UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES
ACATLÁN

ANÁLISIS SEMÁNTICO Y SENTIMENTAL DE LOS MISERABLES

TESINA

QUE PARA OBTENER EL TÍTULO DE
LICENCIADO EN ACTUARÍA

PRESENTA:
FERNANDO SOTO BARAJAS

ASESORA:
DRA. MARÍA DEL CARMEN GONZÁLEZ VIDEGARAY

SANTA CRUZ ACATLÁN, NAUCALPAN, ESTADO DE MÉXICO
FEBRERO 2020

Dedicatorias

A ti, que eres el camino, la verdad y la vida.

A mi mamá, por apoyarme durante mi etapa universitaria y siempre buscar mi seguridad y felicidad, además de estar al pendiente de mis necesidades personales y emocionales durante esta etapa. Te amo, ma.

A mi papá, por haberme inculcado un pensamiento crítico y racional, y actitudes que son parte importante de mi identidad. Te amo, pa.

A mi tía Nena, por ser el pilar de mi educación universitaria, y desinteresadamente alentarme a concluir y motivarme diariamente sobre la importancia de alcanzar esta meta. Este logro es tuyo también, tía. Te quiero mucho, ñeñe.

A mi abuela Kucu, por todo tu cariño que me dio confianza desde muy niño y hasta ahora, siempre has sido parte importante de mí. Te amo, Kucu.

A mis hermanos, Isaac y Ricardo, por inspirarme día a día a conseguir mis sueños, sin ustedes no habría sido igual de feliz mi crecimiento, ni mis experiencias. Los amo mucho.

A Abisaí, mi mejor amigo de la universidad, por escucharme, por compartir momentos *diver*, por todas esas noches de fiesta, y de estrés para prepararnos para los exámenes finales. Y no faltaron los que pusieron el ambiente. Te quiero mucho.

A Blanca, por estar al pendiente de mí, sin importar la distancia, y darme palabras de ánimo, sacándome de la rutina siempre siendo espontánea. Te quiero mucho, amiguita.

A Paola, por estar en los momentos más felices y más tristes de la universidad. Te quiero mucho, beba.

A Uriel, Adriana y Mariana, por estar en estos últimos semestres compartiendo momentos en el servicio social y estando al pendiente de mí.

Muchas gracias por acompañarme en esta gran etapa,

Fernando.

Agradecimientos

A la Universidad Nacional Autónoma de México, con la que siempre le estaré completamente agradecido por su apoyo en esta etapa de mi vida, dotándome de conocimiento, amistades, idiomas y experiencias, todos éstos invaluables. Es, la Universidad, el proyecto más grande de la nación y es fruto del trabajo de todos los mexicanos que generosamente contribuyen a esta causa.

A la Facultad de Estudios Superiores Acatlán, a través del Programa de Actuaría, donde estuve por 5 años de lunes a sábado y siempre me hizo sentir acogido entre sus instalaciones, brindándome profesores que nutrieron mi aprendizaje y me otorgaron una formación integral.

A la Dra. Maricarmen González Videgaray por todo el apoyo brindado en la investigación, siendo la inspiración de este proyecto a través de la materia de Seminario de Investigación, reafirmando su vocación para la enseñanza y la promoción de la vida académica.

A la Act. Luz María Lavín Alanís quien estuvo al pendiente de todo el proceso de titulación y durante mi estancia en la facultad, siempre dispuesta a escucharme y aconsejarme sobre cómo encontrar soluciones. Le agradezco mucho.

Al Mtro. Erick Iván García Reyes por siempre ser un gran amigo, tutor y profesor, y estar cuando más lo he necesitado académicamente, siempre exigiéndome para que obtuviera los mejores resultados.

A todos los profesores que conocí durante la carrera y que me demostraron un cariño leal a la universidad y una labor desinteresada para la educación.

Sinceramente,

Fernando.

“Logic and mathematics are nothing but specialized linguistic structures.”

– Jean Piaget.

Índice

Abstract

Currently there are many algorithmic methods to analyze large sets of text without having to read them one by one, which optimizes the ability to understand. In this research, by using R, a statistical software, it is verified how to carry out the sentimental analysis that allows to know the feelings, emotions, topics and attitudes that describe the classic work "Les Misérables", in its digital version and in english. At the same time, semantic analysis is used to describe the meaning of words. The relationship of the characters and their situations is found using: the correlation from the proximity of the words and LDA using their probabilities. As well as describing the verbs to obtain most of the grammatical elements that illustrate the interpretation of the text. Cloud graphs, bigrams and bar graphs show the whole analysis. Identifying the time of Louis XVIII in France, where Jean Valjean, a convict, Cosette, a sweet girl, and Marius, who fell in love after Cosette, carry out the narration. Which is composed of mostly negative moments, detailed by pictures of poverty, war, prohibition, sadness, lack of love and death; although there is also courage, affection and success too.

Resumen

En la actualidad existen distintos métodos algorítmicos para analizar grandes conjuntos de texto sin tener que leerlos uno por uno, lo que optimiza la capacidad de comprensión. En esta investigación, a través de R, un software estadístico, se verifica cómo llevar a cabo el análisis de sentimientos que permite conocer los sentimientos, emociones, temas y actitudes que describen a la obra clásica "Los Miserables", en su versión digital y en idioma inglés. Al mismo tiempo se utiliza el análisis semántico para describir el significado de las palabras. Se encuentra la relación de los personajes y sus situaciones usando: la correlación a partir proximidad de las palabras y LDA utilizando sus probabilidades. Así como lograr describir los verbos para obtener la mayor parte de elementos gramaticales que ilustran la interpretación del texto. Las gráficas de nube, bigramas y gráficas de barras muestran todo el análisis. Identificando la época de Luis XVIII en Francia, donde Jean Valjean, un convicto, Cosette, una niña dulce, y Marius, quien se enamorara después de Cosette, llevan a cabo la narración. La cual está compuesta de momentos en su mayoría negativos, detallados por cuadros de pobreza, guerra, prohibición, tristeza, desamor y muerte; aunque también se halla la valentía, el cariño y el éxito.

Palabras Clave

Análisis de Sentimientos, Minería de Opinión, Semántica, Lingüística, NLP.

Keywords

Sentimental Analysis, Opinion Mining, Semantics, Linguistics, NLP.

1 Introducción

1.1 Marco Teórico

1.1.1 Género Literario: Novela

La novela (Pérez-Torres, 2014) es un subgénero literario, parte de la narrativa. Es una narración ordenada y completa de sucesos ficticios y humanos, hasta cierto punto creíbles. Una de sus características principales es su extensión, ya que es una de las más largas, a comparación del cuento; su complejidad también es alta debido a que relata varios temas. Suele contar con varios capítulos y aparecen muchos personajes.

1.1.2 Lingüística

Según la RAE (Real Academia Española, 2019a), la lingüística es la ciencia del lenguaje. Para hablar de lingüística es necesario considerar a Noam Chomsky como el precursor de la lingüística moderna. Chomsky (Chomsky, 1955) abre el panorama hacia la relación entre la lingüística y la lógica; en este caso, la lógica proveniente de la sintaxis y las adecuaciones gramaticales.

Chomsky introduce un término fundamental: "La gramática generativa". Brevemente, la gramática generativa es una teoría que establece que sobre una lengua, mediante normas o principios, se pueden predecir, o "generar", combinaciones gramaticalmente correctas. Dentro de estas normas y principios se hallan la sintaxis, la semántica y la fonología. El componente fonológico crea la última fase de la gramática, aunque su función es la emisión de la lengua. Por lo anterior, la fonética no se considera para el análisis literario.

Además, Aguilar (Aguilar, 2004) resume las ideas de la obra de Noam, "Estructuras Sintácticas", donde expone la idea de: "Chomsky objeta que existe un infinito número de oraciones en cada lengua por lo tanto tenemos que asumir que los seres humano están equipados con un mecanismo finito de conocimiento que les permite construir e interpretar un infinito número de oraciones. Este sistema finito de principios es conocido como "la gramática interna del lenguaje".". Es a partir de esta idea donde surge la esperanza de regular el lenguaje, sin subestimar la dinámica natural de la lengua. La gramática es el agente finito del lenguaje.

1.1.2.1 Sintaxis

Para describir la gramática, es importante consultar una referencia estándar. Es por eso, que la investigación se basa en la guía de Eastwood (Eastwood, 1994) que ha sido presentada por la universidad de Oxford. Es útil basarnos en la gramática inglesa, ya que nuestro texto está en dicha lengua.

Existen ocho tipos principales de oración en inglés: Verbo, sustantivo, adjetivo, adverbio, preposición, determinante, pronombre y conjunción. El verbo, sustantivo, adjetivo y adverbio son consideradas como palabras de *vocabulario*, ya que son múltiples y pueden surgir más. A las demás se les conoce como palabras *gramaticales*, puesto que están bien definidas y brindan estructura.

También existe un análisis de modo, persona, tiempo y voz para el verbo, que brinda información temporal y está estrechamente relacionado con el sustantivo. El modo, en inglés, no es problema, ya que el modo subjuntivo, que describe duda o deseo, sólo se reconoce por medio del contexto, pero no implica gran diferencia del modo indicativo. La persona puede ser singular o plural, y se definen 3 casos para cada número; mientras que existen formas impersonales como el infinitivo, participio y gerundio. Los tiempos se establecen como presente, pasado y futuro, cada uno puede estar dotado de un auxiliar para formar tiempos compuestos. Las voces son pasivas y activas, y establecen quién genera la acción y sobre qué o quién. Por último, existen verbos modales tales como "querer", "deber", etcétera, que necesitan dotarse de un infinitivo para cobrar sentido.

Las funciones gramaticales, para las lenguas indoeuropeas, surgen basadas en la morfología grecolatina, que en un principio contaba con una estructura gramatical basada en declinaciones: desinencias que acompañaban al radical. Las declinaciones han sido suprimidas como parte de la evolución de las lenguas, aunque algunas aún las consideran, tal como el alemán; sin embargo, ya no aparecen como desinencias del sustantivo, que era la forma original, sino en aditamentos (clíticos) que aparecen como preposiciones y artículos. Las declinaciones latinas, que son las que mejor describen la gramática europea, son 6: nominativo, genitivo, dativo, acusativo, vocativo y ablativo. Éstas devinieron en sujeto, propiedad del sujeto, objeto indirecto, objeto directo, apelación y complemento circunstancial (modo, tiempo y lugar), respectivamente; en cuanto al sustantivo refiere. E intrínsecamente todo lo que describa a un sustantivo se le catalogará como adjetivo, que es en parte el objeto de estudio.

1.1.2.2 SEMÁNTICA

La semántica es el concepto lingüístico que explica el significado de las palabras en la oración, generalmente su intención es agruparlas o hallar cierta conexión. Goodenough y Rubenstein (Goodenough y Rubenstein, 1965) explican que gracias a la semántica es posible establecer una sinonimia, dependiente de su ubicación en el texto. Cambria y Hussain (Cambria y Hussain, 2012) definen a la semántica como la rama que provee de información conceptual y que relaciona los conceptos del lenguaje natural.

Sobre la semántica (Chomsky, 1955) existen dos clasificaciones: La semántica de sentido y la semántica de referencia.

E. g.

i) Hernán Cortés incursionó en América.

ii) Un admirable conquistador incursionó en América.

iii) Aquel despiadado conquistador incursionó en América.

"Hernán Cortés", "un admirable conquistador" y "aquel despiadado conquistador" pueden expresar la misma referencia: Hernán Cortés. Pero el sentido de las oraciones es diferente, en una el sentido es neutro, mientras que en las otras dos está polarizado.

1.1.2.3 HEURÍSTICA

La heurística (Real Academia Española, 2019b), proveniente del griego εὑρίσκειν, significa hallar, y en una de sus acepciones es considerada como un método empírico que busca hacer un análisis no riguroso, sino manipulando un poco la información hacia las preguntas que deseamos encontrar. Especialistas en el análisis de texto (Cambria y Hussain, 2012; Mirończuk, Protasiewicz y Pedrycz, 2019) concluyen que es necesaria la heurística para la clasificación de las palabras, ya que la precisión absoluta no es aplicable.

1.1.3 ANÁLISIS MATEMÁTICO

Esta sección está destinada a explicar las generalidades matemáticas que aparecerán en la investigación. Entre las definiciones probabilísticas se utilizará la covarianza, y de ahí la correlación. Para formalizar la forma de medir, se ocupará la norma que requiere algunos conceptos adicionales como campo, espacio vectorial y métrica; así como propiedades métricas.

1.1.3.1 ANÁLISIS DE CORRELACIÓN

El análisis de correlación es un buen método para identificar similitudes entre dos elementos, en este sentido pueden verse como variables. De ahí, es posible describir la covarianza como qué tanto difiere una variable de otra, y la correlación, de la que existen muchos tipos, es una medida con un rango limitado de mejor lectura.

$$Cov(X,Y) = E\big((X-\mu_X)(Y-\mu_Y)\big). \quad (1)$$

Ecuación 1: Covarianza. Fuente: (Casella y Berger, 2002).

De la Ecuación 1, se desprende la covarianza muestral.

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{n-1}. (2)$$

Ecuación 2: Covarianza Muestral. Fuente: (Amat Rodrigo, 2016).

Usualmente, el coeficiente de correlación de Pearson es el que tiene propiedades que pueden resultar útles y sencillas de entender.

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}. (3)$$

Ecuación 3: Coeficiente de Correlación. Fuente: (Casella y Berger, 2002).

La Ecuación 3 expresa la correlación poblacional, aunque también existe la fórmula muestral, que es la que se usa con mayor frecuencia.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}. (4)$$

Ecuación 4: Coeficiente muestral de Correlación. Fuente: (Amat Rodrigo, 2016).

La covarianza por sí misma no muestra cuán apegadas puedan ser las variables X y Y; es por eso que la Ecuación 4, el coeficiente de correlación, sí describe si hay dependencia lineal o no, tomando valores entre -1 y 1. Mientras el coeficiente más se aproxime a estas dos fronteras, podemos inferir que existe correlación entre las variables.

1.1.3.2 NORMA

La norma es una medida que sirve para un conjunto de vectores, *espacio vectorial*, y contiene propiedades útiles fundamentadas en la distancia, pero permitiendo el uso de operaciones.

Definición 1 (Campo). Fuente: (Lang, 2000).

Sea *K* un conjunto de números reales $(\mathbb{R})$ o complejos $(\mathbb{C})$. *K* es un *campo* si cumple las siguientes propiedades: (Cambiar a notación escrita)

i) Si *x, y* son elementos de *K*, entonces la adición y el producto también son elementos de *K*.
ii) Si $x \in K$, entonces el elemento inverso (*-x*) es también un elemento de *K*. Si $x \neq 0$, entonces el inverso multiplicativo (x^{-1}) también es un elemento del campo *K*.
iii) Los elementos 0 y 1 son elementos neutros, para la adición y multiplicación respectivamente, de *K*.

Nota bene: A los elementos de *K*, habitualmente, se les llama: *escalares*.

Proposición 1. Fuente (Lang, 2000).

Los números reales $(\mathbb{R})$ y números complejos $(\mathbb{C})$ son campos, respectivamente.

Definición 2 (Espacio Vectorial). Fuente: (Lang, 2000).

Un espacio vectorial *V* sobre un campo *K* es un conjunto de objetos que cuentan con la adición y la multiplicación por escalares, tal que la suma de elementos de *V* pertenece a *V*, el producto de un elemento de *V* por un elemento de *K* pertenece a *V*, y se cumplen las siguientes propiedades:

i) La adición es asociativa.
ii) Existe el elemento neutro para la adición.
iii) Existe el elemento inverso para la adición.
iv) La adición es conmutativa.
v) La multiplicación por escalar distribuye sobre la suma de vectores.
vi) La suma de escalares distribuye sobre un vector.
vii) La multiplicación de dos escalares por un vector es asociativa.
viii) Existencia del elemento neutro para la multiplicación.

Nota bene: A los elementos de *V*, se les conoce como: *vectores*.

Definición 3 (Métrica). Fuente: (Clapp, 2010).

Sea X un conjunto. Una *métrica* (o *distancia*) en X es una función $d: X \times X \rightarrow \mathbb{R}$ que tiene las siguientes propiedades:

i) $d(x,y) = 0$ si y sólo si $x=y$.

ii) $d(x,y) = d(y,x)$ para todos $x, y \in X$.
iii) $d(x,z) \leq d(x,y) + d(y,z)$ para todos $x, y, z \in X$. (Desigualdad del triángulo)

Nota bene: A un conjunto X dotado por una métrica *d*, se le llama *espacio métrico*; y se le caracteriza como (X, *d)*.

Un caso particular de la distancia es la medida de *similiraridad*, que se define a continuación.

Definición 4 (Similaridad). Fuente:(De la Fuente Fernández, 2011).

Sea X un conjunto. Una función $d: X \times X \rightarrow \mathbb{R}$ se denomina *medida de similaridad* si $\forall\, x, y \in X$, y s_0 un real finito arbitrario, cumple las siguientes propiedades:

i) $d(x,y) \leq s_0$
ii) $d(x,x) = s_0$
iii) $d(x,y) = d(y,x)$

Si se considera el valor absoluto de la correlación lineal, Ecuación 4, puede verse a ésta como una *similaridad*. Incluso, puede expresarse como:

$$d_{cor}(x,y) = 1 - |r_{xy}|. \quad (5)$$

Ecuación 5: Medida de similaridad de Correlación. Fuente: (Amat Rodrigo, 2017).

Demostración

i) Dado que la suma de funciones acotadas es acotada y $|r_{xy}| \in [0,1]$, se sigue que $d_{cor}(x,y) \leq 1$.

ii) $$d_{cor}(x,x) = 1 - |r_{xx}| = 1 - \frac{\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(x_i-\bar{x})^2}}$$
$$= 1 - \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}}$$
$$= 1 - \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$
$$= 1 - 1 = 0$$

iii) Es inmediato notar que $|r_{xy}| = |r_{yx}|$, ya que el producto es conmutativo, así $d_{cor}(x,y) = d_{cor}(y,x)$.

Por lo tanto d_{cor} es una medida de similiaridad.

□

La Ecuación 5, es una función que va del espacio de variables aleatorias cuya covarianza existe al intervalo [0,1].

Así, podemos observar las siguientes propiedades:

i) Si r_{xy} es 1, esto implica que están estrechamente relacionados y, por lo tanto, son el mismo elemento, por eso la distancia sería 0. Así también, si r_{xy} es 0, la distancia sería 1, y esto implica que son independientes, o que no existe relación.
ii) El orden de los factores no altera a la correlación.

Proposición 2 (Distancia no negativa). Fuente: (Clapp, 2010).

$$d(x,y) \geq 0 \text{ para todos } x, y \in X.$$

Demostración

Usando las propiedades de métrica,

$$0 = d(x,x) \leq d(x,y) + d(y,x) = 2d(x,y).$$

Se sigue que la distancia es positiva o, a lo menos, nula.

□

Definición 5 (Norma). Fuente: (Clapp, 2010).

Sea V un espacio vectorial sobre el campo $\mathbb{R}$. Una *norma* en V es una función $\|\cdot\|: V \rightarrow \mathbb{R}$ que tiene las siguientes propiedades:

i) $\|v\| = 0$ si y sólo si $v = 0$.
ii) $\|\lambda v\| = |\lambda|\|v\|$ para todos $v \in V$, $\lambda \in \mathbb{R}$.
iii) $\|v + w\| \leq \|v\| + \|w\|$ para todos $v, w \in V$.

Nota bene: A un espacio vectorial *V*, dotado de una norma, se le denomina *espacio normado.* Al espacio normado se le caracteriza como $(V, \|\cdot\|)$.

La norma explica la dirección de los vectores.

Proposición 3. Fuente: (Clapp, 2010).

Todo espacio normado $(V, \|\cdot\|)$ es un espacio métrico con la métrica dada por:

$$d(v,w) = \|v - w\|.$$

Demostración

Sean x, y, z vectores del espacio vectorial V.

i) $\rightarrow d(x,y) = \|x - y\| = 0$
como $\|.\|$ es una norma, y toda norma es igual a 0, sí y sólo si su vector interior es 0. Así $x - y = 0$, entonces $x = y$.
$\leftarrow$Sea $x = y$, entonces $\|x - y\| = 0$.
Siendo la definición de la métrica. Así, $d(x,y) = \|x - y\| = 0$
ii) $d(x,y) = \|x - y\| = \|-1 * (y - x)\| = |-1|\|y - x\| = d(y,x)$
iii) $d(x,z) = \|x - z\| = \|x - z + (y - y)\| \leq \|x - y\| + \|y - z\|$
$= d(x,y) + d(y,z)$

Por lo tanto, es espacio métrico. □

La utilidad de la **Proposición 3** versa en su aplicación a un espacio vectorial, como es el caso de los números reales para n dimensiones.

1.1.3.3 Asignación Latente de Dirichlet

1.1.3.3.1 Teoría de Probabilidad

Definición 6 (Espacio Muestral) Fuente:(Casella y Berger, 2002).

El conjunto, *S*, de todos los posibles resultados de un experimento en particular es llamado *espacio muestral* del experimento.

Definición 7 (Evento) Fuente: (Casella y Berger, 2002).

Un *evento* es cualquier colección de posibles resultados de un experimento, que es, cualquier subconjunto de *S* (Incluido *S*).

Definición 8 (Sigma álgebra) Fuente: (Casella y Berger, 2002).

Una colección de subconjuntos de *S* es llamada una *sigma álgebra*, denotada por $\mathcal{B}$, si satisface las siguientes tres propiedades:

i) $\emptyset \in \mathcal{B}$.
ii) Si $A \in \mathcal{B}$, entonces $A^c \in \mathcal{B}$.
iii) Si $A_1, A_2, \ldots \in \mathcal{B}$, entonces $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$.

