorial es un modelo en que cada texto es representado como vector en término de pesos. Un conjunto de términos $T = \{t_1, t_2, \dots, t_n\}$ que tienen a lo menos una aparición en el documento, sirve como un conjunto de características. Cada documento d_i es representado por un vector $d_i = \{w_{i1}, w_{i2}, \dots, w_{in}\}$ donde w_{ij} es el peso en términos de t_j en el documento i. Entonces, un modelo de espacio vectorial es considerada una matriz M_{ij} donde los términos están en la columna y los documentos en las filas. En la matriz, cada entrada m_{ij} es el peso de los términos t_j en el documento i. Existen distintos enfoques en el modelo de espacio vectorial que es distinto en definición de términos y en el método de peso.

Podemos escoger distintas unidades de texto significativa, tales como términos, tales como patrones de palabras, pero las palabras son usadas ampliamente en enfoques tradicionales. Muchos procesos lingüísticos como corrección automática y normalización, stopwords y lematización, pueden ser aplicado en términos de selección. Especialmente, cuando una palabra sola es usada como término, el conjunto de palabras T puede ser de los N palabras más pesadas con mayor frecuencia de aparición y este modelo es llamado modelo de "Bolsa de Palabras".

Como método de peso, el tf-idf es un método popular, en el cual cada término TF (Term Frequency) representa la importancia de un término en un texto e IDF (Inverse Document Frequency) representa la discriminación de términos para todos los textos. Estas asignaciones de medidas, los mayores pesos de los términos ocurren frecuencia ocurren en un documento determinado pero no en la mayoría de estos. Otro método típico es utilizar un peso Booleano. En este enfoque, un documento es representado como un vector donde las dimensiones poseen un valor booleano, 0 indicando la ausencia y 1 indicando la presencia del diccionario de término correspondiente en el documento.

Por ejemplo, consideramos dos textos de "El perro pequeño ha sido golpeado por un gato grande. A Peter le gusta el perro" y "A María le gustan los animales", en modelo de espacio vectorial, después de quitar las stopwords, todos los lemas que se repiten a lo menos una vez en el texto y las puntuaciones booleneas como términos y métodos de peso, respectivamente. La lista de términos incluye "pequeño", "perro", "golpeado" (lema de golpear), "grande", "gato, "Peter", "como", "Mary", "animal". Nuestra matriz de modelado de espacio vectorial es una matriz M de 9 columnas representando doce términos y dos filas que describen dos textos. Si un término j encuentra una repetición en el Texto i, entonces $M_{ij} = 1$ sino $M_{ij} = 0$. Texto 1 y Texto 2 son representados por [0 1 1 1 1 1 0 1 1] y [1 0 0 0 0 1 1 0 0], respectivamente en el modelo de espacio vectorial.

	animal	*golpear*	*grande*	*gato*	*perro*	*gustar*	*Mary*	*Peter*	*pequeño*
Text 1	0	1	1	1	1	1	0	1	1
Text 2	1	0	0	0	0	1	1	0	0

Figura 4.1 Representación de Espacio Vectorial

4.2 Representación de Texto Mediante Grafos

Los documentos textuales pueden ser representados como grafos de distintas formas. Los nodos denotan características y los arcos representan la relación entre los nodos.

Hay una variedad de información que puede ser utilizada para representar texto como grafo, como sus características morfológicas, sintácticas y semánticas. Algunos tipos básicos incluyen forma de la palabra, lemas, éstos han sido utilizados de forma recurrente en representación de grafos. Mientras que, el orden de las palabras, la ubicación de las palabras, o la estructura sintáctica son consideradas información estructural. En términos de semántica, varios tipos simple de información como sinónimos, hiperónimos son tomados en cuenta. Sin embargo, es difícil capturar el significado semántico más profundo de un texto.

Todo ese tipo de información puede ser combinado de distintas formas con el fin de crear distintas representaciones de textos con grafos.

4.2.1 Grafo

Será la partícula elemental de la representación de texto en este trabajo. Un grafo es una representación de un conjunto de objetos donde algunos pares están conectados. La conexión entre grafos es conocida como vértice (también conocida como nodo, o puntos), y las conexiones entre un par de vértices es conocida como arcos. Éstos pueden ser dirigidos (notar dirección) o no serlo.

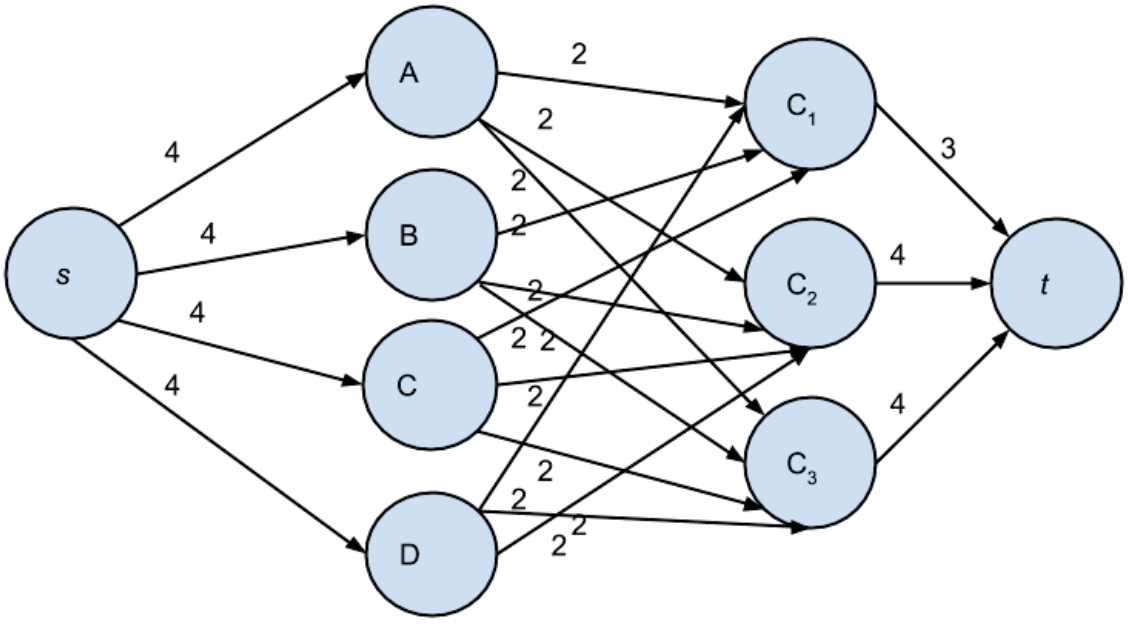

Figura 4.2 Representación de un Grafo

4.2.2 Grafo Sintáctico

El uso de la secuencia en que los términos aparecen en una oración es una forma efectiva de representar la relación e importancia de un término sobre otro en textos donde existe un estricto orden sintáctico (generalmente literatura, ensayos, y textos nuevos).

Formalmente un grafo de secuencia sintáctica es representado por:

- Ges un grafo dirigido
- $V = \{v_{i|i} = 1, \dots, n\}$ es un conjunto finito de vértices que consisten de palabras contenidas dentro del texto
- $E \subseteq V \times V$ es el conjunto finito de arcos que representa que dos vértices están conectados por el significado de su secuencia en un texto
- L es el tag de un arco el cual consiste del número de veces que dos vértices aparecen juntos en un texto
- $\alpha: E \rightarrow L$ es una función que asigna un tag a un par asociado de vértices

Como ejemplo, considera la siguiente oración ξ extraída de un texto T de una base de datos en inglés:

"*The violence on the TV. The article discussed the idea of the amount of violence on the news.*"

Que después de la etapa de pre procesado (detallada más adelante) quedaría de la siguiente forma:

"*the violence on the tv the article discussed the idea of the amount of violence on the news*"

Basado en la representación propuesta, el pre-procesado de la oración ξ puede ser mapeado a una representación de grafo sintáctico como la siguiente:

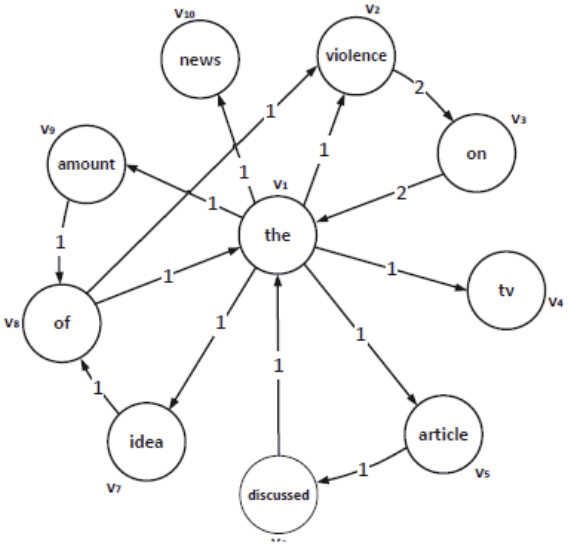

Figura 4.3 Representación Grafo Sintáctico

El grafo mostrado en la figura posee las siguientes características:

- El conjunto de vértices consiste de palabras preprocesadas en una oración ξ considerando que si hay múltiples ocurrencias de una palabra sólo se crea un vértice.
- Un arco entre dos vértices representa que esas palabras aparecen juntas (una seguida de la otra) en una oración, al menos una vez
- La dirección del arco representa el orden en que aparecen en la oración
- El tag puesto al arco representa el número de veces que esa palabra aparecieron juntas en la oración ξ

4.2.3 Grafos de POS Tagging

4.2.3.1 POS Tagging (Etiquetado Gramatical)

Consiste en asignar a cada palabra su significado gramatical en la oración, esto quiere decir, que a cada palabra se le asigna que tipo está siendo dentro de la oración, ya sea sustantivo, verbo, adjetivo, o la categoría léxica con la que se desee etiquetar.

4.2.3.2 Grafos de Coocurrencia basados en POS Tagging

El propósito del POS Tagging es asignar la categoría léxica correcta (Ej: sustantivo, verbo, o artículo) a cada palabra del texto. La principal dificultad con POS Tagging es que la asignación de tipo a una palabra es casi siempre una tarea ambigua ya que la categoría léxica de una palabra usualmente depende en el contexto en que es usada. Para manejar la ambigüedad, POS Taggers usualmente consideran secuencias de n palabras y así derivar el contexto en el que la palabra es utilizada. Éste enfoque evita el uso de conocimiento externo para producir un grafo con etiquetado.

La relación gramatical es usada para encontrar la relevancia de las etiquetas usando un modelo de grafos. El texto primero es etiquetado utilizando información de *part-of-speech* produciendo sustantivo, verbo, adjetivos o vértices de adverbios. El grafo generado toma lugar un pedazo de texto como input y genera un grafo como output. La estructura del grafo es definida utilizando un vector de relevancia el cual es creado usando coincidencias exactas de métricas (e), coincidencias de sub cadenas (s), cadenas distintas (d), o no coincidencias, sinónimos (y), hipónimo (h), y dominio de palabras raras (r)

$$dorg(e, s, d, y, h, r)$$

Una estructura lingüística del párrafo de un texto de un documento es basada en *parse tree* para cada oración de un párrafo. Un *Parse Thicket* es un grafo que contiene *parse tree* por cada oración, como también arcos adicionales por cada relación entre oración entre *parse tree nodes* para palabras como correferencias, relaciones taxonómicas como sub entidades, casos parciales, y predicados de temas, relaciones de estructuras retóricas y actos de discurso. Generalización basada en PT tiene un enfoque cercano al rendimiento humano en términos de buscar similitudes entre textos.

4.2.4 Grafos de Coocurrencia

El grafo de coocurrencia es la técnica en que las palabras son los vértices del grafo y las coocurrencias de pares de palabras se almacenan con un enlace entre ellas.

Se dice que dos palabras co-ocurrencia si entre ellas se encuentra un número determinado de palabras, este número de palabras llamado ventana, donde una ventana de cero palabras serían las palabras que se encuentran una seguida de otra.

El grafo de co-ocurrencia es importante de analizar ya que se ha descubierto la presencia de una distribución invariante en la conectividad de las palabras.

4.2.5 Grafos Semánticos

4.2.5.1 Red Semántica

Una red semántica está basada en la idea de que los objetos o los conceptos pueden ser unidos por alguna relación, ésta relación puede ser representada utilizando una unión que conecte los dos conceptos. Bajo este concepto, una red semántica puede ser representada como un grafo.

4.2.5.2 Grafos Semánticos

El grafo semántico tiene por objetivo construir una representación para cada oración del documento, de manera que capture su estructura semántica y las relaciones de sus términos. Para ello, los conceptos descubiertos para cada oración en la etapa anterior se expanden con los conceptos de niveles superiores en la jerarquía de la base de conocimiento. Por lo tanto, será requisito indispensable para el funcionamiento del algoritmo disponer de una base de conocimiento que contemple las relaciones semánticas, ya sean de hiponimia, hiperonimia, o la meronimia, entre sus conceptos.

En este grafo, cada vértice representa un concepto distinto, mientras las aristas, temporalmente sin etiquetar, representan relaciones semánticas entre dichos conceptos. Finalmente, bajo la hipótesis de que los conceptos que se encuentran en los niveles superior de la jerarquía representan información muy genérica, similar a un mapa conceptual.

4.2.6 Características Estilométricas

Estilometría -el análisis estadístico del estilo literario- complementa literatura tradicional y ofrece capturar los caracteres elusivos del estilo de un autor y cuantificar algunas de sus características. La mayor parte de los estudios estilísticos emplean objetos del lenguaje y la mayoría de estos objetos son basados en el léxico.

La principal suposición bajo los estudios estilométricos es que los autores tienen tanto de forma inconsciente como consciente, aspectos en su estilo. Cada estilo de un autor se cree que tiene ciertas características que son independientes de la voluntad del autor, y que esas características no se pueden manipular de forma consciente por el autor, se considera que son

las más probables de entregar datos fidedignos para un estudio estilométrico. Las dos aplicaciones principales son estudios atribuídos y problemas cronológicos, eso sí, una diferencia en la fecha o autor no es la única posible explicación para peculiaridades estilísticas. Variaciones en estilo pueden ser causadas por diferencias de género o contenido, y similaridad por procesos literarios como la imitación.

4.2.7 Atribución de autoría

En cada problema de atribución de autoría, hay un conjunto de autores candidatos, un conjunto de textos de muestra de un autor conocido que cubre todos los autores candidatos (training corpus), y un conjunto de muestras de textos de autor desconocido (test corpus), cada uno de ellos podría ser atribuido a un autor candidato.

Los avances principales en la atribución de autoría y en la búsqueda de clasificadores de textos que den resultados positivos, son atribuidos a Thomas Corwin Mendenhall (1841-1924), el cual fue un físico autodidacta y un meteorólogo que publicó en el año 1887 los primeros pasos en la estilometría. Influenciado por el matemático inglés Augustus de Morgan, Marshall intentó determinar la autoría de textos basado en la distribución de frecuencias de varias palabras. El siguiente gran estudio de estilometría se atribuye a George Kingsley Zipf (1902–1950) un lingüista americano, filósofo, que estudió la estadística de las repeticiones de palabras en diversos idiomas. George Kingsley Zipf es por quien se nombra la ley de Zipf.

