5 de marzo de 2021.

Dr. Gerardo Maldonado Paz
Jefe de Investigación y Posgrado
Centro de Investigación en Ciencias Cognitivas
Universidad Autónoma del Estado de Morelos
PRESENTE

Por medio de la presente le comunico que he leído la tesis "**Análisis de fractalidad en textos y extracción automatizada de palabras clave**" que presenta el alumno:

David Torres Moreno

para obtener el grado de Maestro en Ciencias Cognitivas. Considero que dicha tesis está terminada por lo que doy mi **voto aprobatorio** para que se proceda a la defensa de la misma.

Baso mi decisión en lo siguiente:

En esta tesis, el alumno hace un análisis del grado de fractalidad de palabras como medio para la extracción automática de *keywords* en un texto. La tesis hace un muy buen trabajo al resumir la investigación reciente en este tema y su relación con el objetivo de la misma. La definición del concepto de fractalidad de una palabra o token es clara, así como el papel que juega este concepto como posible medio para dar cuenta de la importancia de un término dentro de un documento.

A nivel de resultados, la tesis compara la extracción de *keywords* usando el grado de fractalidad contra otros algoritmos usados para este fin (concretamente TextRank y RAKE). La comparación se hace en varios tipos de documentos, concretamente libros, artículos científicos y tesis. En el análisis de artículos científicos, se consideran adicionalmente las *keywords* dadas por los autores como referencia. En este contexto, se concluye que el grado de fractalidad no parece dar cuenta de la importancia de los términos dentro de un documento. En el análisis de textos más largos, como libros y tesis, se encontraron resultados interesantes respecto a la distribución de palabras con grado de fractalidad alto en secciones importantes de los documentos. En este aspecto, se presenta evidencia de que la presencia de palabras con grado de fractalidad alta, las cuales no se encuentran distribuidas de manera uniforme en el texto, sugiere que dicha sección aporta contenido conceptual sustancial al desarrollo del tema del documento.

En conclusión, en este trabajo se explora la pregunta ¿qué significa que una palabra sea importante en un texto? Mediante el uso de la medida de fractalidad de palabras y el concepto de autosimilitud en la distribución de tokens en el texto, el alumno contribuye con un análisis cognitivo interesante sobre esta pregunta.

Sin más por el momento, quedo de usted

A t e n t a m e n t e

Se adiciona página con la e-firma UAEM

_____

Dr. Gerardo Mauricio Toledo Acosta

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

GERARDO MAURICIO TOLEDO ACOSTA | Fecha:2021-03-05 13:50:07 | Firmante

AkzkObU/gn4gTeNi5DGBpwkzr+DXRtfw8KRJXpygZpnuU5tWqM4MmDtFLQc5+Nx32LpxjBLFK33o5ecxQ61wj5Rj0xGypycr+kn/YT6xmV5zmXe5IYNLM5GaNPsrGTSahAHKsu1
PFUjq0JlyvqH/VlYsGkGyj/C57zYwbbDr+6oHKdWlAlHaLOCPsfx4zjfkihEDhJQuMN9xe3aCY5t7LFfOwaLWtEKKJL57EXClNGYbrJSgOepx/wO3vfb09ZCgbZK1ePi6asgVHpXxhAf
ycPHK76B8Fh6eNkkSxMVMephqwpLUD2fIiE/hUOMkCiPAisfRWJtb1mKoEKZHiEQQ9Q==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

8UTHjI

https://efirma.uaem.mx/noRepudio/wofkUtIWC6B1rqARKbs3cRRsa9q4ohxi

Cuernavaca, Morelos a 04 de marzo de 2021

Dr. Gerardo Maldonado Paz
Jefe de Investigación y Posgrado
Centro de Investigación en Ciencias Cognitivas
Universidad Autónoma del Estado de Morelos
PRESENTE

Por medio de la presente le comunico que he leído la tesis "**Análisis de fractalidad en textos y extracción automatizada de palabras clave**" que presenta el alumno:

David Torres Moreno

para obtener el grado de Maestro/a en Ciencias Cognitivas. Considero que dicha tesis está terminada por lo que doy mi **voto aprobatorio** para que se proceda a la defensa de la misma.

Baso mi decisión en lo siguiente:

Históricamente, el estudio del lenguaje y la cognición lingüística se ha visto beneficiado de herramientas matemáticas y computacionales, con el fin de lograr una mejor comprensión de determinados fenómenos lingüísticos. Un fenómeno particular es el proceso de extracción y recuperación de palabras clave, que ha dado lugar a muchos trabajos de investigación en el campo de la Inteligencia Artificial y las Ciencias Cognitivas. Así, existen muchos métodos para la extracción de términos relevantes, simples o compuestos, mediante técnicas basadas en modelos matemáticos y lingüísticos. A pesar de esta abundante literatura, existe un vacío epistémico en cuanto a una noción más precisa de lo que es una palabra clave lo que pone de manifiesto el trabajo realizado por David Torres Moreno. Llenar este vacío es un problema de extrema importancia si se desea indagar sobre la relevancia de estos modelos para dar cuenta de los fenómenos cognitivos, posiblemente subyacentes a la extracción de palabras clave. Considero que la tesis de David plantea este problema a través de un trabajo experimental realizado con diligencia y profundidad.

Sin más por el momento, quedo de usted.

A t e n t a m e n t e

_____
Dr. Jorge Hermosillo Valadez

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

JORGE HERMOSILLO VALADEZ | Fecha:2021-03-04 19:23:28 | Firmante

SQBscPZNtkocrhe9A9C+N+leiStIMd9he9ffRvqIP3JxPt+0rhmjV9IUJZS+qC4of20iC1jLYsO+6+kNNZHEPKwfdHackZwRj7xvi/Sx/oVV5Q++m6INB651xo4yoqsgJ+aWV0ytOErUJ18h3qDWvcUjIrNl7Ar5ZAw/TvLEZ1felLyh96XLIu5VFCB3m0737oNJJYmombaRKFUDG718HC662oV/JERCP0Q6c/KEwG8U9Jz5L/FvU14xENSnnWXHxAVXtdXQvYaeLVHtYxgduu80CcT1C3/J4Io466f5PN9B+bwmlZ71mQyiZgffCxT1zxW/96qgE2d8Z45HiELCmA==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

UxEuZz

https://efirma.uaem.mx/noRepudio/NrVKmBEcVuT5zuyyvo89SGEDQkqo0tXY

Hoja Membretada

10 de marzo del 2021.

Dr. Gerardo Maldonado Paz
Jefe de Investigación y Posgrado
Centro de Investigación en Ciencias Cognitivas
Universidad Autónoma del Estado de Morelos
PRESENTE

Por medio de la presente le comunico que he leído la tesis "**Análisis de fractalidad en textos y extracción automatizada de palabras clave**" que presenta el alumno:

David Torres Moreno

para obtener el grado de Maestro/a en Ciencias Cognitivas. Considero que dicha tesis está terminada por lo que doy mi **voto aprobatorio** para que se proceda a la defensa de la misma.

Baso mi decisión en lo siguiente:

La tesis escrita por el alumno David Torres Moreno está escrita de manera profesional y posee una estructura que permite una lectura fluida. La exposición plasmada en la tesis permite entender los diferentes conceptos de manera sencilla sin perder de vista el objetivo del trabajo de investigación. La metodología y los resultados permiten apreciar cómo fue abordado el problema planteado y la discusión establece claramente el significado de los hallazgos así como de los retos y perspectivas.

Sin más por el momento, quedo de usted

A t e n t a m e n t e

(e.firma UAEM)

_____
Doctor Dan Sidney Díaz Guerrero

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

DAN SIDNEY DIAZ GUERRERO | Fecha:2021-03-10 13:20:38 | Firmante

UFW0BQ9Y4T8A0/rEe8UMWTj2wpU0+S2AD33E/KqCM8Xj6GrIzmmAC8ytAAIMRcafKd12u/mbrpITCjCpW19MTbeirHENyfd+L76vWHACKxKajFtGmQ7K13sxwX+SWjGonUT3NAKV2SlTab4GcG/aFNkH02LN60rWGFoi7Y9fXQgiwZ+yCOmzWQCnPVn6KNIj0mJoZKw3I2gnU36ZrobLcO3IokR/NBhDqzbZMRaj16RUyDv+D/HozPiD7ZqB1+LkHMId7slXCTlYofP2DAuGsTHchMDuAx9kIaTirATqpxNjWv94L/fBypbyKYk4s0CsqKHa2xavPgU3Q8bnt+jvdw==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

khmHAo

https://efirma.uaem.mx/noRepudio/1X7074DD7pbM2Qq8dgnGfu4rnzE2G44a

22 de marzo de 2021

Dr. Gerardo Maldonado Paz
Jefe de Investigación y Posgrado
Centro de Investigación en Ciencias Cognitivas
Universidad Autónoma del Estado de Morelos
PRESENTE

Por medio de la presente le comunico que he leído la tesis "**Análisis de fractalidad en textos y extracción automatizada de palabras clave"** que presenta el alumno:

David Torres Moreno

para obtener el grado de Maestro/a en Ciencias Cognitivas. Considero que dicha tesis está terminada por lo que doy mi **voto aprobatorio** para que se proceda a la defensa de la misma.

Baso mi decisión en lo siguiente:

El trabajo cumple con los estándares de calidad de una tesis de Maestría en cuanto a su relevancia, forma y fondo.

Sin más por el momento, quedo de usted

A t e n t a m e n t e

(e.firma UAEM)

_____
Dr. Alberto Jorge Falcón Albarrán

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

ALBERTO JORGE FALCON ALBARRAN | Fecha:2021-03-22 21:44:43 | Firmante

DKMwDCHEKSRbKnpTbuj2NekhlD5Nz34j2dVIyvQGRbXX+1aR6zbt/iZl7pjCk11OmvZesDLaF/nRP80/Zqz/A1kc9lG/KjQAT/id3HrmZe7BsZvyeLg7v9aIdau91J42szFZK1FN
XbgnJzrcnMrGeFQdyyQRI9HXBDGwlLcstqLZWgok9FcrxPzrIcLd03mWbSrBswy5jus8mdOb6gIZUlemgVy8YataNR2tSvMyTH9IFjfzY7A4rx3sxDUKruPosXpbpRbBzA1FTE
NOECpNVBNTz1zq2A58SaFUZHBEzJrdt9l5cFK4le4zutar7ZDk9IqXPz33OMM8PGFCPJ1aBw==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

utOJpq

https://efirma.uaem.mx/noRepudio/INLKEMHhVEWyOKblqx6wmSHsHoFeuC9Y

8 de marzo de 2021

Dr. Gerardo Maldonado Paz
Jefe de Investigación y Posgrado
Centro de Investigación en Ciencias Cognitivas
Universidad Autónoma del Estado de Morelos
PRESENTE

Por medio de la presente le comunico que he leído la tesis "**Análisis de fractalidad en textos y extracción automatizada de palabras clave**" que presenta el alumno:

David Torres Moreno

para obtener el grado de Maestro en Ciencias Cognitivas. Considero que dicha tesis está terminada por lo que doy mi **voto aprobatorio** para que se proceda a la defensa de la misma.

Baso mi decisión en lo siguiente:

La tesis es un trabajo de investigación original, bien fundamentado, que cumple satisfactoriamente los requisitos y expectativas para una tesis de maestría.

A t e n t a m e n t e

(e.firma UAEM)

_____
Dra. María Asela Reig Alamillo

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

MARIA ASELA REIG ALAMILLO | Fecha:2021-03-08 20:46:15 | Firmante

qdGfl5XdUfaJumq1PALyczPdcwSvqMgWfmewQn8rM8jPH3S4A4cSKdlbX9TMNMY5LPeQa8O3m7ADspc/E1FRyUd5ot+uImq1aGec0tKeN5uJQ6h9pQm2IODihjhHogwKbdm8q
TGmJIDXKJSwaSz38wxXdFRsiG5NzsaJaWHMOYYNpiT6/mgkLFEQEwvx/kbSv515kDDlPsTdI+1e/lvLAPBc8ojAA05YfwQONTmw+fXdnDFfZt4XmRZ6EDFQrCYVfdtGtKEmcfk8
vIXh134yclULHBF8mxdQGIcY8ZmFA8r0Mr1SKlNDNfU2IAeb7R/PcgBVrN3IMyAY/Q+rw9D5rg==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

KghbV3

https://efirma.uaem.mx/noRepudio/Ltu8FgVG53RUOdr3vs569nGcPOp9DezH

Universidad Autónoma del Estado de Morelos

MAESTRÍA EN CIENCIAS COGNITIVAS

ANÁLISIS DE FRACTALIDAD EN TEXTOS Y EXTRACCIÓN AUTOMATIZADA DE PALABRAS CLAVE

TESIS

QUE PARA OBTENER EL GRADO DE MAESTRO/A
EN CIENCIAS COGNITIVAS

P R E S E N T A:

David Torres Moreno

Director de tesis: Dra. María Asela Reig Alamillo

Dr. Jorge Hermosillo Valadez

Comité Tutoral: Dr. Dan S. Díaz Guerrero

Dr. Alberto Falcón Albarrán

Dr. Gerardo Mauricio Toledo Acosta

Cuernavaca, Morelos Marzo, 2021

ÍNDICE

ÍNDICE DE TABLAS Y FIGURAS

Resumen

Benoît Mandelbrot (1982), en su libro Fractal Geometry of Nature, presentó el concepto fractal, mismo que se observa en los objetos de la naturaleza, tales como: las formas de las cadenas montañosas, la morfología de algunas plantas y animales, la configuración de los pulmones y sistemas nerviosos en los vertebrados y el perímetro de las costas del mar que mantienen su dimensión topológica en diferentes escalas; esta característica es llamada autosimilitud. El autor hace un primer acercamiento a la posibilidad de encontrar patrones fractales en el lenguaje, al introducirlo en los árboles lexicográficos. El concepto de fractal ha sido aplicado en diferentes áreas, tales como la cognición, la psicofisiología, el psicoanálisis, la interacción social, la epistemología, etc. En el lenguaje, recientemente Najafi & Darooneh (2015) proponen el análisis del fractal en textos como método para la extracción de palabras clave (*keywords*). La presente investigación, situada en el Procesamiento del Lenguaje Natural (PLN), cuestiona la propuesta del algoritmo de fractalidad como método automatizado de extracción de palabras clave (keywords), principalmente porque en la literatura no existe una noción o definición especifica de "importancia" en Ciencias Cognitivas. Debido a ello, en este trabajo se realiza un análisis comparativo de los resultados del algoritmo de fractalidad y otros algoritmos computacionales que extraen palabras clave, además de mostrar cómo se distingue el algoritmo de fractalidad de estos, analizar la relación de distribución de las palabras con una noción de importancia y, con esto, aportar elementos para contribuir a dar claridad al problema de definición del concepto en estos tipos de métodos automatizados.

Palabras clave: lenguaje, fractal, irregularidad, recuperación y palabra clave.

Introducción

Históricamente, el paradigma "computacionalista" de la mente jugó un papel central dentro de las ciencias cognitivas durante las décadas de 1960 y 1970, esencialmente gracias a los avances tecnológicos en las ciencias computacionales de esa época (Fedorenko et al., 2019). Hoy en día, dicha teoría ha sido puesta en duda por varios paradigmas rivales, debido en gran medida a los descubrimientos y avances en neurociencia cognitiva (Barber & Kutas, 2007). Si bien los conceptos de computación y algoritmo son centrales en matemáticas, no cabe duda de que el estudio de algunos procesos cognitivos complejos del ser humano, como el lenguaje y la cognición lingüística, han sacado provecho de herramientas matemáticas y computacionales, con el fin de lograr una mayor comprensión de ciertos fenómenos de orden pragmático (Walker & Hicko, 2016; Ursino et al., 2010). Es así que a través de la lingüística cuantitativa se han mostrado leyes universales o regularidades en el lenguaje, por ejemplo: la ley de Zipf, ley de Heap y ley de Menzerath-Altmann. Aunado a esto, la teoría del mínimo esfuerzo sostiene que los seres humanos economizamos recursos de procesamiento cognitivo como sucede en la ley de la brevedad, la cual prueba que usamos más palabras cortas que palabras largas.