Definición 9 (Medida de Probabilidad) Fuente: (Casella y Berger, 2002)

Dado un espacio muestral *S* y una sigma álgebra asociada $\mathcal{B}$, una *medida de probabilidad* (ó *función de probabilidad*) es una función *P* con dominio en $\mathcal{B}$ que satisface:

i) $P(A) \geq 0 \, , \forall A \in \mathcal{B}$.
ii) $P(S) = 1$.
iii) Si $A_1, A_2, \ldots \in \mathcal{B}$ son eventos mutuamente excluyentes, entonces
$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

1.1.3.3.2 Teorema de Bayes

Definición 10 (Probabilidad Condicional) Fuente: (Casella y Berger, 2002).

Si *A* y *B* son eventos del espacio muestral *S*, *P* una medida de probabilidad y *P(B)>0*, entonces la *probabilidad condicional de A dado B*, escrita como *P*(*A*|*B*), es

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Teorema 1 (Regla de Bayes) Fuente: (Casella y Berger, 2002).

Sea $A_1, A_2, \ldots$ una partición del espacio muestral, y sea *B* cualquier conjunto. Entonces, para cada $i = 1,2, \ldots,$

$$\mathrm{P}(\mathrm{A_i}|\mathrm{B}) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} P\left(B\middle|\, A_j\right)P\left(A_j\right)}.$$

1.1.3.3.3 MCMC (CADENAS DE MARKOV CON MONTE CARLO)

Definición 11 (Proceso Estocástico) Fuente: (Rincón, 2012).

Un *proceso estocástico* es una colección de variables aleatorias $\{X_t: t \in T\}$ parametrizada por un conjunto *T* (como puede ser el tiempo), llamado espacio parametral, en donde las variables toman valores en un conjunto *S* llamado espacio de estados.

Definición 12 (Propiedad de Markov) Fuente: (Cowles, 2013).

$$P(X_t|X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \dots, X_0 = x_0) = P(X_t|X_{t-1} = x_{t-1})$$

Se dice que un proceso estocástico es una *Cadena de Markov* si cumple la **Definición 12**.

La simulación Monte Carlo (Robert y Casella, 2010) es útil ya que en estadística bayesiana nos ayuda a encontrar una aproximación a la distribución de la que provienen los datos, pero también para encontrar la distribución posterior de los parámetros.

Liu y Clark (Clark, 2013; Liu, 2015) explican que si se desea conseguir una clasificación temática con mayor apego matemático (básandose fuertemente en el Teorema de Bayes y MCMC (Cadenas de Markov con Monte Carlo)), mediante una técnica que permite generalizar, la Asignación Latente de Dirichlet (LDA, por sus siglas en inglés) es el modelo por *default.* Este modelo forma parte de la familia de los modelos no supervisados.

Es una técnica de agrupamiento, *clustering*, que a diferencia de k-medias (o *k-means*), se basa en los patrones que existen, latentes u ocultos, dentro el texto. Mientras que k-medias, sólo utiliza las distancias de los elementos para generar agrupaciones.

Dados un conjunto de palabras, provenientes de un documento (pudiendo ser una oración, o capítulo en nuestro estudio), se puede usar LDA para conocer la distribución de cada palabra asociado a un tema. Así, formalmente, la distribución temática para cada documento se define como:

$$\theta \sim Dirichlet(\alpha)$$

Donde *Dirichlet*(α) representa la distribución Dirichlet de parámetro α.

Las palabras también se distribuirán Dirichlet, únicamente con distintos parámetros:

$$\phi \sim Dirichlet(\eta)$$

El objetivo de LDA es estimar θ y ϕ que es básicamente para conocer qué palabras son importantes para cada tema y cuáles son importantes para cada documento, respectivamente.

La idea principal detrás de los parámetros para la distribución de Dirichlet es que mientras α sea más grande, va a ser más probable que cada documento contenga una mezcla de la mayoría de temas en vez de un tema en particular, buscando un sentido de convergencia para hallar la distribución. Así como para η, donde un valor alto implica que es más probable que cada tema contenga una mezcla de la mayoría de palabras y no una palabra específicamente.

El algoritmo más común de muestreo para implementar LDA es *Gibbs.* Este requiere de las palabras que se van a usar, los documentos y el número de temas propuestos. El algoritmo funciona de la siguiente manera:

1. Pasa por cada documento y aleatoriamente asigna a cada palabra dentro de los documentos a uno de los K temas. Además, crea una matriz que relaciona la palabra con el tema (*word-topic matrix*, **wtm**), y en ésta se cuentan las aparaciones de cada palabra que fue asignadas a cada tema; y una matriz de documento-tema (*document-topic matrix*, **dtm**) donde se encuentra el número de palabras asignadas a cada tema para cada documento.
2. Para cada documento *d*, usa cada palabra *w*. Reasigna un nuevo tema a *w*, donde se elige el tema *t* con la probabilidad de que la palabra *w* dado el tema *t* por la probabilidad del tema *t* dado un documento *d*, cuya fórmula es:

$$P(z_i = j \mid z_{-i}, w_i, d_i) = \frac{C_{wij}^{WT} + \eta}{\sum_{w=1}^{W} C_{wj}^{WT} + W\eta} \times \frac{C_{d_ij}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d_it}^{DT} + T\alpha}$$

Donde, del lado izquierdo del igual, $P(z_i = j)$ es la probabilidad de que el *token i* sea asignada al tema *j*. z_{-i} representa la asignación de temas para todas los demás *tokens*. w_i es la palabra (índice) de el $i_{ésimo}$ *token*. d_i es el documento que contiene al $i_{ésimo}$ *token*.

Para el lado derecho del igual, C^{WT} es la matriz **wtm**, $\sum_{w=1}^{W} C_{wj}^{WT}$ es el número total de *tokens* en cada tema, C^{DT} es la matriz **dtm**, $\sum_{t=1}^{T} C_{d_it}^{DT}$ es el número total de *tokens* del documento *i*. η es el parámetro que define la distribución temática para las palabras, mientras más alta, mayor dispersión habrá entre las palabras en el número de temas especificados por K. α es el parámetro que define la distribución temática para los documentos, mientras más alta, mayor dispersión entre los documentos dentro del número de temas especificados por K. *W* es el número total de palabras en el conjunto de documentos y *T* es el número de tmas, equivalente a la K definida antes.

Para conocer la probabilidad de una palabra dado el tema, se utiliza:

$$\phi_{ij} = \frac{C_{ij}^{WT} + \eta}{\sum_{k=1}^{W} C_{kj}^{WT} + W\eta}$$

Donde ϕ_{ij} es la probabilidad de la palabra *i* para el tema *j*.

También para θ aplica que es la proporción del tema *j* en el documento *d*:

$$\theta_{dj} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} C_{dk}^{DT} + T\alpha}$$

Normalmente lo deseable es conocer las palabras que tenga mayor probabilidad de pertenecer a cierto tema, en la práctica se filtra de esta forma.

1.1.4 Minería de Texto

La minería de texto, o de opinión, comprende una serie de herramientas, en su mayoría estadísticas, que a partir de *datos no estructurados* (en este caso, palabras), brinda un resumen en relación con sentimientos, temas y conexiones.

1.1.4.1 ANÁLISIS DE SENTIMIENTOS

El análisis de sentimientos es materia actual de la minería de opinión, su uso es esencial para la comprensión de grandes contenidos de texto que podrían tomar tiempo siendo analizados de forma manual. Esta propuesta digital recurre a herramientas computacionales y bases de datos de palabras para obtener sus propiedades sintácticas, gramáticas y semánticas. Una interpretación correcta del texto debe partir de qué quiso decir el emisor (He *et al.*, 2015). Si la clasificación del sentido está bien descrita, podrá ser más profunda y detallada.

Alrededor del análisis de sentimientos es importante describir distintos métodos (Stine, 2019): La polaridad del texto, el procesamiento natural del lenguaje (NLP, por sus siglas en inglés), la extracción de entidades (tales como sujetos u objetos), y la identificación del aspecto particular de la entidad; esto está estrechamente relacionado con las redes neuronales y el aprendizaje profundo.

Para el análisis de sentimientos se requieren bases de palabras en inglés que son con las cuales se comparan los adjetivos, existen tres muy importantes: Bing, AFINN y NRC (Ravi y Ravi, 2015); éstas las segmentan en positivas, negativas y neutras, asignándoles un valor numérico a cada palabra, de acuerdo al peso que le otorgan en promedio a cada oración. Son incluidas a menudo en paqueterías para la minería de texto. Asimismo, se han desarrollado paquetes para el análisis de correlación entre los pares de palabras, que proveen de características al análisis semántico, así como paquetes de LDA que ayudan al *Topic Modelling* (Modelado Temático).

1.1.4.2 CORPUS

El corpus es la base extraída del texto a analizar, y su importancia es tal ya que es la que será evaluada en el análisis de sentimientos. Para el análisis adecuado, se consideran únicamente sustantivos, verbos y adjetivos y adverbios, por sus características de palabras de vocabulario, y esa misma naturaleza permite reconocer su sentido.

Hacia una limpieza efectiva del corpus, se suelen utilizar técnicas como *Stemming* o *Lemmatization*, que extraen la raíz de la palabra. El *stemming* funciona retirando el sufijo de las palabras, y buscando relación entre ellas (***studi**-es , **study**-ing)* donde las "negritas" serían el *stem*, o el radical; mientras que en la *lexematización*, se emplea la raíz morfológica. En ***studi**-es , **study**-ing* el lema sería el mismo: *study*. La utilidad de estas técnicas es que reducen verboides y estructuras derivativas (Universidad Complutense de Madrid, 2013) dentro de las oraciones, todas ellas cuyo radical provino de un verbo o una palabra singular producto de la sinonimia. Al proceso automático de esta clasificación se le denomina*: Annotation*.

1.1.5 ANÁLISIS LITERARIO

Para exponer la utilidad intrínseca del análisis de sentimientos, sería prudente exhibir cómo se recomienda tratar a una obra de esta magnitud. Normalmente, una novela suele ser analizada mediante un análisis literario, que tiene la estructura de un ensayo. Es por eso que, a manera de método, McGee (McGee, 2001) enumera los siguientes atributos de un análisis correcto, dirigido a estudiantes:

1. Identificar el género al que pertenece.

2. Conocer a los personajes, ya que cuentan con características morales y psicológicas.
3. Saber si se trata de una ficción o un drama; o ambos.
4. Hallar el contexto, dónde se lleva a cabo la acción.
5. Descifrar los simbolismos, aquellos sellos plasmados por el autor.
6. Encontrar perspectivas culturales e históricas alrededor.
7. Establecer el comentario final, sobre la opinión propia.

1.2 ANTECEDENTES

Se puede situar a los orígenes de la minería de texto cuando se desea convertir a las palabras en valores numéricos: crear esquemas cuantitativos de información cualitativa. En los años 70, Zadeh (Zadeh, 1975b, 1975a) escribe un par de métodos que traerían un análisis mucho más profundo de la información escrita. En estos artículos, expone los problemas que serían los más comunes, tal como la problemática de cómo cuantificar a las palabras, la complejidad de las oraciones y la falta de referencias a contrastar, dada su naturaleza de *datos no estructurados*. Asimismo, categoriza en forma de variables, para establecer una relación semántica; el autor estuvo consciente de la importancia que implicaría, y que sería imposible desconsiderarla. Agrega conceptos lógicos, en forma de tablas de verdad, para plantear planos de oposición.

Ante esto, se consideró importante el correcto manejo y comprensión de la gramática (Elman, 1991) para conocer la distribución habitual del texto. Combinando la lingüística y el área matemática para crear algoritmos mucho más eficaces.

También, los lingüistas fueron trabajando en paralelo para llevar a cabo estos avances, mediante la resolución de bases de datos que clasificaran el sentido de las palabras, más allá de sus significados rigurosos. Un ejemplo claro es WordNet (Miller, 1995) que aparece su fecha de publicación en 1995, y que a la fecha es una herramienta clave para identificar el sentido de las palabras dentro de programas computacionales.

La lingüística es una materia de la que se toman muchos recursos, puesto que dota de estructura al lenguaje, y ésta será la que requerirá mayor trabajo y precisión. Sobre esto, Gibson (Gibson, 1998) explica las dependencias sintácticas, que algunas veces resultarán ambiguas, pero al mismo tiempo podremos usar las estructuras semánticas que apoyará finalmente a la precisión de la predicción.

En el 2000, Selker y Lieberman (Lieberman y Selker, 2000) proponían que el contexto, a través de la inteligencia artificial, establecería el nuevo rumbo del análisis de la información.

Cinco años después, los investigadores de la lengua (Wiebe, Wilson y Cardie, 2005) ya estaban conscientes de la indagación que interesaba al ámbito matemático sobre el Procesamiento Natural del Lenguaje, y que su propósito era automatizar la identificación de opiniones.

Ahora la categorización se habría convertido en la labor primordial a resolver: Entender de qué forma la información podía ser clasificada, para que no se obtuviera sesgo en el análisis. Sobre esto, muchos autores comenzaron a compartir teorías de categorización (Tomasello *et al.*, 2005; Barret, 2006; Niebles, Wang y Fei-Fei, 2008) con el objetivo de interpretar las emociones en un plano más sencillo para el análisis, pero afirmando que la labor computacional tendría que ser ardua.

En el mismo periodo, Chomsky (Chomsky, 2007) introduce el concepto de la biolingüística, la cual define como la rama de estudio actual más importante para la lingüística, alrededor de ella describiría al lenguaje humano y sus estructuras; buscando escenarios que respondan: ¿Qué, cómo y por qué? Teniendo en cuenta que la escritura plantea, como objetivo, comunicar una idea. En principio, el objetivo podría no ser explícito, sino que se encuentra inmerso en las palabras, pero conociendo las propiedades de estas palabras se ve como una tarea únicamente de análisis. A estas palabras se le conoce como funciones gramaticales, que son las herramientas que reconoce Chomsky en la biolingüística, las cuales tienen los mismos valores únicamente para las lenguas de la familia europea (Evans y Levinson, 2009).

Contribuirían los avances psicológicos para la minería de opinión, donde la cita siguiente se puede considerar como un resultado importante: “Los investigadores pueden ligar el uso de una palabra cotidiana con un arreglo externo de comportamientos del mundo real (Tausczik y Pennebaker, 2010)”. Y es tan importante puesto que es el vínculo que propiciará un estudio factible de la opinión, siendo la comunión entre la estadística y la psicología.

Una vez llegado el 2010, las redes sociales comenzaban a tener auge en ciertos grupos de la población mundial, pero los analistas descubrían, al mismo tiempo, los recursos de información que se podían extraer de esas plataformas, y que en un cierto punto serían datos suficientes para la investigación futura. Así es como Pak y Paroubek (Pak y Paroubek, 2010) proponen a Twitter como un base de datos para hacer análisis de sentimientos y minería de opinión. Señalando que en estas fuentes había mucha información, pero que estaba llena de caracteres u objetos que no servirían y que se debían limpiar de alguna forma.

Wing-ki Leung y otros (Wing-ki Leung *et al.*, 2011) mencionan que los tres pasos para concluir un análisis de sentimientos adecuado son: Identificar, Extraer y Clasificar los datos. De esta forma el *corpus* toma una importancia vital para el análisis, puesto que prescindir de él imposibilitaría llevar a cabo la encomienda.

Posteriormente, la minería de opinión (Wing-ki Leung *et al.*, 2011) ya era considerada útil, siendo la estadística su materia prima; y los usuarios fueron teniendo mayor acceso a programas que hicieron posible la aplicación (Collobert *et al.*, 2011) de dichos algoritmos que involucraban precisamente al Procesamiento Natural del Lenguaje. Sin embargo, surgieron autores tales como Liu, Cambria y Hussain (Cambria y Hussain, 2012; Liu, 2012) que regularizaban el conocimiento, a la época, del análisis de sentimiento y la minería de opinión, en ellos establecieron métodos que eran ya la recopilación sustancial de estos temas, confirmando a la semántica como la vía por la cual el análisis tendría que ser desarrollado.

Además, la regularidad del lenguaje, o aquella gramática generativa de la que hablaba Chomsky, iba cobrando sentido tal como lo muestran Mikolov, Yih y Zweig (Mikolov, Yih y Zweig, 2013) sobre un espacio continuo; ya que la característica de la continuidad permitía la idealización vectorial, que introduce el concepto matemático de norma, tomando en cuenta la dirección y sentido para las palabras. Es así como fue posible identificar patrones de relación, como *hombre* es a *mujer* como *rey* es a *reina*, o *tío* a *tía*; o la identificación de plurales. A pesar de estos hallazgos, que aportarían un avance al análisis, se encontraron fallas en los algoritmos de este tipo (Bolukbasi *et al.*, 2016). Mientras que muchas analogías eran resueltas satisfactoriamente y con congruencia, existían otras que estaban dotadas de prejuicios que no precisamente describían una realidad, es decir, una especie de falacia. Un ejemplo fue *hombre* es a *programador* como *mujer* es a *ama de casa*. Es, claramente, la presencia de estereotipos la que revela que no se pueden generalizar las entradas de texto, y la supervisión de la congruencia en los resultados es necesaria.

Otro objetivo de Mikolov et al. (Mikolov *et al.*, 2013) fue desarrollar un modelo que permitiera el aprendizaje de la distribución de los vectores enfocándose en la velocidad de búsqueda de patrones y la calidad de las relaciones.

Mostafa (Mostafa, 2013) evidenció que la investigación coincidia a menudo con otros ámbitos más comerciales. La industria a menudo busca conseguir información desplegada de grandes bases de datos, sin que tome el tiempo de leer uno a uno, y que resulte un análisis preciso para tomar decisiones, o generar ideas más generales a partir de grandes bancos de información; a esta actividad se le conoce como *análisis de mercado*, e. g. filtrando *tweets* sobre un cierto producto, y haciendo trabajo estadístico para saber cuáles son las opiniones naturales de los consumidores sobre este producto.

Everaert y otros (Everaert *et al.*, 2015), de los que sobresale Chomsky, admiten la presencia de las ciencias computacionales cognitivas y, sobre éstas, exponen nuevas hipótesis, en una traducción propuesta sería "Estructuras, no *strings*"; en el *argot* computacional, *string* es el equivalente a una unidad de texto. Esta idea identifica la necesidad del contexto, es decir, no basta con explorar unitariamente, sino debe haber una vecindad de información, cuyas propiedades nutrirán al análisis. Al mismo tiempo se reafirma el concepto de *Gramática Generativa*, que puede verse como otra traducción del artículo, para el desarrollo de las ciencias cognitivas.

En la actualidad, la labor del análisis de sentimientos, las ciencias cognitivas y la minería de opinión está dirigida a las aplicaciones que pudiera tener, en especial a los grandes datos; no sólo en conocer las opiniones, sino en entender comportamientos. Hipson (Hipson, 2019) escribe en su artículo la idea de, a través de la poesía, entender mediante el análisis de sentimientos el afecto de niños y adolescentes, sobre esta investigación halla tendencias de positividad y negatividad en grupos, y ajusta regresiones polinomiales; él mismo afirma que para la psicología del desarrollo estos métodos son sumamente útiles, en términos de investigación cuantitativa.

Por otro lado, también ha servido para la modelación de fenómenos socioeconómicos, Chen y Chen (Chen y Chen, 2019) utilizan información de varios foros para predecir los mercados financieros de Taiwán, a partir de las emociones expresadas por la sociedad, concluyendo que sí guardan relación las vías emocionales para el comportamiento del precio de las acciones. Alrededor del pronóstico del precio de las acciones, existen pocos modelos que lo consiguen, aunque ahora se está abordando por un método *heurístico*, no convencional; se podría deducir que el mercado global está relacionado con el precio de las acciones, pero hasta hace poco sólo el movimiento browniano formaba parte del análisis.

Por último, el análisis también está contemplando las técnicas de detección de rumores en redes sociales, desmitificando si algún hecho que se esté *viralizando* puede ser real. Artículos de este año (Alkhodair *et al.*, 2019; Bondielli y Marcelloni, 2019) estudian este fenómeno a través de la minería de texto.

1.3 JUSTIFICACIÓN

La importancia del análisis de sentimientos (Ravi y Ravi, 2015) se muestra en que favorece el conocer las ideas claves de nuestro emisor, que será el objeto de estudio. A través de sus entradas de texto se pueden clasificar como variables para identificar las actitudes, ánimos temas y sentimientos.

Un reto al que se pueden enfrentar (Mostafa, 2013) para la comprensión de los textos, pueden ser aquellos tecnicismos que nos impidan la fácil lectura del mismo, siendo una particularidad del *corpus*. Se considera la semántica como un punto de partida para comenzar a ahondar en el problema. Cambria y Hussain (Cambria y Hussain, 2012) demuestran que encontrar vecindades de palabras, y comprender sus significados, promueven un análisis realista de los cuadros. Al mismo tiempo, la sintaxis no siempre es aplicada de la forma más sencilla de procesar para un programa. Es por esto que se utilizará la cuantificación del texto y *Annotation* para obtener resultados de una comprensión más sencilla.

Se dedicará el estudio a una obra literaria clásica dado que ofrece muchas bondades todavía no estudiadas, la mayoría de las investigaciones han sido alrededor del análisis de mercado, pero no se ha llevado hacia la comprensión lectora generalizada, específicamente; aunque el estudio de casos coloquiales y particulares, se considerarían de una complejidad distinta dada la particularidad de sus elementos. Los retos que se enfrentan en nuestro objeto de estudio, Los Miserables (Hugo, 1887), podrían ser la época en la que fue escrita, y que podría diferir un poco con un *corpus* no ajustado temporalmente. Sobre este problema, se preferirá la versión traducida al inglés, ya que la mayoría de *corpora* en inglés son los mejor estudiados; además, la traducción de Los Miserables al inglés tiene buena aproximación, que se podría deducir por la familiaridad lingüística entre las lenguas anglosajonas y galas. Además, se dispone del archivo electrónico de la obra, libre de regalías.