Ésta ley dice, en simples palabras, la frecuencia con la que una palabra aparece en un texto. Zipf afirma que un pequeño número de palabras son utilizadas con mucha frecuencia, mientras que frecuentemente ocurre que un gran número de palabras son poco empleadas.

A continuación, una imagen que muestra la frecuencia de los primeros 10 millones de palabras en 30 wikipedias distintas:

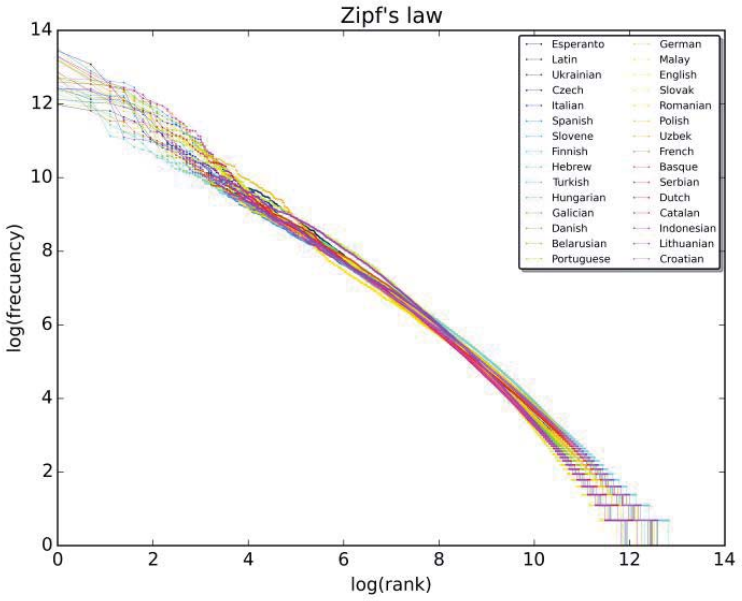

Figura 4.4 Frecuencia 10 primeros millones de palabras en 30 Wikipedias

4.2.8 The Small World of Human Language

En el trabajo de Ferrer i Cancho y Ricard V. Solé se descubre que los grafos formados por los enlaces entre palabras provenientes de frases tienen comportamientos de redes complejas, donde la escala de las palabras, según su conectividad, es invariante, es decir que estos grafos de palabras cumplen propiedades estadísticas.

4.3 Métodos de comparación entre grafos

4.3.1 Similitud de la frecuencia de las palabras entre grafos

Sea $G1$ y $G2$ dos grafos, sean n_1 y n_2 nodos de cada grafo respectivamente. Además decimos que cada nodo n tiene un peso asociado w.

Para cada nodo de G1, se busca uno con el mismo valor en G2. Si este existe se aplica la siguiente diferencia:

$$diferencia = \frac{w_1 - w_2}{w_1 + w_2}$$

Y se suma uno a un contador de los nodos revisados.

Si el nodo de G1 no se encuentra en G2, no se aplica la diferencia y se suma uno al contador de nodos revisados.

Finalmente, la medida de similitud queda dada por:

$$similitud = 1 - \frac{diferencia}{n^{o}\ de\ nodos}$$

La diferencia se divide por la cantidad de nodos con el fin de acotar la medida de similitud en un conjunto entre $[0,1]$ y así hacerlo independiente del tamaño del grafo.

4.3.2 Simrank

SimRank es una medida general de similitud, basa en el simple e intuitivo modelo grafo-tórico. SimRank es aplicable a un dominio con relación objeto a objeto, que mide la similitud del contexto estructural en el que el objeto ocurre, basado en la relación con otros objetos. Efectivamente, SimRank es una medida que dice “dos objetos son considerados similares si son referenciados por objetos similares”

Para un nodo v en un grafo, denotamos por $I(v)$ al conjunto de in-vecinos de v, estos son los nodos que tienen una unión direccionada a v. Los vecinos individuales se denotan como $I_i(v)$, para $1 \leq i \leq |I(v)|$.

Denotamos la similitud entre objetos a y b por $s(a,b) \epsilon\ [0,\infty[$. Dado esto definimos como $s(a,b)$ de la siguiente manera.

$$s(a,b) = \frac{C}{|I(a)| * |I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b))$$

Donde C es una constante entre 0 y 1. Una tecnicidad aquí es que, a o b no pueden no tener ningún in-vecino. Puesto que no se tendría de manera de inferir la similitud entre a y b, para este caso ocurre que $s(a,b) = 0$, por lo que definimos la suma en la ecuación pasa ser 0 cuando $I(a) = 0$ o $I(b) = 0$.

4.3.3 Simrank Adaptado

Sea δ el conjunto de grafos de los autores procesados, sea $G \in \delta$ un grafo de palabras para un autor procesados, sea H como el grafo de palabras al cual se quiere establecer su posible autor.

Para la utilización del SimRank se define como, a un nodo de G, $I_g(a,G)$ como los in-vecinos de a en G, los vecinos individuales se denotan como $Ig_i(a,G)$, para $1 \leq i \leq |Ig(a,G)|$ y se define R1 como una lista de nodos recorridos en G que haya tomado el valor a y R2 como una lista de nodos recorridos en H que haya tomado el valor de b.

Denotemos la similitud entre los nodos $a \in G$ y $b \in H$ por $sa(a,G,b,H) \in [0,1]$. Dado esto definimos como $sa(a,G,b,H)$ de la siguiente manera.

$$sa(a,G,R1,b,H,R2) = \frac{C}{|Ig(a,G)| * |Ig(b,H)|} \sum_{i=1}^{|Ig(a,G)|} \sum_{j=1}^{|Ig(b,H)|} sa(Ig_i(a,G), Ig_j(b,H))$$

Los listados R1 y R2 son utilizados para eliminar los ciclos en la búsqueda del SimRank, impidiendo que el algoritmo termine en un bucle infinito.

Dada esta definición, se establece como el algoritmo de comparación, al promedio del SimRank adaptado para las palabras de H presentes en G:

$$C(G,H) = \frac{\sum^{|G \cap H|} sa(a,G,\{\},b,H,\{\})}{|G \cap H|} \; donde \; a \in G, b \in H \; y \; a = b$$

El valor de $C(G,H) \in [0,\infty[$ donde 0 representa que no es igual.

Los campos {} representan una lista vacía.

4.3.4 Similitud de Vectores Propios

Sea $A1$ y $A2$ matrices adyacentes de los grafos $G1$ y $G2$ respectivamente. Sea $L_1 = D_1 - A_1$ y $L_2 = D_2 - A_2$ los operadores laplaceanos correspondientes a los grafos, donde D_1 y D_2 son las matrices diagonales de grados. Con este método, definimos la similitud de vectores propios de los operadores laplacianos y definimos la similitud entre los grafos como:

$$sim = \sum_{i=1}^{k} (\lambda_{1i} - \lambda_{2i})^{\,2},$$

$$donde\ k\ se\ escoge\ como\ min_j \left\{ \frac{\sum_{i=1}^{k} \lambda_{ji}}{\sum_{i=1}^{n} \lambda_{ji}} > 0.9 \right\}$$

Para $j = 1, 2$ (corresponde a los dos grafos que se están comparando). Esto es, mantenemos el mayor k de los vectores propios que contiene el 90% de la energía.

La valor de similitud que entrega este método es un valor entre $[0, \infty[$, los valores más cercanos a 0 indican que los grafos son más similares entre si, mientras que un valor más alto índica que los grafos son distintos.

4.3.5 Ponderación entre métodos

Sea $A1$ y $A2$ el resultado de similitud entre grafos de los métodos $M1$ y $M2$ respectivamente. Se define la ponderación entre métodos como:

$$pon = (A1 * k) + (A2 * l), con\ k \in [0,1]\ y\ l = 1 - k$$

Se define k y l como los porcentajes de ponderación de cada método.

5 Experimentación

5.1 Set de Datos

5.1.1 Primera Parte

En la primera parte de la experimentación, el criterio principal a la hora de escoger los textos a tratar fue su longitud.

Los textos utilizados fueron los siguientes:

- The Ghostly Village, Autor Desconocido
- The Adventure of the Three Students, Autor Desconocido
- Freud Young, Autor Desconocido
- Discourse On The Method Of Rightly Conducting The Reason And Seeking Truth In The Sciences, Rene Descartes
- Charlie and the Chocolate Factory, Roald Dahl

5.1.2 Segunda Parte

En la segunda parte, el criterio para escoger los datos fue su facilidad de comparación, por lo que los textos a tratar fueron escogidos bajo el criterio de que hayan sido utilizado en estudios similares.

El dataset utilizado se compone de una colección de textos de opinión llamada Guardian Corpues, la cual cuenta con un total de 561 textos con 13 autores distintos.

Tabla 5.1 Número de textos por autor del set de datos

Autor	Cantidad de Textos
Catherine Bennett	27
George Monbiot	35
Hugo Young	36
Jonathan Freedland	84
Martin Kettle	31
Mary Riddell	45
Nick Cohen	32
Peter Preston	61
Polly Toynbee	47
Roy Hattersley	30
Simon Hoggart	77
Will Hutton	31
Zoe Williams	25

5.2 Implementación de los Grafos

5.2.1 Implementación Grafo de Coocurrencia

5.2.1.1 Preprocesado

- Todas las letras que se encuentran en el texto son cambiadas a minúsculas, con el fin de que el software no las considere distintas.
- Se elimina todo carácter que pueda generar un error en la codificación y que no tenga que ver con la palabra.

5.2.1.2 Pruebas

Se probará el algoritmo con las siguientes configuraciones.

- Grafo de Co Ocurrencia con stopwords
- Grafo de Co Ocurrencia sin stopwords
- Grafo de Co Ocurrrencia sólo de stopwords

5.2.1.3 Ejecución del algoritmo

Finalmente se ejecuta el algoritmo y se genera un grafo utilizando la librería *networkx* de Python.

Además, para cada uno de los autores se generó un archivo csv el cual contiene la cantidad de nodos y vértices acumulados a medida que se iban agregando los textos que pertenecían a al autor.

5.2.2 Comparación de Grafos

Para obtener que tan buena es la representación, utilizaremos los distintos algoritmos mencionados que nos entregarán un valor representando la certeza de que un texto pertenezca a un autor dado.

5.3 Resultados

5.3.1 Frecuencia de Palabras

Al generar un histograma de frecuencia de las 100 primeras palabras dentro de los textos, “Charlie and the Chocolate Factory”, “The Adventure of the Three Students”, “Freud Young”, “Discourse On The Method Of Rightly Conducting The Reason And Seeking Truth In The Sciences”, y “The Ghostly Village”, obtenemos los siguientes resultados

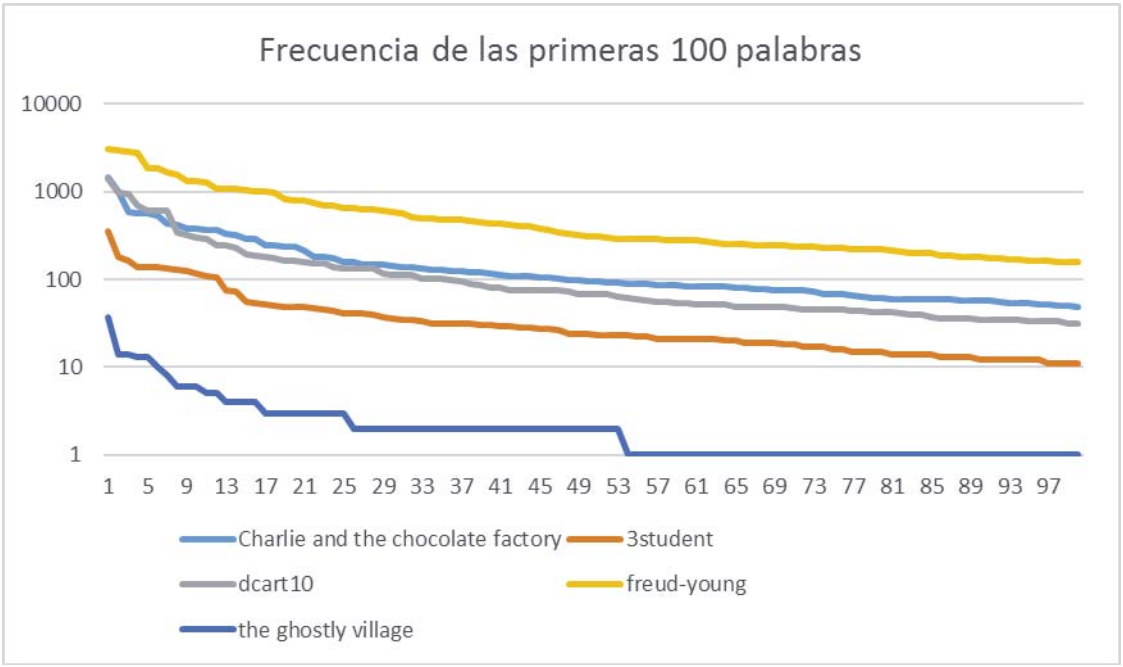

Figura 5.1 Histograma de Frecuencia de Palabras set 1

Como era de esperar, el grafo tiene el comportamiento similar a la ley de Zipf, mostrando que es sólo un cumulo de palabras son las más utilizadas independiente del vocabulario del texto.

Para evaluar la distribución del segundo set de datos se evaluó para cada autor la frecuencia de sus palabras, siendo graficada la frecuencia de las 100 palabras más utilizadas por cada uno de los autores.

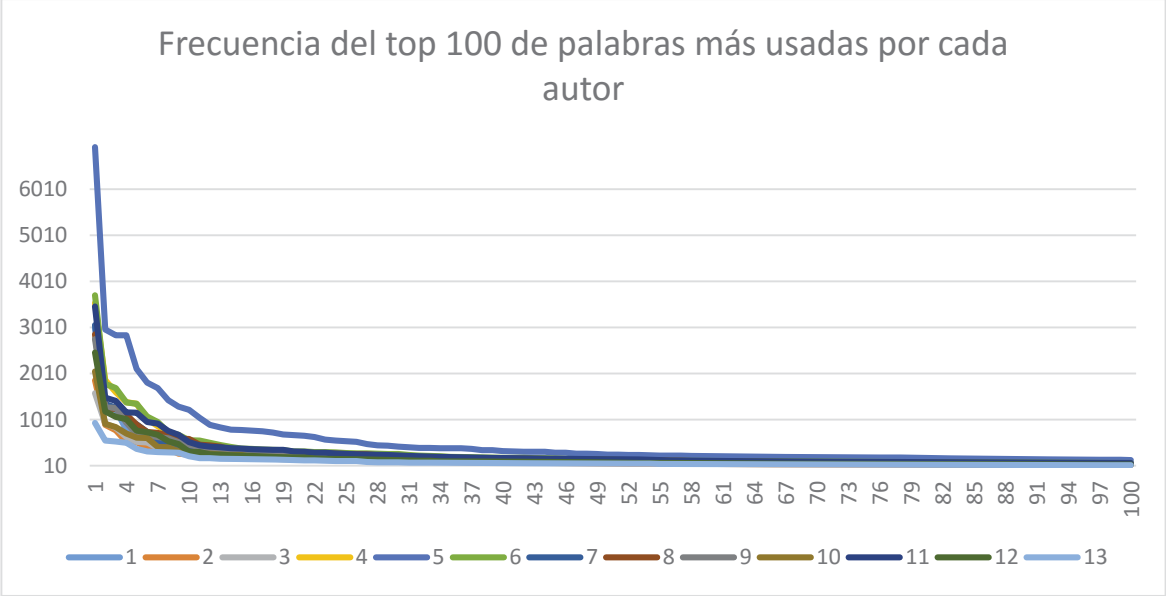

Figura 5.2 Histograma frecuencia palabras set 2

En la Figura 5.2 se puede ver el comportamiento de la frecuencia de las 100 palabras más utilizadas de cada autor. En el eje horizontal se presenta la posición de la palabra, y en el vertical refleja la frecuencia con que el autor utiliza la palabra. El detalle las frecuencias graficadas se encuentra en el anexo.