Un fenómeno particular es el de los procesos de recuperación y extracción de información relevante, o palabras clave (keywords), que han dado lugar a muchos trabajos de investigación en el campo de la Inteligencia Artificial y en las Ciencias Cognitivas en general (Xuebo et al., 2019; Zhang et al, 2016; Ingwersen, 1996). Un antecedente del uso de las keywords se da desde el inicio del internet y el incremento de la cantidad de información digital en bases de datos y servidores, ya que no existía manera de acceder a la información contenida en estos documentos, por lo que surgió la necesidad de desarrollar una herramienta capaz de satisfacer la necesidad de información a los usuarios. Y es ahí cuando se desarrollan los motores de búsqueda de internet. Estas aplicaciones funcionan como interfaces de la interacción entre usuario e internet con el fin de recuperar información específica. Para esto, el usuario introduce términos que requiere que contengan los documentos obtenidos. Google utiliza esto para buscar un conjunto de documentos que los contengan. Este conjunto obtenido es un listado ordenado, en el cual el principal documento es el que define como más significativo de la búsqueda. Pero ¿cómo se pondera esto? Google calcula un índice para cada sitio web

con el algoritmo computacional llamado PageRank. *Grosso modo* mientras más sitios web tengan vínculos hacia uno, entonces es mejor rankeado y en este orden es como se muestran los resultados que contengan la palabra clave definida por el usuario.

A pesar de la abundante literatura sobre el tema, consideramos que existe un vacío epistémico en general sobre la noción de palabra "clave", o "relevante", ya que no logramos encontrar en nuestra investigación, ningún diálogo interdisciplinario o discusión que esclarezca o aporte una definición de este concepto en Ciencias Cognitivas. Llenar este vacío es un problema de extrema importancia, si se desea indagar sobre la pertinencia de los modelos matemáticos que buscan dar cuenta de los fenómenos cognitivos posiblemente asociados con la extracción de palabras clave, siendo esta la principal motivación de nuestro trabajo.

Con el fin de abrir este diálogo e intentar esclarecer el problema, la presente investigación se enfoca en un estudio a profundidad del trabajo reportado por Najafi & Darooneh (2015), quienes proponen el uso del concepto físico del fractal para la extracción de palabras clave de un texto. El fenómeno de fractalidad se ha reportado en la literatura de la Ciencia Cognitiva en relación con la estructura de los fractales con el funcionamiento del cuerpo humano y con la actividad cerebral, y estudiando el comportamiento del ser humano, la solución de problemas y el lenguaje desde la fractalidad (Andres et al., 2019; Pantonia et al., 2019; Payeron, 2007; Jing et al., 2003). Sin embargo, Najafi & Darooneh (2015) no ofrecen una explicación de lo que es una palabra clave, ni proponen cómo este fenómeno de la fractalidad está vinculado a la cognición lingüística, o a la noción de palabra clave. En otras palabras: ¿cómo saber si el fenómeno reportado por estos autores está relacionado con palabras clave?, o ¿cómo saber qué es exactamente lo que su algoritmo está extrayendo? Ante estas preguntas, nuestra investigación se enfoca en un análisis comparativo de este algoritmo contra otros que también se han reportado en la literatura, que cumplen con el mismo propósito, y que emulan el proceso cognitivo del humano.

De esta forma, la presente investigación se enmarca en la metodología del Procesamiento de Lenguaje Natural, una rama de la Inteligencia Artificial que cada vez más busca construir modelos con una base biológica plausible. Mediante un estudio comparativo entre el grado de fractalidad y otros algoritmos que extraen palabras clave, deseamos ahondar en la noción de palabra clave, y con ello, contribuir a esclarecer el problema de cómo definir este

concepto. Para ello, partimos de la hipótesis de que las propiedades distributivas de las palabras están relacionadas con una noción de importancia. Nuestro objetivo es entender la forma en que el grado de fractalidad se distingue de otras medidas, y cómo podemos relacionarla con una noción de importancia. Con ello, esperamos esclarecer el planteamiento del problema, y alimentar la discusión técnica y científica al respecto.

La organización del documento es la siguiente:

En el apartado 1 se aborda los antecedentes teóricos de esta investigación, desde las normas que debe seguir un texto, así como su estructura para comunicar el significado al lector, aunado a esto, se presenta un preámbulo de la aproximación desde el PLN en la recuperación y extracción de información y por último, el concepto fractal y su aplicación en el lenguaje.

En el apartado 2, se cuestiona la propuesta del algoritmo de fractalidad en la extracción de keywords Najafi & Darooneh (2015) y se plantea la hipótesis en cuestión de la distribución de las palabras y el tipo de importancia que brinda esta medida, así como la comparación con otros métodos automatizados de extracción de palabras clave.

En el apartado 3, se describe el método desde el desarrollo del algoritmo hasta su validación, para probar la hipótesis planteada y aplicarlo a diferentes tipos de texto: artículos científicos, libros y tesis de investigación.

En el apartado 4, se muestran los resultados obtenidos para los diferentes experimentos realizados tales como la comparación de extracción de keywords con métodos automatizados y análisis de la relación distribución e importancia en artículos científicos, distribución de palabras por apartados en libros y tesis de investigación.

Por último, en el apartado 5, se realiza una discusión de los resultados obtenidos y cómo la medida del grado de fractalidad es de utilidad en relación con la distribución de las palabras en el escrito y con un tipo de importancia. Este aspecto del algoritmo centrado en regularidades en textos escritos muestra mayores posibilidades de trabajo futuro en línea con trabajos como entropía y *burstiness* de la palabra que pueden ser útiles para analizar distribución en diferentes tipos de textos.

1. Antecedentes

1.1. El lenguaje y el texto como objeto de estudio

El lenguaje es considerado un proceso cognitivo superior, exclusivo del ser humano, por el cual se transmiten intenciones, ideas y pensamientos a través de la lengua natural, con el fin de establecer la comunicación. La comunicación se vuelve entonces una necesidad social, estableciendo el lenguaje como un sistema universal que caracteriza las relaciones sociales.

El lenguaje es el objeto de estudio de la lingüística. Es a través de la lingüística que se describen las lenguas naturales, que son sistemas formados por unidades de diferentes niveles (fonemas, morfemas, palabras, sintagmas, oraciones) para explicitar las reglas que lo constituyen, es decir su gramática. Además de la gramática, parte del estudio consiste en atender a las unidades supraoracionales en discursos o textos, y explicar qué tipos de estructuras, patrones o relaciones subyacen a esas unidades, y qué procesos cognitivos están involucrados en su procesamiento. De la misma forma, existen aportaciones desde la investigación de la psicolingüística y de la psicología cognitiva que han sido relevantes pues consideran lo que sucede en el procesamiento de la información en los textos orales y escritos (Zaldua, 2006).

El lenguaje escrito está asociado a la transmisión de conocimientos a través del tiempo, Beaugrande & Dressler (1997) definieron que un texto "es un acontecimiento comunicativo que cumple siete normas de textualidad"

- Cohesión (dependencias gramaticales).
- Coherencia (relaciones conceptuales).
- Intencionalidad y modalidad (la actitud de productor del texto).
- Aceptabilidad (la actitud del que recibe el texto).
- Informatividad (capacidad de informar).
- Situacionalidad (expresión de un contexto).
- Intertextualidad (posibilidad de conectar significados intrínsecos en los textos).

La escritura entonces es un proceso de elección de palabras (que viene desde la estructura del pensamiento) y de su distribución, es así que se deduce el valor semántico del texto.

El análisis del texto, por tanto, supone el análisis del material lingüístico, pero también es reflejo de la situación de enunciación y los elementos que pueden quedar implícitos gracias a esa situación compartida. Es así que diferentes disciplinas abordan el análisis de textos con objetivos y enfoques diferentes. Mientras que, desde el análisis del discurso escrito se hace hincapié en los elementos relacionados con el contexto, el análisis computacional de textos resulta especialmente relevante para encontrar patrones internos en el texto que iluminen nuestra comprensión de los procesos de creación del mismo, por ejemplo, ¿cómo es que se distribuye la información en el texto? ¿cómo se distribuyen diferentes patrones oracionales o semánticos a lo largo de la unidad textual?

En las últimas décadas, las tecnologías han permitido y contribuido al crecimiento de información expresada en grandes cantidades de textos, y a la par en su forma digital, lo cual permea en distintas líneas de investigación, como es el procesamiento de estos recursos por medio de mecanismos computacionales poniendo a disposición información de manera global en grandes bases de datos para el acceso general.

1.2. Procesamiento de Lenguaje Natural (PLN)

El concepto de procesamiento de la información está relacionado con diferentes teorías que tratan con el problema de comunicación con lenguaje natural. Las teorías del procesamiento de la información suponen que el procesamiento de la información ocurre en etapas, desde recibir la información hasta generar una respuesta, tal como lo realiza el ser humano, con algunas excepciones, ya que la forma como se presenta en la mente aún no se conoce completamente. Para atender esta parte, existen diferentes modelos que se han propuesto y tomado como base en las investigaciones. La metáfora de la computadora, por ejemplo, es una manera de concebir nuestro funcionamiento cognitivo: recibe información, la almacena y la recupera. La mente humana aplica procesos básicos a las representaciones mentales para codificar experiencia, procesarla y almacenarla (Schunk, 2012).

El Procesamiento de Lenguaje Natural (PLN) involucra las áreas de ciencias de la computación, la lingüística y la inteligencia artificial con el fin de investigar mecanismos de interacción entre la máquina y el humano, mediante el uso del lenguaje natural. Estos mecanismos

permiten organizar, clasificar y procesar la información. Para el PLN, la comprensión y generación automática del lenguaje natural es el principal fin (Khurana et al., 2017). La incorporación en la vida cotidiana del PLN es latente en la actualidad y en un futuro cercano: algunos ejemplos son los motores de búsqueda, traductores y resumidores automáticos, los correctores de estilo y ortografía, reconocedores de voz, entre otros. Estas aplicaciones poco a poco son incorporadas en nuestras rutinas y permiten llevar a cabo y facilitar tareas en un menor tiempo. Una máquina puede analizar una gran cantidad de datos de manera rápida, mientras que los humanos lo hacen de manera precisa y exacta pero más lentamente. La tendencia es que las mismas máquinas puedan procesar dinámicamente e interactuar evitando los posibles errores de reconocimiento, traducción a texto y problemas de escritura. Es así que a través del PLN se trata de desarrollar modelos que permiten comprender los procesos involucrados del lenguaje en el ser humano (Cortez et al., 2009).

Las Aplicaciones del PLN son muchas, algunas de estas son: Comprensión del lenguaje, Recuperación de información, Extracción de la información, Búsqueda de respuestas, Generación de discurso, Traducción automática, Reconocimiento del habla y Síntesis de voz. Es así que existen diferentes niveles de análisis dependiendo de su objetivo de estudio:

- Análisis morfológico o léxico. Consiste en el análisis interno de las palabras que forman oraciones para extraer lemas, rasgos flexivos, unidades léxicas compuestas. Es esencial para la información básica: categoría sintáctica y significado léxico.
- Análisis sintáctico. Consiste en el análisis de la estructura de las oraciones de acuerdo con el modelo gramatical empleado (lógico o estadístico).
- Análisis semántico. Proporciona la interpretación de las oraciones, una vez eliminadas las ambigüedades morfosintácticas.
- Análisis pragmático. Incorpora el análisis del contexto de uso a la interpretación final. Aquí se incluye el tratamiento del lenguaje figurado (metáfora e ironía) como el conocimiento del mundo específico necesario para entender un texto especializado.

En cada uno de los niveles existen retos para el PLN. Uno de los principales es la ambigüedad de las palabras, en el nivel semántico, o la ambigüedad sintáctica en el nivel sintáctico; en el nivel pragmático, el significado de la palabra tomando en cuenta el contexto (Vallez & Pedraza-Jiménez, 2007).

Las técnicas que se usan en el PLN para procesar información de los textos son variadas, algunos ejemplos son: Etiquetador de partes del discurso (por sus siglas en inglés, POS, Part of Speech), análisis sintáctico superficial (Chunking / Shallow parsing), palabras funcionales (Stopwords), lematización (Stemming), frases compuestas o estadísticas (Compound or Statistical Phrases) y desambiguación de la palabra (Word Sense Disambiguation). Estas técnicas pueden ser de utilidad al generar un resumen a partir de un texto, construir bases de datos, clasificar texto de acuerdo a categorías, extracción de palabras clave, etc.

La tesis se enmarca en el proceso de detectar y estructurar información a partir de datos no estructurados, es decir, la extracción de la información. El propósito de los sistemas de PLN aplicados a la extracción de información es identificar datos importantes y así aumentar la precisión y eficiencia de una búsqueda (Khurana et al., 2017).

Para iniciar el procesamiento a través del PLN, se requieren de diferentes técnicas para decodificar los símbolos del texto. Una de los principales consiste en la identificación de las palabras: tokenizar el texto, que, en principio, es una simple instrucción de división por espacios (lo más común) del texto que indicarían una unidad llamada token. Otro de ellos es la clasificación de las palabras o grupos de símbolos de acuerdo al tipo de palabra. Esta es una clasificación morfológica, un patrón primordial y útil para realizar tareas de PLN, estadísticas y análisis sintáctico-semántico.

Existe la posibilidad de que el procesamiento de texto para clasificar automáticamente palabras por tipo genere errores o no supere ciertos retos, como por ejemplo, la aparición en el texto de palabras en otro idioma. Este proceso de clasificación de palabras se ve permeado por el hecho de que existen formas más complejas de palabras, como los n-gramas, que en este caso son conjuntos de palabras contiguas en el texto que lingüísticamente forman una unidad semántica. Por ejemplo, en el caso de las palabras en español la unión de algunas entidades tales como “participación ciudadana” es diferente a “participación” y “ciudadana”. Las entidades obtenidas de acuerdo a estos patrones, resulta fácil de identificar para el humano, pero para las maquinas resulta complicado.

Como parte del procesamiento se realiza un análisis del lenguaje, el cual puede ser formal o probabilístico. La primera se basa en el desarrollo de reglas estructurales que se aplican en el

análisis del lenguaje y la segunda genera características de tipo probabilístico en corpus. Un corpus es un "conjunto de textos de materiales escritos y/o hablados, agrupados bajo un conjunto de criterios mínimos" (Bowker & Pearson, 2003, p. 9).

Para esta tesis nos vamos a enfocar en el modelo probabilístico del lenguaje que define una distribución de probabilidad sobre las palabras, a partir del análisis de un corpus. Es decir, a cada palabra se le asocia una probabilidad y esta se genera a partir de un corpus mediante algoritmos o métodos.

1.3 Recuperación de información y extracción de información

En esta tesis nos enfocaremos en la extracción de información, por medio del procesamiento del lenguaje natural. Para esto desarrollaremos los conceptos de recuperación y extracción para aclarar que denotan cosas distintas, aunque son partes de lo mismo (Tolosa & Bordignon, 2007):

- Recuperación de Información (RI): Determinar cuáles son los documentos de una colección que satisfacen una "necesidad de información" de un usuario.
- Extracción de Información (EI): Localizar las porciones de texto que contengan información relevante para unas necesidades concretas de un usuario y proporcionar dicha información de forma adecuada para su procesamiento (de forma manual o automática).

Vallez & Pedraza-Jiménez (2007) define un sistema de recuperación de información textual que lleva a cabo tareas para responder a las consultas de un usuario:

1. Indexación de la colección de documentos: cada documento es descrito mediante el conjunto de términos que representa su contenido.
2. Cuando un usuario formula una consulta el sistema la analiza, y si es necesario la transforma.
3. El sistema compara la descripción de cada documento con la descripción de la consulta, y presenta al usuario aquellos documentos cuyas descripciones sean iguales o se asemejan a la descripción.

4. Los resultados suelen ser mostrados en función de su relevancia, es decir, ordenados en función del grado de similitud entre las descripciones de los documentos y de la consulta.

Según Frakes & Baeza-Yates (1992) casi todos los sistemas de recuperación de información utilizan operadores booleanos o patrones de texto. Los primeros son empleados en sistemas de búsqueda donde existe una gran colección de documentos, como en el internet a través de Google o en una biblioteca digital; en estos sistemas, cada documento es representado por una lista de palabras claves. De igual manera, en los sistemas booleanos, el usuario puede conectar los elementos de la consulta por medio de conectores lógicos. En cambio, en los sistemas basados en patrones, las búsquedas se basan en cadenas de texto o en expresiones regulares, estos sistemas se emplean dentro de la información de documentos, o en colecciones pequeñas de archivos.