El análisis de sentimiento es el método por el cual planeo describir esta obra, ya que es una forma rápida y eficiente, además que ilustra nítidamente las situaciones generales que se podría hallar dentro de la narración. Intentar resolver un conjunto de texto de dimensión semejante, por otro método, es una labor de complejidad mucho más alta; en cambio, el análisis de sentimientos es sencillo e incluso reconoce patrones que manualmente no serían identificables.

Cuando se piensa en saber si el objeto de estudio será rico en variedad de opiniones, la respuesta podría ser afirmativa, pues está confirmada por el premio Nobel de literatura: “El narrador nos cuenta la historia y su historia, nos cuenta qué está sucediendo y lo que el lector debe deducir que está ocurriendo (Vargas-Llosa, 2004)”, quien dedica un libro entero al estudio de la obra maestra creada por Víctor Hugo. Si quedara duda alguna de saber qué tan bien Víctor Hugo plasma sus opiniones, Vargas Llosa añade “…sino una espléndida ficción traída a la realidad y de los ideales, sueños, traumas, angustia y obsesiones…”.

En el caso general, dirigido al público, la ortografía resulta complicada para clasificar, puesto que, muchas veces, el uso del lenguaje coloquial presenta deformaciones a la lengua, pero éstas se tratan para que sean más naturales, si es que se encuentra un patrón. Problemas bien estudiados son las oraciones relativas que estudia Gibson a profundidad (Gibson, 1998), y que dan pie a considerar a la gramática como el problema más cercano, pero también más sencillo de resolver.

1.4 OBJETIVOS E HIPÓTESIS

Se presentará el análisis de una obra literaria clásica a través de un programa estadístico, considerando métodos semánticos y estructuras gramaticales estándares. La búsqueda particular es describir, en la actualidad, cuán sencilla puede ser la comprensión lectora por medio de un algoritmo; no es la intención desmotivar a la lectura que se ha llevado a cabo durante siglos, sino entender que podrían haber nuevos métodos de explicación para la lectura.

En principio, se estudiará la existencia de relación entre la frecuencia de las palabras con la detección de primeros elementos, enfocado a encontrar patrones; después se hará la comprobación con el análisis de sentimientos para describir la polaridad dentro del texto y tener una estructura bien detallada sobre el desarrollo de la información; por último, se utilizará la semántica para interpretar qué relación podrían obtener dichos elementos, y hallar una clasificación temática.

Siguiendo los pasos anteriores se obtendrán: Los personajes principales y secundarios, algunos escenarios clave de la obra, el marco contextual en el que se desarrollan los protagonistas, los sentimientos más ilustrativos de la obra y la concentración de éstos.

2 Material y Métodos

2.1 Les Misérables: Recurso Literario

Les Misérables (Hugo, 1887), provista por *Project Gutenberg*, de acceso y descarga gratuita, en la versión de UTF-8 y en inglés, fue la versión utilizada. El texto cuenta con acentos y caracteres originales del francés, especialmente para el caso de los nombres propios. Se utilizó el texto en UTF-8, retirando los caracteres especiales.

En general, es una obra publicada por primera vez en 1862, la versión original, la traducción es de 1887. Ésta es considerada una pieza clave de la literatura francesa y universal; comprendida por 5 volúmenes:

I. Fantine
II. Cosette
III. Marius
IV. Saint-Denis
V. Jean Valjean

Sus personajes principales son, en orden de relevancia, Jean Valjean, Cosette, Marius, Javert, Fantine, los señores Thénardier, Éponine y Gavroche.

La obra exhorta al lector a experimentar sentimientos compasivos, a través de situaciones de muerte, pobreza, opresión y desigualdad social; que se supone fueron características de la época, el siglo XIX. Sobre su contexto, está bien situada, puesto que la revolución francesa se va desarrollando a la par de la narración, y es ahí donde tiene más relevancia para la historia francesa; que desembocara con la destitución de la monarquía como forma de gobierno, y la institución de la república. Esto propicia tintes coloniales llevados hacia la modernización industrial.

El escenario es, en todo momento, Francia; así que las casas, calles parisinas, fábricas, cantinas, campos y bosques son las locaciones usuales.

2.2 PAQUETES NECESARIOS DE R

R (The R Foundation, 2019) es un software estadístico libre en el que existen paqueterías para distintos propósitos, en cada paquetería se encuentran las funciones que tienen ese propósito en común. El código fue colocado en el anexo: Código en R. Las paqueterías para este estudio utilizadas se mencionan a continuación.

Paquete	**Función General**
tidytext	Minería de Texto.
dplyr	Transformación de los datos.
ggplot2	Visualizaciones.
rcolorbrewer	Brinda gamas de colores más amplias.
wordcloud	Nubes de palabras.
tidyr	Limpieza de datos.
reshape2	Transponer tablas.
igraph	Crea redes para los n-gramas.
ggraph	Evita choques de n-gramas.
widyr	Analiza la correlación.
stringr	Permite operaciones dentro del texto.
ggrepel	Agrega gráficas para ggplot2.
topicmodels	LDA.
gridExtra	Múltiples gráficas en una.
clipr	Agrega tablas al portapapeles.
syuzhet	El diccionario NRC ya no aparece en tidytext
cleanNLP	Describe a las palabras de acuerdo a su estructura

2.2.1 TIDYTEXT

Este paquete puede ser visto como el pilar de la minería de texto, en función del análisis. Permite el uso de *unnest*, quien logra dividir el texto por palabras y los "desanida" convirtiéndolos en *tokens*, o elementos de texto. Dentro de la misma función *unnest* es posible configurar la opción de n-gramas.

Además, contiene las bases de datos de sentimientos (Bing, AFINN y NRC) que son con las que se establecen las comparaciones para determinar los valores de sentimientos encontrados, tanto negativos como positivos. También cuenta con una base llamada *stop-words* que contiene aquellas palabras que podrían ensuciar el análisis, en su mayoría conjunciones o nexos.

2.2.2 DPLYR

Es una paquetería para la manipulación gramatical de los datos, brinda funciones con un lenguaje mucho más familiar, siendo reconocible ya que utiliza verbos explícitos.

Muchas de estas funciones pueden ser aplicadas sin necesidad de la paquetería, pero el nombre de las funciones mejora el conocimiento del manejo de éstas. Ejemplos clave de funciones dentro del paquete son: *Select, filter, rename, mutate.*

Para conectar varias funciones, dentro de este paquete, se usa un comando particular: *Pipe* (*%>%*). Su función es brindar más orden dentro del código.

2.2.3 GGPLOT2

Es la responsable de algunas gráficas que facilitan la interacción con la información, cuenta con opciones especiales basadas en la gramática de las gráficas: Sus funciones fueron creadas con base en nombres fácilmente identificables.

La estructura esencial (Gil Bellosta, 2018) de las gráficas con este paquete es a partir de la combinación de elementos básicos junto a los tipos del gráfico, mediante una sintaxis sencilla. Los elementos son: Datos, estéticas, capas, facetas y temas.

2.2.4 RCOLORBREWER

Extiende la gama de colores y ayuda a crear combinaciones de colores para mejorar su lectura dentro de las gráficas. Los colores (MacArthur, 2010) los divide en tres grupos: Secuenciales, divergentes y cualitativos.

Para los secuenciales, aclara los valores inferiores y oscurece los superiores. En los divergentes, oscurece los valores extremos (inferiores y superiores) y aclara los medios. Los cualitativos es una combinación para diferenciar altamente las clases.

Sólo es necesario especificar el nombre de la paleta que se desee de esta paquetería. E. g. "blues", que configura una gama de azules.

2.2.5 WORDCLOUD

Crea nubes de palabras, a partir del parámetro especificado, para este caso se emplea la frecuencia de palabras, y cuenta con mejoras en sus ajustes basados en HTML. Dentro de esta paquetería es posible ajustar las palabras en un gráfico distinto a la nube, alguno con una estructura más detallada, como un gráfico.

2.2.6 TIDYR

Para algunos *tokens*, especialmente en el caso de los n-gramas, es necesaria esta paquetería, ya que anida y dispersa al texto plano, convirtiendo cada palabra, de nuevo, en una entidad. Sus funciones cruciales son *gather, separate* y *spread*. En este contexto, las funciones siguen siendo descritas explícitamente por sus nombres, tal como *dplyr*.

2.2.7 RESHAPE2

El concepto que aproxima su función es la transposición, aunque este término simboliza que las columnas se hacen filas y las filas, columnas; adicionalmente, para el caso de las nuevas columnas estás agruparán los términos de las filas.

Es necesario incluir dos conceptos: *wide-format* y *long-format* (Anderson, 2013). *Wide-data* tiene una columna para cada variable; *long-data* tiene una columna para los posibles tipos de variables, y otra para sus valores. Análogamente, existen dos funciones que transforman en estos formatos: *melt* y *cast*. *Melt* crea un formato *long;* y *cast,* un formato *wide.* Éstos a partir del formato contrario. Por default, los caracteres numéricos son considerados valores, mientras que a las palabras se le atribuye la propiedad de variables.

2.2.8 IGRAPH

Hace cadenas de relación, redes, entre las palabras, permitiendo gráficas de n-gramas y bigramas, y enlaza mediante las distancias calculadas dentro del texto. Es esencial para la traducción del comportamiento del texto para los n-gramas con relación a otros. En general, a esta aplicación se le denomina análisis de red (Network Analysis).

Los autores (The Igraph Core Team, 2015) lo describen a través de tres características principales:

1. Algoritmos gráficos de implementación sin dolor (pain-free).
2. Manejo rápido de grandes gráficos, con millones de vértices y bordes.
3. Permite hacer prototipos rápidos mediante lenguajes de alto nivel, como R.

2.2.9 GGRAPH

Es otra opción de visualización de n-gramas, o redes. Cuenta con 4 elementos importantes: *layouts*, nodos, bordes y conexiones. Sigue las técnicas tradicionales de graficación, como la selección de color o la dimensión del gráfico.

2.2.10 WIDYR

Cuenta con una función, *pairwise_cor*, que muestra la correlación en los pares de palabras, también contribuye al conteo y la similitud por parejas.

En el paquete (Uryu, 2017) también se hallan funciones como *cast-dual* (para convertir *datasets* en una matriz, manteniendo la tabla), *cors_sparse* (para encontrar la correlación dentro de una matriz) o *pairwise_dist* (para determinar las distancias entre parejas de palabras).

2.2.11 STRINGR

Es una paquetería usual para poder efectuar operaciones con líneas de texto, *strings*. Normalmente, no es posible realizar operaciones con cadenas de texto. La mayoría de sus funciones contienen el prefijo *str_*. Detecta relaciones, extrae *strings*, configura longitudes, mueve *strings*, une-separa y ordena; son las categorías de sus funciones. Grolemund (Grolemund, 2018) creó un pdf que contiene la descripción particular de cada función del paquete, expresada a modo de infografía.

2.2.12 GGREPEL

Cuando se está creando una gráfica y se desea mejorar la visualización en ggplot2, este paquete debe estar presente.

2.2.13 TOPICMODELS

Es el que brinda las funciones de LDA, con los parámetros expuesto en la sección de Asignación Latente de Dirichlet.

2.2.14 GRIDEXTRA

Algunas veces sirve mejor utilizar varias gráficas al mismo tiempo, y este paquete permite agregarlas sin dificultad.

2.2.15 CLIPR

Es un paquete útil, ya que evita que se tenga que estar copiando un *dataframe*, de forma manual, en combinación con un *pipe*, y sólo se debe colocar "Ctrl + V", o el símil de acuerdo al sistema operativo, y en Excel se pega con el formato de origen. Tal vez, vale la pena agregar *head(n)*, siendo *n* el número de filas superiores que se deseen copiar.

2.2.16 SYUZHET

En la última actualización de *tidytext*, ya no se encuentra el diccionario NRC, por lo que se tiene que optar por este paquete.

2.2.17 CLEANNLP

Implementa la *annotation*, materia prima para el modelado temático. Provee de la lematización, y da las categorías gramaticales, así como su posible función dentro de la oración. El principal producto aparece en el apartado de *token*. Es importante mencionar que requiere de un método preciso y ordenado entre sus funciones, además de que la función que realiza la *annotation* demora varios minutos.

2.3 CREACIÓN DEL PROYECTO

Para iniciar el proceso, es una buena recomendación tener un proyecto en R y ahí ir depositando los archivos que servirán para la manipulación del texto. En otro caso, sólo hay que identificar la ubicación de Directorio de Trabajo con *getwd*().

2.4 PREPARACIÓN DE LOS DATOS

Se descargó la obra completa de Los Miserables del *Gutenberg Project*, y se guardó en *.txt* en la carpeta del proyecto homónimo. Es muy importante, una vez descargado, cerciorarse que está codificado en UTF-8, ya que R puede presentar problemas si está en alguna otra codificación. En el código aparecen funciones para arreglar las contracciones , (Liske, 2018b) y remover caracteres especiales; como parte de la limpieza del texto. Se retiró todo lo que estaba antes del primer capítulo y las notas finales; para evitar confusiones provenientes de comentarios del editor. Ya limpio el texto, se usó la función *unnest*, para convertirlo en *tokens* (se nombran así a las palabras, y cada una equivale a un elemento, ahora *token*). También se ocupó la opción de *stop words*, que excluye la mayoría de nexos y artículos, puesto que no brindan una representación semántica ni aportación alguna al análisis de sentimientos. Una vez hecho lo anterior, el texto estuvo listo y almacenado en un nuevo *data frame*, que es el que fue analizado.

2.5 DESCRIPCIÓN GRÁFICA DE LOS DATOS

Se buscó caracterizar gráficamente cuáles fueron los hallazgos a partir de las gráficas de nube (Liske, 2018c), que concentran las palabras con mayor aparición, y las representa con un mayor tamaño. Las nubes de palabras muestran qué palabras fueron utilizadas para tener una idea general del contenido.

2.6 COMPOSICIÓN LÉXICA

Mediante la diversidad y densidad léxica se reveló qué clase de obra es la que se está analizando, y cómo se va comportando a lo largo del libro. La diversidad léxica fueron aquellas palabras distintas que existan en cada apartado, mientras que la densidad es la proporción. La composición también se extrajo de la versión *annotated* provista por *cleanNLP*.

2.7 ANÁLISIS DE SENTIMIENTOS

Utilizando *tidytext*, se describe cómo en cada oración de cada volumen son expresados los sentimientos, de acuerdo a la categorización proveniente de ésta. En la paquetería, podemos usar resultados de varias fuentes de léxico (Robinson y Silge, sin fecha), las utilizadas fueron NRC, AFINN y Bing. Para una cuantificación inmediata, se empleó Bing que originalmente sólo está clasificada en *positive* y *negative*, pero se convirtió en puntaje. Para el resto, hay que ejecutar un conteo. En los casos donde existe una categoría grupal de adjetivos se crean conjuntos semánticos. La descripción se hace gráficamente.

Posteriormente, para obtener una verificación de los adjetivos, se analizaron los casos subordinados como aquellas estructuras de voz pasiva y las precedidas por *not.*

2.8 ANÁLISIS SEMÁNTICO POR CORRELACIÓN

Para esta sección se ocuparon *n-gramas* (Silge y Robinson, 2019) que se representan en mapas, muestran las uniones de pares de palabras. Es importante crear un paso de *unnest*, para que ahora aparezcan los *tokens* como *n-gramas*. Se empleó el conteo para conocer la frecuencia, los niveles de correlación y la representación gráfica.

2.9 ANÁLISIS SEMÁNTICO POR LDA

Liske (Liske, 2018a) documenta este proceso. Creando una versión *annotated* del *corpus*, se llevó a cabo la implementación del modelo LDA, usando el método *Gibbs*. Se decidió, en esta parte, cambiar la agrupación inicial, y segmentar el texto en 6 secciones; con la intención de hallar un tema para cada conjunto. Se experimentó la forma de *lemma* para nombres propios, sustantivos, verbos y adjetivos.

3 Resultados

3.1 Frecuencias de Sustantivos y Adjetivos

Para dar inicio a tener una idea de cómo se habrían de desarrollar las ideas sobre los protagonistas, se decidió medir la frecuencia de las palabras, como se aprecia en el Cuadro 1.

Cuadro 1: Palabras con mayor frecuencia. Fuente: Elaboración propia.

Palabra	Frecuencia
marius	1358
jean	1227
valjean	1112
cosette	1012
day	794
time	760
father	601
moment	562
hand	542
eyes	537

Ya que en la consulta anterior se retiraron las *stop_words*, es oportuno aclarar cuáles fueron agregándolas al anexo.

Para conocer qué consultas son posibles mediante NRC, se identifican las clasificaciones, del Cuadro 2.

Cuadro 2: Clasificaciones de la base de datos NRC. Fuente: Elaboración propia.

Clasificación	Frecuencia
negative	3,324
positive	2,312
fear	1,476
anger	1,247
trust	1,231
sadness	1,191
disgust	1,058
anticipation	839
joy	689
surprise	534

Asimismo para saber cómo se habrían de dividir la disyuntiva natural de la felicidad, se decidió encontrar la intersección de la obra con las clasificaciones NRC de tipo *positive*, *negative* (Cuadro 3 y Cuadro 4, respectivamente).

Cuadro 3: Palabras *positive* con mayor frecuencia (NRC). Fuente: Elaboración propia.

Palabra	**Frecuencia**
child	468
saint	398
god	356
mother	356
love	353
found	275
white	243
garden	237
happy	145
joy	143

Cuadro 4: Palabras *negative* con mayor frecuencia (NRC). Fuente: Elaboración propia.

Palabra	**Frecuencia**
rue	661
mother	356
black	269
death	240
terrible	223
fell	196
grave	173
revolution	162
lost	157
darkness	138

De acuerdo a Bing, los resultados para *negative* y *positive,* son los observados en el Cuadro 5 y Cuadro 6.

Cuadro 5: Palabras *negative* con mayor frecuencia (Bing). Fuente: Elaboración propia.

Palabra	Frecuencia
rue	661
poor	292
dead	245
death	240
terrible	223
fell	196
cold	170
lost	157
darkness	138
fall	136

Cuadro 6: Palabras *positive* con mayor frecuencia (Bing). Fuente: Elaboración propia.

Palabra	Frecuencia
saint	398
love	353
grand	145
happy	145
joy	143
fine	139
smile	130
charming	110
profound	109
beautiful	108

3.2 Nube de Palabras

En la nube de palabras de la Figura 1 se decidió seleccionar las 20 palabras de mayor frecuencia, ya que tienen una mejor representación gráfica.

Figura 1: Nube de palabras con mayor frecuencia. Fuente: Elaboración propia.

3.3 Diversidad y Densidad Léxica

La diversidad léxica, impresa en la Figura 2, se utiliza para saber cuántas palabras distintas existen en cada *Index*, para este ejercicio en particular se reagrupan por 200 líneas, por temas de visualización.

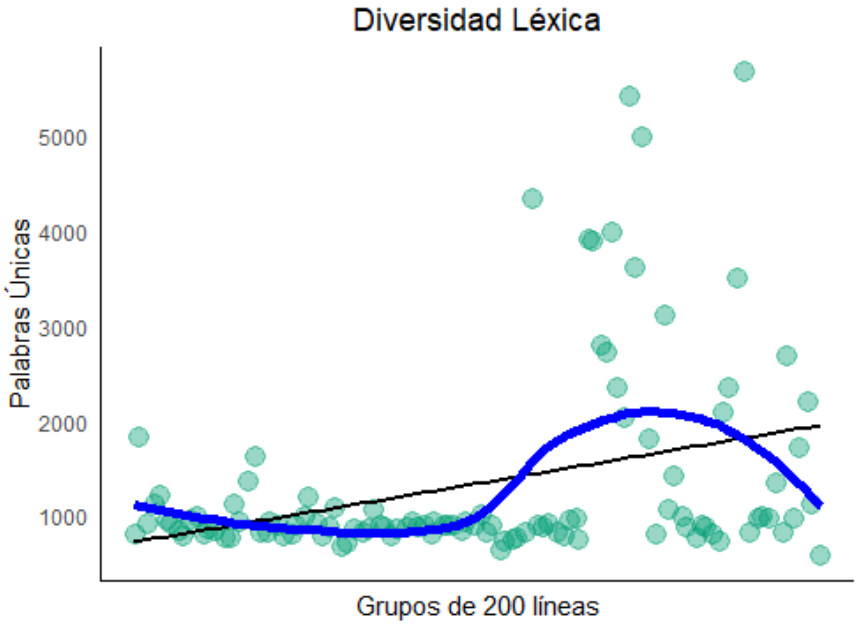

Figura 2: Diversidad Léxica (Modelo lineal y Suavizamiento). Fuente: Elaboración propia.

En cambio, en la Figura 3, la densidad léxica es el número de palabras distintas por *Index*, pero divididas por el número de palabras totales en el subconjunto.