5.3.2 Resultados Obtenidos del Grafo de Co-Ocurrencia

A continuación, se muestra el grafo generado a partir de la representación de grafo de co-ocurrencia es la siguiente

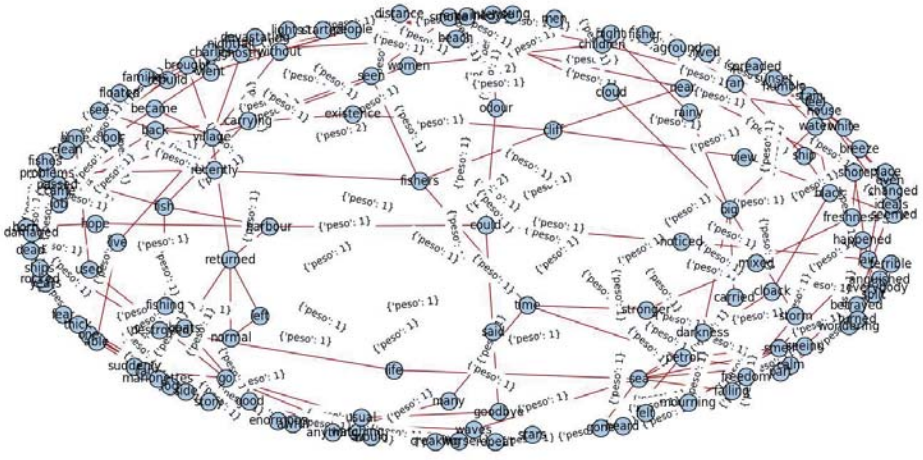

Figura 5.3 Grafo de Co-Ocurrencia del cuento “The Ghostly Village”

Como se puede observar, los arcos del grafo representan el peso el cual aumenta en uno cada vez que existe una co-ocurrencia de las palabras. Los vértices denotan las palabras del texto.

5.3.3 Crecimiento de los nodos y aristas de un grafo de Co Ocurrencia

A continuación, se muestra como el grafo de Co Ocurrencia va generando nodos y aristas a medida que se le introducen más textos al grafo.

5.3.3.1 Grafo de Co-Ocurrencia Con Stopwords

A continuación, se muestra de forma gráfica la cantidad de nodos y aristas que se van generando de algunos autores, representando los textos como un grafo de co-ocurrencia con stopwords.