Para el caso de internet y bibliotecas digitales, debido al enorme volumen de datos, que además va en incremento, es muy complicado poder recuperar información relevante. Y se consideran relevantes los documentos que son capaces de satisfacer una necesidad de información del usuario. Para que los buscadores de información sean capaces de recuperar páginas relevantes han de identificar las palabras clave o keywords. Si en el proceso de recuperación no existen estas palabras clave o descriptores, el mismo proceso deberá involucrar un proceso de extracción de información para poder identificar y discriminar los documentos que satisfacen la búsqueda y los que no. Por lo tanto, la recuperación y la extracción de la información están relacionadas, no pudiéndose recuperar páginas relevantes de las que no se haya extraído antes las palabras clave o términos necesarios para valorar el documento.

Un tipo de textos que, por su constitución y la necesidad de localizarlos en amplias bases de datos se prestan al estudio de la recuperación y extracción de información son los artículos científicos.

1.3.1. Palabras clave (keywords) en artículos científicos

Para la publicación de artículos científicos en revistas, se solicitan una serie de normas específicas para su correcta indización en las bases de datos. La indización es un procedimiento aplicado a publicaciones mediante el cual “se seleccionan los conceptos que mejor

representan su contenido para su almacenamiento y recuperación respectivamente" (Gil & Alonso, 2005, p. 63). Esta también puede ser realizada por una persona especializada o de forma automática a través de un programa especial.

Las referencias de búsquedas para la recuperación de información se estructuran en dos tipos de campos, aquellos que ayudan a identificar el documento y aquellos que describen su contenido; el título, resumen, las palabras clave (keywords). El resumen y las palabras clave están construidas en lenguaje natural y son decididos por el autor. Las búsquedas por palabras clave, "…se emplearán para la recuperación de información en el título, resumen o en otros campos que la propia base de datos haya elaborado para la búsqueda de contenido en lenguaje natural" (Muñoz-Martín, 2016, p. 181).

Una palabra clave (keyword) es una "palabra o frase corta significativa que describen el contenido de un trabajo en lenguaje natural, el mismo que se utiliza en la comunicación humana. Son términos libres y variados que dependen de la riqueza del vocabulario de quien los utilice." (Muñoz-Martin, 2016, p 1). A pesar de que las palabras clave son establecidas por el autor, estas son imprescindibles para facilitar, identificar y recuperar información en búsquedas bibliográficas puesto que ayudan a los indexadores y motores de búsqueda a encontrar artículos relevantes por temática. La incorrecta selección de palabras clave en los artículos tendrá como consecuencia la no difusión, no accesibilidad y olvido de artículos relacionados con un tema de interés.

En resumen, para los artículos científicos los autores eligen las palabras clave que, de acuerdo a ellos, describen el contenido del texto, en algunos otros casos utilizan herramientas que extraen estas palabras de acuerdo a una metodología establecida. La importancia de una buena elección de palabras clave es que, en la nueva era tecnológica y en la indización en base de datos permite encontrar los artículos con temas específicos. Uno de los grandes problemas radica en cuando estas palabras clave no concentran la información referente al texto, vuelve este proceso aún más difícil. Aunado a esto, en la búsqueda de bibliografía de un proceso de investigación aumenta el tiempo de la fase de recopilación de la información.

El concepto "palabra clave" se puede utilizar en muchos contextos, donde se le otorga diferentes matices a su significado. Por ejemplo, los términos que permiten describir el contenido

de un sitio web, ya que tener una lista de palabras significativas permite hacerse una idea rápida de su contenido y ayuda a acceder a él.

Para los artículos científicos, las palabras clave, mayormente las escoge el autor y la elección está sujeta a la subjetividad; en otros tipos de textos que son muy frecuentes no sucede lo mismo, por ejemplo, las "páginas web". El uso de motores de búsqueda proporciona una herramienta de filtrado de los temas que se busquen. Esta aplicación es muy utilizada por las diferentes herramientas digitales pues con esto, se otorgan documentos relevantes (o información sobre ellos) a un usuario, según la consulta que se haya realizado (Kageura & Umino, 1998). Esta tarea es una de las más comunes que existen, para llevar a cabo una recuperación de información, es necesaria una consulta o *query*, donde se proporcionan términos (keywords) para buscar la información y es la única pista que da el usuario para llevar a cabo la búsqueda necesaria.

Por otro lado, desde el PLN se han desarrollado diferentes métodos para la extracción de términos automatizada (keywords), que además son de gran utilidad en el desarrollo de herramientas lexicográficas, creación de glosarios, análisis diacrónico de la lengua y para dar soporte al web semántico.

1.3.2. Métodos para extracción de palabras clave (keywords)

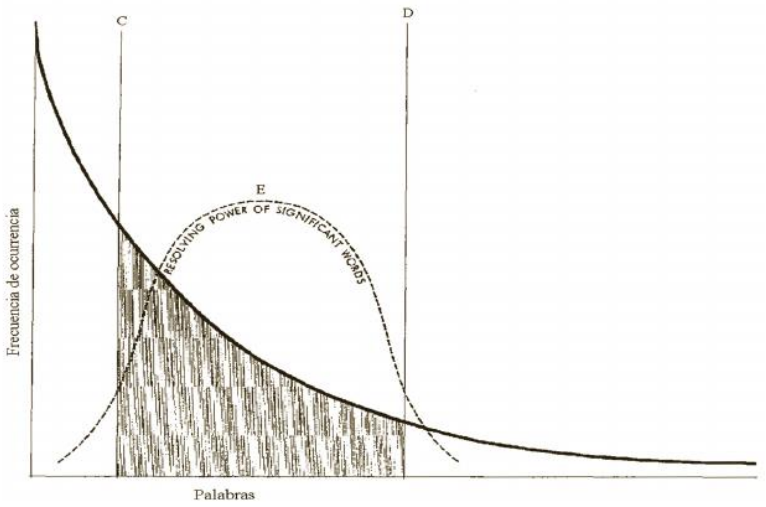
Existen diferentes métodos para la extracción de términos, la clasificación, la agrupación de documentos y la identificación de relaciones. En específico, los métodos para la extracción de términos tienen como objetivo la extracción de términos simples o compuestos por medio de técnicas basadas en modelos matemáticos (probabilísticos), lingüísticos o la combinación de ambos, para identificar palabras clave. Flach (2012) define métricas de rendimiento para evaluación de estos métodos tales como: precisión, sensibilidad (recall) y medida F.

En un primer momento, uno de los principales métodos para obtener estas palabras clave era el uso de la frecuencia de las palabras de un texto, es decir las palabras que se utilizaban más en el escrito. Esto supone un problema, puesto que las palabras más frecuentes son las palabras gramaticales o funcionales, como lo son los artículos o las preposiciones. En PLN son

denominadas “stopwords” ya que estas palabras no tienen significado conceptual y, por tanto, no capturan el contenido del texto. Es así que estos métodos han ido mejorando con el tiempo.

Urbizagástegui & Restrepo (2011) realizaron un experimento exploratorio de identificar las palabras clave de un artículo científico aplicando la ley de Zipf y el punto de transición de Goffman. La ley de Zipf muestra las altas ocurrencias de stopwords, así como las bajas ocurrencias de las palabras menos conocidas (no tan usadas). El punto de transición de Goffman deriva en una ecuación para localizar el rango de las frecuencias donde se encuentran las palabras de mayor contenido semántico (ver figura 1). En este experimento se realizaron cuatro exploraciones con el texto: 1) Todo el texto, 2) Eliminación de stopwords en el texto 3) Lematización del texto y 4) Lematización del texto y eliminación de stopwords. La conclusión para todas las exploraciones fue que los resultados obtenidos en las cuatro exploraciones coinciden en la identificación de las palabras clave.

Figura 1. Zona de ocurrencia de palabras más significantes



Recuperado en: http://www.scielo.org.mx/img/revistas/ib/v25n54/a4f1.jpg

Un método donde se incluyen técnicas de lingüística y estadística, es el C-Value / NC-Value y fue desarrollado por Frantzi (2000). Este método consiste en dos algoritmos C-Value y NC-Value. El algoritmo de C-Value consiste en un filtro lingüístico; reglas para la elección de candidatos a palabras clave y una asignación de rankeo, a través de cuatro factores: Frecuencia de los candidatos en el texto, frecuencia de candidatos con mayor longitud, el número de

los candidatos más grandes y la longitud del candidato en palabras. Con este algoritmo se obtiene una lista de rankeo de los términos candidatos a palabras clave. El siguiente paso es el algoritmo NC-Value, que genera un listado rankeado de las palabras clave tomando en cuenta los términos vecinos (con una unidad de factor de peso) que se relacionan con estas a través del texto, es decir el contexto. Estos métodos han sido probados para el idioma inglés.

Barron-Cedeño et al. (2009) proponen la adaptación para el procesamiento de textos en español, realizando tres modificaciones a los algoritmos: aumentar el número de términos reales como lista de candidatos, considerar palabras con longitud uno y el cambio de procesamiento por lematización de las palabras. Los términos obtenidos son específicos del corpus que se procesa, el problema es que se tiene el factor humano incluido.

Un método para extracción de palabras clave es Rapid Automatic Keyword Extraction (RAKE) el cual recibe como entrada una lista de stopwords, una lista de delimitadores de frase y un conjunto de delimitadores de palabras. RAKE utiliza las stopwords y los delimitadores de frase para dividir el texto en palabras clave candidatas tomando en cuenta el delimitador de palabras. La coocurrencia de las palabras dentro de las palabras candidatas a palabras clave es significativa. Las asociaciones de las palabras son una manera de adaptar automáticamente el estilo y contenido del texto. De la misma forma el factor humano es impredecible, aunque en menor gravedad, debido a que los parámetros de entrada son importantes para obtener buenos resultados al extraer palabras clave.

Por otro lado, existe el método TextRank, el cual se basa en el algoritmo creado por Larry Page and Sergey Brinque llamado PageRank. TextRank fue inspirado en la referencia bibliográfica de los artículos científicos referente a un tema (Dode & Hasani, 2017), y PageRank en los vínculos que contienen los sitios de internet y es utilizado actualmente por Google, estos algoritmos se basan en el desarrollo de grafos. Tomando la estructura léxica y la semántica a partir de un texto, se realiza una analogía, las palabras son tomadas como vértices y las conexiones se dan por la adyacencia de palabras, es decir, en el texto "La música mejora la vida del ser humano", los vértices son: *música*, *vida* y *ser humano*; si se define en el algoritmo como palabras candidatas a los sustantivos y pronombres. Al procesar la palabra *vida*, la adyacencia de esta son *música* y *ser humano,* es decir se crea una arista hacia ese vértice. Esto se procesa con todo el texto construyendo un grafo, lo que computacionalmente se

procesa como una matriz de filas de palabras por columnas de palabras y el índice obtenido es la suma de sus conexiones con otras palabras. Uno de los factores a considerar para Textrank es que se tiene que elegir el tamaño de la ventana, que no es más el número de palabras contiguas, pero de la misma forma se tiene que realizar de manera manual.

Los autores Mihalcea & Tarau (2004) concluyen que TextRank es competitivo con los algoritmos de última generación validado con la evaluación de la técnica de *precision* y *recall*. Al comparar los resultados entre el RAKE y TextRank por precision los resultados son de 33.7% vs 31.2%, respectivamente. Para recall 41.5% y 43.1%, respectivamente. Finalmente, para F-measure de 37.2%y 36.2%, respectivamente (Rose et al., 2010).

Para obtener computacionalmente palabras clave a partir de un corpus, existe un procesamiento estándar previo en los textos: a) se eliminan las stopwords; que son preposiciones o funcionales y b) las palabras se lematizan; es decir se toma la palabra que representa las formas flexionadas, por ejemplo, si la palabra es "comiendo" su lematización de la palabra sería "comer" y c) la intervención del humano o al postular candidatos a palabras clave. Los métodos descritos anteriormente evalúan la relevancia de las palabras en función de criterios que: 1) han sido propuestos y racionalizados por los diseñadores de cada algoritmo, y 2) están dados bajo el supuesto de que la frecuencia y la coocurrencia de las palabras en el texto determinan su relevancia. Recientemente se ha propuesto un método para extracción de keywords que no comparte estas características y trata todo el texto sólo en función de una cierta autosimilitud en las propiedades distributivas de las palabras, asignando un índice de importancia a todas ellas, siendo así una propuesta menos fundada en conceptos racionales (subjetivos), que podría estar dando cuenta de algunas propiedades estructurales del lenguaje (o propiedades distributivas de palabras consideradas importantes).

1.4. Fractalidad y extracción de palabras clave

Una propuesta reciente y novedosa es un método diferente de extracción de keywords utilizando el concepto de fractal. Por medio de técnicas matemáticas se calcula a cada palabra del texto una medida denominada grado de fractalidad, con el cual se ordenan las palabras de mayor importancia (palabras clave) a menor importancia (Najafi & Darooneh, 2015). El enfoque de los autores, es que un texto es un arreglo de palabras consecutivas (espacio unidimensional) que arrastran contenido semántico y está potenciado por los patrones de

distribución de las palabras; siguiendo las reglas gramaticales del lenguaje. Si las palabras del texto son barajadas, es decir, las palabras que lo componen se desorganizan aleatoriamente, el texto pierde significado. Esto es, si tenemos una sucesión de palabras como “México es un país soberano”, esta constituye una oración, pues tiene el patrón correcto de estructura lingüística. Mientras que, si desorganizamos las palabras, por ejemplo “País es México soberano un”, la relación de las palabras carece de sentido porque no se ajusta a la estructura sintáctico-semántica del español, aunque nuestro cerebro busque darle una interpretación.

Es así que el significado refleja un tipo de regularidad en el texto, es decir, el orden de las palabras que está determinado por reglas gramaticales y el orden semántico son importantes para representar significado, y esta regularidad se manifiesta en los patrones de ocurrencia de cada palabra del texto. Por otro lado, los patrones de ocurrencia irregular reflejan la característica del concepto fractal y esta irregularidad es capturada a través de un método para calcular a cada palabra una dimensión fractal. Debido a la problemática de calcular un rango especifico para obtener esta dimensión, se propone una medida derivada llamada grado de fractalidad.

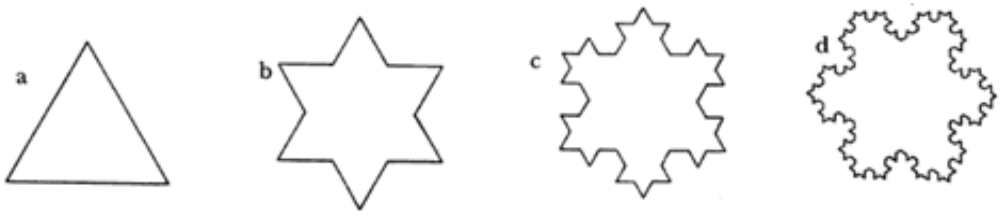
1.4.1 Fractales

Mandelbrot (1982) es el creador de la geometría fractal, donde la palabra fractal es derivada del latín *fractus* que significa fracturado o irregular. En su libro *Fractal Geometry of Nature* muestra que existen varios objetos de la naturaleza que son fractales, tales como, las formas de las cadenas montañosas, la morfología de algunas plantas y animales, la configuración de los pulmones y sistemas nerviosos en los vertebrados, el perímetro de las costas del mar, donde al medir con escalas pequeñas su longitud tiende a infinito.

En este orden de ideas, el concepto de fractal pertenece al área de matemáticas y física. Un fractal es a grandes rasgos un objeto cuya estructura se repite en diferentes escalas, es decir estos elementos tienen una estructura recursiva. Un ejemplo ilustrativo es la siguiente figura, que es la Triángulo de Koch, o también llamado el copo de nieve. La figura base es un triángulo equilátero (parte a), el siguiente paso del proceso iterador es: para cada arista del triángulo se divide en tres iguales tomando la del medio y sustituyéndola por nuestra figura base, dando como resultado (parte b), se continúa realizando la operación anterior en cada arista

que exista, dando como resultado (parte c), se continua nuevamente la operación anterior y da como resultado (parte d).

Figura 2. Triángulo de Koch



Recuperado de: https://upload.wikimedia.org/wikipedia/commons/f/fd/Von_Koch_curve.gif

Es notable que la característica principal de la fractalidad es la autosimilitud, es decir, los patrones que se encuentran repetidamente en cualquier escala y son semejantes a la forma total del conjunto. Otra propiedad importante de la fractalidad son las iteraciones: como en la figura anterior, se debe realizar el proceso un número infinito de veces.