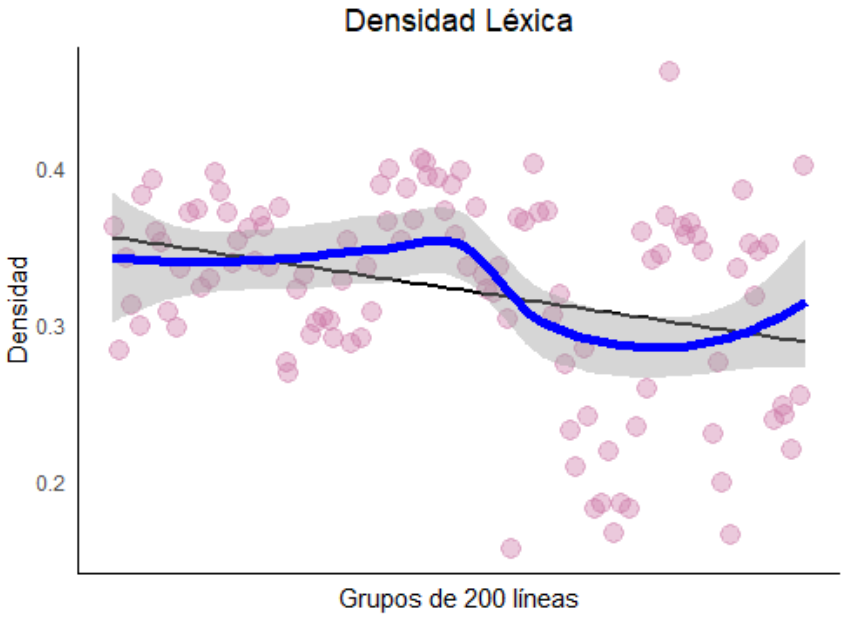

Figura 3: Densidad Léxica (Modelo lineal y Suavizamiento). Fuente: Elaboración propia.

3.4 Distribución de Sentimientos

Una vez más, la Figura 4 describe en la parte inferior las palabras positivas en color naranja, y las negativas en oscuro. Es proporcional la frecuencia al tamaño de la tipografía.

Figura 4: Distribución de palabras negativas y positivas con mayor frecuencia. Fuente: Elaboración propia.

En la Figura 5 y Figura 6 el *index*, o índice, está relacionado con la línea que ocupa en el texto original, así que cada línea fue medida de acuerdo con su proporción sentimental. Si la barra tiene dirección hacia abajo, implica que la oración tuvo sentido negativo, y viceversa.

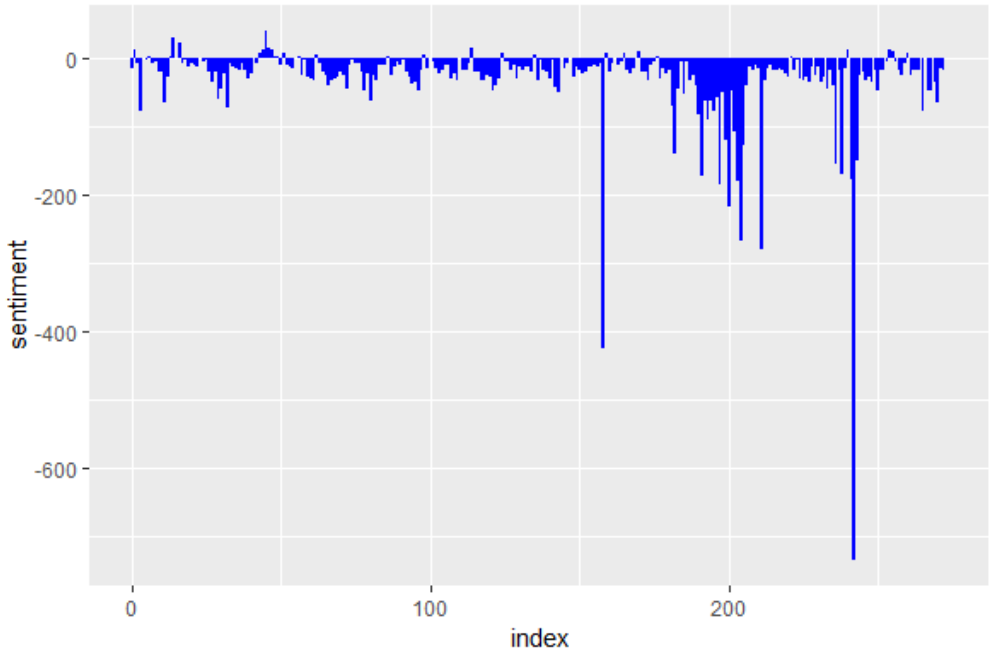

Figura 5: Distribución de sentimientos por índex de texto. (Bing) Fuente: Elaboración propia.

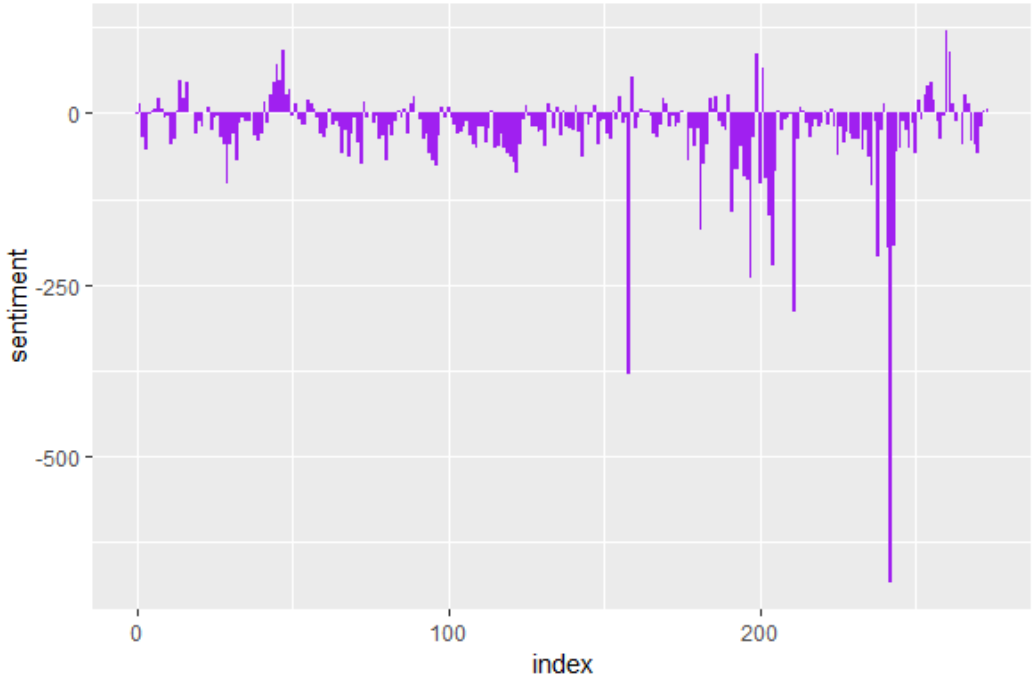

Figura 6: Distribución de sentimientos por índex de texto. (AFINN) Fuente: Elaboración propia.

En ambas gráficas existe el mismo mínimo, y haciendo una exploración resulta ser el conjunto 242, cuya nube de palabras corresponde a la Figura 7.

Figura 7: Index más negativo. Fuente: Elaboración propia.

En el caso de NRC, se decidió identificar las clasificaciones para conocer cuál era su participación dentro de la obra, siendo la provista por la Figura 8.

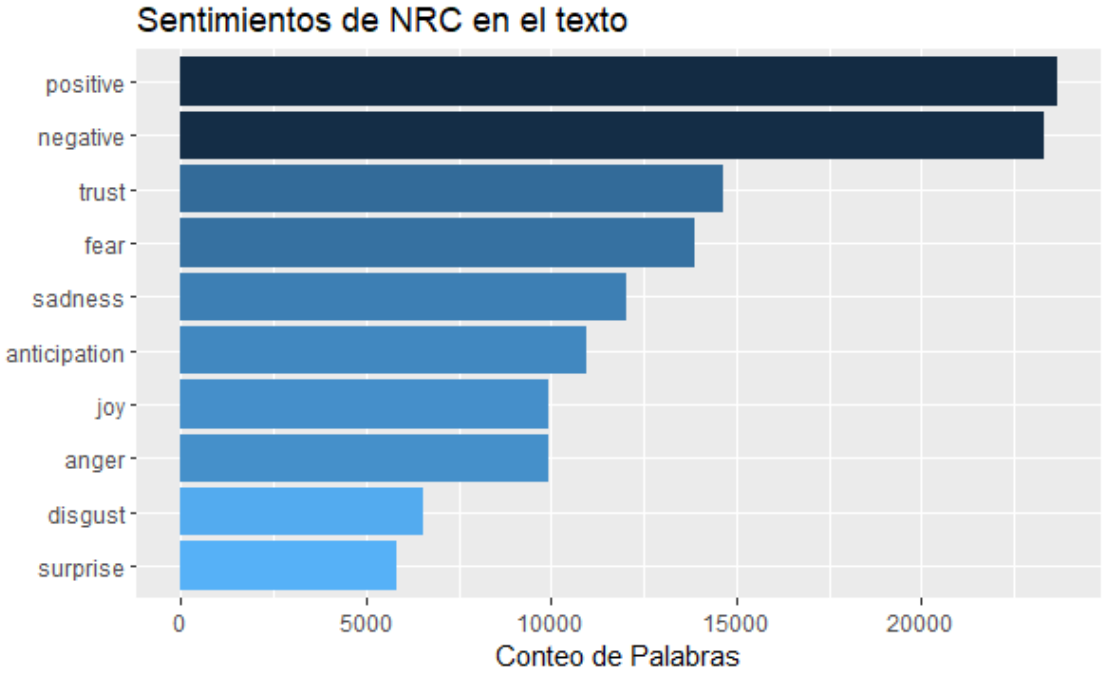

Figura 8: Frecuencia de clasificaciones NRC en el texto. Fuente: Elaboración propia.

3.5 Descripción Verbal y Adjetiva

Para la Figura 9 se consideraron los casos en los que la conjunción negativa no precede verbos y se decidió segmentarlos de acuerdo con el sentimiento expresado. Es decir, los puntajes negativos son aquellos que describen una idea más positiva, o contraria; este análisis sólo es aplicable debido a *not*. En el mismo sentido, los puntajes positivos representan situaciones negativas: El análisis de esta gráfica es *sui generis*.

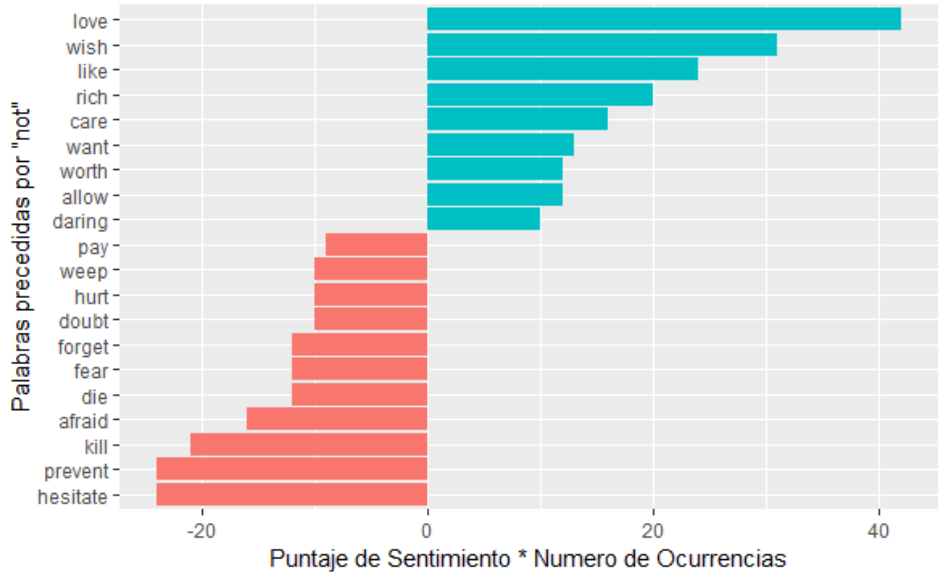

Figura 9: Palabras con mayor frecuencia precedidas por “not”. Fuente: Elaboración propia.

Para conocer el presente, pasado y futuro dentro de la historia es útil consultar las palabras con los auxiliares *was*, *is* y *will;* se muestran ilustrados en la Figura 10, Figura 11, Figura 12, respectivamente.

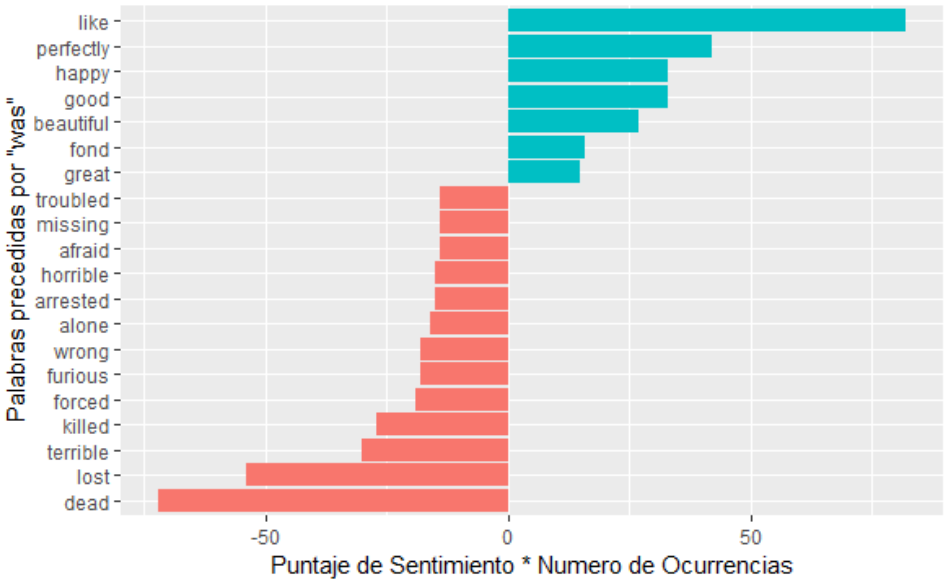

Figura 10: Palabras con mayor frecuencia precedidas por “was”. Fuente: Elaboración propia.

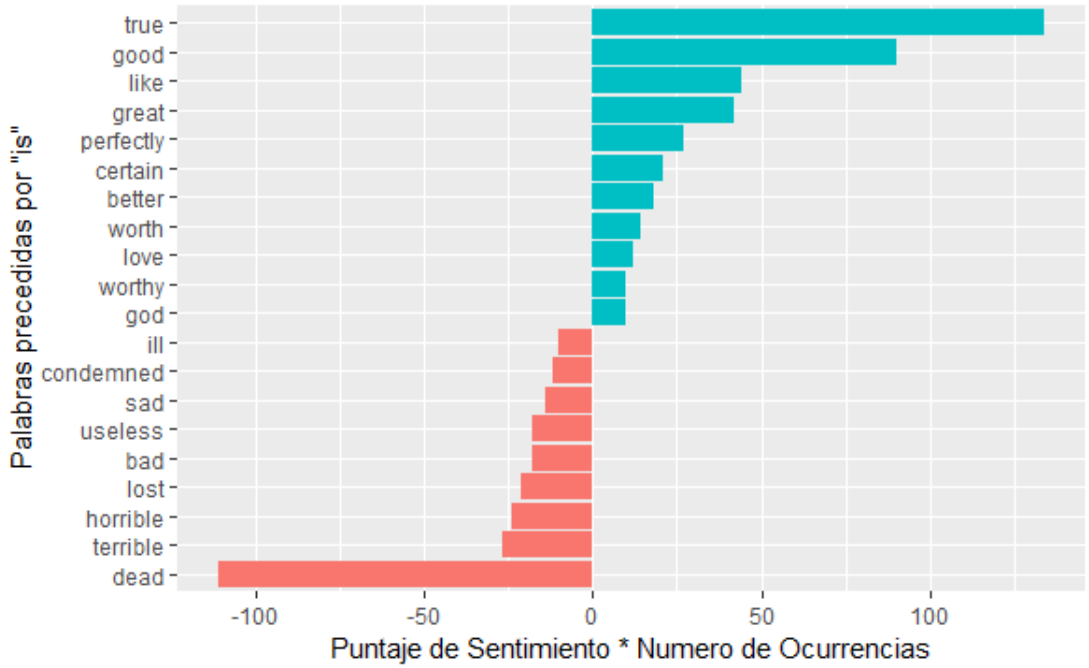

Figura 11: Palabras con mayor frecuencia precedidas por “is”. Fuente: Elaboración propia.

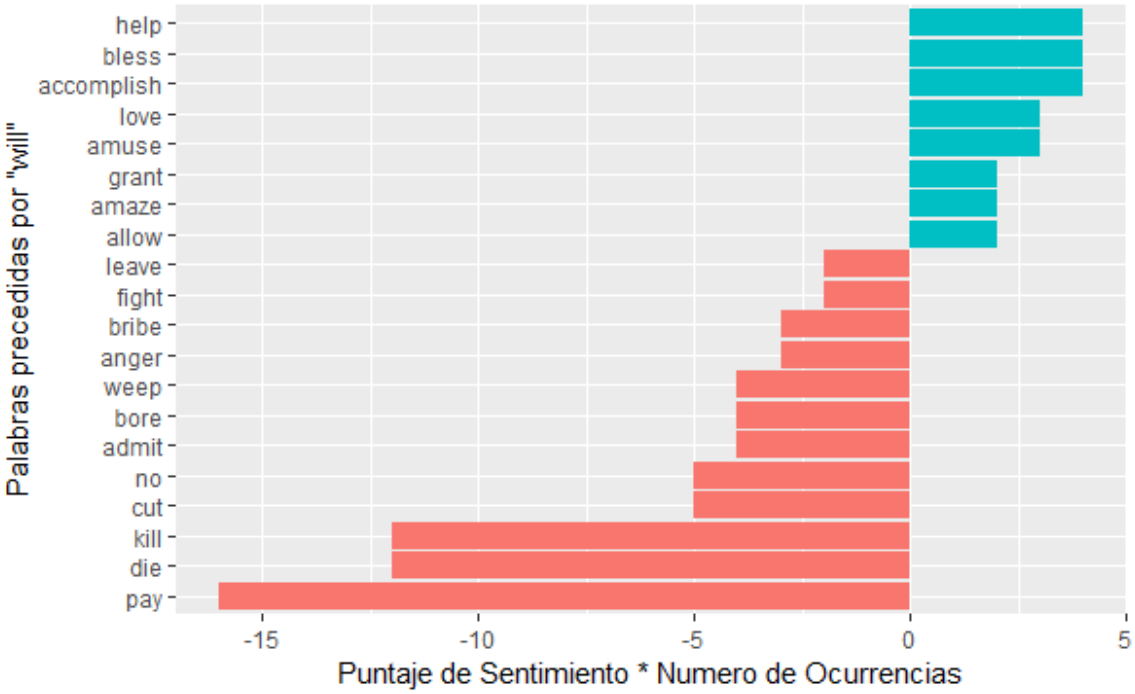

Figura 12: Palabras con mayor frecuencia precedidas por “will”. Fuente: Elaboración propia.

3.6 Análisis Semántico por Correlación

Dado que son ocho personajes los más emblemáticos, se dividieron en dos segmentos para establacer el contexto particular de cada uno. El criterio del orden se debió a su aparición dentro de la obra. En la Figura 13 están los personajes que se encuentran antes del nacimiento de Cossete, y en la Figura 14 los subsecuentes.

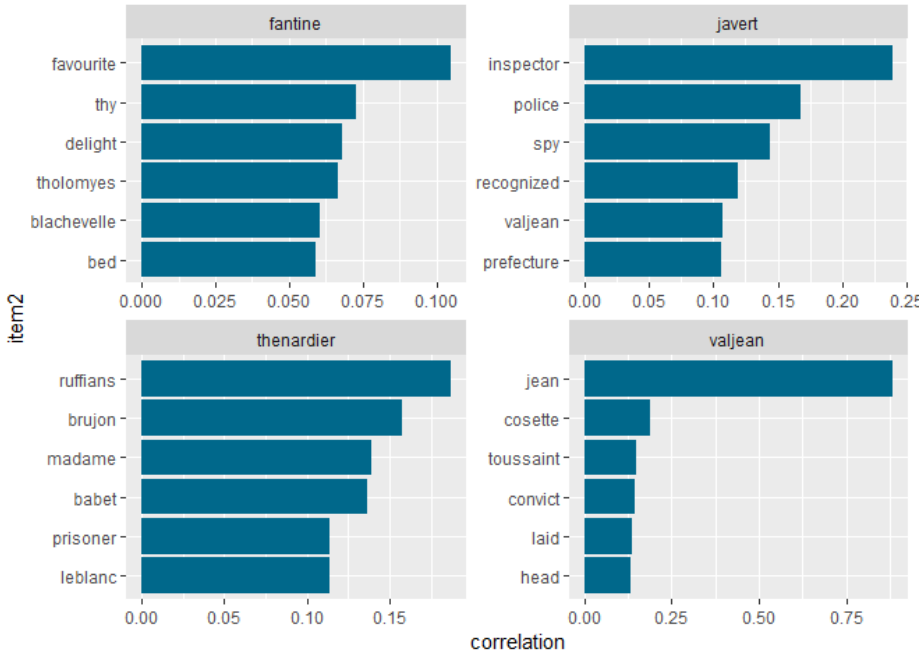

Figura 13: Correlación de personajes principales (1ra parte). Fuente: Elaboración propia.