5.3.3.2 Catherine Bennett

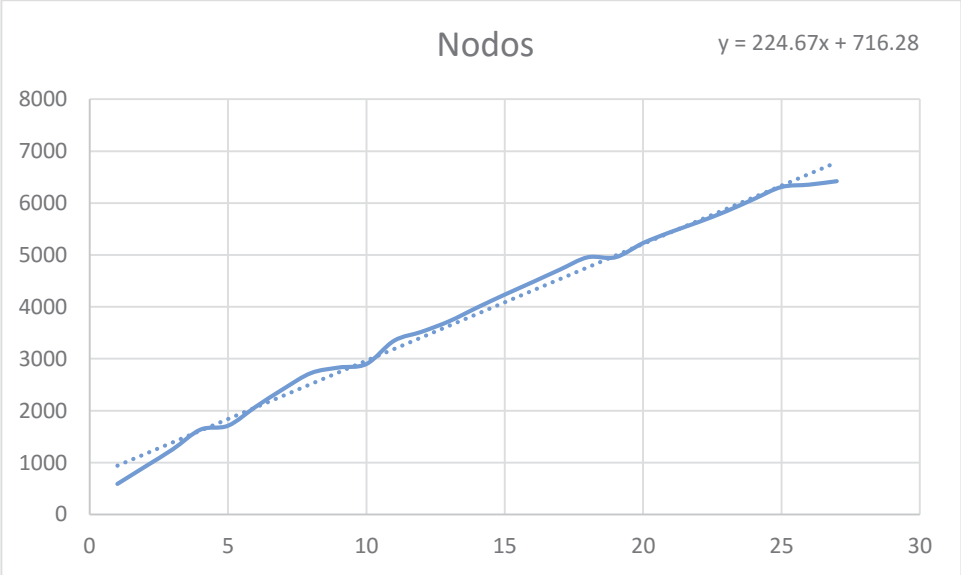

Figura 5.4 Nodos v/s Número de Textos

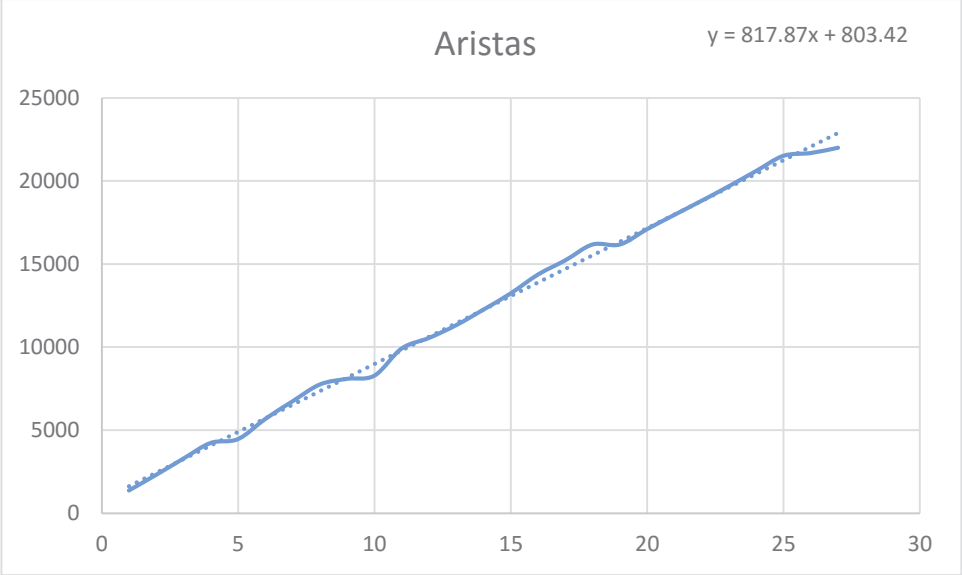

Figura 5.5 Aristas v/s Número de Textos

5.3.3.3 George Monbiot

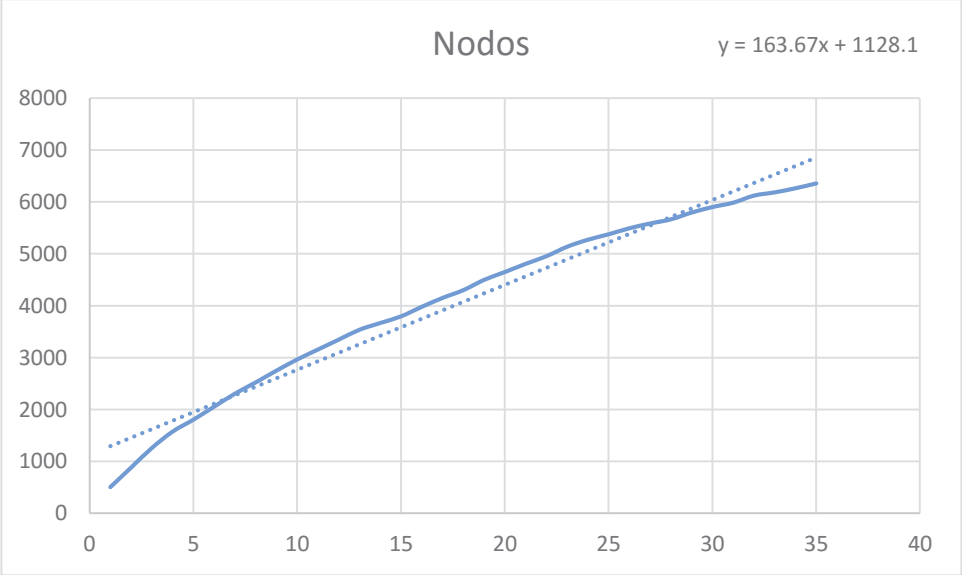

Figura 5.6 Nodos v/s Número Textos

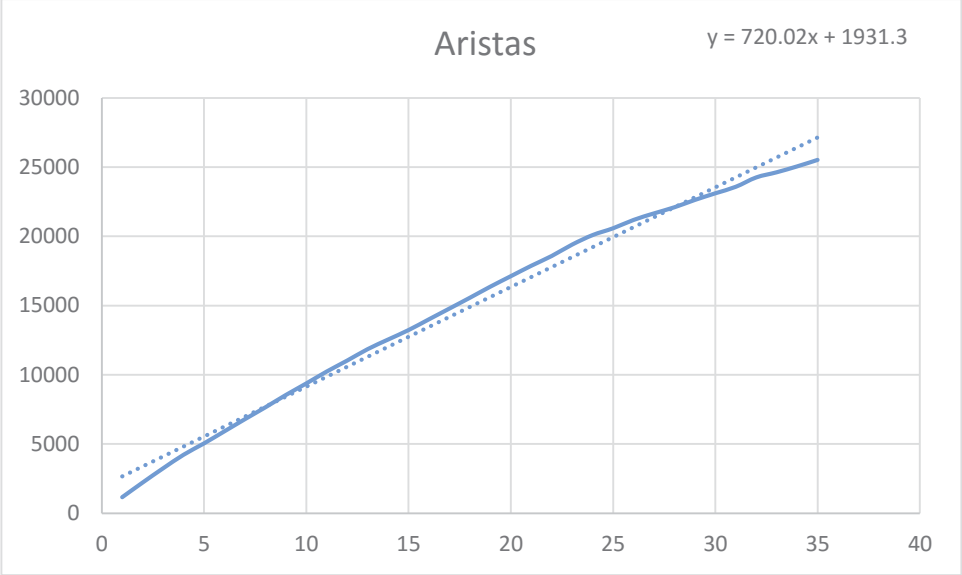

Figura 5.7 Aristas v/s Número Textos

5.3.3.4 Hugo Young

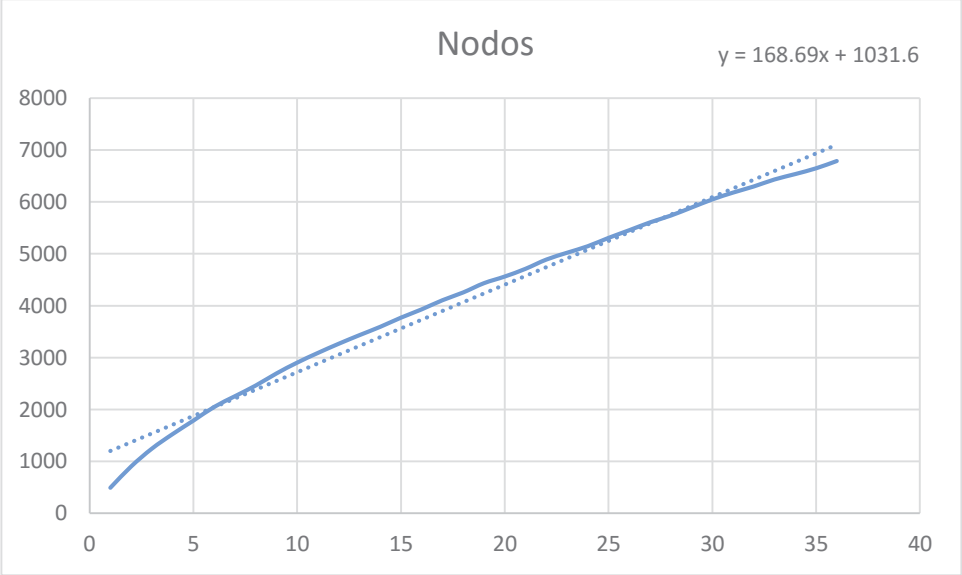

Figura 5.8 Nodos v/s Número de Textos

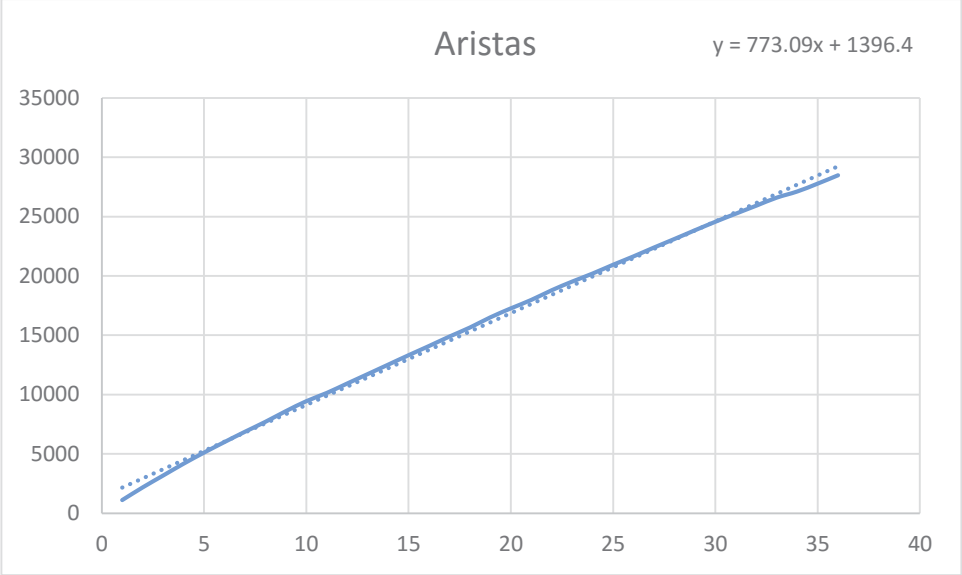

Figura 5.9 Aristas v/s Número de Textos

5.3.3.5 Jonathan Freedland

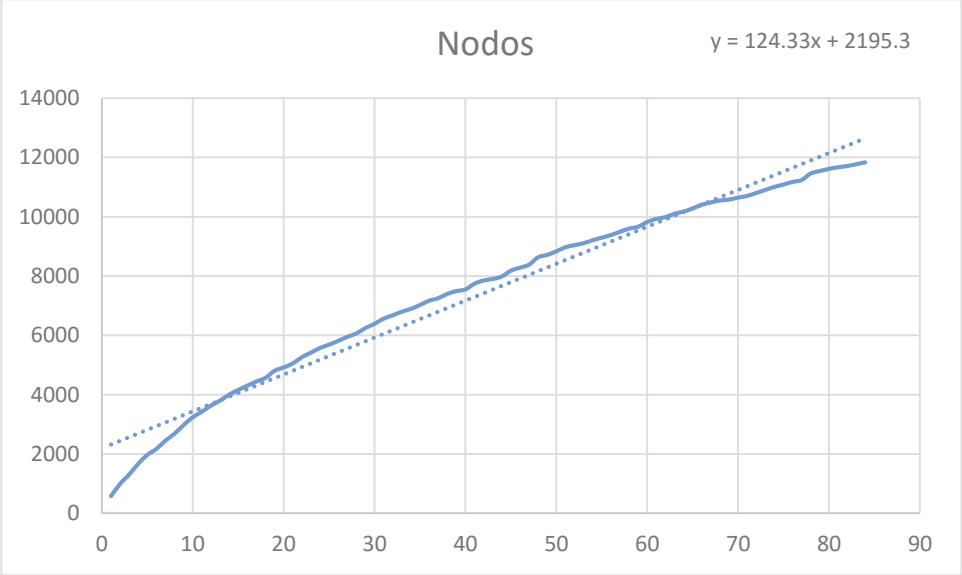

Figura 5.10 Nodos v/s Número de Textos

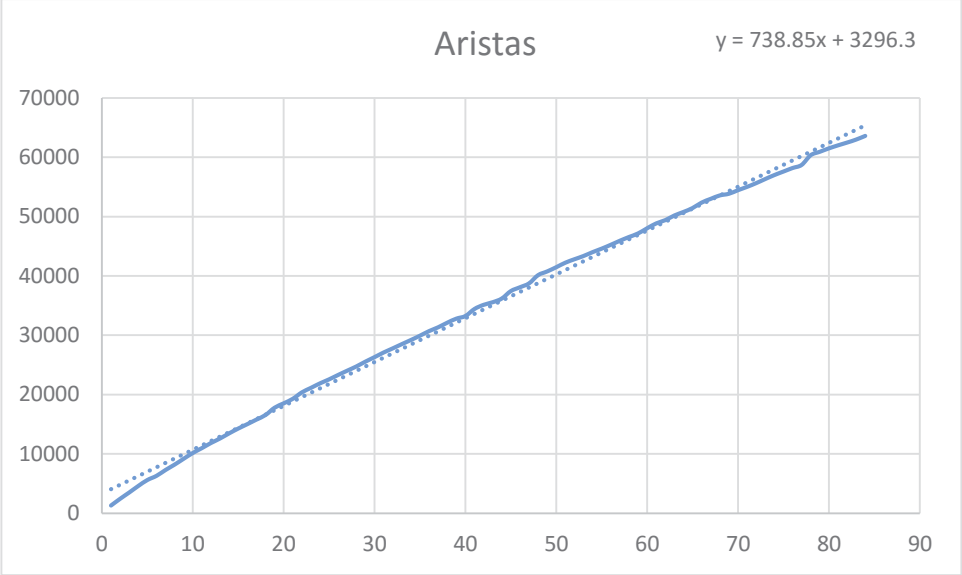

Figura 5.11 Aristas v/s Número de Textos

5.3.3.6 Grafo de Co-Ocurrencia Sin Stopwords

A continuación, se muestra de forma gráfica la cantidad de nodos y vértices que se van generando de algunos autores, represeentando los textos como un grafo de co-ocurrencia sin stopwords.