La perspectiva desde la geometría fractal, es analizar el fenómeno no en su forma única y simple, sino en toda su complejidad, y ha sido aplicado en diferentes áreas. Para poder describir el patrón y estructura por el que se genera el fractal, se define una dimensión fractal del objeto.

Es así que se han desarrollado diferentes técnicas para analizar la dimensión fractal de los fenómenos naturales y artificiales. La dimensión Minkowski es una de estas, para determinar la dimensión fractal de un conjunto S en un espacio euclidiano R^n. Para este ejemplo la siguiente imagen muestra el uso de la técnica para calcular la dimensión fractal de la costa de Inglaterra. La imagen es cubierta por una malla conformada por cuadrados (cajas) donde se cuenta el número de cajas que contienen parte de la costa de Inglaterra. Para el primer paso con un número de cajas N con lados de longitud l, se requieren de 22 cajas para contener toda la costa.

Figura 3. Método Minkowski

Recuperado de: https://upload.wikimedia.org/wikipedia/commons/thumb/2/28/Great_Britain_Box.svg/2560px-Great_Britain_Box.svg.png

Si el tamaño de los cuadros es de longitud l_2, donde $l_1 > l_2$, se requieren de 53 cajas y si se continua el proceso con cajas más pequeñas, el número de cajas que se requerirá para contener la costa tenderá a infinito. La fórmula para la dimensión fractal con esta técnica es:

$$\lim_{\varepsilon \to 0} \frac{\log N(\varepsilon)}{\log \frac{1}{\varepsilon}}$$

Donde N(ε) es el número de cajas de longitud ε que cubren el conjunto.

1.4.2 Fractal - Lenguaje

Pestaña (1999) realizó un análisis del uso del concepto de fractalidad en el comportamiento psicosocial a través de los resúmenes (abstracts) de artículos de la base de datos PsycLIT de los años 1988 a 1998. Los resúmenes permitieron identificar las áreas donde se ha utilizado el concepto: a) Psicofisiología; relacionando la estructura de los fractales con el funcionamiento del cuerpo humano y con la actividad cerebral, b) Cognición; estudio del comportamiento del ser humano desde la fractalidad, en el lenguaje y la solución de problemas. c) Psicoanálisis; analogía entre los mecanismos de la vida mental profunda y la estructura de

los fractales, d) Interacción Social; uso de modelo fractal en carácter teórico y experimental, e) Organizaciones; teorías de la complejidad como marco de referencia en la asesoría de grupos corporativos y f) Epistemología.

Mandelbrot (1982) hace un primer acercamiento del uso de concepto de fractal en el lenguaje con una analogía de autosimilitud en árboles lexicográficos regulares en donde cada rama representa una versión reducida del árbol completo.

Posteriormente, Shannon (1993) hace una analogía de patrones similares en el dominio físico de los objetos, así como en el lenguaje a través de sus niveles de resolución, la pragmática, la semántica, lo sintáctico y morfológico, propiciando la investigación cognitiva-lingüística. Esto abrió la posibilidad de ver al lenguaje desde un enfoque a través del concepto fractal.

Pareyon (2007, p. 376) menciona que "Al buscar una afinidad equivalente entre los fractales y los lenguajes no verbales, podemos observar que su nivel de complejidad estructural coincide con su desarrollo potencial como objetos fractales. Esto significa que, en una construcción a largo plazo, aquí también surgen cualidades fractales". Estas cualidades se han encontrado en sistemas de comunicación animal, gestos en la comunicación de los simios y los patrones de sonido en ballenas y delfines, así como el empleo de hápticos, hastics, oculesics o paroxemics utilizados por diferentes comunidades vivas (Pareyon, 2007).

Es así que surge una propuesta de un marco general de investigación, de relación entre el lenguaje y los fractales naturales a través de la Teoría fractal. La lingüística tiene sus diferentes niveles de análisis: discurso, sintaxis, semántica, léxico, morfología y fonología y cada uno de estos niveles le corresponden diferentes tipos de unidades que lo conforman: textos o discursos, oraciones y sintagmas; unidades de significado, palabras, morfemas y fonemas, respectivamente.

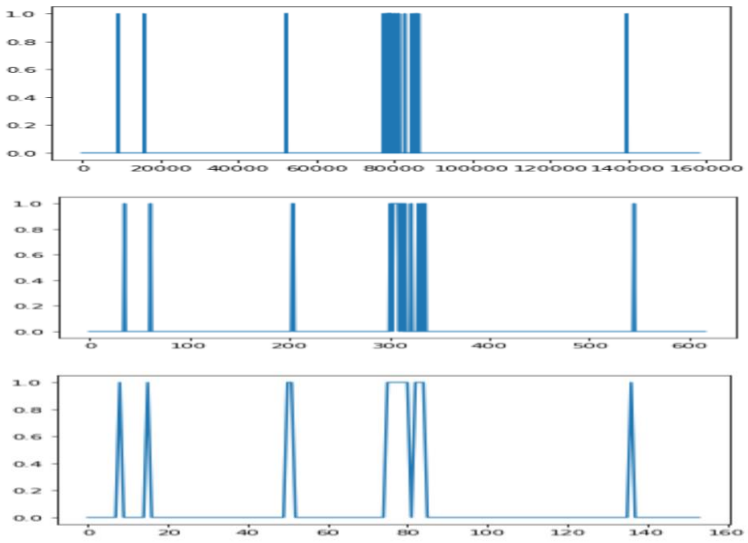
Otro acercamiento en este campo de la investigación han sido Andres et al. (2019), donde muestran resultados preliminares de un análisis fractal en el lenguaje de señas, basados en las conjeturas débil y fuerte de Hřebíček sobre la estructura fractal de lenguajes. La versión débil dice que la ley Menzerath-Altmann se aplica a todos los niveles de lenguaje hasta construcciones semánticas y la versión fuerte se basa en que existe un isomorfismo entre la forma logarítmica de la ley Menserath-Altmann y la fórmula de Moran-Hutchison para calcular la dimensión fractal de objetos con autosimilitud.

En esta misma línea, Najafi & Darooneh (2015) proponen que el fractal se encuentra también en el lenguaje, específicamente en el escrito. Aseveran que un texto puede entenderse mediante regularidades en la distribución espacial de las palabras y sus frecuencias. Su propuesta considera el lenguaje escrito como una secuencia de palabras ordenadas en un espacio unidimensional y que las posiciones de las palabras forman un patrón fractal con una dimensión específica llamada dimensión fractal.

Una vez conocido el término de fractal y conociendo sus dos características principales: irregularidad y autosimilitud, surge la pregunta ¿Cómo es que se representa en el lenguaje escrito? La autosimilitud hace énfasis en que un objeto se verá reflejado en diferentes escalas de un subconjunto del mismo. Esto es, si tomamos una parte específica del objeto y se hace un acercamiento se reflejará el objeto completo o gran parte de este. De acuerdo a esta característica de *autosimilitud*, se establece una dimensión fractal que representa su irregularidad.

Para ver esto, vamos a graficar la distribución de las palabras a través del texto como se muestra en la siguiente figura, donde se logra ver la autosimilitud en diferentes escalas, a saber 1, 256 y 1024:

Figura 4. Distribución de *hybrid* en diferentes escalas

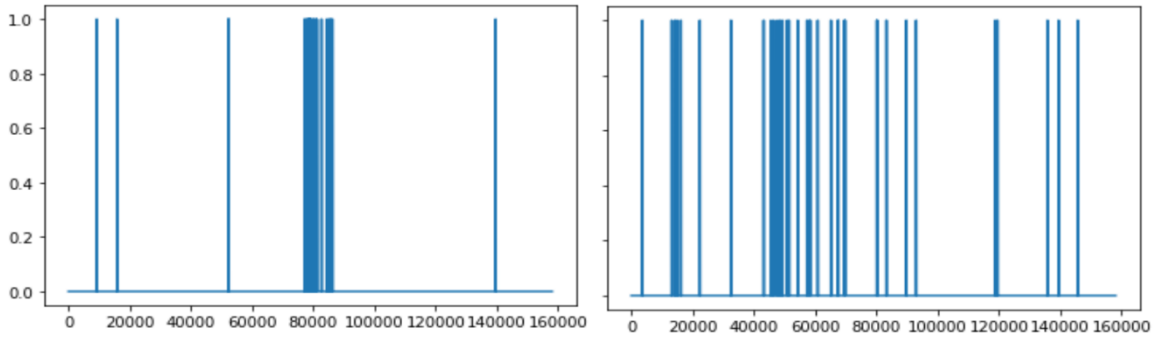


Fuente: Elaboración propia

La escala, cuando es igual a 1, significa que las posiciones son equivalentes a una caja con una palabra, para 256 y 1024 se tiene una de 256 y 1024 palabras, respectivamente. Una

comparación de palabras distribuidas uniformemente y no se muestra a continuación. Las palabras tienen la misma frecuencia (36) pero tienen diferente distribución espacial.

Figura 5. Comparación de la distribución de *hybrid* y *rarely*



Fuente: Elaboración propia

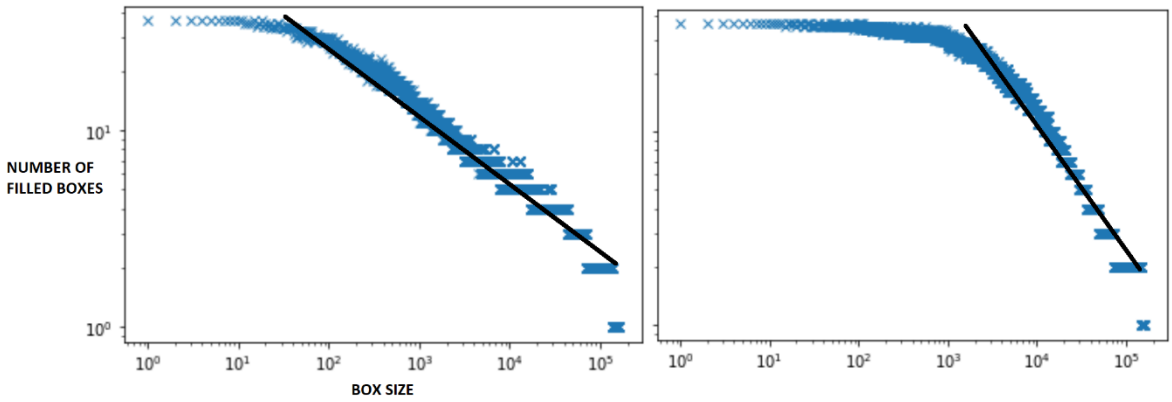
Entonces, el texto tiene N palabras y cada ocurrencia de una palabra tiene una posición en el mismo. Es así que la irregularidad de las palabras en un texto escrito se verá reflejada por las posiciones que ocupen. Para calcular esta irregularidad, los autores utilizan le método de *box counting* que es una relación de cajas llenas vs tamaños de caja. Pero ¿qué significa? Este método captura de acuerdo a diferentes escalas (tamaños de caja) cuántas cajas (cajas llenas) contienen a la palabra través del texto. Es así que se tiene una curva donde: si las cajas tienen un tamaño pequeño, el número de cajas llenas se aproxima al número de ocurrencias de la palabra. Mientras que, si la caja tiene tamaño cercano a N, entonces el número de cajas llenas tiende a 1. La dimensión fractal forma una ley de potencias entre el número de cajas llenas y el tamaño de las cajas.

El número de cajas se denota por $\frac{N}{s}$ donde N es la longitud del texto. Entonces N_b es el número de cajas llenas con la siguiente relación con la ley de potencias:

$$N_b(s) \ \alpha \ s^{-D}$$

Donde D es la dimensión fractal de la palabra, la cual es obtenida por la pendiente de la gráfica en escala logarítmica de N_b y s. Para poder calcular la pendiente se debe escoger un rango apropiado. Como se muestra en la siguiente figura la línea punteada denota la pendiente de la curva; para el caso de las palabras *hybrid* y *rarely* se utilizan diferentes rangos para el cálculo de la pendiente, por lo que es muy importante e influye en el valor de la dimensión.

Figura 6. Resultados de *boxcounting*



Fuente: Elaboración propia

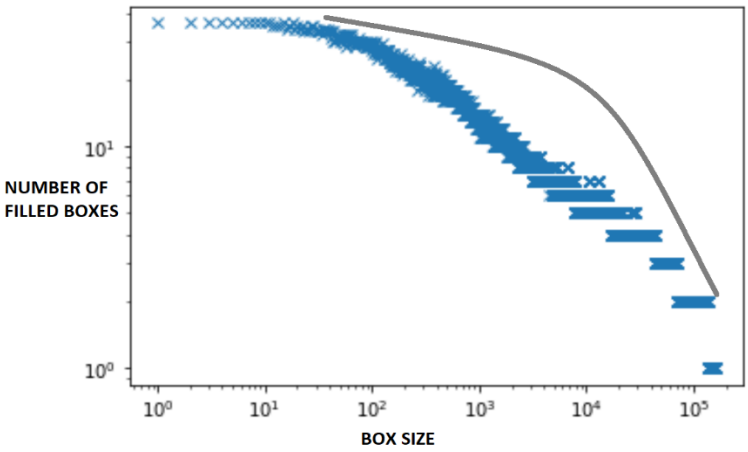
La irregularidad entonces se vería en la curva de la gráfica, mientras el valor de la pendiente esté más cerca de 1 existe una menor irregularidad. En caso contrario, mientras más se aleje de 1 existe una mayor irregularidad. Es decir, las irregularidades en el texto se expresan por medio de la posición de las palabras, misma que no es al azar y no es distribuida uniformemente en el texto.

Debido a la dificultad de obtener la dimensión fractal, los autores proponen una fórmula para calcular un índice a cada palabra llamado grado de fractalidad, que, según su propuesta, clasifica las palabras según su importancia en el texto. Para esto, comparan los patrones de distribución de las palabras de un texto original y el mismo pero barajado. Si las ocurrencias de una palabra son distribuidas uniformemente en el texto, todas las cajas tienen la misma probabilidad de contener la palabra procesada. Sin embargo, el número máximo de cajas llenas que se puede obtener es igual a la frecuencia de la palabra. En caso de que una caja contenga más ocurrencias de palabra, se contabiliza como una caja llena única, es decir, el número de cajas llenas será menor a su frecuencia. Al incrementar el tamaño de la caja el número de cajas llenas decrementa, por lo que se da la siguiente conjetura para un texto barajado siendo M la frecuencia de la palabra w:

$$N_b^{sh}(s,\omega) = \frac{M}{1 + (\frac{M-1}{N-1})(s-1)}$$

.

Figura 7. Resultados de *hybrid* con *boxcounting* de la conjetura



Fuente: Elaboración propia

En el texto barajado, todas las palabras son distribuidas uniformemente y es menos probable la concentración de las palabras. Con la gráfica anterior, se muestra que la conjetura de los autores se aproxima a los resultados del texto barajado. Lo cual genera un cálculo normalizado sin generar un texto barajado para cada procesamiento.

Este índice mide la diferencia entre el patrón fractal de una palabra en el texto original en relación con una versión barajada. Un valor grande del grado de fractalidad significa que la distribución del patrón de una palabra tiene más diferencias con una distribución uniforme. El cual se define como sigue, donde w es una palabra:

$$\delta(\omega) = \sum_s \log\left(\frac{N_b^{sh}(s,\omega)}{N_b(s,\omega)}\right)$$

Los autores introducen una ponderación del índice, al multiplicarlo por el *log(M)*, siendo *M* la frecuencia de la palabra, esto con el fin de tener una medida que incorpore la frecuencia de las palabras. Esta modificación genera mayormente cambios en el rankeo de las palabras que se encuentran en medio de la lista.

Una aseveración que realizan los autores es que, si una palabra se encuentra uniformemente distribuida en el texto, su grado de fractalidad es menor y esto quiere decir que son palabras no importantes. Las palabras más concentradas en un apartado en el texto entonces, tienen

un grado de fractalidad mayor, es decir, son palabras importantes (keywords). Esto es capturado por el grado de fractalidad, y de mejor forma por la medida combinada.

A pesar de que no tener una mejor manera de evaluación del método, los mismos autores evalúan el algoritmo contra otros resultados de métodos computacionales, a través de las técnicas mencionadas anteriormente: *precision* y *recall*. Los resultados del rankeo obtenido se comparan contra el glosario del texto en cuestión. El análisis muestra que el método del grado de fractalidad tiene el mayor valor en el *recall* que los otros dos métodos y en *precision* el valor obtenido es mayor que el de C-Value y menor que el de Entropía. Esto indica que el algoritmo de fractalidad obtiene más coincidencias de las palabras extraídas que los otros métodos.