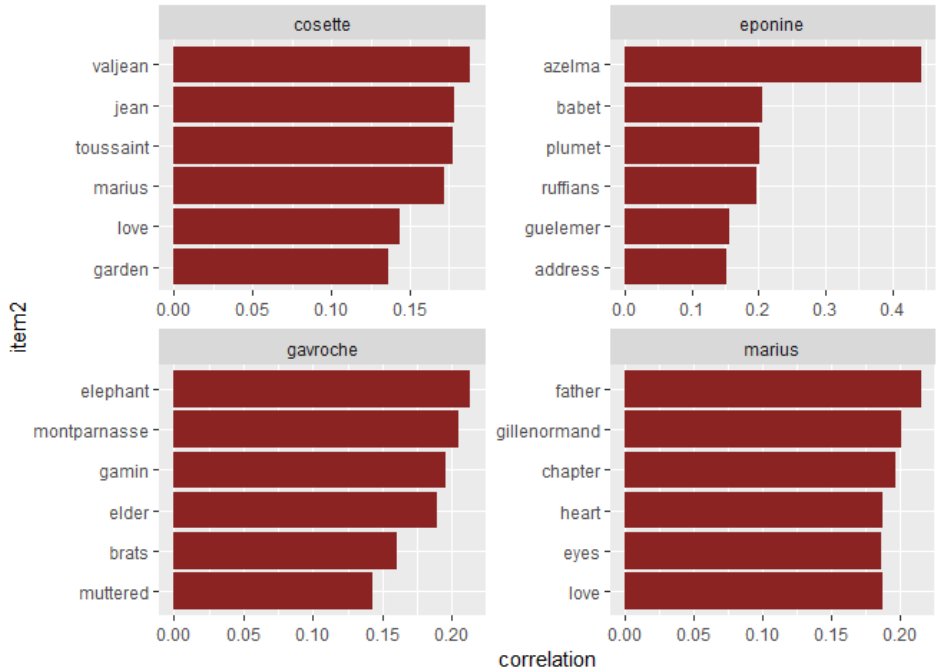

Figura 14: Correlación de personajes principales (2da parte). Fuente: Elaboración propia.

A partir de palabras que no pudieran contar con una semántica diversa, en la Figura 15 se optó por mostrar una correlación generalizada.

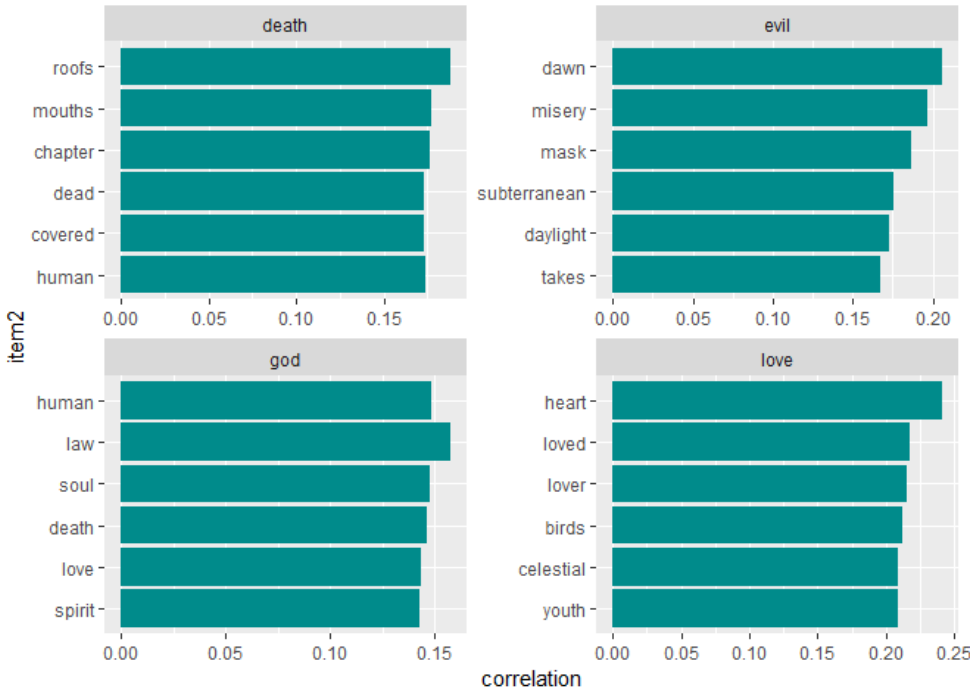

Figura 15: Palabras descriptivas generales. Fuente: Elaboración propia.

Por último, en la Figura 16, a través de bigramas, se establece la relación de las parejas de palabras con mayor frecuencia dentro de la obra.

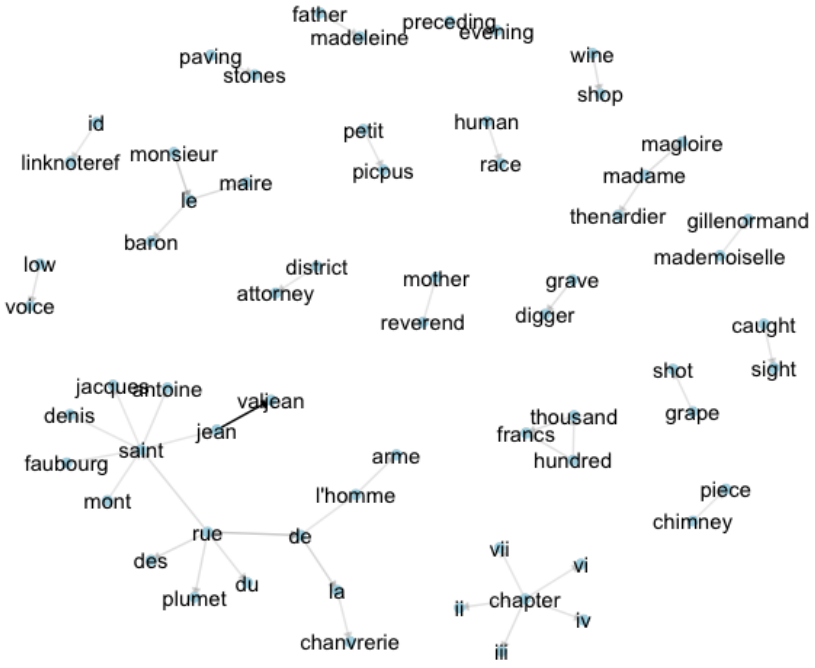

Figura 16: Bigramas con frecuencia y relación. Fuente: Elaboración propia.

3.7 Análisis Semántico por LDA

Se calcularon los temas sugiriendo que podían existir 6 temas más importantes, uno en cada volumen y uno adicional. La primera extracción surgió de los sustantivos, presentados en la Figura 17.

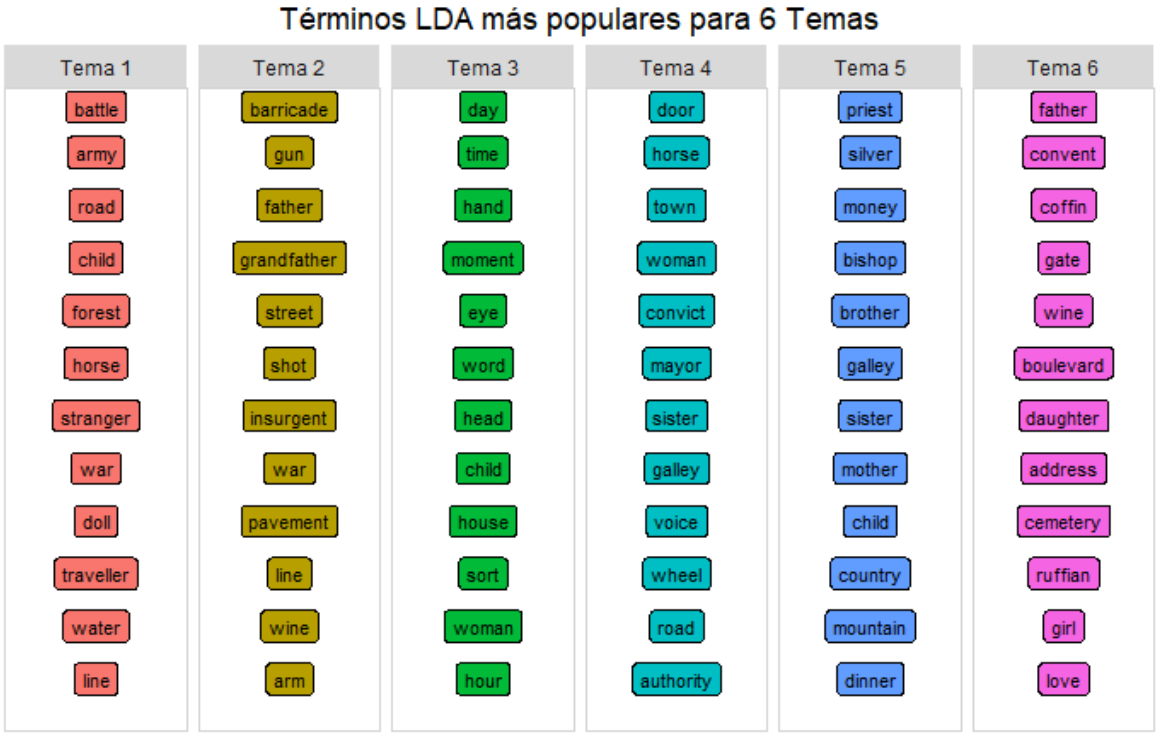

Figura 17: Sustantivos más populares para 6 temas. Fuente: Elaboración propia.

Una vez ligados los sustantivos a seis temas, en la Figura 18, se escribieron nombres arbitrarios que pudiesen englobar estas ideas, y se les ubicó dentro del texto.

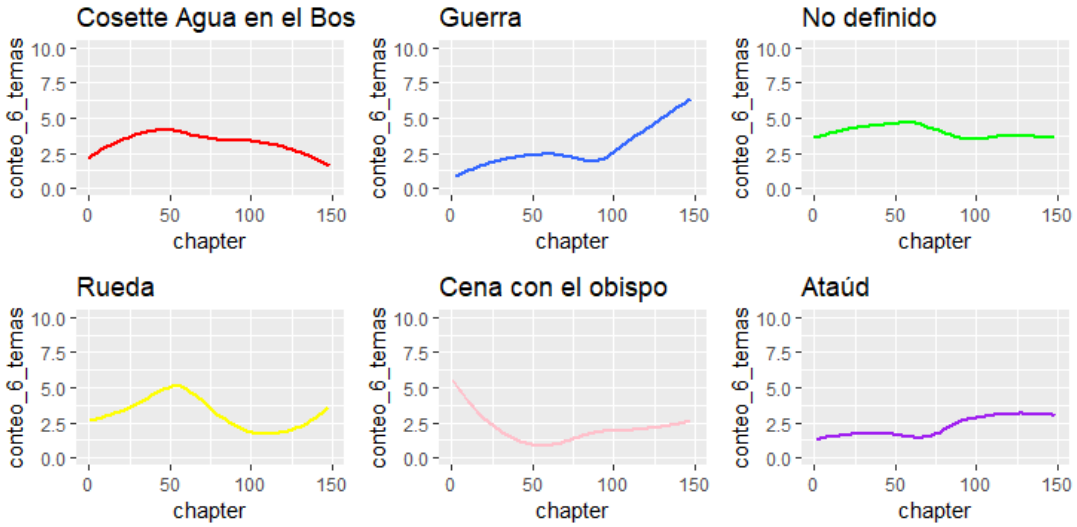

Figura 18: Ubicación temática de sustantivos en el texto. Fuente: Elaboración propia.

Se aplicó el mismo ejercicio para los verbos, destacando que no se esperaría una relación inmediata entre los posibles temas de sustantivos y verbos. Los verbos de la Figura 19 aparecen en forma de lexema.

Figura 19: Verbos más populares para 6 temas. Fuente: Elaboración propia.

Se clasificaron los nombres de las gráficas, de la Figura 20, de acuerdo a los verbos más representativos empíricamente.

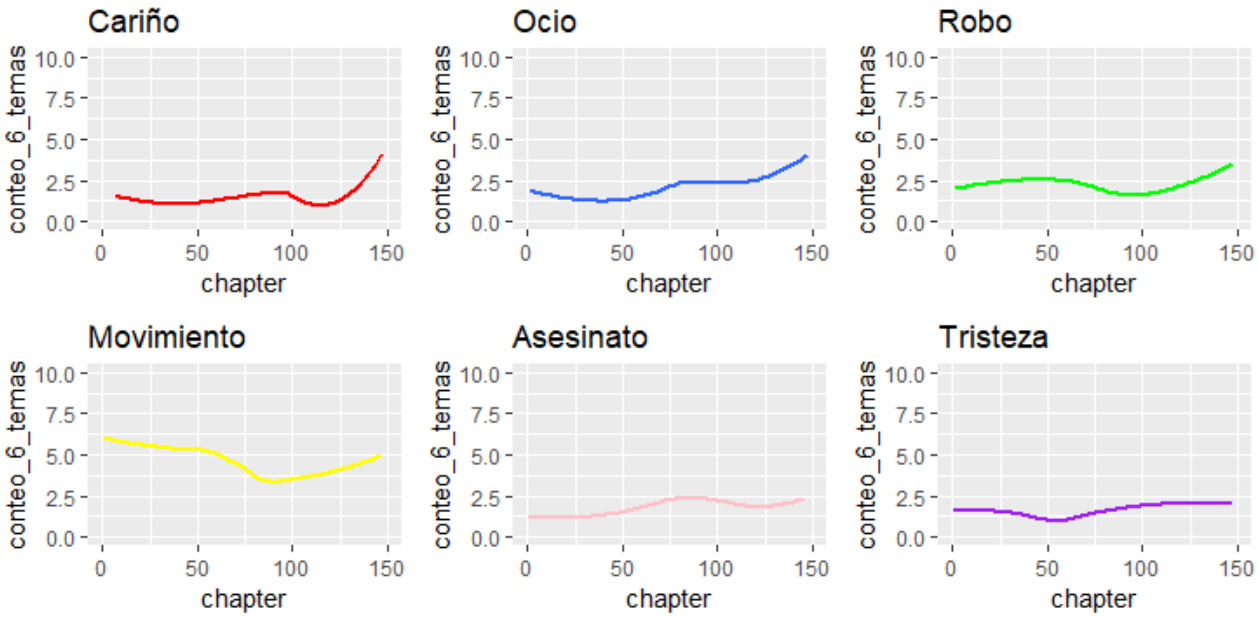

Figura 20: Ubicación temática de verbos en el texto. Fuente: Elaboración propia.

Los adjetivos, de nuevo, se clasificaron en seis secciones en la Figura 21.

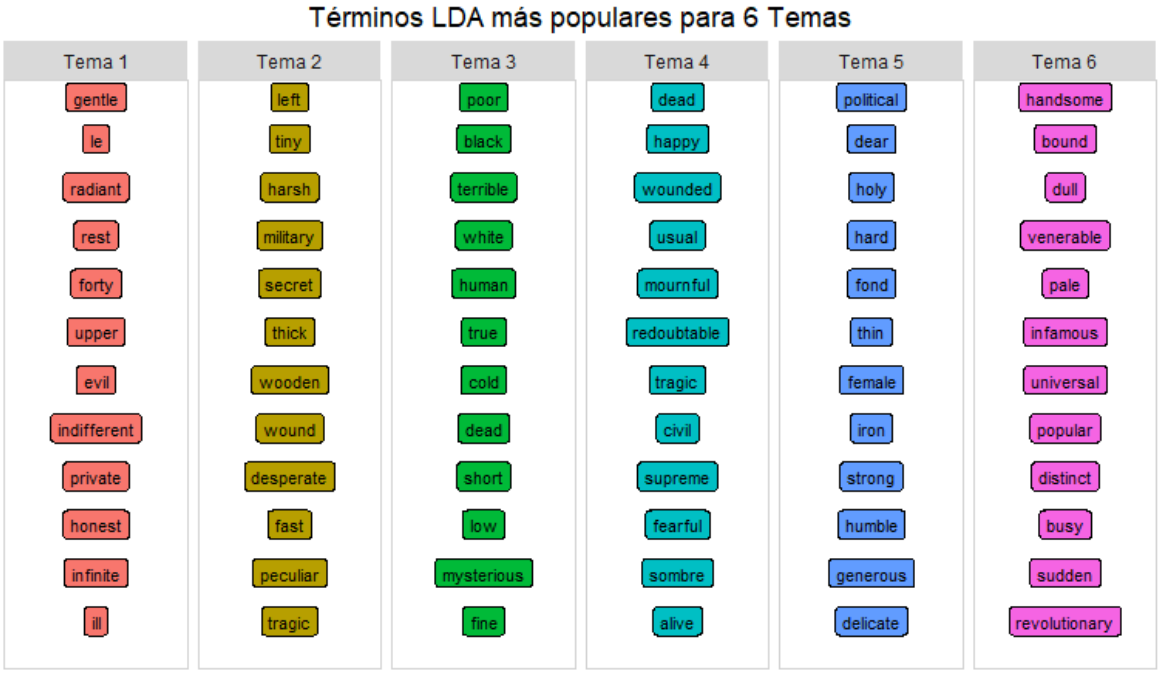

Figura 21: Adjetivos más populares para 6 temas. Fuente: Elaboración propia.

Se ajustaron, en la Figura 22, *suavizamientos* a los conteos de las palabras, para describir patrones.

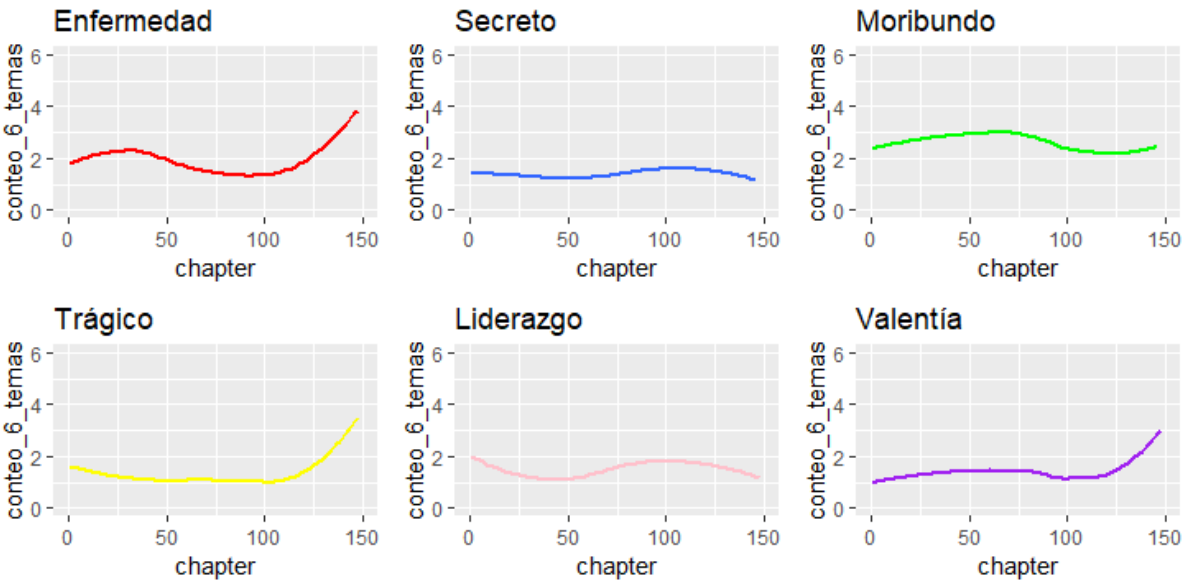

Figura 22: Ubicación temática de verbos en el texto. Fuente: Elaboración propia.

4 Discusión

4.1 Interpretación Léxica

El análisis que se obtiene del Cuadro 1 implica que Jean Valjean, Marius y Cosette, al encabezar las apariciones, serán considerados como los personajes principales, sin embargo Jean Valjean no fue la palabra más frecuente, sino Marius, la razón puede deberse a que Jean Valjean en muchos casos es el interlocutor y Marius resulta más bien un personaje secundario: del que se habla. La Figura 1, en su calidad de nube de palabras, alude más palabras que podrían tener sentido, y es aquí donde se encuentra Javert, que no había aparecido en un principio, con el cuadro. Otra palabra que resulta significativa es *rue*; haciendo partícipe al contexto francés, *rue* es una palabra francesa que significa "calle", y es el potencial escenario de los personajes. También Thernardier es la décima palabra más frecuente, y será parte de los recursos antagónicos, aunque es apresurado suponerlo.

Una nota que no se debe desconsiderar es que para el análisis de sentimientos se excluyeron las *stop_words*, pero se retoman para el resto de la investigación.

4.2 Análisis de Sentimiento

Sabiendo que usamos Bing, AFINN y NRC, las consultas resultaron sencillas; sin embargo, en el caso particular de NRC se deben contemplar que sus clasificaciones son nominales no binarias, *id est,* si se desea hacer una consulta real es necesario determinar los posibles casos de categorización, resolviendo esto, en el Cuadro 2 se encuentran las 10 clasificaciones disponibles, aunque también se reconoce que el conjunto *negative* y *positive* son lo más extensos. Por lo tanto, en la primera consulta de NRC (Cuadro 3) la palabra más positiva que hallan dentro de la obra es *child*, y hay que tomar en cuenta que Cosette, la mitad de la historia, es referida como una niña. Hasta la antepenúltima palabra de dicho cuadro, los hallazgos pertenecen a la categoría gramatical (Elman, 1991) de sustantivo, nótese que *love* puede tener propiedades verbales también; en mi opinión, a los posibles sustantivos se les debe de dar un tratamiento con la mayor neutralidad posible, con la intención de no caer en falacias inmediatas. Sin embargo, *saint* y *god* son palabras relacionadas con la divinidad, y éste es un tema que, si bien suele simular ser positivo, puede estar circunscrito por una situación sumamente negativa.

De nuevo, un escenario recurrente (Cuadro 4) es *rue,* del francés "calle", que aparece en las palabras *negative,* y es justamente un problema que se podía tener, dado que se traslaparon los idiomas. En francés, que es la lengua de la obra, esta palabra tiene todo el sentido, puesto que es una locación; sin embargo, en cuanto al análisis en inglés no hay error, puesto que *rue*, en dicha lengua, es un verbo que significa "lamentar" (Miller, 1995), y gracias a la semántica, se puede aclarar la diferencia, pero es sustancial esta parte del análisis: la comprobación general de los resultados, a partir del contexto.