5.3.3.7 Catherine Bennett

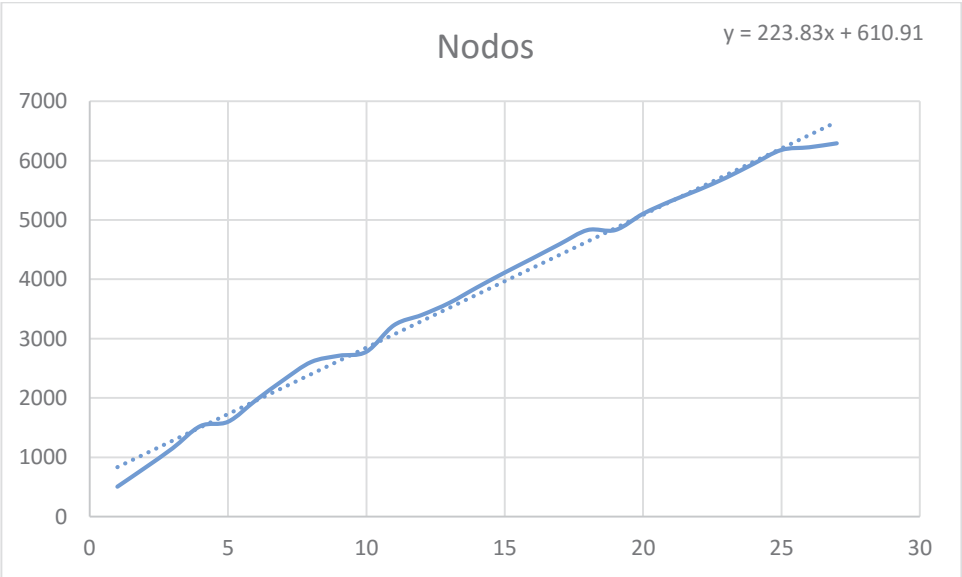

Figura 5.12 Nodos v/s Número de Textos

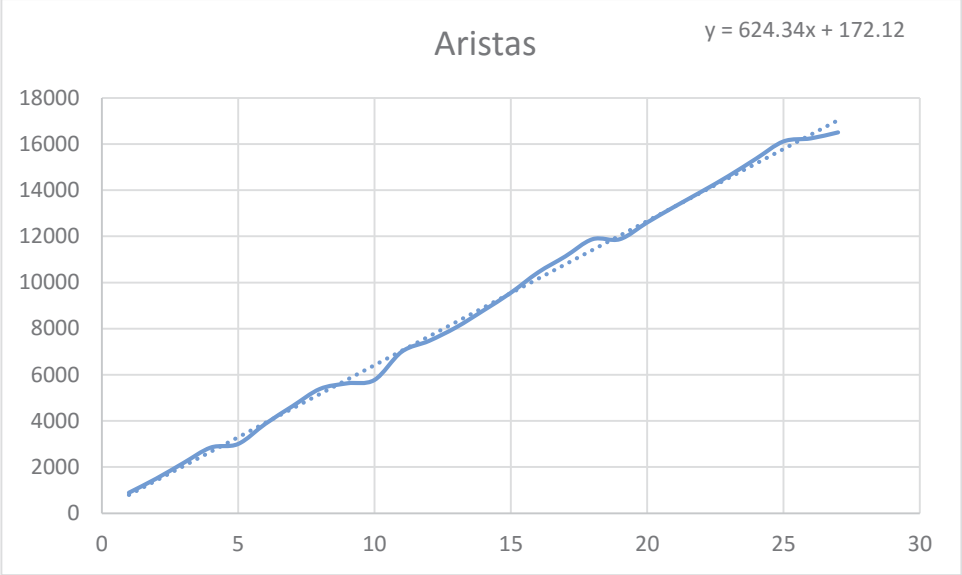

Figura 5.13 Aristas v/s Número de Textos

5.3.3.8 George Monbiot

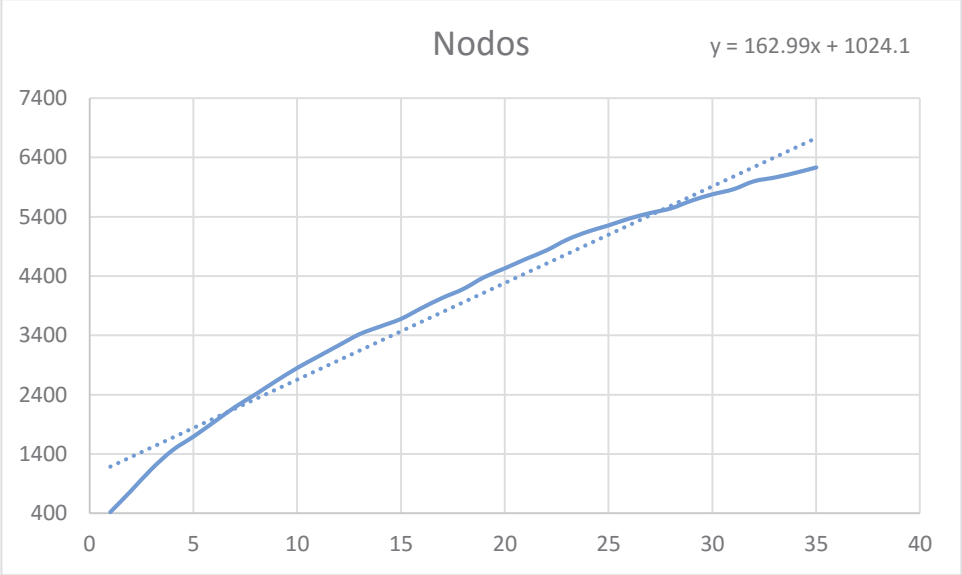

Figura 5.14 Nodos v/s Número de Textos

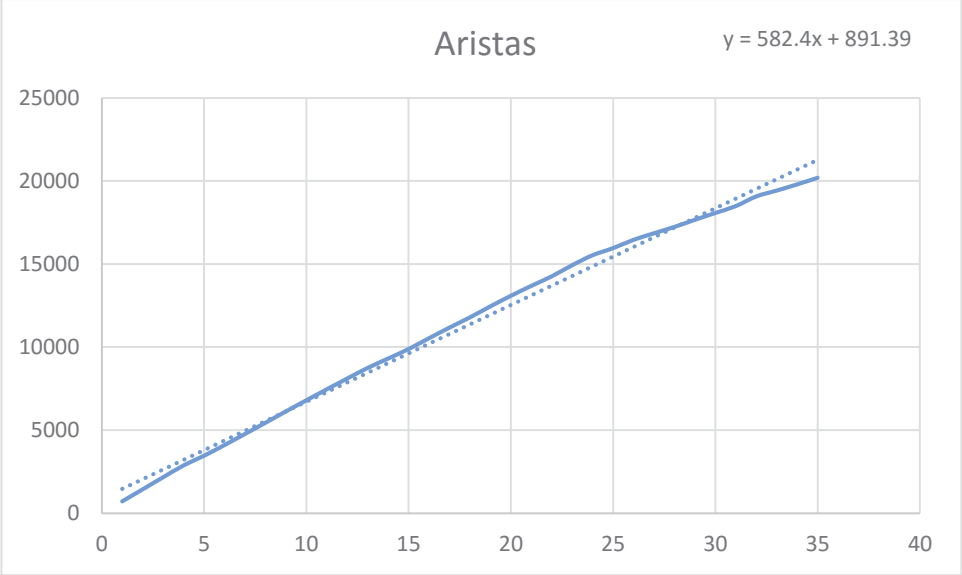

Figura 5.15 Aristas v/s Número de Textos

5.3.3.9 Hugo Young

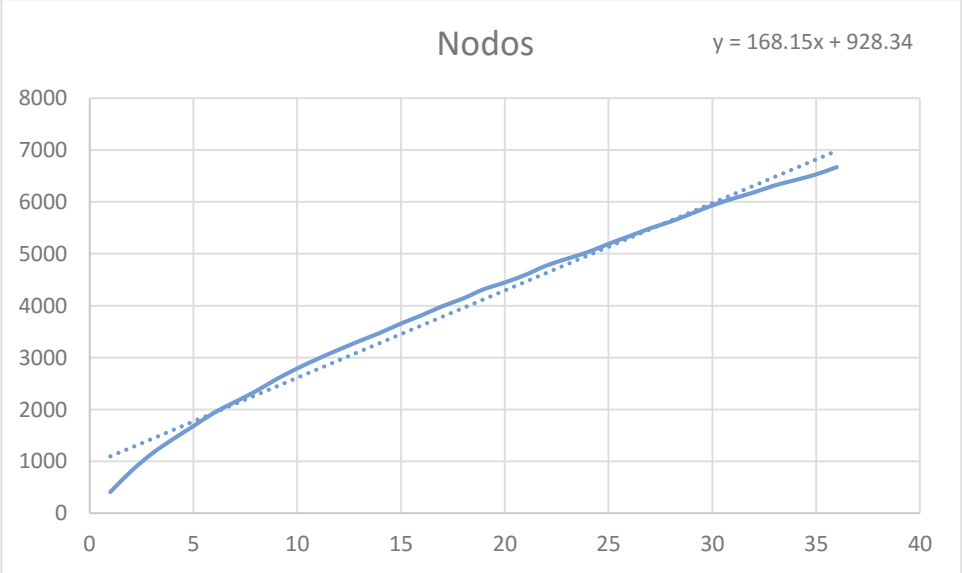

Figura 5.16 Nodos v/s Número de Textos

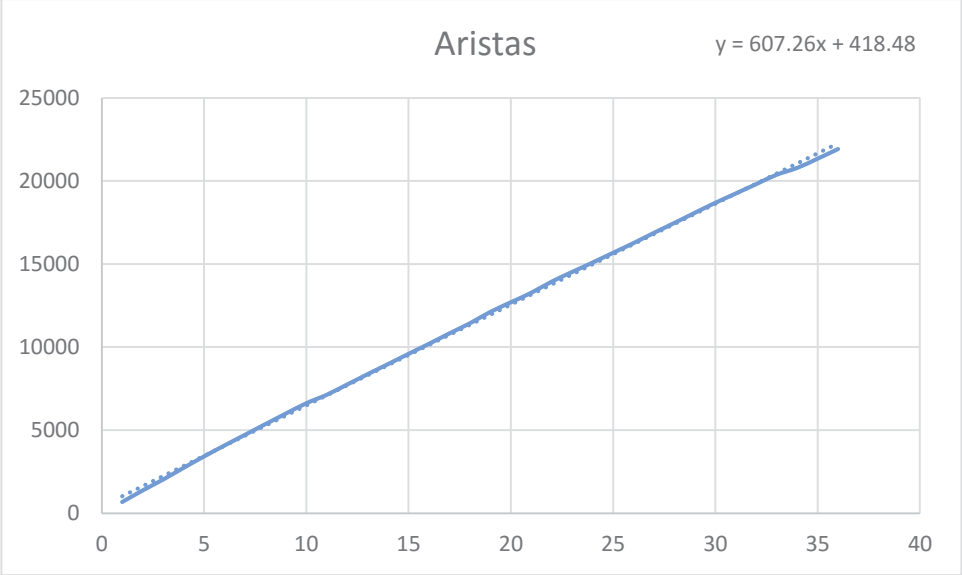

Figura 5.17 Aristas v/s Número de Textos

5.3.3.10 Jonathan Freedland

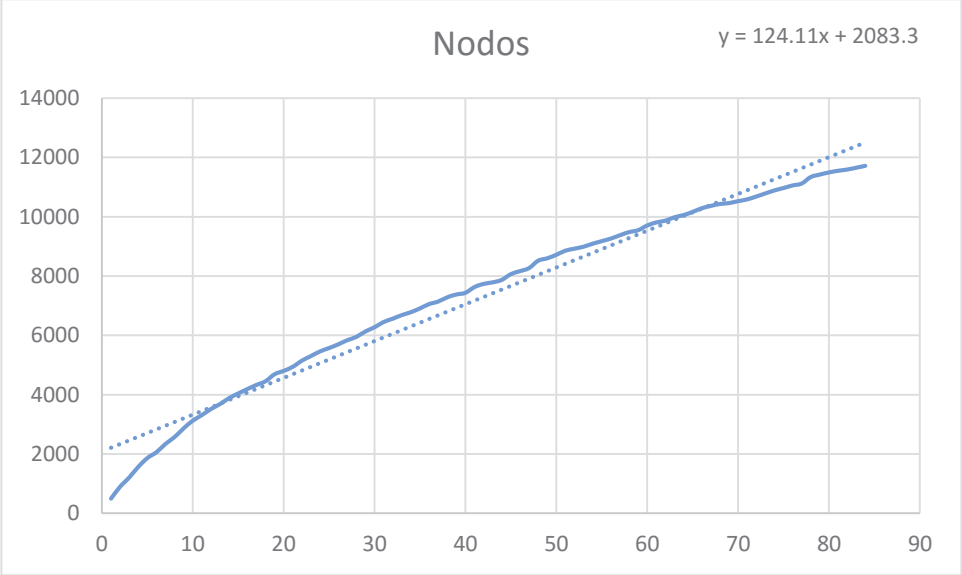

Figura 5.18 Nodos v/s Número de Textos

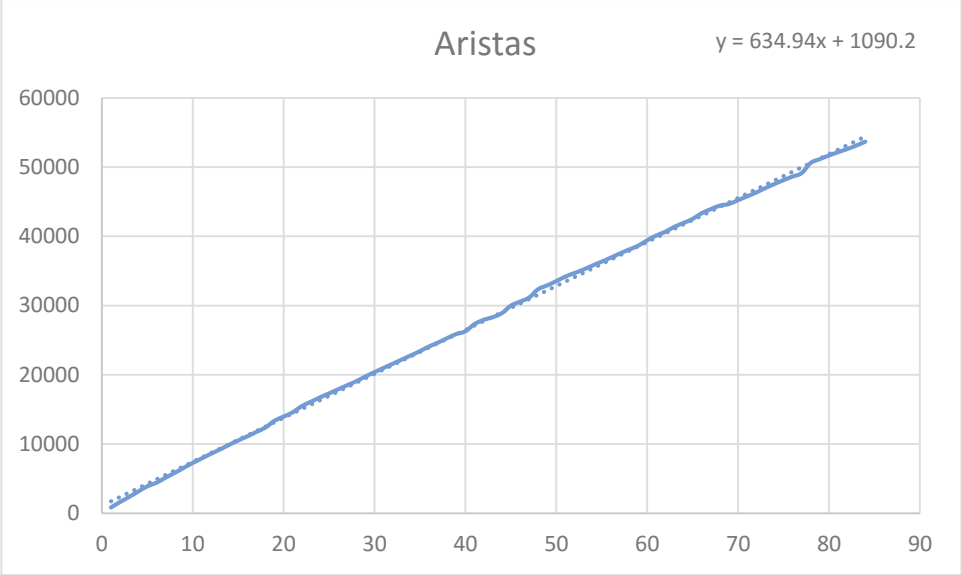

Figura 5.19 Aristas v/s Número de Textos

5.3.3.11 Grafo de Co-Ocurrencia De Stopwords

A continuación, se muestra de forma gráfica la cantidad de nodos y vértices que se van generando de algunos autores, representando los textos como un grafo de co-ocurrencia de stopwords.