Como parte del proceso de comparación, es relevante mencionar que en el glosario no solo tiene palabras sino también frases. Para tratar con las keywords multipalabra del glosario se separan en palabras. Por ejemplo, si en el glosario existe la palabra *ganoid fishes* se separan para tener *ganoid* y *fishes*. Una mención importante es que en el rankeo si dos palabras o más tienen el mismo grado de fractalidad o la medida combinada se rankean con el mismo lugar, y la siguiente palabra continua hasta la siguiente posición.

La conclusión de los autores es: "En comparación con otros dos métodos representativos en esta área, la Entropía y C-Value, nuestro enfoque es más eficaz como método para la extracción automática de palabras clave" (Najafi & Darooneh, 2015, p. 16).

Como hemos afirmado, Najafi & Darooneh (2015) no ofrecen una explicación de lo que es una palabra clave, ni proponen cómo este fenómeno de la fractalidad está vinculado a la cognición lingüística, o a la noción de palabra clave. Lo anterior es aún más problemático, en tanto que existe un vacío epistémico en general sobre la noción de palabra "clave", o "relevante" en Ciencias Cognitivas. Llenar este vacío es un problema de extrema importancia, si se desea indagar sobre la pertinencia de los modelos matemáticos que buscan dar cuenta de los fenómenos cognitivos posiblemente asociados con la extracción de palabras clave. Si bien nuestro trabajo no pretende llenar este vacío, si busca aportar para esclarecer el problema. Como se ha visto, Najafi & Darooneh (2015) atribuyen un valor de importancia a aquellas palabras que están distribuidas de una manera no uniforme; es decir: la distribución irregular de una palabra en una parte del texto hace que su grado de fractalidad aumente, y

con ello su importancia o relevancia, a decir de los autores. En la presente investigación indagamos sobre esta propiedad de distribución irregular, o concentración de las palabras en ciertas partes del texto, con el fin de esclarecer esta noción de importancia: ¿En qué sentido es importante una palabra con distribución irregular (grado de fractalidad alto)? ¿Qué información nos da esta medida de grado de fractalidad, y cómo se compara lo extraído por este algoritmo frente a lo que otros algoritmos, que se han reportado como mecanismos de recuperación de palabras clave, extraen?

2. Planteamiento del problema

La idea de que la importancia de las palabras en el significado global de un texto se puede capturar por medio del grado de fractalidad es sugerente pero no está claramente sustentada. Esta hipótesis parece asumir una noción estándar del concepto de importancia o de lo que es una palabra importante en un texto. De acuerdo a nuestra investigación, no se encuentra en la bibliografía una discusión ni una definición aceptada del concepto de importancia de una palabra para entender el contenido o el significado global de un texto. Es así que no existe un criterio claro y único para medir la importancia de una palabra en el texto y parece que este proceso de identificar palabras clave tiene una carga subjetiva de quién las elige. Esto nos lleva a preguntarnos qué tipo de "relevancia" o de "importancia" (o en relación con qué) conllevan las palabras que extrae el algoritmo propuesto por Najafi & Darooneh y cómo se comparan estas palabras con las que extraen otros algoritmos cuyo propósito es la extracción de keywords. En este sentido y con la definición establecida de keyword, surgen las siguientes preguntas de investigación:

Preguntas de investigación

- ¿En qué sentido es importante una palabra con distribución irregular señalado por el grado de fractalidad alto?
- ¿Cómo se compara lo extraído por este algoritmo frente a lo que otros algoritmos, reportados como mecanismos de recuperación de palabras clave, extraen?
- ¿En qué forma el algoritmo del grado de fractalidad es útil para la extracción de keywords?
- ¿Cómo acercarnos a una noción de palabra clave, mediante la caracterización de la distribución de las palabras obtenidas por el algoritmo de grado de fractalidad, en comparación con otros algoritmos?

Objetivos

- Entender la forma en que el grado de fractalidad se distingue de otras medidas, y cómo podemos relacionarla con una noción de importancia.

- Comparar resultados del algoritmo del grado de fractalidad para extraer palabras clave (keywords) de textos con los resultados de otros algoritmos de extracción de keywords (TextRank y RAKE).

Hipótesis:

Las propiedades distributivas de las palabras están relacionadas con una noción de importancia. La extracción de keywords por medio del algoritmo de fractalidad conlleva un diferente tipo de importancia en comparación con otros algoritmos computacionales que emulan el proceso cognitivo del humano.

3. Método

Para poder aplicar y validar el algoritmo, se desarrolló con la información del artículo base y en lenguaje de programación Python, mismo que se eligió por su practicidad y librerías disponibles para PLN. Una vez desarrollado, se procedió a validar los resultados del libro "On The Origin of Species by Means of Natural Selection" de Charles Darwin que es el que en el artículo utilizan como insumo.

3.1. Desarrollo del algoritmo

Para el desarrollo del algoritmo se eligió codificarlo en Python, de acuerdo a su ventaja, practicidad, documentación y disponibilidad de librerías útiles para el PLN. Se desarrollaron métodos específicos para cargar archivos de texto plano, para el preprocesamiento del texto, para ejecución del algoritmo de grado de fractalidad y finalmente para guardar los resultados en archivos csv y txt.

La codificación del método para obtener el grado de fractalidad fue con base al artículo y la conjetura propuesta. Una vez desarrollado el algoritmo fue necesario definir criterios para medir su rendimiento y comportamiento. Estos criterios se centran principalmente en su simplicidad y en el uso eficiente de los recursos. El uso eficiente de los recursos suele medirse en función de dos parámetros: la memoria y el tiempo de ejecución.

Para realizar el cálculo del grado de fractalidad de cada palabra, se aplica el método de *Box Counting*, el cual se calculó con tamaños de caja de 1 a N (siendo N el número de palabras del texto). Es importante notar que en el artículo solo se ocupan potencias de 2, reduciendo el costo de procesamiento considerablemente. Al utilizar *Box Counting* se incrementa la complejidad del algoritmo, convirtiéndola en cuadrática (n^2), es decir, si n se duplica, el tiempo de ejecución del algoritmo aumenta cuatro veces, es decir, de acuerdo al tamaño del texto será el tiempo de procesamiento.

Una alternativa para reducir los tiempos es que, del vocabulario del texto, no se calcule el grado de fractalidad para las palabras que tienen frecuencia uno en el texto porque la fórmula devuelve un valor igual a cero. Otra alternativa es no calcular el grado de fractalidad a las stopwords, debido a que son palabras sin contenido semántico.

En el procesamiento de los datos se identificó una mejora al algoritmo, la cual identifica las posiciones de las palabras y las almacena para su posterior operación y obtener las cajas llenas para cada uno de los tamaños de caja.

El proceso codificado es el siguiente:

- Preprocesamiento: eliminación de apartados del índice, contenidos, referencias, etc. Eliminación de caracteres no alfabéticos, es decir, números, signos de puntuación, etc.
- Procedimiento del método *Box counting* para cada palabra. Se utilizan cajas de tamaño en potencias de 2, es decir, valores de 2, 4, 8, etc donde s< N.
- Cálculo de la medida combinada, al obtener el grado de fractalidad se multiplica por log(*M*), donde *M* es la frecuencia.
- Rankear la lista con el índice obtenido.

3.2. Preparación de documento

El insumo que los autores procesaron es el libro *On The Origin of Species by Means of Natural Selection* de Charles Darwin. En el artículo se especifica un preprocesamiento al documento, el cual es mantener el cuerpo de texto sin índice ni resúmenes de contenidos, además de la eliminación de caracteres no alfabéticos. El número de tokens finalmente obtenidos fue de 191,740 y un vocabulario (tokens no repetidos) de 8,842. Como mencionan en el artículo existe una versión del libro de Darwin en https://www.gutenberg.org/, donde se obtuvo una copia con las siguientes observaciones: el número total de palabras es de 149,730 y el vocabulario es de 7,395 y no coincide con lo reportado por los autores, a pesar de realizar el preprocesamiento establecido. A pesar de estas diferencias se procesó el documento con el algoritmo desarrollado. El tiempo de ejecución del algoritmo fue de un poco más de 2 horas, con una maquina con un Intel Xeon. Este procesamiento se realizó para tamaños de cajas de 1 a N de acuerdo a su distribución de cada palabra.

Al obtener los resultados del procesamiento se identificaron dos cosas:

1. Los valores del grado de fractalidad publicados contra los obtenidos no coinciden.
2. Los valores publicados solo son positivos, mientras que en los obtenidos se tienen números grandes, negativos y positivos.

3.3. Validación del algoritmo

Derivado de la diferencia en los resultados obtenidos y los publicados, se contactó a los autores para identificar y resolver estas diferencias. Los autores sugirieron realizar tres cambios al algoritmo desarrollado, a saber:

1. Realizar un barajeo al texto antes de la ejecución del algoritmo.
2. Invertir el numerador y denominador de la fórmula publicada.
3. Calcular el valor absoluto al resultado.

La aplicación de estos tres cambios en el algoritmo, así como las combinaciones de estos, permitió aproximar los resultados. Se desechó el primer cambio del barajeo del texto antes de la ejecución del algoritmo, el segundo no afectó en los resultados por ser el recíproco del logaritmo y solo se agregó el cálculo del valor absoluto. Para tratar de equiparar los valores solo se calculó el *Box Counting* con valores en potencias de 2 y no para todos los tamaños de caja.

Los intentos de verificación, en comparación con los resultados obtenidos y los del artículo, se aproximan, pero no son iguales, esto es, debido a que no se cuenta con el mismo insumo del libro y el número total de palabras ni el vocabulario no son los mismos. Los resultados obtenidos para esta nueva versión del algoritmo son los siguientes:

Tabla 1. Resultados del algoritmo de fractalidad al procesar libro de Charles Darwin

	Top 20 resultados propios			**Top 20 resultados del artículo**		
Word	**Frequency**	**Fractality**	**Rank**	**Frequency**	**Fractality**	**Rank**
slaves	33	20.96404145	1	34	17.42	1
wax	39	17.92299221	2	42	15.65	8
floated	18	17.51152158	3	18	15.98	6
dried	8	17.20822497	4	9	15.11	12
masters	16	17.19388397	5	17	15.52	10
deg	11	16.60081543	6			
hammock	5	16.2089712	7			
cell	24	15.88130904	8			
pupae	13	15.68054994	9	13	15.72	7
sanguinea	12	15.59556888	10			
spheres	20	14.93197892	11			
electric	6	14.70140559	12			
forelimbs	4	14.67867923	13			

ripe	5	14.59953329	14			
twigs	5	14.37638974	15			
shoulderstripe	7	14.24706308	16			
pond	8	13.98893256	17			
neuters	11	13.86041927	18	12	14.93	16
cells	47	13.73188542	19			
dun	7	13.72826929	20	8	14.60	19

Fuente: Elaboración propia con datos del artículo de Najafi & Darooneh (2015)

De lado izquierdo se muestran los resultados obtenidos de las 20 principales palabras rankeadas y de lado de derecho se muestran los resultados publicados. Las palabras marcadas son las que coinciden con lo publicado, a pesar de que los valores son distintos. Comparando los 20 valores que muestran en el artículo contra los resultados obtenidos se identifica que hay 8 palabras que coinciden en listado de las primeras 20, 11 no se encuentran en el documento procesado y una es la palabra "cuckoo" la cual tiene diferente frecuencia (13).

De acuerdo al tema del libro, las palabras: *slaves, illegitimate, saliva* y *pedicellariae* son palabras importantes y por ende tienen un alto valor de fractalidad. Las palabras *the, of, and* y *in* son palabras irrelevantes sin contenido importante y tienen un grado de fractalidad bajo.

Se realizó el mismo ejercicio de comparación con la medida combinada, que es la multiplicación del grado de fractalidad obtenido por el logaritmo de la frecuencia de cada palabra. Hay 12 palabras de las primeras 20 rankeadas, que coinciden con lo que se publicó en el artículo, aunque los valores de este indicador no coinciden. Cinco palabras no se encuentran el texto procesado, la palabra *nest* con frecuencia de 40 se encuentra en la posición 23, la palabra *clover* con una frecuencia de 8 se encuentra en la posición 137 y *cuckoo* tiene una frecuencia de 13 y se encuentra en la posición 44. Este rankeo de las últimas palabras se puede ver afectado por su frecuencia al combinarla con el grado de fractalidad.

Tabla 2. Comparación de resultados con medida combinada

	Top 20 resultados propios			**Top 20 resultados del artículo**		
Word	**Frequency**	**Combined measured**	**Rank**	**Frequency**	**Combined measured**	**Rank**
slaves	33	31.83418918	1	34	26.68	1
wax	39	28.51663856	2	42	25.40	2
cells	47	22.96105619	3	58	17.36	17
floated	18	21.98173156	4	18	20.07	8
cell	24	21.91956127	5	28	18.39	13
masters	16	20.70349926	6	17	19.10	10

spheres	20	19.42695246	7	19	17.22	19
pupae	13	17.46724437	8	13	17.51	16
deg	11	17.28796776	9			
hybrids	120	17.10134905	10	135	23.20	3
workers	27	17.01661662	11			
sanguinea	12	16.83044545	12			
sterility	78	16.33259243	13	100	22.53	5
instincts	72	16.32610858	14	87	23.00	4
bees	56	16.30401601	15			
dried	8	15.54057566	16			
pollen	73	15.3632113	17			
f	44	15.30270742	18	46	17.66	15
formations	88	14.98677659	19			
neuters	11	14.43413924	20			

Fuente: Elaboración propia con datos del artículo de Najafi & Darooneh (2015)

Puesto que la validación de la información no es al 100%, se procesó además el libro *The First Three Minutes* de Steven Weinberg, el cual de la misma forma arrojó valores cercanos de grado de fractalidad comparado con lo publicado en el artículo "Long range dependence in texts: A method for quantifying coherence of text" de los mismos autores.

Una vez concluido el algoritmo, se realizaron los siguientes procesamientos de diferentes tipos de textos:

1. Artículos científicos; ya que con estos existe una relación de palabras clave definidas por el autor del artículo para comparar con lo obtenido del algoritmo de fractalidad.
2. Se analiza la distribución de las palabras clave definidas por el autor del escrito y las obtenidas por el algoritmo de grado de fractalidad.
3. Se comparan los resultados del algoritmo RAKE y TextRank al procesar los mismos artículos. La intención es identificar qué tan bueno es el algoritmo de fractalidad contra estos métodos.
4. Se procesa el libro de Charles Darwin con el algoritmo de TextRank, esto con el fin de identificar cuáles palabras son las que identifica como palabras clave y compararlas con los resultados del algoritmo del grado de fractalidad.
5. Análisis de distribución de las palabras mejor rankeadas a través de los capítulos de los textos.

4. Resultados

A pesar de la que la validación de los resultados entre lo desarrollado y lo publicado no es idéntico, al realizar la gráfica de distribución y la gráfica de cajas llenas vs tamaños de caja, estas representan lo mismo, por lo que se adjudica esta diferencia a los insumos del procesamiento. Es así que se realizó un nuevo análisis al procesar diferentes tipos de textos: artículos científicos, libros y tesis de investigación. El proceso que se realizó se explicó en el apartado de método.

4.1. Extracción de keywords en artículos científicos

El primer procesamiento de textos fue con artículos científicos. Se eligieron más de 30 artículos escritos en español de los cuales se encontró el mismo patrón. De manera representativa se muestran los siguientes:

1. "La importancia de no menospreciar las palabras clave" de Manuel Molina Arias, con las siguientes keywords: Descriptores, Lenguaje controlado, Lenguaje natural, Palabras clave, Tesauro.
2. "La problemática fractal: un punto de vista cognitivo con interés didáctico" de Sabrina Garbin, con las siguientes keywords: fractal, procesos infinitos, visualización, percepción.