Así, también, *mother* es una palabra *positive* y *negative*, *ergo* la deduciremos neutral. *Death* es un escenario interesante, cuenta con una frecuencia significativa. En el contexto, tanta muerte puede implicar una crisis, algo trágico, tal como una guerra, y por ende es negativo, hese aquí la primer prueba de la revolución, estando después *revolution*. También se cuenta con adjetivos lúgubres: *black, terrible, darkness*. Aunque la aparición que más detalla el rumbo de la historia es la de *lost* y *grave*.

En comparación, con Bing, *poor* lo describe en los primeros registros del Cuadro 5. Aunque *dead* es el tercer más importante de acuerdo a esta escala. La sinonimia de "Los miserables" con pobreza, puede comenzar a cobrar sentido. Por otro lado, en el Cuadro 6 se descarta a *child* y *god,* y se le da mayor peso a *love.*

Sobre la diversidad léxica, inmersa en la obra de Víctor Hugo, la pendiente de la Figura 2 explica que conforme el texto avanza, la diversidad léxica aumenta; y el suavizamiento explica cómo en los últimos grupos de 200 líneas se encuentra la mayor diversidad. Es decir, hay una mayor variedad en las últimas páginas. Una hipótesis lógica sería tener como la inclusión de nuevos personajes, escenarios o actividades.

Para tener una segunda referencia sobre la distribución léxica, la Figura 3 ajusta una recta que marca una tendencia negativa en cuanto a su densidad; esto implica que a pesar de haber gran variedad de palabras en los últimos capítulos, también hay mucho texto. Sin embargo, el pico de mayor densidad está en la misma sección que la hipótesis del párrafo anterior. La densidad promedio del texto es .32, o 32 palabras diferentes por cada 100. Las *stopwords* estuvieron presentes para los ejercicios de distribución léxica. Una vez dicho esto, concluiría que Víctor Hugo fue un escritor creativo con respecto al empleo de las palabras.

En la Figura 4 existe, gráficamente, una mayor cantidad palabras alegres, aunque, a pesar de que las palabras tristes no tienen todo el peso en frecuencia, su intensidad sí es representativa y es la que se observa en la Figura 5, que ilustra mejor este fenómeno. Es importante recordar que cada palabra tiene distinta puntuación sentimental (Madani, Erritali y Bengourram, 2018), es por eso que existen palabras tristes que contrarrestan y rebasan a las palabras alegres en la mayoría de enunciados.

De nuevo, la Figura 5, con las barras azules hacia abajo confirma que se trata de una obra sumamente triste, donde las palabras que la conforman son en su mayoría negativas. La distribución no es tan fácil de reconocer, pero con un poco de detalle, podemos observar, situándonos en el índex 200 y 240, dos valles negativos relativamente marcados que se traducirían como situaciones de alta carga negativa, incluso podría tratarse del clímax de la obra. En todos los casos, las líneas negativas superan, en número y dimensión, a las positivas.

Se puede efectuar una comparación con la Figura 6, que está hecha con base en AFINN; aunque los resultados resumen varias conclusiones de forma similar, reafirmando que AFINN pondera con un mayor intervalo a sus elementos.

En ambas gráficas hay una caída, en el mismo punto: el *index* 242. Es inmediato cuestionarnos lo que dice dicho conjunto, y es por eso que se realizó una nube con esta intención, que es la que aparece en la Figura 7. Palabras como alcantarilla, Jean Valjean, Paris, calle, Marius, barricada, Enjolras, tiempo, luz, muerte. De hecho, en esta sección muere Gavroche, y es donde la guerra verdaderamente pues hay sangre y la batalla es descrita minuciosamente.

NRC, en la Figura 8, sin considerar *positive* y *negative,* expresa a los sentimientos próximos a la confianza como los más frecuentes, y en segundo lugar *fear* que podría representar al miedo.

4.3 ANÁLISIS SEMÁNTICO

El análisis para la descripción verbal (Figura 9) es diferente, puesto que se puede caer en falacias. Sin embargo, la semántica (Mikolov, Yih y Zweig, 2013) apoya a este tarea considerablemente; es decir, es claro que verbos como *wish* o *love* podrían representar emociones tales como esperanza, deseo, en un tono positivo, aunque precedidos por la conjunción *not* este significado se invierte, e incluso acentúa una connotación negativa.

Es la representación de la falta de deseo, desamor, marginación social y descuido, la que se identifica con *wish, love, rich* y *care.* En sentido opuesto, *hesitate, kill, afraid, die* junto a *not* significan precisión, misericordia y esperanza.

Hacia el reconocimiento de adjetivos y situaciones particulares, mediante el uso una técnica, parecida a la anterior, que reduce el sesgo. Si se utilizan auxiliares verbales como *was, is* o *will* es posible hallar mejores y más precisos resultados explicativos, además de temporalidad o dirección.

Si se comparan, en la Figura 10, *was lost, was dead* versus *was happy* y *was good*, la proporción es casi el doble en las composiciones negativas. La muerte impera, convirtiéndose así en una tragedia, basándome en la definición de la Universidad de Oxford, donde la define como "Una obra seria con un final triste, especialmente en el que el personaje principal muere… (Oxford University Press, 2019)". Situándonos con la temporalidad que aporta *was*, hace una descripción de algo pasado, marca del pretérito, incluso puede implicar una acción ya terminada. Hay escenarios felices, buenos, hermosos, alegres y de riqueza; también asesinato, forcejeo, arresto, condena y sufrimiento. La distribución de los sentimientos negativos es variada, subraya encarcelamiento, o delitos, y temor.

En la Figura 11, donde se utiliza *is* como auxiliar, la verdad es de la misma magnitud de la muerte; la bondad, proveniente de *good*, contra *terrible*, refiere a una postura de esperanza pero sustentada en un marco funesto. El presente, cuyo representante es *is*, explica cómo establece, Víctor Hugo, escenarios vigentes en la historia, igualmente descriptiva. Por otro lado, esta figura está más equilibrada, en el grado positivo y negativo, que la hecha con *was*.

En resumen, el pasado es negativo, el presente es neutro.

Lo que sigue es el análisis del futuro. El inglés es práctico si del futuro queremos hablar, usa un auxiliar que indudablemente redirige al futuro, a lo que habrá de ser o suceder. Sin embargo, éste no cuenta con magnitudes tan amplias como el presente o el pasado, el rango del futuro se encuentra en -15 y 5. La palabra que estuvo más combinada con *will* es *pay,* la cual cuenta con un sentido literal que implica tal vez el pago de una fianza, o el sentido figurado que expresa venganza. Siguiendo, *kill* y *die* tienen intención de amenaza o peligro de muerte literal. "Ayudará, bendecirá o logrará", son símbolos esperanzadores, y es aquí donde el futuro puede atribuir ilusión, aunque el fatalismo sigue imperando en mayor medida y con mayor frecuencia. Por consiguiente, es propio concluir que el futuro es negativo.

El análisis verbal que sustenta Eastwood (Eastwood, 1994) en su gramática, dicta que constituye 4 elementos. El tiempo está sustentado de forma evidente, en presente, pasado y futuro. El modo, indicativo y subjuntivo, también está intrínseco, cuando *is* incluye a ambas categorías. La persona es la tercera del singular, sin embargo, el equilibrio lo da *will* que abarca todas las personas verbales. El uso del auxiliar fue, precisamente, para que tanto voces activas como pasivas tuvieran lugar, otra bondad de los verbos copulativos como *be*.

Conocer las relaciones de los personajes es sencillo usando los cuadros de correlación, ya que la distribución de las palabras sí es proporcional con su relación en la obra (Mikolov *et al.*, 2013).

Alfabéticamente en la figura Figura 13, Fantine es el primer elemento a estudiar. Fantine tiene una correlación fuerte con Tholomyes, quien fue el amante de Fatine y padre biológico de Cosette. Favourite era el nombre de una de las amigas de Fantine, y Blachevelle su novio. Si nos decidimos a analizar *thy*, ésta se emplea comúnmente en textos ingleses clásicos, y representa a la tercera persona del singular, pero ya está en desuso y es muy formal; regularmente se designaba para invocar a Dios.

Javert es la palabra que continúa, y está estrechamente relacionado con Valjean y también con inspector, aunque como Valjean tiene relación con *convict*, es inmediato inferir: Javert es un inspector, espía, policía, que busca a Valjean, un convicto. Javert, además de inspector, es policía, pero tiene relación con un alcalde.

Thenardier y *madame* están correlacionadas, entonces sería correcto describir a *Madame Thenardier*, de quien se pueden apreciar bastantes características, por el tratamiento de *madame*, podemos inferir que se le merece respeto, o cuenta con autoridad; pero es necesario comentar *ruffians* para otorgarle una figura antagónica, además su marido está presente. Éponine también aparece con los Thénardiers, por lo que no sería correcto atribuirla a su familia. Leblanc es uno de los sobrenombres que se adjudica Jean Valjean, lo que explica la relación de estas dos figuras.

Valjean, o Jean Valjean, será un convicto, perseguido por Javert, pero que tiene la misma relación con Cosette. Cosette está rodeada de amor, descrito por la Figura 14. Toussaint es la sirvienta de Jean Valjean y Cosette.

Éponine tiene una hermana llamada Azelma, convive con el ladrón Babet, vive con los Thenardiers y tiene un sacudidor.

Gavroche también vive en Montparnasse, hay un elefante en la importancia de su personaje, *gamin* tiene un sentido de pilluelo, atribuible a un niño.

Así, Marius Pontmercy, aparece continuamente con su abuelo, y Cosette. *Heart, eyes, love* promueven una personalidad enamoradiza.

Analizando palabras con un sentido extremo en la Figura 15, obtenemos conclusiones más detalladas. La muerte, el mal, Dios y el amor son el escenario. La muerte está en los tejados, en las desembocaduras de los ríos, o literalmente las bocas, y están cubiertas. El mal está en el alba, en la miseria, cubierto por una máscara. Dios define la ley, el alma, la muerte y al amor. El amor está en el corazón, las aves, lo celestial y la juventud.

De estas tres tablas las palabras, que presentan una correlación alta, son *love-heart, eponine-azelma, jean-valjean, javert-inspector,* donde su índice está encima de .20.

Preguntas de relación, que no hubieran podido haber quedado claras, pueden ser resueltas a través de la Figura 16. Es decir, dentro de la historia, Madeleine, sí es un personaje, de hecho es un padre. *Plumet* que no tenía sentido junto a Éponine, ahora es más sencillo concluir que es el nombre de una calle. La moneda son los francos. Y Gillenormand es un apellido, así como Magloire.

Si se presenta el error "No tidy method for objects of class LDA_Gibbs", cuando se haya cerrado la sesión anteriormente y se vuelva abrir, entonces tendremos que dar "Terminate Session" y correr el código de nuevo; hay un bug en *topicmodels* que ocasiona esto.

Los hallazgos más subjetivos son, definitivamente, los que surgieron del LDA. Los seis temas de la Figura 17, son capítulos fáciles de reconocer, excepto uno que es el que contiene los sustantivos más generales; cuando se agrupan por el significado que pudieron haber tenido, y se etiquetan, es posible crear las gráficas de la Figura 18, que es un *suavizamiento* del conteo. La primera gráfica (extremo superior izquierdo) es cuando Valjean conoce a Cosette en el bosque, que ésta escena se sitúa cerca del capítulo 50, las palabras filtros para este caso fueron *child, forest, doll, water*. El segundo tema, de línea azul, es claro que representa la guerra. El verde, al ser tan general, no se le pudo dar una descripción congruente. La escena de la rueda que es levantada por Jean Valjean, que es emblemática, porque es cuando Javert reconsidera que el alcalde podría ser aquel ex convicto que conoció en prisión. La cena con el obispo es el inicio del libro. Para el tema 6, había considerado el capítulo cuando Jean Valjean encuentra el *convento* y cuando una *monja* fallece, puesta en un *ataúd*, deciden trasladarla a un *cementerio,* pero con el propósito de que intercambiar el contenido del ataúd.

Por otro lado, la Figura 19 es la relación de los seis temas hacia los verbos más *probables*, literalmente. Escenas de cariño, de robo, de movimiento y de sedentarismo, de asesinato y tristeza fueran las clasificaciones interpretadas ilustradas por la Figura 20. La gráfica de cariño es interesante que tiene una trayectoria parecida a la de guerra de la Figura 18.

Respondiendo a cómo se comportan los adjetivos, la Figura 21 contiene a los adjetivos más populares para cada tema. Los temas fueron etiquetados de la forma siguiente: enfermedad, secreto, moribundo, trágico, liderazgo y valentía, respectivamente. Dotar de un título a un grupo de adjetivos es más complicado, porque pueden existir conexiones disparatadas. Aunque cuando graficamos la distribución, en la Figura 22, enfermedad es la que mejor descrita, así como tráfico y valentía; casi todas apuntando al final de la historia como el más característico.

4.4 ANÁLISIS LITERARIO

Basándome en el método de McGee (McGee, 2001), se siguen los siguientes pasos:

1. Identificar el género al que pertenece.

Los Miserables de Víctor Hugo, pertenece al género literario de Novela.

2. Conocer a los personajes, ya que cuentan con características morales y psicológicas.

Jean Valjean es un exconvicto que busca escapar de la justicia, acompañado por una niña: Cosette. Cosette rodeada de amor, de Jean Valjean y Marius. Javert figura como el inspector que desea atrapar a Jean Valjean durante la trama. Los Thénardiers, cuyo aspecto asemeja ser el de unos rufianes, que tienen dos hijas llamadas Éponine y Azelma. Fantine una mujer siempre devota a la misericordia de Dios.

3. Saber si se trata de una ficción o un drama; o ambos.

Es un drama, rodeado de esperanza en Dios.

4. Hallar el contexto, dónde se lleva a cabo la acción.

Transcurre en Francia, en el margen de la revolución, la guerra. Hay amor, muerte, frío, felicidad, alegría y libertad. Robo, cariño, enfermedad, tragedia y valentía.

5. Descifrar los simbolismos, aquellos sellos plasmados por el autor.

La muerte es un proceso agonizante, inquisitivo pero valiente. El gobierno malvado conspirará para crear sufrimiento, pero con la ayuda de Dios la felicidad, el amor y las almas honestas triunfarán. Aunados a las infidelidades y el amor juvenil.

6. Encontrar perspectivas culturales e históricas alrededor.

Guerra: La revolución francesa como un proceso, políticamente, regenerativo de la perspectiva social.

7. Establecer el comentario final, sobre la opinión propia.

Una tragedia, rodeada de guerra, que involucra elementos divinos, pero que exhorta al lector a ser compasivo, y confiar en la misericordia de los actos.

4.5 CONCLUSIÓN

Se verificó que el análisis sentimental y semántico dota al usuario de nociones tanto generales como particulares para comprender la obra clásica. Es cierto, como propone Vargas Llosa (Vargas-Llosa, 2004), que la obra muestra contrastes sentimentales bien definidos, en su mayoría negativos.

El análisis sentimental y el análisis semántico se deben usar en conjunto para la minería de opinión, ya que proveen al analista de sujeto y predicado. Ésto confirma la proposición ofrecida por Chomsky (Chomsky, 1955) sobre la gramática generativa.

La semántica tiene buena aproximación mediante la correlación, además de que el intervalo en el que está definido la correlación aporta practicidad al momento de interpretar la información, y así se les puede dotar de un sinónimo por dicha característica, tal como diversos autores proponen (Chomsky, 1955; Goodenough y Rubenstein, 1965; Cambria y Hussain, 2012). Sobre la sinonimia, LDA crea temas que ayudan englobar características latentes, que sirven para obtener una referencia más generalizada.

La heurística tiene bien apego cuando para tratar el análisis literario.

Pinker y Jackendoff (Pinker y Jackendoff, 2005) argumentan que la gramática no es recursiva, lo cual es falso. Hay varios patrones gramaticales (voz pasiva, negación, pasado y participio) que sirven para el análisis, como lo demostró este estudio.

Existen grandes avances de *software* para facilitar la tarea, aunque el contexto es imprescindible para determinar si existen excepciones.

El Procesamiento Natural del Lenguaje (Stine, 2019) es una labor que en esta novela no presenta dificultad, por ser un texto de minuciosa edición.

4.6 INVESTIGACIÓN FUTURA

El estudio se debe extender a lenguas distintas al inglés, tal vez comenzar con aquellas que sea de su misma familia lingüística, o haciendo la interpretación de prefijos y sufijos que están altamente relacionados de entre las lenguas indoeuropeas. El uso del acento gráfico de las lenguas romances puede servir como un diferenciador expreso para el análisis.

Asimismo, este estudio se puede replicar en obras importantes de la literatura, o dirigirlo al análisis de las redes sociales, donde Twitter es una empresa que está brindando muchas facilidades en torno a la minería de opinión.

Este código se podría adecuar a un paquete de R, o a una interfaz interactiva como una Shiny.

5 Referencias

Aguilar, M. Á. (2004) "Chomsky, La Gramática Generativa", *Investigación y educación*, 3(7).

Alkhodair, S. A. *et al.* (2019) "Detecting breaking news rumors of emerging topics in social media". doi: 10.1016/j.ipm.2019.02.016.

Amat Rodrigo, J. (2016) *RPubs - Correlación lineal y regresión lineal simple en R, RPubs*. Disponible en: https://rpubs.com/Joaquin_AR/223351 (Consultado: el 20 de julio de 2019).

Amat Rodrigo, J. (2017) *RPubs - Clustering y heatmaps: aprendizaje no supervisado con R, RPubs*. Disponible en: https://rpubs.com/Joaquin_AR/310338 (Consultado: el 20 de julio de 2019).

Anderson, S. C. (2013) *An Introduction to reshape2 - Reshaping data easily with the reshape2 R package. - seananderson.ca*. Disponible en: https://seananderson.ca/2013/10/19/reshape/ (Consultado: el 26 de julio de 2019).

Barret, L. F. (2006) *Solving the Emotion Paradox: Categorization and the Experience of Emotion, Personality and Social Psychology Review*.

Bolukbasi, T. *et al.* (2016) *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*. Disponible en: http://cs.cmu.edu/directory/csd. (Consultado: el 11 de abril de 2019).

Bondielli, A. y Marcelloni, F. (2019) "A survey on fake news and rumour detection techniques", *Information Sciences*, 497, pp. 38–55. doi: 10.1016/j.ins.2019.05.035.

Cambria, E. y Hussain, A. (2012) *Sentic Computing*. Springer. Editado por Springer. Springer. doi: 10.1007/978-3-319-23654-4 ISBN.

Casella, G. y Berger, R. L. (2002) *Statistical Inference*. Duxbury.

Chen, M.-Y. y Chen, T.-H. (2019) "Modeling public mood and emotion: Blog and news sentiment and socio-economic phenomena", *Future Generation Computer Systems*, 96, pp. 692–699. doi: 10.1016/j.future.2017.10.028.

Chomsky, N. (1955) "Logical Syntax and Semantics: Their Linguistic Relevance", *Language*, 31(1), pp. 36–45.

Chomsky, N. (2007) "Biolinguistic explorations: Design, development, evolution", *International Journal of Philosophical Studies*, 15(1), pp. 1–21. doi: 10.1080/09672550601143078.

Clapp, M. (2010) *Introducción al Análisis Real*. Editado por Universidad Nacional Autónoma de México. Universidad Nacional Autónoma de México.

Clark, S. (2013) *Topic Modelling and Latent Dirichlet Allocation*.

Collobert, R. *et al.* (2011) "Natural Language Processing (Almost) from Scratch", *Journal of Machine Learning Research*, 12, pp. 2493–2537. Disponible en: http://www.jmlr.org/papers/volume12/collobert11a/collobert11a.pdf (Consultado: el 28 de marzo de 2019).

Cowles, M. K. (2013) *Applied Bayesian statistics: with R and OpenBUGS examples*. New York: Springer Science & Bussiness. doi: 10.1016/j.peva.2007.06.006.

Eastwood, J. (1994) *Oxford Guide To English Grammar*. Oxford University Press.

Elman, J. L. (1991) "Distributed Representations, Simple Recurrent Networks, And Grammatical Structure", *Machine Learning*, 7(2), pp. 195–225. doi: 10.1023/A:1022699029236.

Evans, N. y Levinson, S. C. (2009) "The myth of language universals: Language diversity and its importance for cognitive science", *BEHAVIORAL AND BRAIN SCIENCES*, 32, pp. 429–492. doi: 10.1017/S0140525X0999094X.

Everaert, M. B. H. *et al.* (2015) "Structures, Not Strings: Linguistics as Part of the Cognitive Sciences", *Trends in Cognitive Sciences*. Elsevier Ltd, 19(12), pp. 729–743. doi: 10.1016/j.tics.2015.09.008.

Gibson, E. (1998) "Linguistic complexity: locality of syntactic dependencies", *Cognition*, 68, pp. 1–76.

Gil Bellosta, C. J. (2018) *R para profesionales de los datos: una introducción*. Disponible en: https://www.datanalytics.com/libro_r/elementos-de-un-grafico-en-ggplot2.html (Consultado: el 24 de julio de 2019).

Goodenough, J. B. y Rubenstein, H. (1965) "Contextual correlates of synonymy", *Communications of the ACM*, 8(10). doi: 10.1145/365628.365657.

Grolemund, G. (2018) *Strings - Cheatsheets*. Disponible en: https://github.com/rstudio/cheatsheets/blob/master/strings.pdf (Consultado: el 29 de julio de 2019).

He, W. *et al.* (2015) "A novel social media competitive analytics framework with sentiment benchmarks", *Information & Management*. North-Holland, 52(7), pp. 801–812. doi: 10.1016/J.IM.2015.04.006.

Hipson, W. E. (2019) "Using sentiment analysis to detect affect in children's and adolescents' poetry", *International Journal of Behavioral Development*, 43(4), pp. 375–382. doi: 10.1177/0165025419830248.

Hugo, V. (1887) *Les Misérables*. Disponible en: http://www.gutenberg.org/files/135/135-0.txt (Consultado: el 10 de abril de 2019).