5.3.3.12 Catherine Bennett

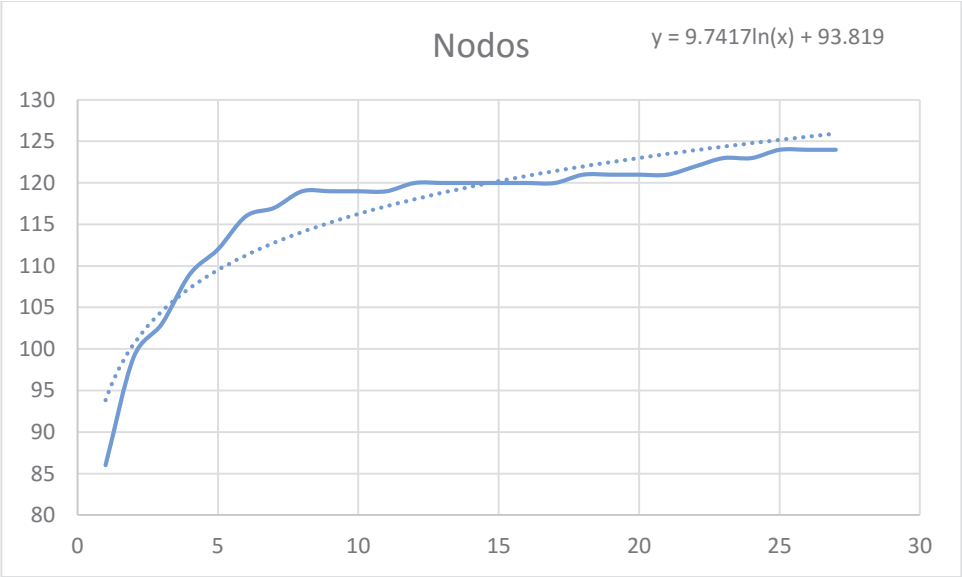

Figura 5.20 Nodos v/s Número de Textos

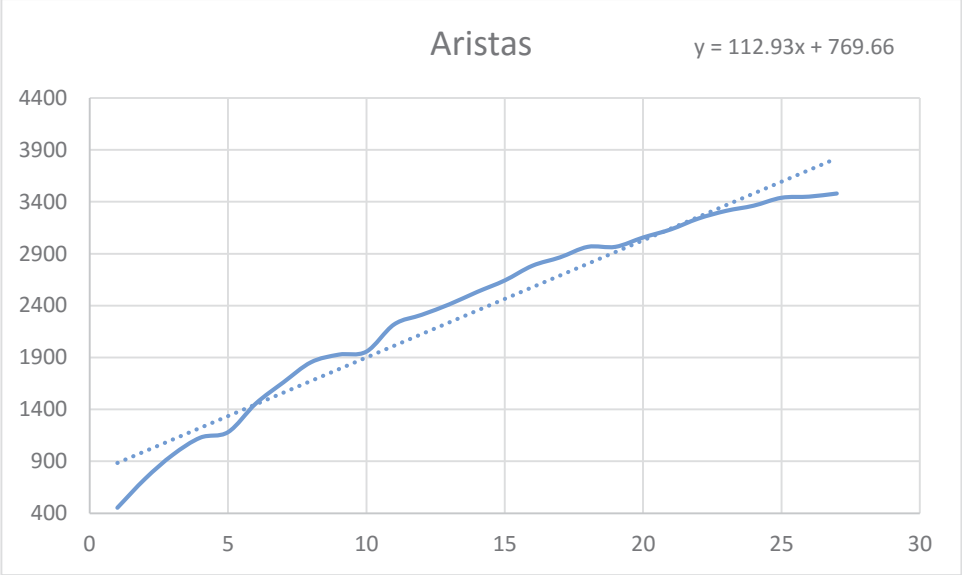

Figura 5.21 Aristas v/s Número de Textos

5.3.3.13 George Monbiot

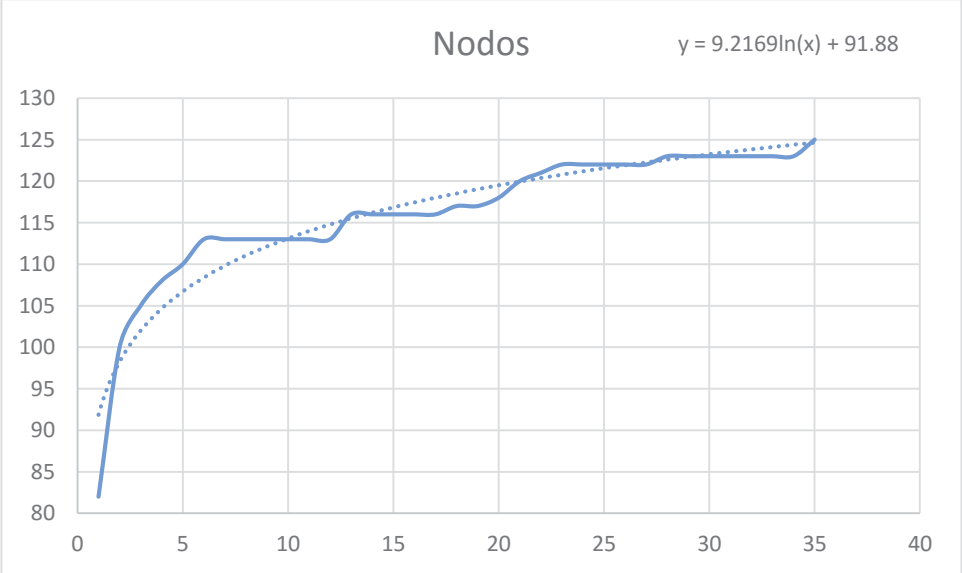

Figura 5.22 Nodos v/s Número de Textos

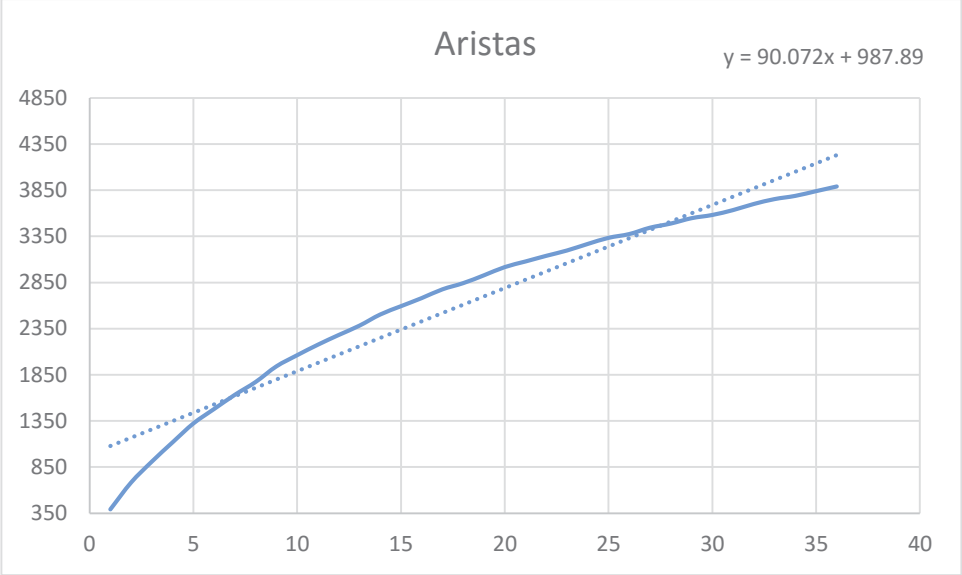

Figura 5.23 Aristas v/s Número de Textos

5.3.3.14 Hugo Young

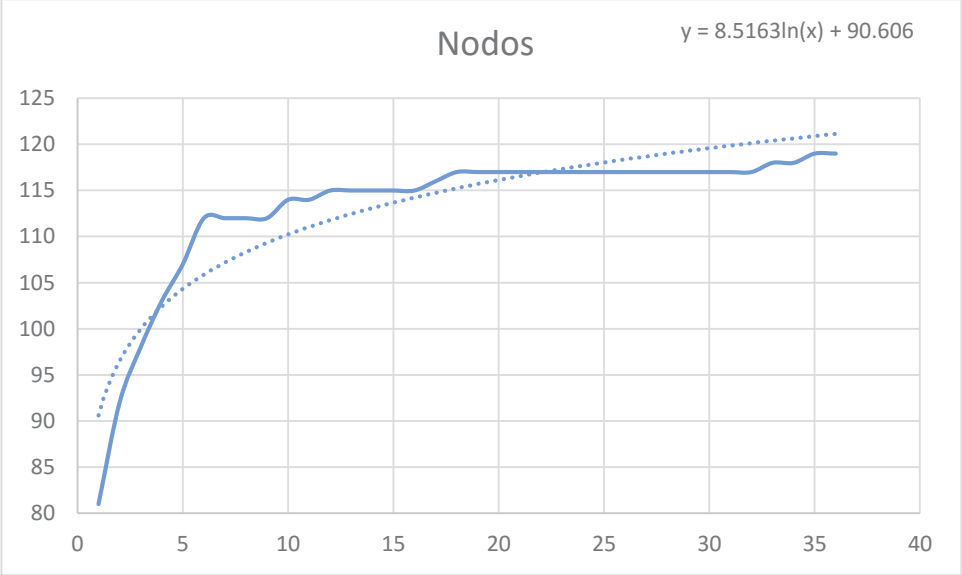

Figura 5.24 Nodos v/s Número de Textos

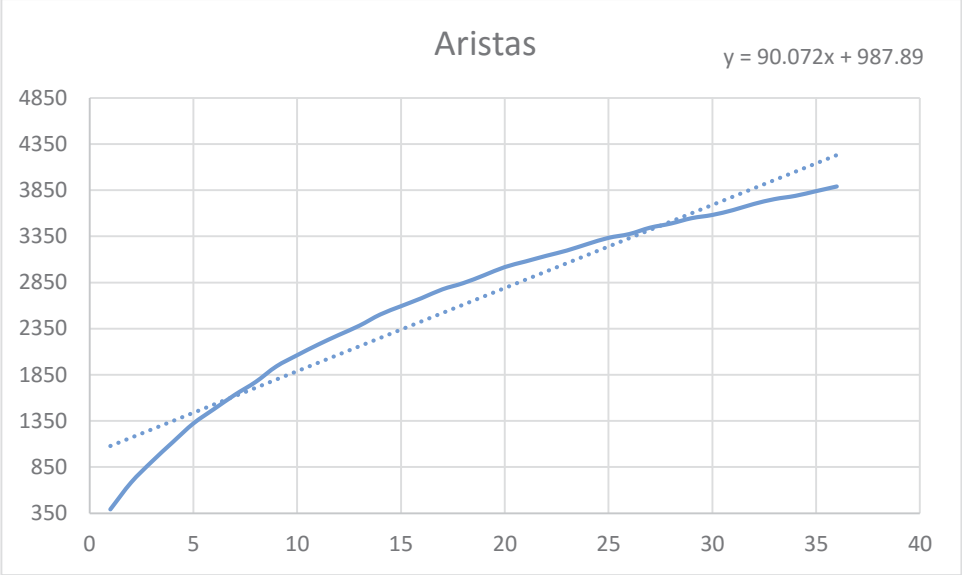

Figura 5.25 Aristas v/s Número de Textos

5.3.3.15 Jonathan Freedland

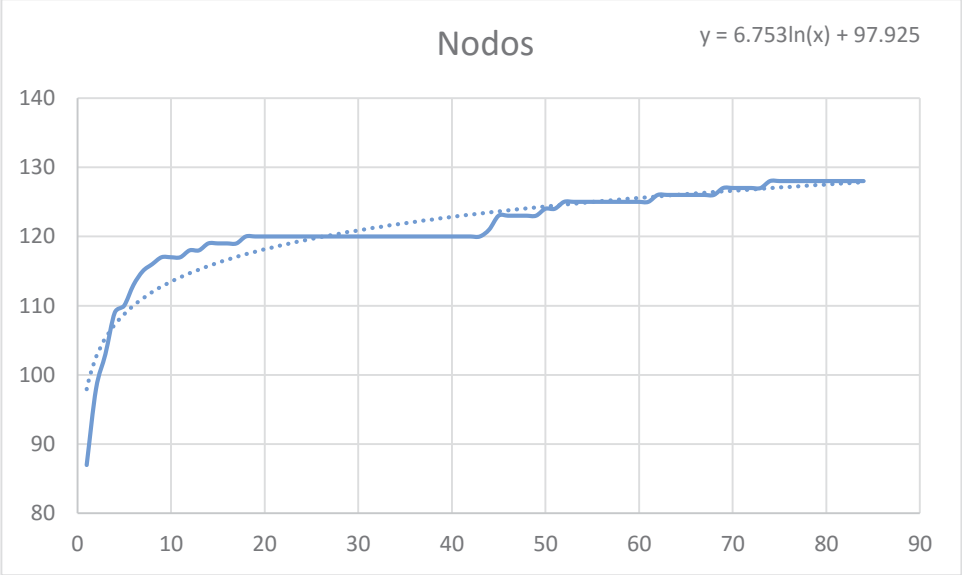

Figura 5.26 Nodos v/s Número de Textos

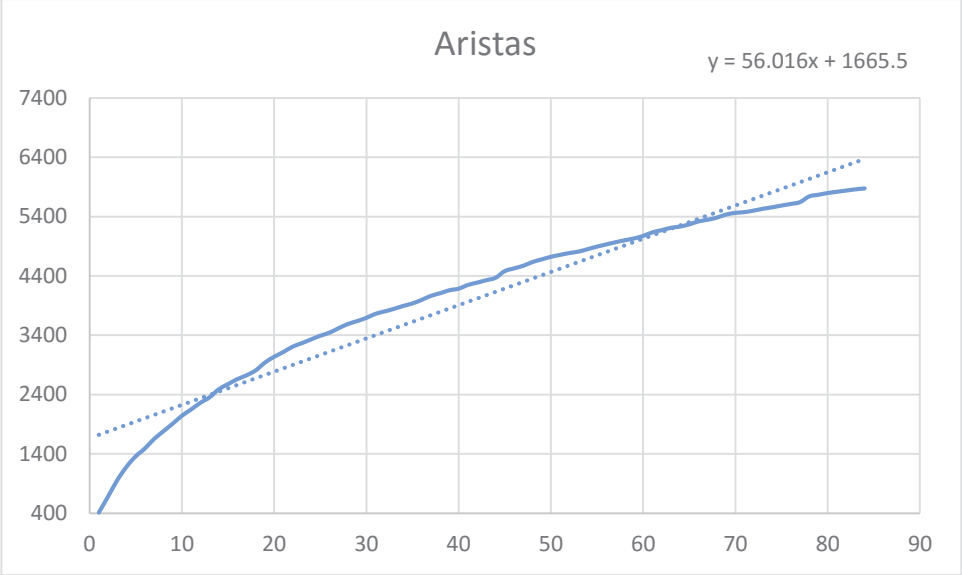

Figura 5.27 Aristas v/s Número de Textos

5.3.4 Análisis de Resultado del Crecimiento del Grafo de Co Ocurrencia

Para la representación utilizando grafos de co ocurrencia sin stopword y con stopword, tanto los gráficos para los nodos, como los para los vértices se presentan de forma similar a una recta, variando un poco su inclinación dependiendo de la cantidad de textos que se le entrega al autor.

Para la representación utilizando grafos de co ocurrencia de stopwords notamos una mayor variación, mostrando que los nodos se comportan de una forma similar a la función de logaritmo natural, que mostramos a continuación.

$$\ln(x) = \int_1^x \frac{dt}{t}, x > 0$$

Para los vértices, notamos un comportamiento similar al anterior, en forma de recta.

5.3.5 Atribución de Autoría Grafo de Co Ocurrencia utilizando similitud de la frecuencia de las palabras entre grafos

A continuación, se muestra un resumen de la atribución de autoría utilizando un análisis sobre las frecuencias de las palabras que se representan en el grafo de coocurrencia.

5.3.5.1 Similitud de la frecuencia de las palabras entre grafos con Grafo de Co Ocurrencia con stopwords.

En este caso se muestra el resultado al evaluar el grafo de coocurrencia conformado por todas las palabras del texto. En total, el porcentaje de acierto es de un 19,3% de los textos evaluados.

Tabla 5.2 Resumen de atribución mediante frecuencias de palabras.

Autor	Correctos	Incorrectos
georgemonbiot	3	6
royhattersley	8	0
catherinebennett	0	7
maryriddell	3	12
jonathanfreedland	2	21
peterpreston	0	16
nickcohen	0	8
simonhoggart	1	20
hugoyoung	0	9
martinkettle	8	0
pollytoynbee	0	12
willhutton	2	6
Total	28	117

5.3.5.2 Similitud de la frecuencia de las palabras entre grafos con Grafo de Co Ocurrencia sin stopwords

La siguiente tabla, corresponde a los resultados obtenidos al evaluar el grafo de coocurrencia generado al quitar del texto los stopwords. En este caso, el porcentaje de acierto es de un 17,9% de los textos evaluados.

Tabla 5.3 Resumen de atribución mediante frecuencias de palabras con stopwords.

Autor	Correctos	Incorrectos
georgemonbiot	2	7
royhattersley	8	0
catherinebennett	0	7
maryriddell	3	12
jonathanfreedland	2	21
peterpreston	0	16
nickcohen	0	8
simonhoggart	1	20
hugoyoung	0	9
martinkettle	8	0
pollytoynbee	0	12

willhutton	1	7
Total	26	119

5.3.5.3 Similitud de la frecuencia de las palabras entre grafos con Grafo de Co Ocurrencia de stopwords

En el tercer caso de esta serie se midió el resultado sobre un grafo conformado solo por las stopwords, las mismas palabras que el caso previo excluía. En este caso el porcentaje llega al 29,9% de los textos.

Tabla 5.4 Resumen de atribución mediante frecuencias de palabras de stopwords.

Autor	Correctos	Incorrectos
catherinebennett	2	5
georgemonbiot	0	9
hugoyoung	8	1
jonathanfreedland	0	21
martinkettle	7	1
maryriddell	0	12
nickcohen	4	4
peterpreston	2	14
pollytoynbee	4	8
royhattersley	2	6
simonhoggart	6	14
willhutton	2	6
zoewilliams	5	2
Total	42	103

5.3.6 Atribución de Autoría Grafo de Coocurrencia utilizando Simrank Adaptado

A continuación, se muestra un resumen de la atribución de autoría utilizando un grafo de coocurrencia y simrank, con sus distintas variantes en los stopwords.

5.3.6.1 Simrank Adptado Grafo de Co Ocurrencia sin stopword

A continuación, se muestra una tabla con el resumen de la atribución de autoría para el grafo de co ocurrencia sin stopword y utilizando SimRank Adaptado

Tabla 5.5 Resumen Atribución de Autoria

Autor	Correctos	Incorrectos
catherinebennett	0	7
georgemonbiot	1	8
hugoyoung	0	9
jonathanfreedland	17	4
martinkettle	1	7
maryriddell	1	11
nickcohen	0	8
peterpreston	1	15
pollytoynbee	0	12
royhattersley	0	8
simonhoggart	5	15
willhutton	0	8
zoewilliams	0	7
Total	26	119

6 Trabajo futuro

En un futuro trabajo para la determinación de la autoría mediante grafos de palabras resulta interesante profundizar en la creación de un indicador fruto de la mezcla de más de una técnica, ponderando la influencia de cada una según que tan buenos resultados dan por sí solas.

Para poder aplicar más de una técnica es también importante tener como una meta futura optimizar los códigos utilizados, ya que con las tecnologías que se trabajó los tiempos de procesamiento fueron bastante elevados.

7 Conclusiones

Atribuir la autoría de un texto anónimo al identificar la forma en que el autor se expresa es un acto que podría dar solución a muchos dilemas actuales, como los casos de acoso o la atribución de textos anónimos que circulan actualmente.

Este problema se abordó desde los grafos de palabras, pero los porcentajes de aciertos obtenidos no son suficientes para tener un método fiable para atribuir autoría, sin hay aspectos que quedan claros tras esta investigación.

Los grafos de palabras tienden a crecer en tamaño de nodos con una cota superior, esto se debe a que un autor maneja un cierto lenguaje que será independiente a la extensión del texto, pero no es así con las aristas del grafo de coocurrencia, que su cantidad crecerá en forma directamente proporcional al lago del texto que representa.

En trabajos previos los textos se analizaban sin tener en cuenta las stopwords, palabras que se consideran vacías, pero se ha detectado en ellas una importancia a la hora de detectar autoría, ya que son estas palabras las que representan la forma de escribir de un autor.

Al manejar grafos que en sus nodos tienen asignado un peso se puede trabajar otra característica de la Estilometría, la frecuencia con que un autor usa las palabras. En este caso los mejores resultados se obtuvieron al evaluar la frecuencia con que un autor utiliza stopwords en el texto ya que como se mencionó, las stopwords nos entregan información sobre la forma de escritura.

Finalmente, los resultados obtenidos con los métodos y experimentación descrita en este informe no son suficientes para poder atribuir autoría de forma automática mediante grafos, ya que los porcentajes de acierto son muy bajos.

8 Referencias

- Fernando Püschel Araya, Nicolás Zárate Guzmán (2009). DETERMINACIÓN DE AUTORÍA POR MEDIO DE REDES DE PALABRAS
- The State of Authorship Attribution Studies: (1) The History and the Scope; (2) The Problems -- Towards Credibility and Validity.
- Minería textual, Ricardo Eíto Brun y Jose A. Senso
- A Survey of Modern Authorship Attribution Methods, Efstathios Stamatatos.
- Graph based Representation and Analysis of Text Document: A Survey of Techniques, S.S. Sonawane, Dr. P.A. Kuklnari (2014)
- Algorithms for Graph Similarity and Subgraph Matching, Danai Koutra, Ankur Parikh, Aaditya Ramdas, Jing Xiang (2011)

Anexos

A: The Ghostly Village

The night was rainy; a big storm was falling on the sea. The waves were enormous and the fog was thick. The ships rocked one side to the other as marionettes.

Suddenly, an awful creaking was heard in the darkness. A big cloud of smoke was seen in the distance and an intense odour could be noticed in the air. Everybody was wondering what had happened.

A ship had ran aground near the shore and had split part of the petrol it carried. A big black stain spreaded on the water, as a big black cloack which had the sea gone into mourning. The smell of petrol was each time stronger and mixed with the freshness of the breeze each sunset near the beach. Charles and Anne used to go watching the stars. When they felt that freedom that only those who have not betrayed their ideals feel. They were the children of a fisher and lived in a humble white house very near from the cliff.

The fishers had recently had problems to fish, fishing was not very good. Now, it would be worse, there would not be anything in many time. Fishers will not be seen carrying fish to the harbour. They could not be said goodbye as it was usual. Now they will have to go far, to be able to live.

The village became a village without people. A ghostly village. Just a few women and children remained there. Men and young people went to look for a job and came back once in a while to see their families. At nightfall, a few lights, brought the village back to existence.

But from the cliff the view was not the same, it seemed that even the breeze had changed of place. The air smell of petrol and the sea's calm had turned to a terrible anguished seeing how all the sea life was being destroyed. Dead fishes floated and all was devastating. The few people who remained, started to rebuild and clean all that had been damaged.

Some years passed until the village returned to normal. Some of who had left returned and the boats returned to the harbour. Hope was born again with the fear that the story would repeat.

B: Stopwords

Lista de stopwords obtenidas de Ranks.nl

"a"	"did"	"herself"	"not"	"the"	"we've"
"about"	"didn't"	"him"	"of"	"their"	"were"
"above"	"do"	"himself"	"off"	"theirs"	"weren't"
"after"	"does"	"his"	"on"	"them"	"what"
"again"	"doesn't"	"how"	"once"	"themselves"	"what's"
"against"	"doing"	"how's"	"only"	"then"	"when"
"all"	"don't"	"i"	"or"	"there"	"when's"
"am"	"down"	"i'd"	"other"	"there's"	"where"
"an"	"during"	"i'll"	"ought"	"these"	"where's"
"and"	"each"	"i'm"	"our"	"they"	"which"
"any"	"few"	"i've"	"ours"	"they'd"	"while"
"are"	"for"	"if"	"ourselves"	"they'll"	"who"
"aren't"	"from"	"in"	"out"	"they're"	"who's"
"as"	"further"	"into"	"over"	"they've"	"whom"
"at"	"had"	"is"	"own"	"this"	"why"
"be"	"hadn't"	"isn't"	"same"	"those"	"why's"
"because"	"has"	"it"	"shan't"	"through"	"with"
"been"	"hasn't"	"it's"	"she"	"to"	"won't"
"before"	"have"	"its"	"she'd"	"too"	"would"
"being"	"haven't"	"itself"	"she'll"	"under"	"wouldn't"
"below"	"having"	"let's"	"she's"	"until"	"you"
"between"	"he"	"me"	"should"	"up"	"you'd"
"both"	"he'd"	"more"	"shouldn't"	"very"	"you'll"
"but"	"he'll"	"most"	"so"	"was"	"you're"
"by"	"he's"	"mustn't"	"some"	"wasn't"	"you've"
"can't"	"her"	"my"	"such"	"we"	"your"
"cannot"	"here"	"myself"	"than"	"we'd"	"yours"
"could"	"here's"	"no"	"that"	"we'll"	"yourself"
"couldn't"	"hers"	"nor"	"that's"	"we're"	"yourselves"

C: Frecuencias graficadas en la Figura 5.2

Autores:

1 georgemonbiot
2 royhattersley
3 catherinebennett
4 maryriddell
5 jonathanfreedland
6 peterpreston
7 nickcohen
8 simonhoggart
9 hugoyoung
10 martinkettle
11 pollytoynbee
12 willhutton
13 zoewilliams

Datos graficados:

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	3013	1861	1585	3539	6921	3706	3064	2859	2756	2059	3463	2466	939
2	1236	896	927	1871	2967	1784	1348	1250	1282	922	1486	1188	559
3	1230	778	838	1607	2839	1698	1270	1244	1266	846	1421	1073	535
4	785	532	634	1389	2839	1384	1016	1122	947	713	1172	1022	512
5	667	521	533	1353	2111	1359	911	905	706	619	1160	774	378
6	577	408	513	1024	1811	1078	744	745	671	612	958	737	325
7	516	388	397	869	1694	959	547	725	633	422	927	684	310
8	502	375	344	753	1434	739	418	671	608	412	768	547	302
9	444	271	300	586	1295	684	404	593	525	332	684	480	292
10	311	249	300	558	1223	565	375	591	470	316	524	316	216
11	311	235	296	458	1048	556	372	471	415	271	453	305	179
12	309	210	257	432	900	507	327	457	397	271	429	257	175
13	300	208	249	377	844	464	317	424	361	250	412	251	161
14	275	196	217	372	794	418	300	358	324	248	394	242	159
15	256	196	204	360	789	385	291	344	303	248	387	229	156
16	249	168	203	347	776	380	285	326	299	200	371	222	153
17	228	165	184	342	761	363	274	314	284	189	366	208	147
18	227	163	179	309	729	356	270	312	277	183	357	205	146
19	225	157	171	308	687	348	267	308	274	182	357	197	143
20	220	156	167	288	674	320	244	308	265	174	322	191	136
21	207	154	154	276	659	320	224	301	256	174	316	188	130
22	207	150	147	273	632	310	219	281	240	166	296	170	129
23	200	138	144	271	576	305	215	281	230	157	292	160	124
24	195	136	139	271	556	302	211	276	216	131	276	156	114
25	184	136	138	267	546	290	211	269	206	127	271	153	112
26	184	120	124	247	531	278	191	266	196	127	262	153	111
27	179	116	119	242	483	278	190	257	195	125	252	146	91