De acuerdo con Najafi & Darooneh: el grado de fractalidad coloca las palabras más importantes en las primeras posiciones en el rankeo por la medida combinada. Aunado a esto, estas palabras importantes tienen una distribución no uniforme a través del texto. Los resultados obtenidos al procesar el primer artículo son los siguientes:

Tabla 3. Resultados del procesamiento artículo 1

	Text size	1476		
	Vocabulary size	370		
	Processing time (s)	0.05186129		
Rank	**Word**	**Frequency**	**Degree_of_fractality**	**Combined_measured**
1	campos	4	4.28277448	2.57848717
2	términos	21	1.68829371	2.23229452
3	salud	3	4.50200748	2.14800346

4	lenguaje natural	5	2.68482629	1.87661304
5	mesh	10	1.86452019	1.86452019
6	búsqueda	11	1.70821599	1.77892363
7	descriptores	15	1.38517776	1.62909545
8	elementos	3	3.403395191	1.623832184
9	título	5	2.174000661	1.519561252
10	resumen	5	2.136876393	1.493612502

Fuente: Elaboración propia

La comparación de esta tabla con las palabras clave definidas por el autor del artículo, muestra que solo hay dos coincidencias en el top ten obtenido por el algoritmo de grado de fractalidad.

Tabla 4. Resultados del procesamiento artículo 2

	Text size	8077		
	Vocabulary size	1310		
	Processing time(s)	0.30504227		
Rank	**Word**	**Frequency**	**Degree_of_fractality**	**Combined_measured**
1	longitud	24	4.90885483	6.77525662
2	instrumento	7	6.88977064	5.82253166
3	f	11	5.5471488	5.77676018
4	concepto	13	5.0612475	5.63794301
5	proceptos	6	7.24203993	5.63540242
6	topológica	7	5.90894139	4.99363479
7	dimensión	26	3.26312608	4.61723643
8	área	20	3.53791492	4.60293343
9	acepción	4	7.1151819	4.28376635
10	aleatorio	8	4.4983242	4.06239154

Fuente: Elaboración propia

Estos resultados de la tabla corresponden al segundo artículo, muestran que no hay coincidencias entre las palabras clave incluidas en la publicación y las palabras en el top ten del grado de fractalidad.

A continuación, se muestran los resultados obtenidos al procesar el artículo 1 por los algoritmos de TextRank y RAKE:

Tabla 5. Resultados de TextRank y RAKE del procesamiento del artículo 1

TextRank		**RAKE**	
Palabra	Rank	**Palabra**	Rank
Términos	1	Términos	17
Descriptores	2	Palabras clave	70
Palabras clave	3	Descriptores	98
Lenguaje controlado	23	Lenguaje natural	111
Lenguaje natural	24	Lenguaje controlado	117

Fuente: Elaboración propia

El TextRank tiene mejores resultados que el algoritmo del Grado de Fractalidad, colocando tres de cinco palabras clave de la publicación en el *top ten* y las otras palabras restantes en posiciones no tan lejanas. Para el caso de los resultados del RAKE podríamos decir que coloca estas keywords en otras posiciones. Para el caso del segundo artículo, los resultados son los siguientes:

Tabla 6. Resultados de TextRank y RAKE del procesamiento del artículo 2

TextRank		**RAKE**	
Palabra	Rank	**Palabra**	Rank
Fractal	7	Visualización	36
Visualización	33	Percepción	38
Procesos infinitos	34	Fractal	202
Percepción	250	Procesos infinitos	-

Fuente: Elaboración propia

Comparando los resultados de los tres algoritmos, el TextRank coloca las palabras clave elegidas por el autor en mejores posiciones que los otros dos algoritmos, incluido el algoritmo de fractalidad. Recordemos que la intención de usar el concepto de grado de fractalidad es modelar el proceso de extracción de palabras con contenido semántico de un texto, tal como lo haría el ser humano. Al comparar con otros algoritmos computacionales parece ser que no se tienen buenos resultados tomando como referencia las palabras que el mismo autor propone como importantes.

Como parte del análisis consistente en comparar las keywords obtenidas con el grado de fractalidad con las obtenidas con otras métricas, se procesó el libro de Charles Darwin con el

algoritmo computacional de TextRank. En la siguiente tabla se muestran los resultados del *top ten*:

Tabla 7. Resultados de TextRank del procesamiento del libro de Charles Darwin

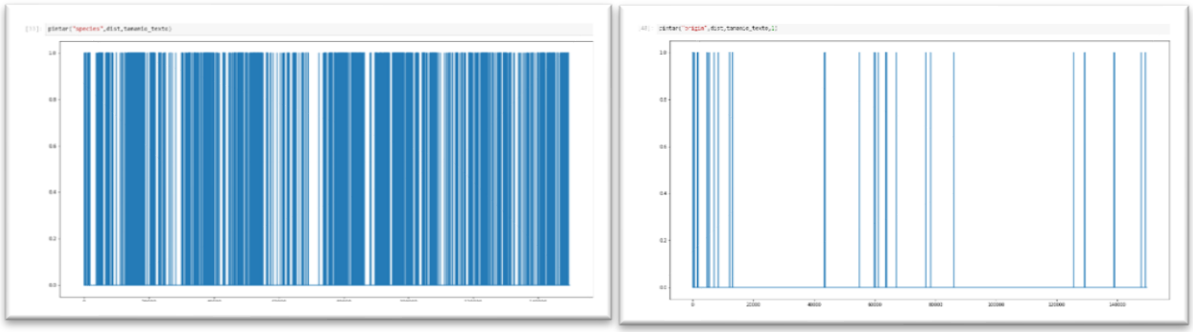
Word	Rank	Index
species	1	47.21429215
plants	2	23.75807429
case	3	20.9676611
animals	4	20.42369454
varieties	5	19.75799763
selection	6	19.689987
forms	7	18.07785369
instance	8	16.95769469
cases	9	15.95784828
nature	10	15.69190958

Fuente: Elaboración propia

Con los resultados anteriores, parece que los resultados de TextRank se asemejan más a lo que, intuitivamente, el humano selecciona como palabras más importantes o keywords de estos textos. Como podemos apreciar en los resultados, el TextRank rankea la palabra *species* como la principal keyword. Si los resultados del algoritmo de fractalidad son comparables con los resultados del TextRank podría ser un indicio de que se están recuperando las keywords del texto, esto es asumiendo que el TextRank lo haga. Al analizar estos resultados se muestra que las palabras que rankean cada algoritmo son muy diferentes.

Podríamos intuir que es necesario más de una ocurrencia de una palabra, para poder elegirla como clave para el contenido semántico del texto, pero esto no se pensaría de un texto explicativo de un nuevo concepto. En el libro de Charles Darwin, la palabra *species* tiene un gran número de ocurrencias a través del texto. Si se eligiera manualmente (humano) las palabras clave del texto, esto es, cuáles son las palabras más importantes de ese libro, muy probablemente la elegirían, así como la palabra *Origin*. En la siguiente gráfica se muestra su distribución.

Figura 8. Distribución de la palabra *species* y *origin* del libro de Charles Darwin

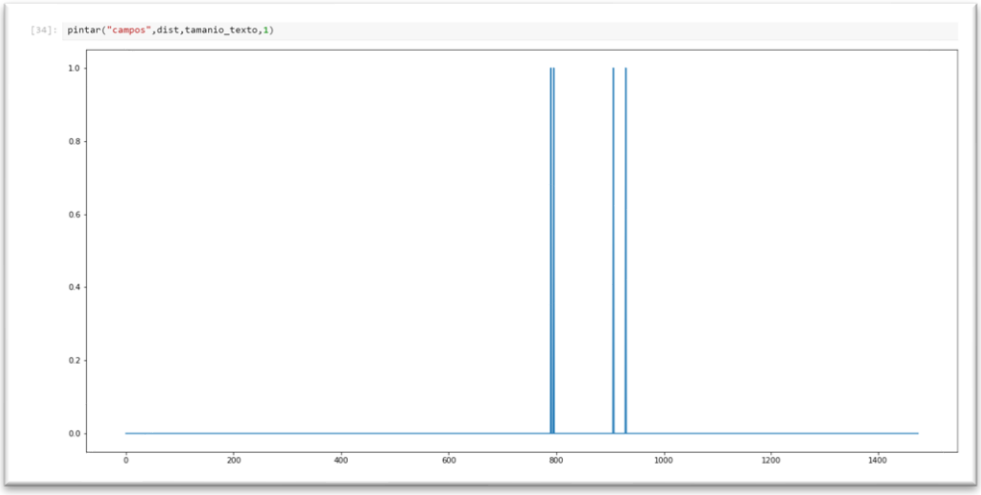


Fuente: Elaboración propia con Python 3.7

4.2 Relación distribución – importancia: la distribución de las keywords

Una de las relaciones que según afirman los autores, se encuentra entre grado de fractalidad y distribución de la palabra, por lo que al terminar el procesamiento se generan las gráficas de la distribución de las palabras mejor rankeadas según el grado de fractalidad. El eje de las abscisas (x): es el número de tokens en el texto, y el eje de las ordenadas (y): la ocurrencia de la palabra en esa posición. La siguiente gráfica representa la distribución de la primera palabra obtenida al procesar el artículo 1: *campos*, la cual tiene una frecuencia de 4.

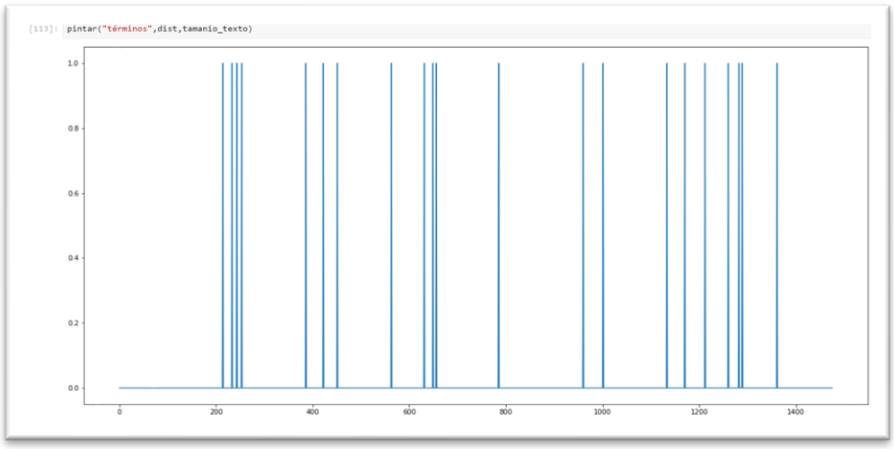
Figura 9. Distribución de la palabra *campos* en el artículo 1



Fuente: Elaboración propia con Python 3.7

Cada barra vertical refiere a que la palabra se encuentra en esa posición, si se torna de un color más obscuro significa que existen varias ocurrencias contiguas. Para la siguiente gráfica se muestra la distribución de la palabra rankeada por el grado de fractalidad en la segunda posición: *términos* con una frecuencia de 21.

Figura 10. Distribución de la palabra *términos* en el artículo 1

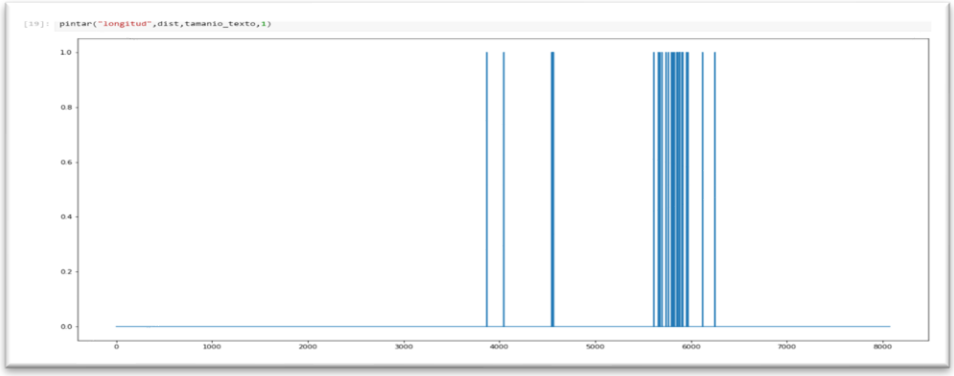


Fuente: Elaboración propia con Python 3.7

Esta gráfica da cuenta de que la relación entre el grado de fractalidad y la distribución concentrada de las ocurrencias de la palabra, no se da. Parece ser que la palabra se encuentra distribuida a través de todo el texto, a pesar de tener un grado alto de fractalidad. Los resultados del segundo artículo corroboran esta conclusión.

Las siguientes gráficas dan cuenta de la distribución de las palabras que tienen las mejores posiciones al procesar el artículo 2. La siguiente gráfica muestra la distribución de la primer palabra mejor rankeada por el grado de fractalidad: *longitud* con una frecuencia de 24.

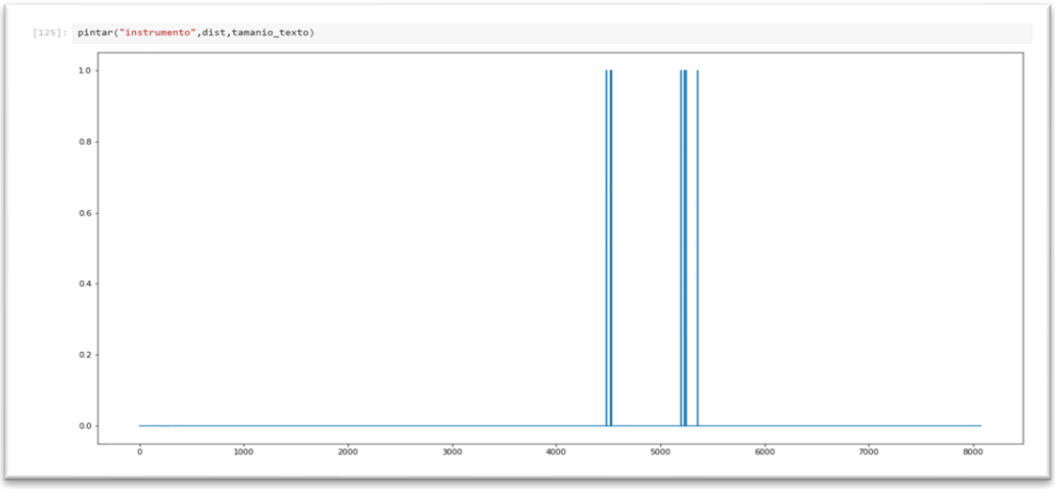
Figura 11. Distribución de la palabra *longitud* en el artículo 2



Fuente: Elaboración propia con Python 3.7

La siguiente grafica pertenece a la segunda palabra mejor rankeada: *instrumento*, con una frecuencia de 7.

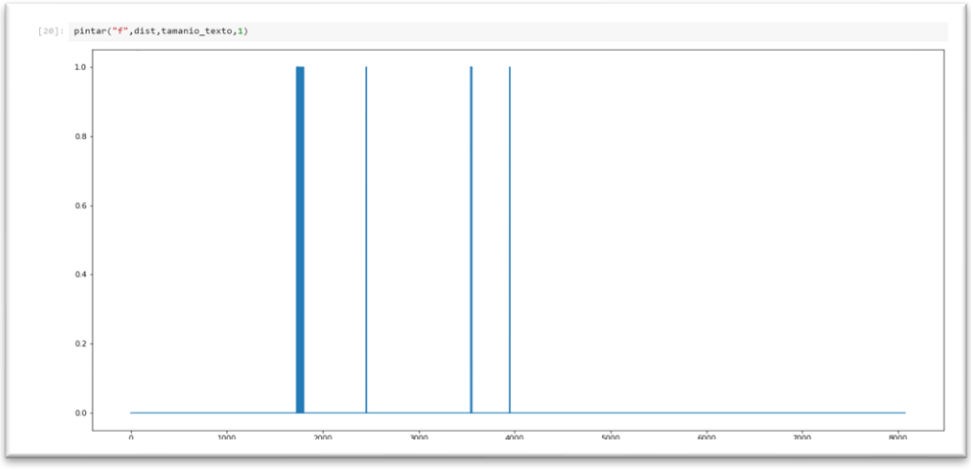
Figura 12. Distribución de la palabra *instrumento* en el artículo 2



Fuente: Elaboración propia con Python 3.7

Por último, se muestra la gráfica con los datos de la tercer palabra mejor rankeada: “f”, con una frecuencia de 11. Analizando esta palabra en el texto, se encuentra en un apartado donde se utiliza para hacer referencia a un ejemplo, por lo que el número de ocurrencias suceden en el mismo párrafo o en párrafos contiguos, además de que en los siguientes se hace uso nuevamente.