De la Fuente Fernández, S. (2011) *Análisis Conglomerados*. Disponible en: http://www.fuenterrebollo.com/Economicas/ECONOMETRIA/SEGMENTACION/CONGLOMERADOS/conglomerados.pdf (Consultado: el 29 de enero de 2020).

Lang, S. (2000) *Linear Algebra*. 3rd ed. Springer.

Lieberman, H. y Selker, T. (2000) "Out of context: Computer systems that adapt to, and learn from, context", *IBM systems journal*, 39(3.4), pp. 617–632. Disponible en: https://pdfs.semanticscholar.org/4407/6c914d6de7dcc7edcc6abd562a740ef854ee.pdf (Consultado: el 25 de junio de 2019).

Liske, D. (2018a) *Machine Learning and NLP using R (article) - DataCamp*. Disponible en: https://www.datacamp.com/community/tutorials/ML-NLP-lyric-analysis (Consultado: el 23 de abril de 2019).

Liske, D. (2018b) *R NLP & Machine Learning: Lyric Analysis (article) - DataCamp*. Disponible en: https://www.datacamp.com/community/tutorials/R-nlp-machine-learning (Consultado: el 11 de abril de 2019).

Liske, D. (2018c) *Tidy Sentiment Analysis in R (article) - DataCamp*. Disponible en: https://www.datacamp.com/community/tutorials/sentiment-analysis-R (Consultado: el 23 de abril de 2019).

Liu, B. (2012) *Sentiment Analysis(Introduction and Survey) and Opinion Mining, Morgan &Claypool*. doi: 10.1162/COLI.

Liu, E. (2015) *Latent Dirichlet Allocation Using Gibbs Sampling*. Disponible en: http://ethen8181.github.io/machine-learning/clustering_old/topic_model/LDA.html#content (Consultado: el 5 de enero de 2020).

MacArthur, S. (2010) *R: Using RColorBrewer to colour your figures in R | R-bloggers*. Disponible en: https://www.r-bloggers.com/r-using-rcolorbrewer-to-colour-your-figures-in-r/ (Consultado: el 24 de julio de 2019).

Madani, Y., Erritali, M. y Bengourram, J. (2018) "Sentiment analysis using semantic similarity and Hadoop MapReduce", *Knowledge and Information Systems*. Springer London, pp. 1–24. doi: 10.1007/s10115-018-1212-z.

McGee, S. J. (2001) *Analyzing Literature: A Guide for Students.* Addison-Wesley Longman. Disponible en: http://wps.ablongman.com/wps/media/objects/327/335558/AnalyzingLit.pdf (Consultado: el 31 de julio de 2019).

Mikolov, T. *et al.* (2013) "Distributed Representations of Words and Phrases and their Compositionality", *Advances in neural information processing systems*, pp. 3111–3119. Disponible en: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf (Consultado: el 11 de abril de 2019).

Mikolov, T., Yih, W.-T. y Zweig, G. (2013) *Linguistic Regularities in Continuous Space Word Representations*. Association for Computational Linguistics. Disponible en: http://research.microsoft.com/en- (Consultado: el 11 de abril de 2019).

Miller, G. A. (1995) *WordNet Search - 3.1*. Disponible en: http://wordnetweb.princeton.edu/perl/webwn?s=Rue&sub=Search+WordNet&o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&h= (Consultado: el 21 de mayo de 2019).

Mirończuk, M. M., Protasiewicz, J. y Pedrycz, W. (2019) "Empirical evaluation of feature projection algorithms for multi-view text classification", *Expert Systems with Applications*, 130, pp. 97–112. doi: 10.1016/j.eswa.2019.04.020.

Mostafa, M. M. (2013) "More than words: Social networks' text mining for consumer brand sentiments", *Expert Systems With Applications*, 40, pp. 4241–4251. doi: 10.1016/j.eswa.2013.01.019.

Niebles, J. C., Wang, H. y Fei-Fei, L. (2008) "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words", *Int J Comput Vis*, 79, pp. 299–318. doi: 10.1007/s11263-007-0122-4.

Oxford University Press (2019) *tragedy noun - Definition, pictures, pronunciation and usage notes | Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.com*. Disponible en: https://www.oxfordlearnersdictionaries.com/definition/english/tragedy (Consultado: el 31 de julio de 2019).

Pak, A. y Paroubek, P. (2010) "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", *Analysis*, pp. 1320–1326. doi: 10.1371/journal.pone.0026624.

Pérez-Torres, A. (2014) *Los Géneros Literarios*. Disponible en: https://www.uaeh.edu.mx/docencia/VI_Lectura/bachillerato/documentos/2014/LECT120.pdf (Consultado: el 25 de junio de 2019).

Pinker, S. y Jackendoff, R. (2005) "The faculty of language: what's special about it?", *Cognition*, 95, pp. 201–236. doi: 10.1016/j.cognition.2004.08.004.

Ravi, K. y Ravi, V. (2015) "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications", *Knowledge-Based Systems*. Elsevier B.V., 89, pp. 14–46. doi: 10.1016/j.knosys.2015.06.015.

Real Academia Española (2019a) *«Diccionario de la lengua española» - Edición del Tricentenario, Real Academia Española*. Disponible en: https://dle.rae.es/srv/fetch?id=NNPFPOl (Consultado: el 22 de julio de 2019).

Real Academia Española (2019b) *heurístico, ca | Definición de heurístico, ca - «Diccionario de la lengua española» - Edición del Tricentenario*. Disponible en: https://dle.rae.es/?id=KHdGTfC (Consultado: el 25 de junio de 2019).

Rincón, L. (2012) *Introducción a los procesos estocásticos, Departamento de Matemáticas, Facultad de Ciencias UNAM*. doi: 10.1016/j.anifeedsci.2007.06.033.

Robert, C. P. y Casella, G. (2010) *Introducing Monte Carlo Methods With R*. New York: Springer. doi: 10.1007/978-0-387-78171-6.

Robinson, D. y Silge, J. (sin fecha) *Sentiment lexicons from three sources*. Disponible en: https://juliasilge.github.io/tidytext/reference/sentiments.html (Consultado: el 4 de mayo de 2019).

Silge, J. y Robinson, D. (2019) *4 Relationships between words: n-grams and correlations | Text Mining with R*. Disponible en: https://www.tidytextmining.com/ngrams.html (Consultado: el 4 de mayo de 2019).

Stine, R. A. (2019) "Sentiment Analysis", *Annual Reviews*, 6, pp. 287–308. doi: 10.1146/annurev-statistics.

Tausczik, Y. R. y Pennebaker, J. W. (2010) "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods", *Journal of Language and Social Psychology*, 29(1), pp. 24–54. doi: 10.1177/0261927X09351676.

The Igraph Core Team (2015) *R igraph manual pages*. Disponible en: https://igraph.org/r/doc/aaa-igraph-package.html (Consultado: el 26 de julio de 2019).

The R Foundation (2019) "R Program". Disponible en: https://www.r-project.org/.

Tomasello, M. *et al.* (2005) "Understanding and sharing intentions: The origins of cultural cognition", *BEHAVIORAL AND BRAIN SCIENCES*, 28, pp. 675–735. doi: 10.1017/S0140525X05000129.

Universidad Complutense de Madrid (2013) *La derivación, Proyecto de Innovación y Mejora de la Calidad Docente Nº 131-2013*. Disponible en: https://www.ucm.es/plataformaele/la-derivacion (Consultado: el 11 de enero de 2020).

Uryu, S. (2017) *widyr | R Package Showcase*. Disponible en: http://uribo.github.io/rpkg_showcase/data_handling/widyr.html (Consultado: el 29 de julio de 2019).

Vargas-Llosa, M. (2004) *The Temptation of the Impossible*. Princeton University Press.

Wiebe, J., Wilson, T. y Cardie, C. (2005) "Annotating expressions of opinions and emotions in language", *Language Resources and Evaluation*, 39(2–3), pp. 165–210. doi: 10.1007/s10579-005-7880-9.

Wing-ki Leung, C. *et al.* (2011) "A probabilistic rating inference framework for mining user preferences from reviews", *World Wide Web*, 14, pp. 187–215. doi: 10.1007/s11280-011-0117-5.

Zadeh, L. A. (1975a) "The Concept of a Linguistic Variable and its Application to Approximate Reasoning-I", *Information Sciences*, 8, pp. 199–249. Disponible en: https://pdf.sciencedirectassets.com/271625/1-s2.0-S0020025500X02186/1-s2.0-0020025575900365/main.pdf?x-amz-security-token=AgoJb3JpZ2luX2VjEHsaCXVzLWVhc3QtMSJHMEUCIQDu%2FvnpligiNBJpM2avKjcAMrH%2BdiNvRmOwVucVwrK19gIgCeU%2B%2BWlOJSY6RBZDmWmHM%2FTRRDEDVz8NsRK (Consultado: el 28 de marzo de 2019).

Zadeh, L. A. (1975b) "The Concept of a Linguistic Variable and its Application to Approximate Reasoning-II*", *Information Sciences*, 8, pp. 301–357. Disponible en: https://pdf.sciencedirectassets.com/271625/1-s2.0-S0020025500X02198/1-s2.0-0020025575900468/main.pdf?x-amz-security-token=AgoJb3JpZ2luX2VjEHsaCXVzLWVhc3QtMSJHMEUCIQC3j0g%2Bo%2BqOkRfbI10acluyLNraQIontRH3ATRk54EYqwIgL%2F%2BoEC5Q0BoCXYGfhgD7dkLKReIQYiy014ETH (Consultado: el 28 de marzo de 2019).