28	176	115	118	236	456	275	186	241	195	118	249	140	89
29	161	114	116	200	446	265	183	229	184	115	246	137	86
30	160	109	115	193	427	264	182	204	180	115	234	132	83
31	156	107	113	184	410	247	179	196	168	111	224	127	82
32	152	102	112	169	396	234	170	193	167	110	215	123	81
33	151	98	107	167	396	228	168	189	159	108	209	122	79
34	129	93	106	166	393	217	164	189	149	107	206	106	76
35	127	93	102	165	391	207	162	177	147	106	204	105	75
36	127	92	101	157	389	202	152	175	146	106	196	105	75
37	122	83	97	153	379	195	151	172	142	106	190	105	74
38	121	79	95	143	348	190	150	172	141	102	186	104	69
39	117	77	95	141	347	188	146	166	139	100	180	102	69
40	112	76	95	136	328	184	140	166	135	90	177	102	69
41	105	76	91	135	319	184	134	165	126	90	175	100	67
42	105	76	91	135	318	180	127	161	124	82	172	98	66
43	104	75	89	133	314	178	120	158	124	81	171	94	64
44	104	73	88	133	312	177	117	157	122	80	168	92	63
45	97	70	87	130	296	165	112	154	118	79	165	91	63
46	94	69	86	127	292	164	106	152	117	79	161	91	63
47	92	66	85	123	274	161	103	152	114	76	161	90	62
48	91	65	81	122	271	149	103	145	114	76	157	90	62
49	90	63	79	122	263	149	102	144	111	75	156	87	61
50	89	61	77	117	254	147	101	141	110	75	153	80	60
51	87	61	76	115	251	134	100	139	110	75	151	79	59
52	87	59	74	111	248	128	95	131	108	74	146	75	57
53	87	59	74	110	248	122	87	130	107	74	145	71	56
54	84	58	70	109	242	119	86	129	105	71	142	70	56
55	84	57	69	106	235	118	83	128	100	69	132	70	54
56	82	57	67	105	233	118	81	124	99	68	129	70	54
57	82	56	67	101	231	118	79	119	97	67	123	70	53
58	80	53	66	100	222	117	76	116	95	66	123	65	52
59	77	53	65	100	221	116	74	116	95	65	121	64	50
60	73	50	63	99	221	115	74	113	91	63	121	64	50
61	69	49	62	99	220	114	73	108	88	59	115	64	49
62	65	48	59	96	214	114	73	100	85	57	113	63	49
63	64	46	58	96	212	112	70	97	83	57	111	61	48
64	64	45	56	96	210	110	70	97	82	57	108	58	47
65	61	43	56	94	209	110	69	95	82	56	105	58	47
66	61	43	55	92	207	108	68	94	80	55	104	57	44
67	61	41	52	91	203	106	68	91	79	55	103	56	43
68	61	40	52	91	201	105	68	91	77	54	103	56	43

69	60	39	52	89	199	103	67	89	77	53	103	55	42
70	59	39	50	89	197	101	67	87	76	52	103	55	42
71	56	39	49	87	197	101	66	84	76	52	102	54	39
72	56	39	48	86	196	99	64	83	74	51	101	53	38
73	55	39	48	84	195	98	64	82	74	51	100	53	37
74	55	38	47	84	193	97	63	82	71	51	100	52	37
75	55	37	47	83	188	97	61	81	70	50	96	52	35
76	53	36	46	82	188	96	59	81	69	50	96	52	34
77	51	35	46	81	188	95	59	79	69	50	95	52	33
78	51	35	44	78	188	95	58	77	68	50	95	50	32
79	51	34	44	77	184	94	58	77	68	49	94	49	31
80	50	34	44	74	183	94	57	76	68	48	93	49	31
81	49	33	43	74	173	94	56	75	64	47	90	48	31
82	48	33	43	72	173	91	56	74	64	45	87	48	31
83	48	32	42	70	172	87	55	74	63	45	87	47	30
84	48	32	41	69	168	85	55	73	62	44	86	46	30
85	47	31	41	68	167	85	54	72	62	44	85	46	29
86	47	31	41	67	164	83	54	71	59	44	83	46	29
87	46	30	41	67	160	83	54	71	58	44	83	46	29
88	45	30	41	67	158	82	53	70	57	44	82	46	29
89	45	30	40	66	157	82	52	70	57	43	81	45	28
90	45	30	40	65	156	81	52	66	56	42	80	45	27
91	45	29	40	64	155	80	50	66	56	42	78	44	27
92	44	29	39	61	151	80	49	65	56	42	78	41	27
93	44	29	38	61	148	80	49	65	56	41	78	41	26
94	44	29	38	61	147	79	48	64	54	41	78	40	26
95	44	28	37	61	144	76	48	64	53	41	78	39	26
96	43	28	37	61	142	76	44	63	53	41	77	39	26
97	43	28	37	60	141	75	42	62	53	41	74	39	26
98	43	28	37	60	139	75	42	62	53	40	74	39	25
99	43	27	37	59	139	73	42	59	52	40	73	38	25
100	42	27	37	58	135	73	41	59	52	39	71	38	25

D: Porcentaje de stopwords en un texto

- **Porcentaje de stopwords Catherine Bennett**

Textos	Cantidad de Palabras	Sin Stopwords	Vocabulario	% de Stopwords en el Texto
Society_003.txt	1367	723	585	47
Society_004.txt	1014	507	459	50
UK_001.txt	1155	598	516	48
UK_002.txt	1071	581	562	45
UK_003.txt	256	135	164	47
UK_004.txt	1399	753	657	46
UK_005.txt	1318	681	635	48
UK_006.txt	1230	641	603	47
UK_007.txt	432	221	260	48
UK_008.txt	240	128	175	46
UK_009.txt	2262	1224	898	45
UK_010.txt	885	464	444	47
UK_011.txt	1046	566	514	45
UK_012.txt	1269	668	611	47
UK_013.txt	1379	708	625	48
UK_014.txt	1642	891	674	45
World_001.txt	1135	637	591	43
World_002.txt	1323	709	639	46
World_003.txt	558	282	309	49
World_004.txt	1235	676	643	45
World_005.txt	1121	622	578	44
World_006.txt	1268	642	585	49
World_007.txt	1184	653	588	44
World_008.txt	1331	725	608	45
World_009.txt	1293	709	645	45
World_010.txt	247	126	155	48
World_011.txt	516	246	273	52

- **Porcentaje de stopwords George Monbiot**

Textos	Cantidad de Palabras	Sin Stopwords	Vocabulario	% de Stopwords en el Texto
World_007.txt	1218	612	499	49
World_008.txt	1183	605	523	48
World_009.txt	1205	625	580	48
World_010.txt	1210	634	533	47
World_011.txt	1162	561	486	51
World_012.txt	1177	581	490	50
World_013.txt	1204	654	530	45
World_014.txt	1196	652	516	45
World_015.txt	1200	637	563	46
World_016.txt	1205	652	562	45
World_017.txt	1187	637	517	46
World_018.txt	1204	651	518	45
World_019.txt	1204	637	531	47
World_020.txt	1218	649	509	46
World_021.txt	1203	622	469	48
World_022.txt	1153	652	517	43
World_023.txt	1226	655	545	46
World_024.txt	1243	642	527	48
World_025.txt	1192	660	560	44
World_026.txt	1212	671	538	44
World_027.txt	1210	639	551	47
World_028.txt	1214	606	549	50
World_029.txt	1211	643	584	46
World_030.txt	1199	660	526	44
World_031.txt	778	416	401	46
World_032.txt	923	522	410	43
World_033.txt	755	411	379	45
World_034.txt	782	414	374	47
World_035.txt	766	413	407	46
World_036.txt	772	414	371	46
World_037.txt	809	442	391	45
World_038.txt	1131	649	513	42
World_039.txt	752	430	366	42
World_040.txt	725	390	355	46
World_041.txt	792	450	414	43

- **Porcentaje de stopwords Hugo Young**

Textos	Cantidad de Palabras	Sin Stopwords	Vocabulario	% de Stopwords en el Texto
UK_007.txt	1128	554	490	50
World_001.txt	1217	599	551	50
World_002.txt	1201	595	560	50
World_003.txt	1209	625	557	48
World_004.txt	1218	635	537	47
World_005.txt	1172	623	551	46
World_006.txt	1250	621	537	50
World_007.txt	1186	597	533	49
World_008.txt	1217	605	566	50
World_009.txt	1193	610	555	48
World_010.txt	988	488	483	50
World_011.txt	1200	622	557	48
World_012.txt	1159	619	567	46
World_013.txt	1223	617	554	49
World_014.txt	1204	608	565	49
World_015.txt	1177	588	544	50
World_016.txt	1219	632	587	48
World_017.txt	1176	647	557	44
World_018.txt	1309	670	589	48
World_019.txt	1232	614	542	50
World_020.txt	1167	619	565	46
World_021.txt	1197	661	591	44
World_022.txt	1177	610	537	48
World_023.txt	1162	624	546	46
World_024.txt	1135	602	552	46
World_025.txt	1166	626	565	46
World_026.txt	1160	638	541	45
World_027.txt	1156	642	575	44
World_028.txt	1182	641	585	45
World_029.txt	1125	627	573	44
World_030.txt	1155	589	545	49
World_031.txt	1185	611	547	48
World_032.txt	1211	628	578	48
World_033.txt	822	418	410	49
World_034.txt	1161	606	565	47
World_035.txt	1144	601	535	47

- **Porcentaje de stopwords Jonathan Freedland**

Textos	Cantidad de Palabras	Sin Stopwords	Vocabulario	% de Stopwords en el Texto
World_017.txt	1277	685	571	46
World_018.txt	1269	708	610	44
World_019.txt	1285	673	567	47
World_020.txt	1308	691	610	47
World_021.txt	1258	676	597	46
World_022.txt	916	477	455	47
World_023.txt	1256	656	577	47
World_024.txt	1250	664	572	46
World_025.txt	1288	664	596	48
World_026.txt	1270	654	575	48
World_027.txt	1237	676	568	45
World_028.txt	1210	649	569	46
World_029.txt	1210	643	543	46
World_030.txt	1203	618	558	48
World_031.txt	1221	640	545	47
World_032.txt	1042	564	491	45
World_033.txt	1206	655	555	45
World_034.txt	1310	675	557	48
World_035.txt	1851	965	782	47
World_036.txt	1221	619	531	49
World_037.txt	1225	654	554	46
World_038.txt	1689	886	710	47
World_039.txt	1210	654	576	45
World_040.txt	1215	649	560	46
World_041.txt	1223	641	527	47
World_042.txt	1223	637	511	47
World_043.txt	1218	637	557	47
World_044.txt	1221	615	522	49
World_045.txt	1213	662	593	45
World_046.txt	1284	684	595	46
World_047.txt	1184	578	561	51
World_048.txt	1189	654	559	44
World_049.txt	1241	651	569	47

World_050.txt	1254	673	542	46
World_051.txt	1243	660	562	46
World_052.txt	1249	686	605	45
World_053.txt	1240	671	546	45
World_054.txt	1234	668	605	45
World_055.txt	1276	646	559	49
World_056.txt	855	459	415	46
World_057.txt	2061	1085	839	47
World_058.txt	1183	640	562	45
World_059.txt	795	416	390	47
World_060.txt	1228	680	549	44
World_061.txt	2171	1118	897	48
World_062.txt	1229	666	554	45
World_063.txt	1227	671	555	45
World_064.txt	2397	1287	949	46
World_065.txt	1201	639	574	46
World_066.txt	1259	670	601	46
World_067.txt	1289	683	603	47
World_068.txt	1154	604	531	47
World_069.txt	1195	619	509	48
World_070.txt	1181	652	557	44
World_071.txt	1205	649	572	46
World_072.txt	1244	666	578	46
World_073.txt	1243	671	586	46
World_074.txt	1209	619	559	48
World_075.txt	1232	647	564	47
World_076.txt	1488	850	683	42
World_077.txt	1561	814	621	47
World_078.txt	1230	637	543	48
World_079.txt	1491	791	663	46
World_080.txt	1259	680	588	45
World_081.txt	1268	656	590	48
World_082.txt	1841	952	772	48
World_083.txt	1256	664	592	47
World_084.txt	1191	599	523	49
World_085.txt	561	269	278	52
World_086.txt	1210	648	566	46
World_087.txt	1209	664	581	45
World_088.txt	1222	667	577	45
World_089.txt	1359	718	587	47
World_090.txt	1193	659	598	44

World_091.txt	1189	645	568	45
World_092.txt	1213	609	517	49
World_093.txt	1134	598	519	47
World_094.txt	3273	1670	1179	48
World_095.txt	1220	652	554	46
World_096.txt	1237	638	572	48
World_097.txt	1196	612	513	48
World_098.txt	1246	643	529	48
World_099.txt	1205	632	545	47
World_100.txt	1264	652	575	48

- **Porcentaje de stopwords Martin Kettle**

Textos	Cantidad de Palabras	Sin Stopwords	Vocabulario	% de Stopwords en el Texto
World_006.txt	1166	563	507	51
World_007.txt	1087	667	444	38
World_008.txt	366	198	211	45
World_009.txt	1258	675	549	46
World_010.txt	1198	602	510	49
World_011.txt	1154	615	533	46
World_012.txt	1323	664	556	49
World_013.txt	1157	606	515	47
World_014.txt	1226	610	521	50
World_015.txt	1280	672	535	47
World_016.txt	1176	631	506	46
World_017.txt	567	292	274	48
World_018.txt	858	426	410	50
World_019.txt	1211	666	550	45
World_020.txt	1259	654	539	48
World_021.txt	1150	608	537	47
World_022.txt	814	424	413	47
World_023.txt	812	435	412	46
World_024.txt	901	476	439	47
World_025.txt	791	425	373	46
World_026.txt	889	466	413	47
World_027.txt	752	410	349	45
World_028.txt	901	491	433	45
World_029.txt	850	469	416	44

World_030.txt	783	417	382	46
World_031.txt	772	434	368	43
World_032.txt	723	388	338	46
World_033.txt	750	428	400	42
World_034.txt	878	517	408	41
World_035.txt	696	378	337	45
World_036.txt	893	456	414	48