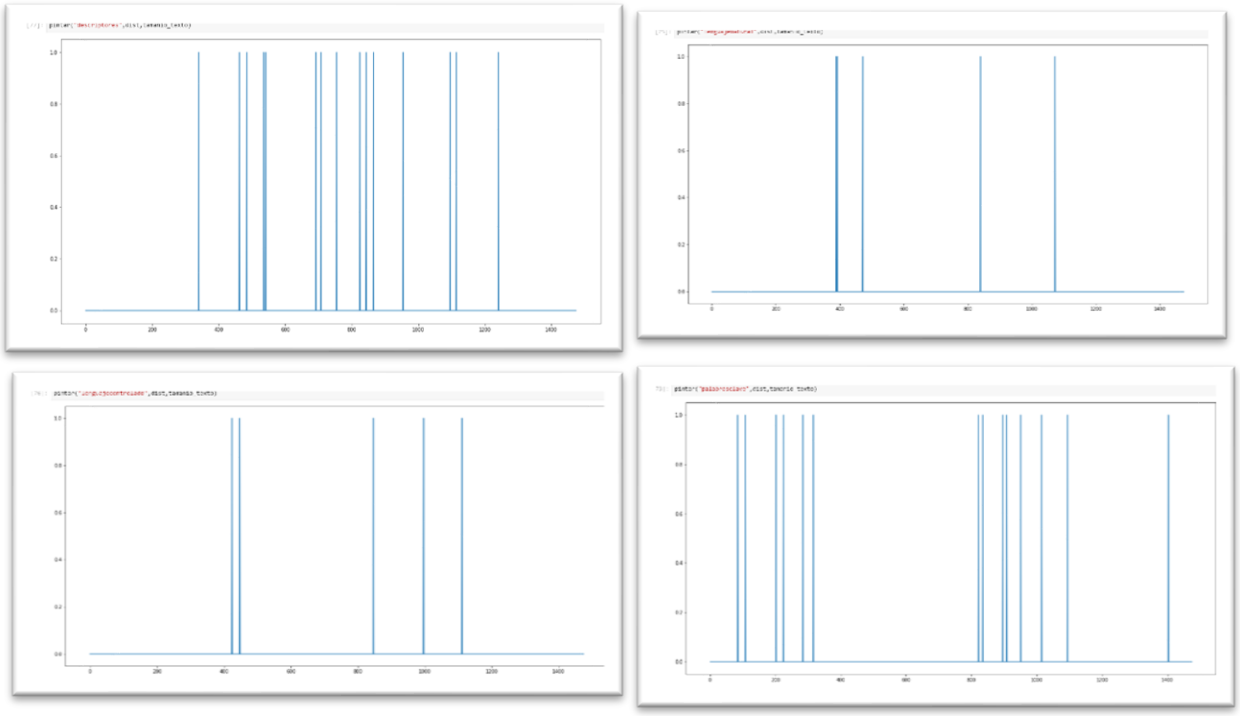
Figura 13. Distribución de la palabra *f* en el artículo 2



Fuente: Elaboración propia con Python 3.7

Los resultados muestran que en la distribución de la palabra se encuentra una concentración de ésta en ciertos apartados del texto, aunque se desagregan en otros más, por lo que no está clara la conclusión. Para poder analizar la relación entre distribución – importancia, ahora esta última en términos de la elección del humano sobre su propio texto, las siguientes gráficas representan la distribución de las palabras clave elegidas por los autores de los artículos científicos mencionados.

Figura 14. Distribución de las palabras *descriptores, lenguaje natural, lenguaje controlado y palabras clave* en el artículo 1



Fuente: Elaboración propia con Python 3.7

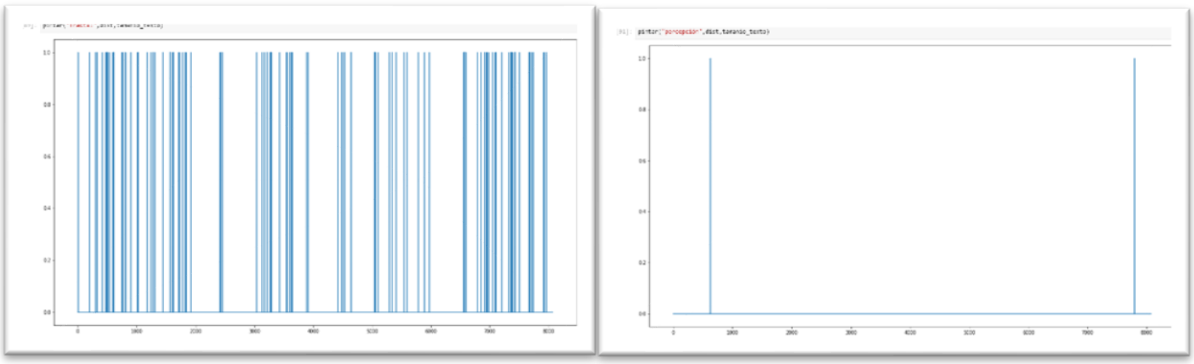
Para el procesamiento de la palabra clave: Lenguaje controlado, se realizó un preprocesamiento en el texto uniendo estas palabras de tal forma que el token resultante es: lenguajecontrolado. Este preprocesamiento se realizó con todas las palabras clave de esta forma, a fin de capturar un único índice del grado de fractalidad para esa palabra.

Con este análisis se pone a prueba si hay una relación entre las palabras que el humano escogió como keywords y su distribución en el texto. Los autores afirman que las palabras que se

concentran más son las más importantes. Tal como se muestra, esta relación no se da en la distribución de las palabras clave de los artículos: las palabras escogidas como palabras clave aparecen distribuidas uniformemente y no concentradas, lo cual puede ser por la subjetividad plasmada en la elección del autor. Recordemos que el proceso del autor es propio, es quien conoce el tema del artículo y bajo su criterio elige palabras que define como importantes para el texto. No existe un número o estadística referente al uso de una palabra, pero la distribución de esta en el texto nos podría dar otra información relevante de la importancia. El autor hará uso de una palabra para reforzar el tema principal, cuantas veces sea necesario y a través de los apartados de un texto.

La siguiente gráfica da cuenta de esto, donde las palabras elegidas como keywords por el autor en el segundo artículo, muestra que pueden existir muchas ocurrencias de la misma o solo unas pocas.

Figura 15. Distribución de las palabras *fractal y percepción* en el artículo 2



Fuente: Elaboración propia con Python 3.7

De lado izquierdo se representa la distribución de la palabra clave *fractal* y de lado derecho la de *percepción*. La palabra *fractal*, como se muestra, tiene una frecuencia alta y se encuentra distribuida por todo el texto, mientras que *percepción* solamente tiene dos ocurrencias en el texto; una al inicio y una al final. A pesar de esto, el autor carga de representación de contenido a estos términos, con diferencias en la distribución y en su frecuencia. Es decir, que la importancia que asigna el autor al elegir las keywords no parece tener que ver con la frecuencia de la palabra, ni con su distribución, al menos no consistentemente. Por otro lado, las

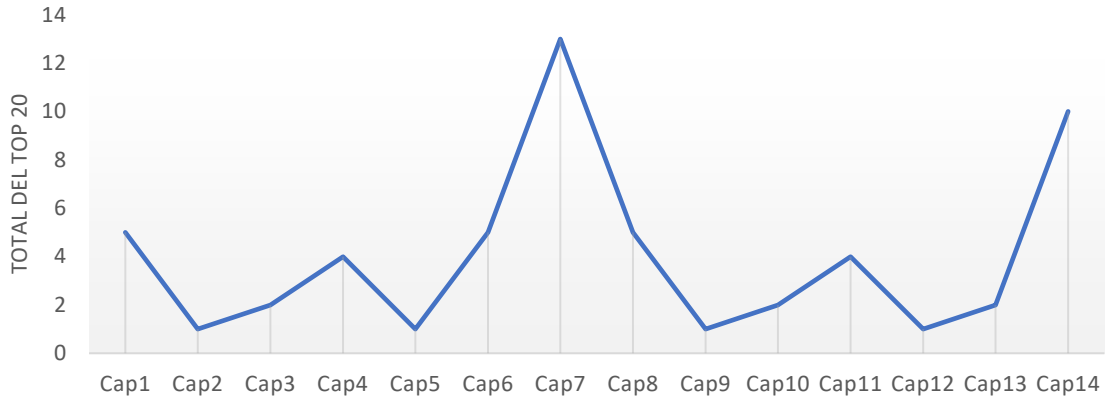
métricas automatizadas sí se apegan a alguno de estos criterios. Es así que el tipo de importancia en cada metodología corresponde con la subjetividad por parte del autor y de objetividad por métodos automatizados.

Los resultados no muestran una relación grado de fractalidad – distribución contundente en los artículos científicos analizados, tanto si se toma como "importancia" el resultado del análisis por el grado de fractalidad como si se pone a prueba la distribución de las palabras escogidas por el humano como "palabras clave" de su propio texto. Se aprecia que existen concentraciones en apartados, pero se utilizan estos términos en otros apartados. La importancia de la palabra (keywords) no tiene relación con la distribución de la misma a través del texto o al menos el mismo tipo de importancia.

4.3 Ubicación de las palabras obtenidas por el algoritmo de fractalidad en libros.

Como se demostró, el algoritmo de grado de fractalidad no extrae las keywords de forma comparable a como lo hace el humano en el caso de los artículos científicos, o de forma comparable a como lo hacen los algoritmos TextRank y RAKE. Por otro lado, la relación distribución – importancia tampoco se valida. ¿Qué información nos muestra el grado de fractalidad entonces? Un análisis derivado de los resultados fue la distribución de las palabras obtenidas por el algoritmo a través de capítulos o apartados. Estos apartados se construyen típicamente como unidades temáticas, en los que se trata de uno o varios subtemas relacionados entre sí y diferenciados de otros apartados. Tomando como una unidad los apartados y capítulos, se procesan los textos y grafica la distribución de las palabras. Se analiza primero como las primeras 20 palabras obtenidas al procesar el libro completo de Charles Darwin con el algoritmo de fractalidad, se distribuyen por capítulos a través del texto.

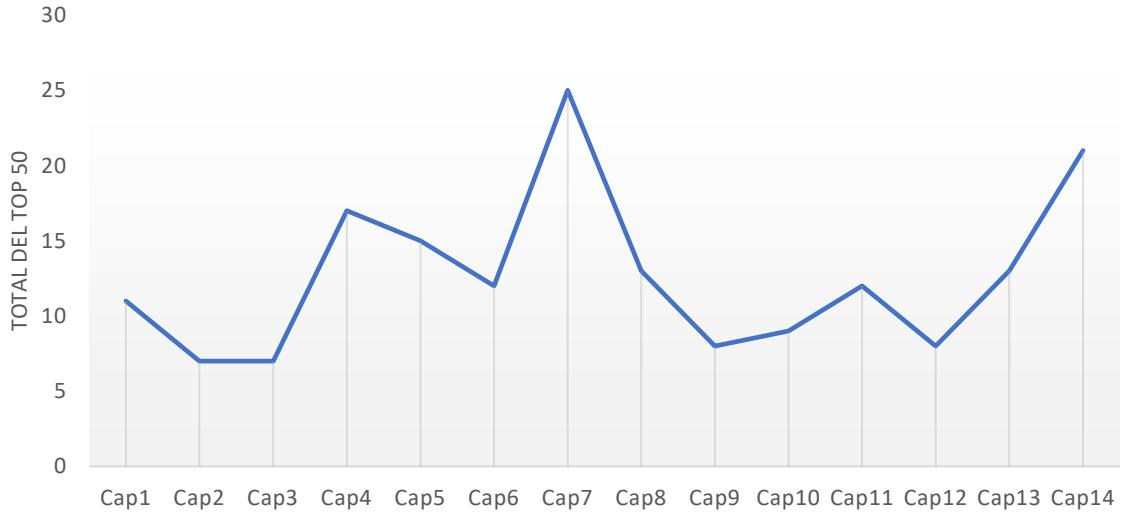
Figura 16. Distribución por capítulos del top 20 del libro de Charles Darwin



Fuente: Elaboración propia.

Es interesante cómo la mayoría de las palabras obtenidas del top 20 se concentran en el capítulo siete, el cual se titula “Instinct”. Esto puede dar cuenta que ese capítulo es importante. A continuación, se realizó el mismo ejercicio, pero procesando el top 50, 100 y 200 palabras obtenidas por el rankeo.

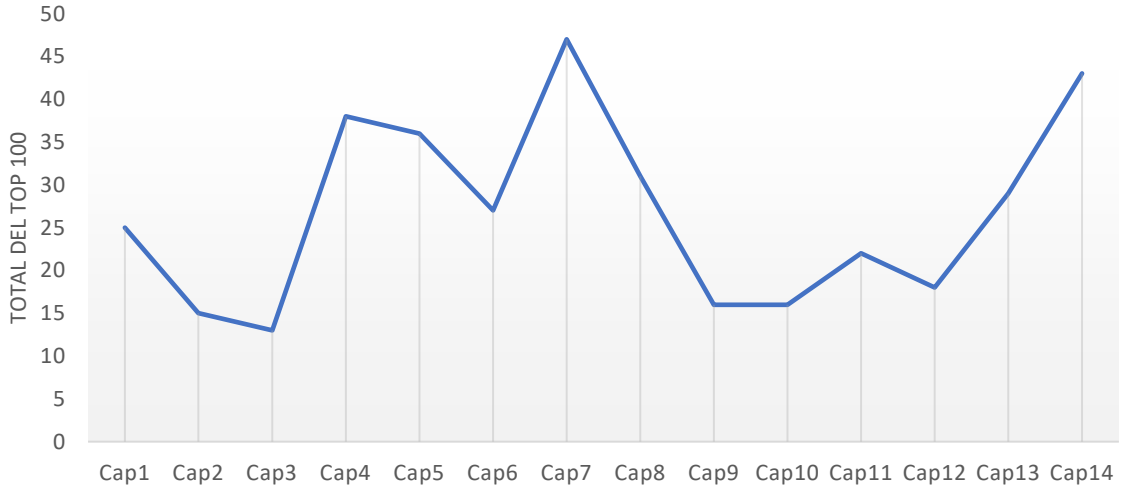
Figura 17. Distribución por capítulos del top 50 del libro de Charles Darwin



Fuente: Elaboración propia.

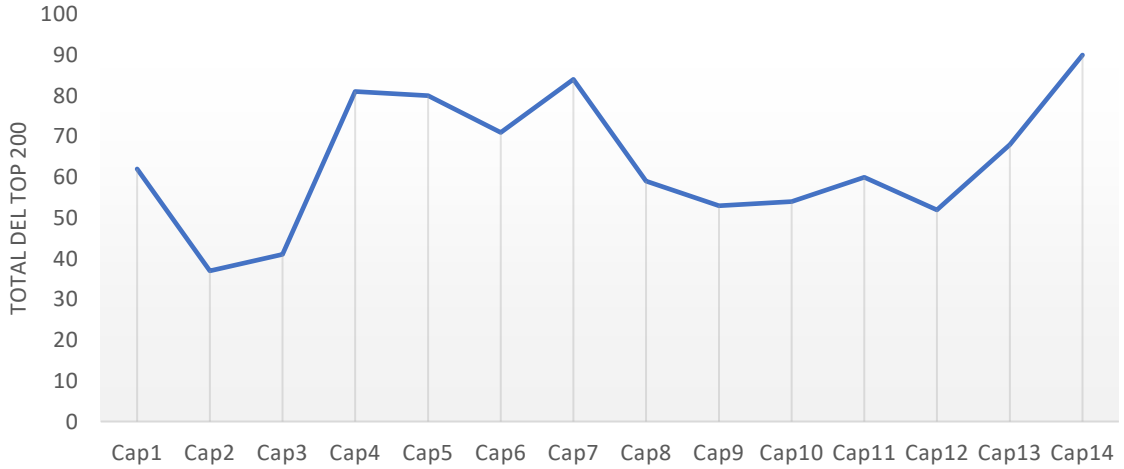
Comparando las dos gráficas anteriores, notamos que se mantienen los picos de los apartados que concentran las primeras palabras del rankeo. Una de las propiedades del fractal es la autosimilitud, la cual se refleja en estas graficas manteniendo los picos de los capítulos.

Figura 18. Distribución por capítulos del top 100 del libro de Charles Darwin



Fuente: Elaboración propia.

Figura 19. Distribución por capítulos del top 200 del libro de Charles Darwin.