6 Anexo

6.1 CÓDIGO EN R

```
getwd() #Hay que asegurarnos que el archivo esté en la dirección adecuada, que en su defecto será el mismo que el Proyecto


# Librerías -----------------------------------------------------------------


library(tidytext)

library(dplyr)

library(tidyr)

library(wordcloud)

library(RColorBrewer)

library(stringr)

library(reshape2)

library(igraph)

library(ggplot2)

library(ggraph)

library(widyr)

library(topicmodels)

library(ggrepel)

library(gridExtra)

library(clipr)

library(syuzhet) #Para utilizar NRC, ya no aparece en la nueva versión de tidytext

library(cleanNLP)


# Limpieza de Texto ---------------------------------------------------------

libro <- read.delim(file="lesmis.txt",encoding = 'UTF-8') #Retiré manualmente la última parte, donde se agregan las notas de pie de página, y el anexo

kleen <- libro #Se carga el txt de los miserables, la versión descargada del Gutenberg Project.


fix.contractions <- function(doc) { # Expande las contracciones del inglés

  doc <- gsub("won't", "will not", doc)

  doc <- gsub("can't", "can not", doc)
```

```
  doc <- gsub("n't", " not", doc)
  doc <- gsub("'ll", " will", doc)
  doc <- gsub("'re", " are", doc)
  doc <- gsub("'ve", " have", doc)
  doc <- gsub("'m", " am", doc)
  doc <- gsub("'d", " would", doc)
  doc <- gsub("'s", "", doc)
}

#Arreglamos las contracciones
kleen <- kleen %>% sapply(fix.contractions)

kleen <- kleen[-c(1:499),] %>% as.data.frame() #Es importante quitar la primera parte, ya que ahí se halla el índice, y si no va a fallar el código más adelante.

colnames(kleen) <- "text"

#Para limpiar el texto de caracteres especiales
kleen <- sapply(kleen, function(x) iconv(x,from="UTF-8",to="ASCII//TRANSLIT")) %>% as.data.frame()

#Los acentos los separa con una apostrofe, por lo que se debe hacer una función especial para quitarla
apostrofe <- function(x) gsub("'", "", x)
kleen <- sapply(kleen,apostrofe) %>% as.data.frame()

#Quita signos de puntuación y demás basura no alfabetica
removeSpecialChars <- function(x) gsub("[^a-zA-Z0-9 ]", " ", x)
kleen <- sapply(kleen, removeSpecialChars) %>% as.data.frame()

kleen <- kleen[-c(223,12367),] %>% as.data.frame() #Retiraremos estas dos filas, ya que están vacías

colnames(kleen) <- "text"

kleen <- kleen %>% mutate(text=as.character(kleen$text))
```

```
tidykleen <- kleen %>%
  mutate(linenumber = row_number(),
        chapter = cumsum(str_detect(text, regex("^chapter [\\divxlc]",
                                    ignore_case = TRUE)))) %>%
  ungroup() %>%
  unnest_tokens(word, text)%>%
  anti_join(stop_words)

# tidykleen <- tidykleen[ which(tidykleen$chapter>0)]
#Si se decide incluir el código con la introducción, esta línea sirve para excluir toda la información hasta el primer capítulo

# Análisis de Frecuencia ----------------------------------------------------

conteo <- tidykleen %>% #para conocer los resultados preliminares por frecuencia, en una tabla
  count(word, sort = TRUE) %>%
  head(12) %>%
  write_clip()

# Bases de Sentimientos -----------------------------------------------------

bingset <- get_sentiments("bing") #La librería bing contiene una calificación de palabras por su sentido. Sólo existen positivas o negativas.

nrcset <- get_sentiment_dictionary(dictionary = "nrc") %>% select(-lang) #NRC hace una descripción más precisa, a partir del sentimiento general que describe

afinnset <- get_sentiments("afinn") #Afinn da un puntaje numérico a cada palabra en un intervalo de -5 a 5, sólo números enteros.

loughranset <- get_sentiments("loughran")

nrcset %>% count(sentiment,sort=TRUE) %>% write_clip() #Existen 10 clasificaciones: Negative, positive, fear, anger, trust, sadness, disgust, anticipation, joy, surprise
```

```
loughranset %>% count(sentiment,sort=TRUE) #Existen 6 clasificaciones: nefative, litigious, positive, uncertainty, constraining, superfluous.


bingsentiment <- tidykleen %>%
  inner_join(bingset) %>%
  count(index = linenumber %/% 80, sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = positive - negative)


nrc_joy <- nrcset %>%
  filter(sentiment == "joy")


nrc_sad <- nrcset %>%
  filter(sentiment == "sadness")


tidykleen %>%
  inner_join(nrc_joy) %>%
  count(word, sort = TRUE) %>%
  head(10) %>%
  write_clip()


tidykleen %>%
  inner_join(nrc_sad) %>%
  count(word, sort = TRUE)%>%
  head(10) %>%
  write_clip()


tidykleen %>%
  inner_join(bingset %>% filter(sentiment == 'negative')) %>%
  count(word, sort = TRUE)%>%
  head(10) %>%
  write_clip()


tidykleen %>%
```

```
  inner_join(bingset %>% filter(sentiment == 'positive')) %>%
  count(word, sort = TRUE)%>%
  head(10) %>%
  write_clip()

# Nubes de Palabras ----------------------------------------------------

tidykleen %>%
  anti_join(stop_words) %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 20, colors = blues9))

# Lexical Diversity ----------------------------------------------------------

trial <- kleen %>%
  mutate(index = row_number() %/% 200)

lex_diversity_per_line <- trial %>%
  unnest_tokens(word, text) %>%
  group_by(index) %>%
  summarise(lex_diversity = n_distinct(word)) %>%
  arrange(desc(lex_diversity))

mean(lex_diversity_per_line$lex_diversity)

my_colors <- c("#E69F00", "#56B4E9", "#009E73", "#CC79A7", "#D55E00")

theme_lyrics <- function()
{
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_blank(),
        axis.ticks = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
```

```
      legend.position = "none")
}

diversity_plot <- lex_diversity_per_line %>%
  ggplot(aes(index, lex_diversity)) +
  geom_point(color = my_colors[3],
             alpha = .4,
             size = 4,
             position = "jitter") +
  stat_smooth(color = "black", se = FALSE, method = "lm") +
  geom_smooth(aes(x = index, y = lex_diversity), se = FALSE,
              color = "blue", lwd = 2) +
  ggtitle("Diversidad Léxica") +
  xlab("Grupos de 200 líneas") +
  ylab("Palabras Únicas") +
  scale_color_manual(values = my_colors) +
  theme_classic() +
  theme_lyrics()

diversity_plot

lex_density_per_line <- trial %>%
  unnest_tokens(word, text) %>%
  group_by(index) %>%
  summarise(lex_density = n_distinct(word)/n()) %>%
  arrange(desc(lex_density))

density_plot <- lex_density_per_line %>%
  ggplot(aes(index, lex_density)) +
  geom_point(color = my_colors[4],
             alpha = .4,
             size = 4,
             position = "jitter") +
  stat_smooth(color = "black",
```

```
          se = FALSE,
          method = "lm") +
  geom_smooth(aes(x = index, y = lex_density),
          se = T,
          color = "blue",
          lwd = 2) +
  ggtitle("Densidad Léxica") +
  xlab("Grupos de 200 líneas") +
  ylab("Densidad") +
  scale_color_manual(values = my_colors) +
  theme_classic() +
  theme_lyrics()

density_plot

mean(lex_density_per_line$lex_density)

# Distribución de Sentimientos ----------------------------------------

tidykleen %>%
  inner_join(bingset) %>%
  count(word, sentiment, sort = TRUE) %>%
  acast(word ~ sentiment, value.var = "n", fill = 0) %>%
  comparison.cloud(colors = c("black", "orange"),
                   max.words = 50)

afinnsentiment <- tidykleen %>%
  inner_join(afinnset) %>%
  group_by(index = linenumber %/% 80) %>%
  summarise(sentiment = sum(value))

ggplot(bingsentiment, aes(index, sentiment))+
  geom_col(show.legend = FALSE, fill = "Blue")
```

```
ggplot(afinnsentiment, aes(index, sentiment))+
  geom_col(show.legend = FALSE, fill = "Purple")

bingsentiment %>%
  filter(sentiment==min(sentiment))

tidykleen %>% filter(linenumber %in% (242*80):((243*80)-1)) %>% distinct(linenumber)

cap_mas_triste <- tidykleen %>% filter(linenumber %in% c(19360:19439)) %>% select(-chapter)

pal = brewer.pal(9,"Purples")

cap_mas_triste %>%
  anti_join(stop_words) %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 30, colors = pal))

nrcsentiment %>%
  group_by(sentiment) %>%
  summarise(word_count = n()) %>%
  ungroup() %>%
  mutate(sentiment = reorder(sentiment, word_count)) %>%
  ggplot(aes(sentiment, word_count, fill = -word_count)) +
  geom_col() +
  guides(fill = FALSE) +
  labs(x = NULL, y = "Conteo de Palabras")+
  ggtitle("Sentimientos de NRC en el texto") +
  coord_flip()
#N-gramas----------------------
kleen_bigrams <- kleen %>%
  mutate(linenumber = row_number(),
         chapter = cumsum(str_detect(text, regex("^chapter [\\divxlc]",
                                         ignore_case = TRUE)))) %>%
  ungroup() %>%
```

```
  unnest_tokens(bigram, text, token = "ngrams", n=2)

kleen_bigrams %>%
  count(bigram, sort = TRUE)

bigrams_separated <- kleen_bigrams %>%
  separate(bigram, c("word1", "word2"), sep = " ")

bigrams_filtered <- bigrams_separated %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word)

# new bigram counts:
bigram_counts <- bigrams_filtered %>%
  count(word1, word2, sort = TRUE) %>%
  print()

trigramas <- kleen %>% #Hace trigramas
  unnest_tokens(trigram, text, token = "ngrams", n = 3) %>%
  separate(trigram, c("word1", "word2", "word3"), sep = " ") %>%
  filter(!word1 %in% stop_words$word,
         !word2 %in% stop_words$word,
         !word3 %in% stop_words$word) %>%
  count(word1, word2, word3, sort = TRUE)

# Precesión de Not, Was, Is, Will ----------------------------------------------------
bigrams_separated %>%
  filter(word1 == "not") %>%
  count(word1, word2, sort = TRUE)

not_words <- bigrams_separated %>%
  filter(word1 == "not") %>%
  inner_join(afinnset, by = c(word2 = "word")) %>%
  count(word2, value, sort = TRUE) %>%
```

```
  print()

not_words %>%
  mutate(contribution = n * value) %>%
  arrange(desc(abs(contribution))) %>%
  head(20) %>%
  mutate(word2 = reorder(word2, contribution)) %>%
  ggplot(aes(word2, n * value, fill = n * value > 0)) +
  geom_col(show.legend = FALSE) +
  xlab("Palabras precedidas por \"not\"") +
  ylab("Puntaje de Sentimiento * Numero de Ocurrencias") +
  coord_flip()

was_words <- bigrams_separated %>%
  filter(word1 == "was") %>%
  inner_join(afinnset, by = c(word2 = "word")) %>%
  count(word2, value, sort = TRUE)

was_words %>%
  filter(word2 != 'no') %>%
  mutate(contribution = n * value) %>%
  arrange(desc(abs(contribution))) %>%
  head(20) %>%
  mutate(word2 = reorder(word2, contribution)) %>%
  ggplot(aes(word2, n * value, fill = n * value > 0)) +
  geom_col(show.legend = FALSE) +
  xlab("Palabras precedidas por \"was\"") +
  ylab("Puntaje de Sentimiento * Numero de Ocurrencias") +
  coord_flip()

is_words <- bigrams_separated %>%
  filter(word1 == "is") %>%
  inner_join(afinnset, by = c(word2 = "word")) %>%
  count(word2, value, sort = TRUE)
```

```
is_words %>%
  filter(word2 != 'no') %>%
  mutate(contribution = n * value) %>%
  arrange(desc(abs(contribution))) %>%
  head(20) %>%
  mutate(word2 = reorder(word2, contribution)) %>%
  ggplot(aes(word2, n * value, fill = n * value > 0)) +
  geom_col(show.legend = FALSE) +
  xlab("Palabras precedidas por \"is\"") +
  ylab("Puntaje de Sentimiento * Numero de Ocurrencias") +
  coord_flip()

will_words <- bigrams_separated %>%
  filter(word1 == "will") %>%
  inner_join(afinnset, by = c(word2 = "word")) %>%
  count(word2, value, sort = TRUE)

will_words %>%
  mutate(contribution = n * value) %>%
  arrange(desc(abs(contribution))) %>%
  head(20) %>%
  mutate(word2 = reorder(word2, contribution)) %>%
  ggplot(aes(word2, n * value, fill = n * value > 0)) +
  geom_col(show.legend = FALSE) +
  xlab("Palabras precedidas por \"will\"") +
  ylab("Puntaje de Sentimiento * Numero de Ocurrencias") +
  coord_flip()


# Gráficas con N-gramas ----------------------------------------------------

# conteo original
bigram_counts
```

```
bigram_graph <- bigram_counts %>%
  filter(n > 20, word1 != 'NA') %>%
  graph_from_data_frame()

set.seed(2019)

ggraph(bigram_graph, layout = "fr") +
  geom_edge_link() +
  geom_node_point() +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1)

#Gráfica 2
set.seed(2016)

a <- grid::arrow(type = "closed", length = unit(.05, "inches"))

ggraph(bigram_graph, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n), show.legend = FALSE,
                 arrow = a, end_cap = circle(.02, 'inches')) +
  geom_node_point(color = "lightblue", size = 2) +
  geom_node_text(aes(label = name), vjust = .5, hjust = .5) +
  theme_void()


# Correlación ----------------------------------------------------------

word_pairs <- tidykleen %>%
  pairwise_count(word, linenumber, sort = TRUE) %>%
  print()

word_pairs %>%
  filter(item1 == "valjean")
```

```
word_cors <- tidykleen %>%
  group_by(word) %>%
  filter(n() >= 20) %>%
  pairwise_cor(word, linenumber, sort = TRUE)

#Correlaciones con conjuntos de palabra

word_cors %>%
  filter(item1 %in% c("fantine", "javert", "thenardier", "valjean")) %>%
  group_by(item1) %>%
  top_n(6) %>%
  ungroup() %>%
  mutate(item2 = reorder(item2, correlation)) %>%
  ggplot(aes(item2, correlation)) +
  geom_bar(stat = "identity",fill = 'deepskyblue4') +
  facet_wrap(~ item1, scales = "free", ) +
  coord_flip()

word_cors %>%
  filter(item1 %in% c("cosette", "eponine", "gavroche", "marius")) %>%
  group_by(item1) %>%
  top_n(6) %>%
  ungroup() %>%
  mutate(item2 = reorder(item2, correlation)) %>%
  ggplot(aes(item2, correlation)) +
  geom_bar(stat = "identity",fill = 'brown4') +
  facet_wrap(~ item1, scales = "free") +
  coord_flip()

word_cors %>%
  filter(item1 %in% c("death", "evil", "love", "god")) %>%
  group_by(item1) %>%
  top_n(6) %>%
  ungroup() %>%
```

```
  mutate(item2 = reorder(item2, correlation)) %>%
  ggplot(aes(item2, correlation)) +
  geom_bar(stat = "identity",fill = 'cyan4') +
  facet_wrap(~ item1, scales = "free") +
  coord_flip()

# Funciones LDA visualización ----------------------------------------------

word_chart <- function(data, input, title) {
  data %>%
    #set y = 1 to just plot one variable and use word as the label
    ggplot(aes(as.factor(row), 1, label = input, fill = factor(topic) )) +
    #you want the words, not the points
    geom_point(color = "transparent") +
    #make sure the labels don't overlap
    geom_label_repel(nudge_x = .2,
                     direction = "y",
                     box.padding = 0.1,
                     segment.color = "transparent",
                     size = 3) +
    facet_grid(~topic) +
    theme_lyrics() +
    theme(axis.text.y = element_blank(), axis.text.x = element_blank(),
          #axis.title.x = element_text(size = 9),
          panel.grid = element_blank(), panel.background = element_blank(),
          panel.border = element_rect("lightgray", fill = NA),
          strip.text.x = element_text(size = 9)) +
    labs(x = NULL, y = NULL, title = title) +
    #xlab(NULL) + ylab(NULL) +
    #ggtitle(title) +
    coord_flip()
}

top_terms_per_topic <- function(lda_model, num_words) {
```

```
  #tidy LDA object to get word, topic, and probability (beta)
  topics_tidy <- tidy(lda_model, matrix = "beta")

  top_terms <- topics_tidy %>%
    group_by(topic) %>%
    arrange(topic, desc(beta)) %>%
    #get the top num_words PER topic
    slice(seq_len(num_words)) %>%
    arrange(topic, beta) %>%
    #row is required for the word_chart() function
    mutate(row = row_number()) %>%
    ungroup() %>%
    #add the word Topic to the topic labels
    mutate(topic = paste("Tema", topic, sep = " "))
  #create a title to pass to word_chart
  title <- paste("T?rminos LDA m?s populares para", k, "Temas")
  #call the word_chart function you built in prep work
  word_chart(top_terms, top_terms$term, title)
}
# NLP ----------------------------------------------------------------------

cnlp_init_udpipe() #para inicializar el nlp

annotation <- cnlp_annotate(input = kleen) #Hay que tener paciencia, porque este proceso tarda varios minutos

kleen_annotated <- annotation$token #extraemos de la lista únicamente token

names(kleen_annotated)

tidykleen_sections <- tidykleen %>%
  mutate(section = linenumber %/% 4370)
```

```
#Sustantivos

source_tidy <- kleen_annotated %>%
  select(linenumber = doc_id, word= token, lemma, upos) %>%
  filter(upos == "NOUN") %>% #choose only the nouns
  inner_join(tidykleen_sections, by = c("word", "linenumber")) %>%
  select(linenumber,chapter,section, word, lemma, upos) %>%
  distinct()

source_dtm <- source_tidy %>%
  count(section,lemma,sort = T) %>%
  ungroup() %>%
  cast_dtm(section,lemma,n)

k <- 6
num_words <- 12
seed = 4321
lda <- LDA(source_dtm, k = k, method = "GIBBS",control = list(seed = seed))

top_terms_per_topic(lda, num_words)

p1 <-  source_tidy %>%
  filter(lemma %in% c("child","forest","horse","doll","water")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE, col = "red")  +
  ggtitle("Cosette Agua en el Bosque")+
  ylim(0,10)+
  xlim(0,150)

p2 <-  source_tidy %>%
```

```
  filter(lemma %in% c("barricade","insurgent","war","street","shot")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE)  +
  ggtitle("Guerra")+
  ylim(0,10)+
  xlim(0,150)

p3 <-  source_tidy %>%
  filter(lemma %in% c("hand","moment","child","house")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE, col="green")  +
  ggtitle("No definido")+
  ylim(0,10)+
  xlim(0,150)
p4 <-  source_tidy %>%
  filter(lemma %in% c("horse","mayor","wheel","convict","town")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE,col="yellow")  +
  ggtitle("Rueda")+
  ylim(0,10)+
  xlim(0,150)
p5 <-  source_tidy %>%
```

```
  filter(lemma %in% c("priest","silver","bishop","dinner")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE,col="pink")  +
  ggtitle("Cena con el obispo")+
  ylim(0,10)+
  xlim(0,150)
p6 <-  source_tidy %>%
  filter(lemma %in% c("convent","coffin","cemetery","daughter")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE,col="purple")  +
  ggtitle("Ata?d")+
  ylim(0,10)+
  xlim(0,150)


grid.arrange(p1,p2,p3,p4,p5,p6, ncol = 3)

#Verbos

source_tidy <- kleen_annotated %>%
  select(linenumber = doc_id, word= token, lemma, upos) %>%
  filter(upos == "VERB") %>% #choose only the verbs
  inner_join(tidykleen_sections, by = c("word", "linenumber")) %>%
  select(linenumber,chapter,section, word, lemma, upos) %>%
  distinct()
```

```
source_dtm <- source_tidy %>%
  count(section,lemma,sort = T) %>%
  ungroup() %>%
  cast_dtm(section,lemma,n)

k <- 6
num_words <- 12
seed = 4321
lda_verbos <- LDA(source_dtm, k = k, method = "GIBBS",control = list(seed = seed))

top_terms_per_topic(lda_verbos, num_words)

p1 <-  source_tidy %>%
  filter(lemma %in% c("ejaculate","charm","regret","contente")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE, col = "red")  +
  ggtitle("Cari?o")+
  ylim(0,10)+
  xlim(0,150)

p2 <-  source_tidy %>%
  filter(lemma %in% c("play","win","light","study")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE)  +
  ggtitle("Ocio")+
  ylim(0,10)+
```

```
  xlim(0,150)

p3 <-  source_tidy %>%
  filter(lemma %in% c("steal","condemn","summon","deny")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE, col="green")  +
  ggtitle("Robo")+
  ylim(0,10)+
  xlim(0,150)
p4 <-  source_tidy %>%
  filter(lemma %in% c("pass","enter","leave","return","hold","find")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE,col="yellow")  +
  ggtitle("Movimiento")+
  ylim(0,10)+
  xlim(0,150)
p5 <-  source_tidy %>%
  filter(lemma %in% c("kill","plunge","thrust","advance","fire")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE,col="pink")  +
  ggtitle("Asesinato")+
  ylim(0,10)+
```

```
  xlim(0,150)
p6 <-  source_tidy %>%
  filter(lemma %in% c("weep","acquire","renounce","enlighten")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE,col="purple")  +
  ggtitle("Tristeza")+
  ylim(0,10)+
  xlim(0,150)



grid.arrange(p1,p2,p3,p4,p5,p6, ncol = 3)

#Adjetivos

source_tidy <- kleen_annotated %>%
  select(linenumber = doc_id, word= token, lemma, upos) %>%
  filter(upos == "ADJ") %>% #choose only the verbs
  inner_join(tidykleen_sections, by = c("word", "linenumber")) %>%
  select(linenumber,chapter,section, word, lemma, upos) %>%
  distinct()

source_dtm <- source_tidy %>%
  count(section,lemma,sort = T) %>%
  ungroup() %>%
  cast_dtm(section,lemma,n)

k <- 6
num_words <- 12
seed = 4321
lda_adj <- LDA(source_dtm, k = k, method = "GIBBS",control = list(seed = seed))
```

```
top_terms_per_topic(lda_adj, num_words)

p1 <-  source_tidy %>%
  filter(lemma %in% c("gentle","radiant","evil","indifferent","ill")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE, col = "red")  +
  ggtitle("Enfermedad")+
  ylim(0,6)+
  xlim(0,150)

p2 <-  source_tidy %>%
  filter(lemma %in% c("tiny","harsh","secret","wooden","tragic")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE)  +
  ggtitle("Secreto")+
  ylim(0,6)+
  xlim(0,150)

p3 <-  source_tidy %>%
  filter(lemma %in% c("poor","terrible","low","cold","dead")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
```

```
  geom_smooth(se = FALSE, col="green")  +
  ggtitle("Moribundo")+
  ylim(0,6)+
  xlim(0,150)
p4 <-  source_tidy %>%
  filter(lemma %in% c("tragic","civil","fearful","sombre","alive","mournful")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE,col="yellow")  +
  ggtitle("Tr?gico")+
  ylim(0,6)+
  xlim(0,150)
p5 <-  source_tidy %>%
  filter(lemma %in% c("political","holy","strong","humble","generous")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE,col="pink")  +
  ggtitle("Liderazgo")+
  ylim(0,6)+
  xlim(0,150)
p6 <-  source_tidy %>%
  filter(lemma %in% c("handsome","venerable","popular","distinct","revolutionary")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE,col="purple")  +
```

```
  ggtitle("Valent?a")+
  ylim(0,6)+
  xlim(0,150)



grid.arrange(p1,p2,p3,p4,p5,p6, ncol = 3)

#Adjetivos

source_tidy <- kleen_annotated %>%
  select(linenumber = doc_id, word= token, lemma, upos) %>%
  filter(upos == "PROPN") %>% #choose only the verbs
  inner_join(tidykleen_sections, by = c("word", "linenumber")) %>%
  select(linenumber,chapter,section, word, lemma, upos) %>%
  distinct()

source_dtm <- source_tidy %>%
  count(section,lemma,sort = T) %>%
  ungroup() %>%
  cast_dtm(section,lemma,n)

k <- 6
num_words <- 12
seed = 4321
lda_propn <- LDA(source_dtm, k = k, method = "GIBBS",control = list(seed = seed))

top_terms_per_topic(lda_propn, num_words)

p1 <-  source_tidy %>%
  filter(lemma %in% c("gentle","radiant","evil","indifferent","ill")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
```

```
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE, col = "red")  +
  ggtitle("Enfermedad")+
  ylim(0,6)+
  xlim(0,150)

p2 <-  source_tidy %>%
  filter(lemma %in% c("tiny","harsh","secret","wooden","tragic")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE)  +
  ggtitle("Secreto")+
  ylim(0,6)+
  xlim(0,150)

p3 <-  source_tidy %>%
  filter(lemma %in% c("poor","terrible","low","cold","dead")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE, col="green")  +
  ggtitle("Moribundo")+
  ylim(0,6)+
  xlim(0,150)
p4 <-  source_tidy %>%
  filter(lemma %in% c("tragic","civil","fearful","sombre","alive","mournful")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
```

```
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE,col="yellow")  +
  ggtitle("Tr?gico")+
  ylim(0,6)+
  xlim(0,150)
p5 <-  source_tidy %>%
  filter(lemma %in% c("political","holy","strong","humble","generous")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE,col="pink")  +
  ggtitle("Liderazgo")+
  ylim(0,6)+
  xlim(0,150)
p6 <-  source_tidy %>%
  filter(lemma %in% c("handsome","venerable","popular","distinct","revolutionary")) %>%
  group_by(chapter) %>%
  mutate(conteo_6_temas = n()) %>%
  select(chapter, conteo_6_temas) %>%
  distinct() %>%
  ggplot(aes(chapter, conteo_6_temas)) +
  geom_smooth(se = FALSE,col="purple")  +
  ggtitle("Valent?a")+
  ylim(0,6)+
  xlim(0,150)


grid.arrange(p1,p2,p3,p4,p5,p6, ncol = 3)
```

6.2 STOP WORDS

Palabra	Frecuencia	Palabra	Frecuencia	Palabra	Frecuencia
the	79,632	you	9,912	be	5,070
of	38,901	not	8,838	are	4,914
to	28,269	on	8,559	from	4,743
a	28,005	this	8,502	all	4,677
and	23,415	which	8,403	who	4,512
in	21,132	with	8,040	by	4,416
he	18,246	at	7,932	one	4,368
was	16,947	have	6,420	what	4,314
that	16,080	him	6,273	they	4,143
it	13,731	for	6,135	no	4,128
is	13,353	there	5,655	an	4,059
his	11,811	as	5,526	but	3,522
had	11,637	her	5,352	were	3,465
i	10,296	she	5,085	me	3,177

Palabra	Frecuencia	Palabra	Frecuencia	Palabra	Frecuencia
been	3,111	over	1,101	shall	582
my	3,102	once	1,077	even	572
would	2,876	how	1,068	take	570
do	2,874	before	1,065	m	558
said	2,712	again	1,041	went	554
we	2,607	well	1,004	never	546
when	2,592	nothing	990	every	542
their	2,349	after	987	four	538
has	2,280	same	984	because	537
them	2,271	down	940	few	537
more	2,232	still	924	having	528
will	2,190	come	910	took	524
himself	2,187	right	909	something	514
out	2,130	being	897	while	513
then	2,103	each	876	made	507
these	2,085	under	861	going	502
did	2,037	know	858	seemed	464
your	2,025	say	852	away	462
into	1,992	see	852	last	458
up	1,863	such	852	much	458
very	1,809	first	840	whole	454
than	1,770	now	814	may	444
other	1,740	must	810	herself	438
two	1,610	us	810	own	438
or	1,566	go	762	men	436
here	1,536	does	753	whom	434
only	1,521	our	738	might	430
so	1,520	just	718	itself	423
man	1,500	without	698	everything	420
some	1,485	off	693	always	410
if	1,476	little	678	another	400
those	1,434	let	676	myself	399
its	1,416	why	672	taken	390
could	1,392	can	664	came	380
about	1,341	three	642	both	378
should	1,230	most	630	get	376
old	1,202	too	615	almost	368
like	1,180	upon	610	behind	364
any	1,170	way	610	certain	362
where	1,164	against	600	themselves	362
through	1,149	between	594	saw	358
am	1,112	good	589	done	356

Palabra	Frecuencia	Palabra	Frecuencia	Palabra	Frecuencia
though	344	better	206	full	135
name	329	five	205	large	133
anything	326	felt	204	often	132
thought	323	among	200	sometimes	131
think	322	became	200	oh	130
new	318	rather	198	seems	130
many	316	side	198	quite	128
thus	314	order	193	look	123
above	309	near	192	below	122
great	309	put	190	doing	120
since	308	yourself	190	case	119
longer	307	ever	184	keep	118
alone	304	high	180	knew	116
place	304	yet	180	several	116
long	293	end	177	use	114
back	291	fact	177	becomes	108
young	285	however	176	part	108
become	278	necessary	176	following	107
seen	274	along	174	turn	107
second	272	six	172	ask	106
make	270	others	170	sure	106
during	264	together	170	state	102
also	262	further	168	nevertheless	100
later	244	whose	168	saying	100
nor	244	around	166	asked	98
point	244	perhaps	166	find	98
years	244	open	164	got	98
cannot	243	really	162	course	95
possible	242	given	160	small	93
room	242	knows	156	either	92
yes	239	tell	156	forth	92
already	236	work	156	present	92
thing	235	says	154	placed	91
want	234	certainly	150	making	86
far	230	least	150	gave	85
turned	230	give	148	across	84
towards	224	whether	148	moreover	84
face	220	known	146	wish	83
until	219	need	144	year	83
things	218	began	143	ourselves	82
less	212	enough	142	able	81
within	208	opened	138	seven	81

Palabra	Frecuencia	Palabra	Frecuencia	Palabra	Frecuencia
best	80	clearly	46	clear	29
everywhere	80	kept	46	facts	29
neither	80	self	46	instead	29
beside	79	whence	46	sees	29
ought	78	merely	43	whither	28
followed	77	turning	43	afterwards	27
although	76	beyond	42	show	27
next	76	please	42	thoroughly	27
outside	76	points	42	indeed	26
comes	74	goes	41	kind	26
hardly	74	opening	41	therefore	26
seem	70	d	40	wanted	26
soon	70	according	39	wants	26
gone	68	allow	39	asking	24
used	68	former	38	help	24
except	67	looking	38	interest	24
general	67	nine	38	elsewhere	23
third	65	thanks	38	finds	23
eight	64	becoming	37	liked	23
besides	63	meanwhile	37	ones	23
thoughts	63	cause	36	un	23
probably	62	latter	36	greater	22
none	59	sides	36	indicated	22
big	58	somewhere	36	toward	22
different	56	yourselves	36	try	22
nearly	56	entirely	35	useful	22
et	55	getting	35	ways	22
presented	55	places	35	apart	21
aside	54	early	34	anywhere	20
yours	54	needs	34	ex	20
else	52	o	34	faces	20
hence	52	seeing	34	groups	20
mean	52	tried	34	hers	20
serious	52	v	34	inward	20
mr	51	appear	33	per	20
sent	51	ended	33	que	20
re	50	pointed	33	x	20
beings	49	orders	32	exactly	19
gives	48	formerly	31	member	19
number	48	group	31	parts	19
whatever	48	everybody	30	somewhat	19
believe	47	otherwise	30	gets	18

Palabra	Frecuencia	Palabra	Frecuencia	Palabra	Frecuencia
ours	18	greatest	11	plus	7
particularly	18	problems	11	pointing	7
problem	18	puts	11	shows	7
etc	17	showed	11	usually	7
fifth	17	t	11	containing	6
sorry	17	willing	11	g	6
turns	17	b	10	goods	6
working	17	c	10	higher	6
contain	16	cases	10	lets	6
contains	16	consider	10	p	6
ends	16	inner	10	r	6
example	16	looks	10	sub	6
follows	16	parted	10	wherever	6
fully	16	seeming	10	ending	5
rooms	16	thank	10	f	5
twice	16	thence	10	grouped	5
various	16	unless	10	lest	5
needed	15	works	10	numbers	5
opens	15	accordingly	9	secondly	5
seconds	15	indicate	9	smallest	5
truly	15	seriously	9	tries	5
younger	15	smaller	9	wanting	5
actually	14	thereupon	9	backed	4
brief	14	trying	9	corresponding	4
changes	14	welcome	9	definitely	4
hither	14	anybody	8	everyone	4
keeps	14	especially	8	hello	4
nobody	14	highest	8	likely	4
ordered	14	interests	8	presenting	4
particular	14	members	8	thereby	4
therein	14	n	8	unfortunately	4
whoever	14	non	8	uses	4
worked	14	novel	8	wherein	4
generally	13	nowhere	8	y	4
happens	13	older	8	youngest	4
important	13	theirs	8	zero	4
presents	13	throughout	8	allows	3
described	12	u	8	anyway	3
thinks	12	value	8	associated	3
whenever	12	consequently	7	concerning	3
backs	11	immediate	7	downwards	3
causes	11	l	7	hereafter	3

Palabra	Frecuencia	Palabra	Frecuencia	Palabra	Frecuencia
ignored	3	j	2	backing	1
interesting	3	lately	2	co	1
obviously	3	latest	2	differently	1
regards	3	mostly	2	downs	1
sensible	3	q	2	e	1
showing	3	regarding	2	longest	1
anyone	2	s	2	oldest	1
appropriate	2	sup	2	ordering	1
asks	2	tends	2	parting	1
beforehand	2	thereafter	2	selves	1
cant	2	thorough	2	via	1
considering	2	today	2	wonder	1
don't	2	wells	2	z	1
indicates	2	area	1		
interested	2	awfully	1		