- **Porcentaje de stopwords Mary Riddell**

Textos	Cantidad de Palabras	Sin Stopwords	Vocabulario	% de Stopwords en el Texto
UK_003.txt	1210	660	619	45
UK_004.txt	1710	900	778	47
UK_005.txt	1196	660	609	44
UK_006.txt	1275	756	696	40
UK_007.txt	1219	704	697	42
UK_008.txt	1230	719	669	41
UK_009.txt	1215	687	652	43
UK_010.txt	1213	661	594	45
UK_011.txt	1230	720	698	41
UK_012.txt	1256	701	650	44
UK_013.txt	1165	619	600	46
UK_014.txt	3204	1788	1322	44
UK_015.txt	1215	677	660	44
UK_016.txt	1116	581	527	47
UK_017.txt	1213	697	612	42
UK_018.txt	1216	658	629	45
UK_019.txt	1217	683	625	43
UK_020.txt	1261	698	635	44
UK_021.txt	1234	656	627	46
UK_022.txt	1278	702	649	45
UK_023.txt	1367	758	660	44
UK_024.txt	1249	700	639	43
World_001.txt	1140	674	578	40
World_002.txt	1181	674	611	42
World_003.txt	1225	693	622	43
World_004.txt	1189	655	589	44
World_005.txt	1210	641	618	47

World_006.txt	1192	660	622	44
World_007.txt	1183	644	587	45
World_008.txt	1204	677	628	43
World_009.txt	1225	699	653	42
World_010.txt	1346	773	689	42
World_011.txt	1224	680	654	44
World_012.txt	1211	677	616	44
World_013.txt	1218	683	640	43
World_014.txt	1202	663	639	44
World_015.txt	1240	699	660	43
World_016.txt	1150	675	652	41
World_017.txt	1236	728	674	41
World_018.txt	1218	704	668	42
World_019.txt	1208	667	608	44
World_020.txt	1586	908	767	42
World_021.txt	1391	781	727	43
World_022.txt	1221	707	673	42
World_023.txt	1285	741	703	42

- **Porcentaje de stopwords Nick Cohen**

Textos	Cantidad de Palabras	Sin Stopwords	Vocabulario	% de Stopwords en el Texto
Politics_017.txt	1564	803	677	48
Politics_018.txt	1501	815	648	45
Politics_019.txt	1279	686	593	46
Politics_020.txt	1437	771	684	46
Politics_021.txt	1484	799	705	46
Politics_022.txt	1553	830	707	46
Politics_023.txt	1485	794	710	46
Politics_024.txt	1464	829	689	43
Politics_025.txt	1544	815	711	47
Politics_026.txt	1615	876	763	45
Politics_027.txt	1520	862	703	43
Politics_028.txt	1489	788	659	47
Politics_029.txt	1490	808	681	45
Politics_030.txt	1693	912	774	46
Society_001.txt	1291	706	549	45
Society_002.txt	1483	839	714	43

UK_001.txt	1418	733	634	48
UK_002.txt	1534	746	673	51
UK_003.txt	341	192	224	43
UK_004.txt	1398	756	671	45
UK_005.txt	1399	757	672	45
UK_006.txt	1500	794	677	47
UK_007.txt	1626	873	767	46
World_001.txt	1304	676	621	48
World_002.txt	1093	553	529	49
World_003.txt	1533	794	656	48
World_004.txt	1176	633	565	46
World_005.txt	402	206	230	48
World_006.txt	1584	858	716	45
World_007.txt	1523	850	714	44
World_008.txt	1157	632	584	45
World_009.txt	434	233	252	46

- **Porcentaje de stopwords Peter Preston**

Textos	Cantidad de Palabras	Sin Stopwords	Vocabulario	% de Stopwords en el Texto
World_006.txt	785	467	460	40
World_007.txt	964	538	535	44
World_008.txt	979	548	523	44
World_009.txt	957	535	505	44
World_010.txt	971	555	522	42
World_011.txt	1011	537	511	46
World_012.txt	999	540	499	45
World_013.txt	997	531	498	46
World_014.txt	1133	621	547	45
World_015.txt	1003	545	524	45
World_016.txt	1303	741	667	43
World_017.txt	975	557	525	42
World_018.txt	813	443	449	45
World_019.txt	974	530	468	45
World_020.txt	982	547	532	44
World_021.txt	993	508	486	48
World_022.txt	769	405	416	47
World_023.txt	995	564	541	43

World_024.txt	966	519	515	46
World_025.txt	921	508	509	44
World_026.txt	949	538	516	43
World_027.txt	814	453	440	44
World_028.txt	953	519	485	45
World_029.txt	961	516	478	46
World_030.txt	1005	535	528	46
World_031.txt	1013	529	546	47
World_032.txt	1002	538	523	46
World_033.txt	1120	598	561	46
World_034.txt	872	460	454	47
World_035.txt	1161	621	560	46
World_036.txt	1181	622	591	47
World_037.txt	960	526	508	45
World_038.txt	1162	641	590	44
World_039.txt	1164	640	594	45
World_040.txt	1163	599	551	48
World_041.txt	900	447	463	50
World_042.txt	1174	610	582	48
World_043.txt	1179	633	583	46
World_044.txt	1160	621	570	46
World_045.txt	1148	636	581	44
World_046.txt	1172	658	582	43
World_047.txt	1152	580	497	49
World_048.txt	1172	664	555	43
World_049.txt	1156	573	542	50
World_050.txt	1163	633	575	45
World_051.txt	1872	1015	810	45
World_052.txt	705	406	395	42
World_053.txt	1139	585	544	48
World_054.txt	1632	963	771	40
World_055.txt	1189	609	554	48
World_056.txt	1337	739	606	44
World_057.txt	688	392	408	43
World_058.txt	637	363	383	43
World_059.txt	526	297	325	43
World_060.txt	593	325	344	45
World_061.txt	1158	659	591	43
World_062.txt	1132	612	571	45
World_063.txt	1144	641	582	43
World_064.txt	1141	579	531	49

World_065.txt	1141	607	558	46
World_066.txt	1076	562	512	47

- **Porcentaje de stopwords Polly Toynbee**

Textos	Cantidad de Palabras	Sin Stopwords	Vocabulario	% de Stopwords en el Texto
Society_007.txt	1235	704	599	42
Society_008.txt	1170	660	578	43
Society_009.txt	997	556	486	44
Society_010.txt	995	566	459	43
Society_011.txt	1195	704	525	41
Society_012.txt	1238	697	573	43
Society_013.txt	1133	639	556	43
Society_014.txt	1191	659	517	44
Society_015.txt	1150	661	539	42
Society_016.txt	1237	688	550	44
Society_017.txt	1170	641	551	45
Society_018.txt	1219	709	555	41
Society_019.txt	1188	640	538	46
Society_020.txt	1158	625	504	46
Society_021.txt	1163	674	564	42
Society_022.txt	1121	673	548	39
Society_023.txt	1151	653	547	43
Society_024.txt	1151	653	547	43
Society_025.txt	1117	643	537	42
Society_026.txt	968	516	476	46
Society_027.txt	1180	649	557	45
Society_028.txt	1259	690	561	45
Society_029.txt	1197	656	539	45
Society_030.txt	1224	664	511	45
Society_031.txt	1152	659	570	42
Society_032.txt	1278	671	513	47
Society_033.txt	1196	635	493	46
Society_034.txt	1153	661	518	42
Society_035.txt	1154	632	517	45
Society_036.txt	1182	672	570	43
UK_001.txt	1213	665	543	45
UK_002.txt	1211	649	558	46

UK_003.txt	1159	633	536	45
UK_004.txt	1375	807	653	41
UK_005.txt	1226	646	571	47
World_001.txt	1177	649	539	44
World_002.txt	1178	659	617	44
World_003.txt	4550	2453	1569	46
World_004.txt	1179	654	599	44
World_005.txt	1159	644	570	44
World_006.txt	1179	656	595	44
World_007.txt	1185	690	596	41
World_008.txt	1171	680	597	41
World_009.txt	1211	633	557	47
World_010.txt	1233	687	565	44
World_011.txt	1205	654	555	45
World_012.txt	1228	636	538	48

- **Porcentaje de stopwords Roy Hattersley**

Textos	Cantidad de Palabras	Sin Stopwords	Vocabulario	% de Stopwords en el Texto
Politics_015.txt	825	421	389	48
Politics_016.txt	832	429	420	48
Politics_017.txt	827	415	397	49
Politics_018.txt	820	429	402	47
Politics_019.txt	812	425	406	47
Politics_020.txt	822	426	426	48
Politics_021.txt	745	408	374	45
Politics_022.txt	729	388	364	46
Society_001.txt	817	412	392	49
Society_002.txt	795	409	398	48
Society_003.txt	784	412	419	47
Society_004.txt	797	423	390	46
UK_001.txt	825	409	404	50
UK_002.txt	823	402	393	51
UK_003.txt	819	421	439	48
UK_004.txt	829	423	404	48
UK_005.txt	816	401	388	50
UK_006.txt	929	460	442	50
UK_007.txt	826	426	441	48

UK_008.txt	744	382	366	48
UK_009.txt	1751	1047	775	40
UK_010.txt	835	396	396	52
UK_011.txt	869	415	441	52
UK_012.txt	816	424	412	48
UK_013.txt	727	349	357	51
UK_014.txt	878	437	432	50
UK_015.txt	871	435	443	50
World_001.txt	849	407	393	52
World_002.txt	831	428	418	48
World_003.txt	807	419	407	48

- **Porcentaje de stopwords Simon Hoggart**

Textos	Cantidad de Palabras	Sin Stopwords	Vocabulario	% de Stopwords en el Texto
Politics_040.txt	1047	510	516	51
Politics_041.txt	515	272	276	47
Politics_042.txt	507	269	276	46
Politics_043.txt	1019	517	535	49
Politics_044.txt	498	255	282	48
Politics_045.txt	604	302	297	50
Politics_046.txt	419	238	262	43
Politics_047.txt	1054	520	505	50
Politics_048.txt	580	297	304	48
Politics_049.txt	469	240	248	48
Politics_050.txt	551	261	271	52
Politics_051.txt	432	241	264	44
Politics_052.txt	538	264	296	50
Politics_053.txt	537	251	281	53
Politics_054.txt	482	242	264	49
Politics_055.txt	489	255	273	47
Politics_056.txt	490	238	256	51
Politics_057.txt	508	251	284	50
Politics_058.txt	586	302	312	48
Politics_059.txt	497	272	274	45
Politics_060.txt	550	277	288	49
Politics_061.txt	556	298	294	46
Politics_062.txt	788	433	446	45
Politics_063.txt	469	244	270	47

Politics_064.txt	1046	555	525	46
Politics_065.txt	514	275	283	46
Politics_066.txt	574	288	317	49
Politics_067.txt	618	322	348	47
Politics_068.txt	666	346	340	48
Politics_069.txt	582	310	324	46
Politics_070.txt	486	257	268	47
Politics_071.txt	544	282	301	48
Politics_072.txt	530	280	278	47
Politics_073.txt	1037	544	551	47
Politics_074.txt	587	315	304	46
Politics_075.txt	987	538	520	45
Politics_076.txt	561	290	321	48
Politics_077.txt	572	305	314	46
Politics_078.txt	933	480	468	48
Politics_079.txt	548	282	303	48
Politics_080.txt	683	336	355	50
Politics_081.txt	724	366	365	49
Politics_082.txt	612	316	339	48
Politics_083.txt	571	315	306	44
Politics_084.txt	510	270	303	47
Politics_085.txt	549	279	311	49
Politics_086.txt	536	277	297	48
Politics_087.txt	618	318	311	48
Politics_088.txt	1009	525	502	47
Politics_089.txt	543	269	299	50
Politics_090.txt	514	262	297	49
Politics_091.txt	1025	561	550	45
Politics_092.txt	512	268	292	47
Politics_093.txt	532	287	294	46
Politics_094.txt	589	312	300	47
Politics_095.txt	474	246	288	48
Politics_096.txt	1079	584	538	45
Politics_097.txt	1033	561	554	45
Politics_098.txt	885	460	483	48
Politics_099.txt	479	261	289	45
Politics_100.txt	527	287	306	45
Society_001.txt	645	312	326	51
Society_002.txt	680	344	341	49
Society_003.txt	634	334	335	47
Society_004.txt	599	329	335	45

Society_005.txt	599	329	335	45
UK_001.txt	1143	603	574	47
UK_002.txt	638	337	331	47
UK_003.txt	599	329	335	45
UK_004.txt	714	384	380	46
UK_005.txt	654	345	342	47
UK_006.txt	655	328	329	49
World_001.txt	448	237	266	47
World_002.txt	488	260	279	46
World_003.txt	605	303	323	49
World_004.txt	628	342	366	45
World_005.txt	634	336	347	47

- **Porcentaje de stopwords Will Hutton**

Textos	Cantidad de Palabras	Sin Stopwords	Vocabulario	% de Stopwords en el Texto
Society_001.txt	1181	592	501	49
Society_002.txt	1343	694	554	48
Society_003.txt	1053	534	483	49
Society_004.txt	1211	649	481	46
Society_005.txt	1151	620	520	46
Society_006.txt	1186	631	491	46
UK_001.txt	1124	578	506	48
UK_002.txt	1162	612	539	47
UK_003.txt	1167	641	533	45
UK_004.txt	1155	602	535	47
UK_005.txt	1166	611	536	47
World_001.txt	1095	546	459	50
World_002.txt	904	496	447	45
World_003.txt	1186	688	551	41
World_004.txt	1178	649	573	44
World_005.txt	1188	649	566	45
World_006.txt	1165	630	550	45
World_007.txt	1169	630	538	46
World_008.txt	1174	637	531	45
World_009.txt	1303	721	563	44
World_010.txt	1214	638	566	47
World_011.txt	1207	623	496	48

World_012.txt	1131	637	567	43
World_013.txt	1066	544	479	48
World_014.txt	999	553	485	44
World_016.txt	1253	661	553	47
World_017.txt	1221	640	539	47
World_018.txt	1328	737	561	44
World_020.txt	1257	707	609	43
World_021.txt	1090	592	484	45
World_022.txt	1390	763	661	45

- **Porcentaje de stopwords Zoe Williams**

Textos	Cantidad de Palabras	Sin Stopwords	Vocabulario	% de Stopwords en el Texto
Society_010.txt	862	400	390	53
Society_011.txt	14	5	13	64
Society_012.txt	821	408	408	50
Society_013.txt	895	437	427	51
Society_014.txt	914	450	429	50
UK_001.txt	538	268	297	50
UK_002.txt	681	308	342	54
UK_003.txt	705	338	359	52
UK_004.txt	809	375	413	53
UK_005.txt	914	454	445	50
UK_006.txt	921	415	421	54
World_001.txt	527	255	255	51
World_002.txt	673	333	337	50
World_003.txt	855	419	410	50
World_004.txt	912	456	448	50
World_005.txt	862	403	416	53
World_006.txt	785	393	394	49
World_007.txt	854	432	432	49
World_008.txt	859	407	412	52
World_009.txt	923	443	437	52
World_010.txt	829	417	432	49
World_011.txt	825	403	389	51
World_012.txt	913	452	450	50
World_013.txt	807	422	410	47
World_014.txt	893	438	437	50