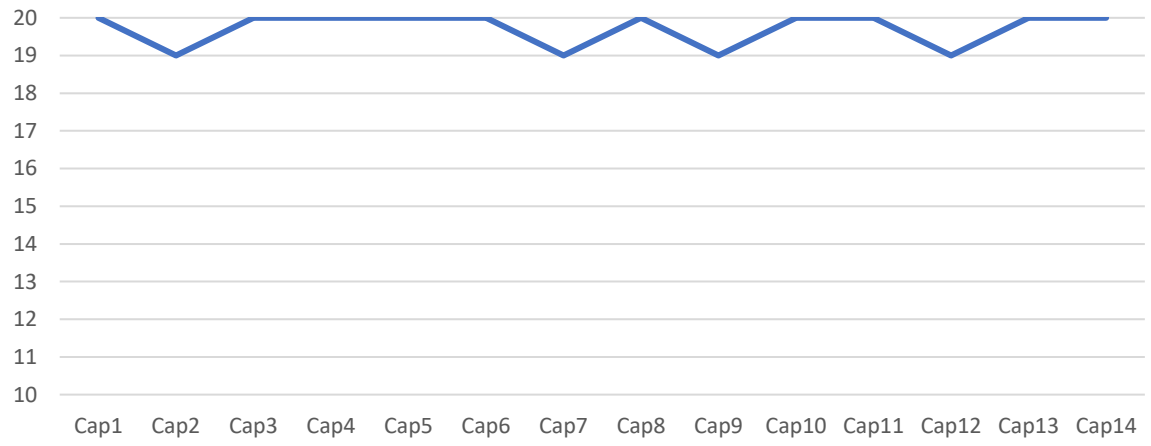


Fuente: Elaboración propia.

Esto se logra apreciar conforme se grafica el número de palabras mejor rankeados por capítulos y se va aumentando. Mientras se agregan palabras del rankeo los picos se mantienen, esto es que hay más palabras concentradas en estos capítulos que son importantes, según refleja su grado de fractalidad y se puede suponer que el autor colocó estas palabras ahí por alguna razón relativa a la estructura deseada en su texto, o quizá que esto pertenezca a su patrón de escritura.

Antes de mostrar los resultados de otros libros del mismo autor y para continuar contrastando la medida de grado de fractalidad y la de otros algoritmos de extracción de keywords, se muestra la distribución por capítulos del top 20 obtenido por el algoritmo TextRank en la siguiente gráfica.

Figura 20. Distribución por capítulos del top 20 obtenido con TextRank del libro de Charles Darwin



Fuente: Elaboración propia.

La comparación de la distribución por capítulos ahonda en el resultado de que la valoración de “palabras clave” e “importancia” es muy diferente en los dos algoritmos. Para el caso de TextRank se refleja en que las palabras distribuidas uniformemente son las más importantes, y estas se encuentran en varios apartados. En caso contrario, lo que promueve el algoritmo de fractalidad es que a mayor concentración de la palabra en el texto es más importante y se refleja en ciertos apartados.

4.4 Ubicación de las palabras obtenidas por el algoritmo de fractalidad en tesis de investigación.

Los textos analizados hasta el momento fueron libros y artículos científicos. Estos difieren en dos aspectos clave: su longitud, y su estructura más o menos constreñida por el tipo de texto: los artículos son más breves y también tienen una estructura más definida, a priori. Para completar el análisis y poner a prueba si las características encontradas en el procesamiento del libros y artículos tienen que ver con la mayor rigidez de la estructura de los artículos o la longitud de los textos, se procesaron tesis de investigación de posgrado; completamente y por capítulos. La idea de este análisis es identificar si en un documento con una estructura definida (de tamaño menor que un libro, pero longitud mayor típicamente que un artículo científico), la extracción de las palabras clave por el algoritmo se compara con las palabras obtenidas del resumen (abstract) y/o las palabras clave definidas por el autor.

Para este proceso se usaron 14 tesis de la siguiente manera:

- Se procesaron las tesis completas, obteniendo las palabras por el algoritmo.
- Se procesaron los resúmenes de cada documento por el algoritmo, solo en el caso de que existía el apartado.
- Se recuperaron las palabras clave definidas en los documentos por parte del autor.
- Se graficó la distribución de las palabras obtenidas por el algoritmo por capítulos.

La comparación de estos tres conjuntos mostró que en 10 de las tesis de posgrado no se encontró una relación entre: palabras obtenidas al procesar toda la tesis, palabras de procesar el resumen y las palabras clave definidas por el autor. Para los cuatro documentos faltantes sí hubo coincidencias y se describen a continuación:

1. Archivo: HEBYJR00T

   Coincidencia en dos keywords de cuatro, una en el procesamiento del resumen y una en el procesamiento de todo el texto.

2. Archivo: RAAMLC01T

   Coincidencia en tres keywords de cinco, tres en el procesamiento del resumen y una en el procesamiento de todo el texto.

3. Archivo: FIBVHC07T

   Coincidencia de cuatro keywords de nueve, cuatro en el procesamiento del resumen.

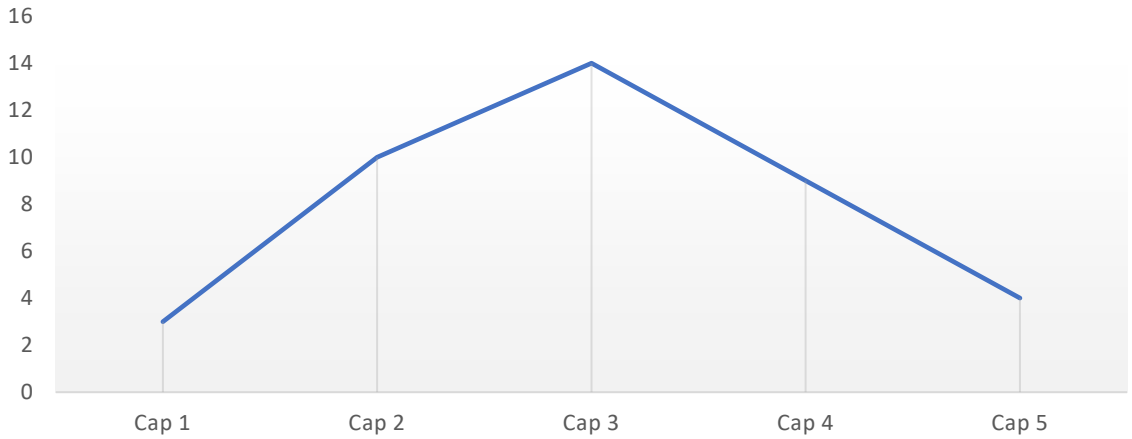
4. Archivo: VAGSVN06T

   Coincidencia de tres keywords de seis, tres en el procesamiento del resumen y una en el procesamiento de todo el texto.

La comparación se realizó con un corte de las primeras top 20 palabras obtenidas por el grado de fractalidad del texto completo y del resumen. Al analizar las palabras clave definidas por el autor no se encontraron en el resumen. De manera intuitiva, se esperaba que el autor al elegir palabras clave del tema global de las tesis estas las retomaría en el resumen. Esto no paso por lo que se descartó diez de las 14 tesis procesadas. Esto nuevamente apunta a que el grado de fractalidad está señalando palabras importantes no coincide en la mayoría de los casos con lo que el autor señala como contenido importante del texto.

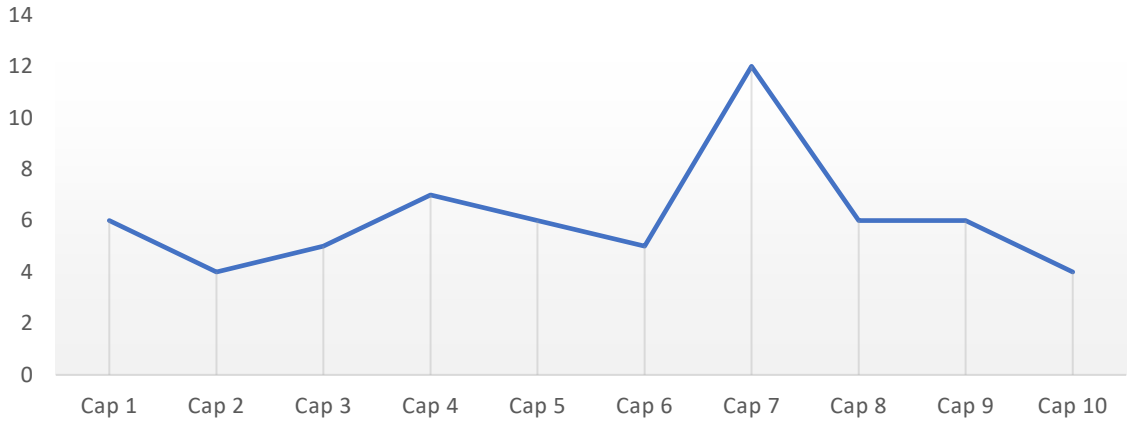
Aunado a esto, se analizó la distribución por capítulo del top 20 de las palabras mejor rankeadas por el algoritmo. Para esto se procesaron cinco tesis de las 14 tesis mencionadas de la siguiente forma: cada tesis se dividió por capítulos para identificar en cuáles se encontraban esas principales palabras obtenidas en el rankeo.

Figura 21. Distribución por capítulos del top 20 obtenido con TextRank de la tesis MABCTR02T



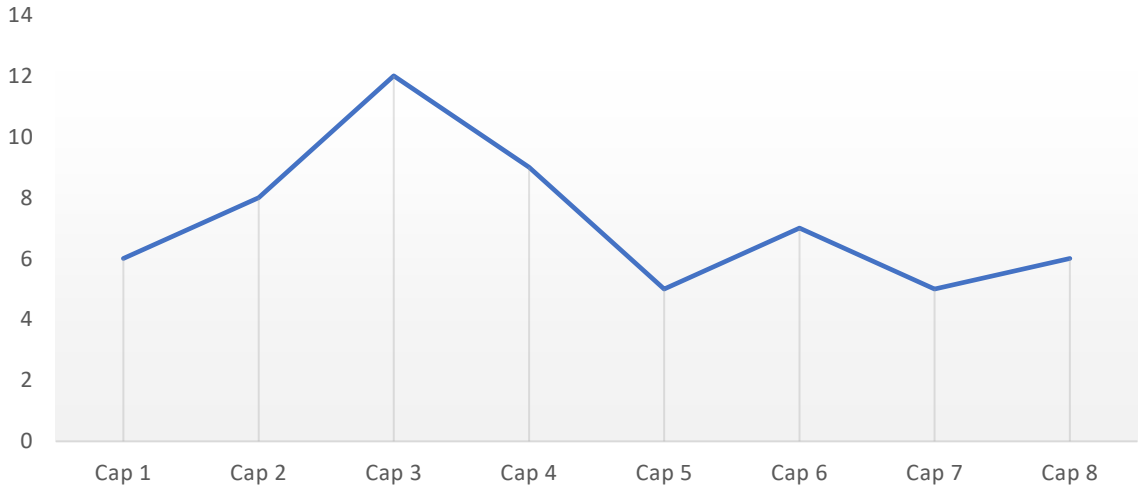
Fuente: Elaboración propia.

Figura 22. Distribución por capítulos del top 20 obtenido con TextRank de la RXDC00T



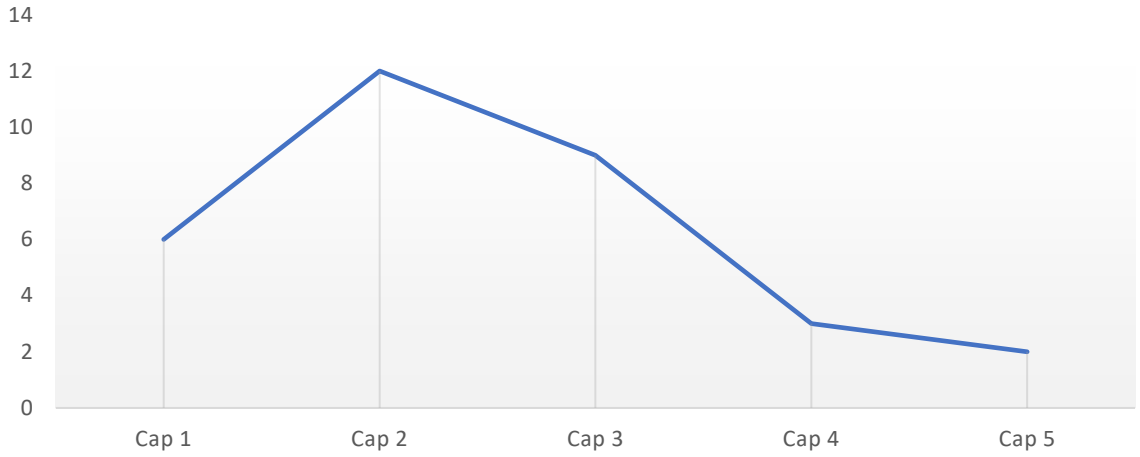
Fuente: Elaboración propia.

Figura 23. Distribución por capítulos del top 20 obtenido con TextRank de la RAPACD04T



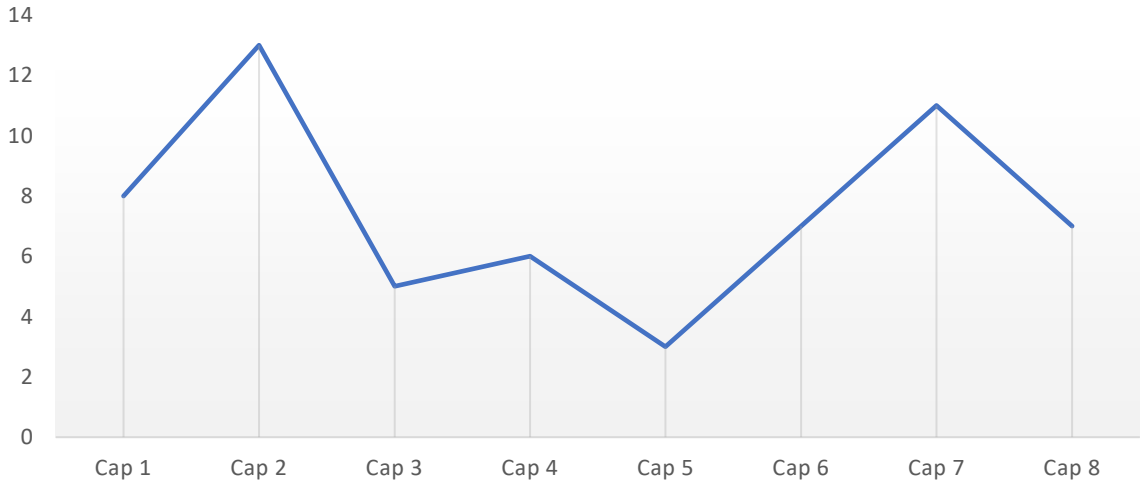
Fuente: Elaboración propia.

Figura 24. Distribución por capítulos del top 20 obtenido con TextRank de la FIBVHC07T



Fuente: Elaboración propia.

Figura 25. Distribución por capítulos del top 20 obtenido con TextRank de la VAGSVN06T



Fuente: Elaboración propia.

Desde la primera tesis hasta la cuatro, en las gráficas de la distribución por capítulos se muestra un pico más pronunciado en el marco teórico o el método, mientras que T5 tiene dos picos y justamente son los apartados de marco teórico y de resultados. Aunque estos datos no son concluyentes, las palabras del top 20 extraídas por el algoritmo se concentran en apartados importantes para el tema global del texto, en el sentido de que en este apartado se introducen conceptos y por tanto, términos que aportan altamente contenido conceptual al documento

completo de la tesis y que, específicamente deben aparecer concentrados en este punto del documento.

En resumen, los resultados de los análisis presentados a lo largo del capítulo muestran que el algoritmo de grado de fractalidad en la extracción de keywords no es tan eficiente como el TextRanK si se toma como referencia de la eficacia las palabras clave escogidas por el humano, autor del texto. Esto fue comprobado principalmente con artículos científicos. Para el caso de la relación entre distribución-importancia, para el algoritmo de fractalidad se comprobó que no hay una relación intrínseca, esto se comprobó en artículos científicos por el grado de fractalidad. Para el caso de la elección por parte del autor, y el algoritmo de TextRank la relación se da de la siguiente forma: con una distribución uniforme mayor importancia.

Aunado a esto, se identificó que la elección de keywords en artículos científicos no está sujeta a número determinado de ocurrencias de la palabra en el texto. De los análisis derivados de estos resultados se muestra que el rankeo obtenido del algoritmo del grado de fractalidad, coloca a las principales palabras en apartados específicos del texto, esto se validó en libros. Más aún, cuando se integran más palabras rankeadas (en orden) se mantienen los picos donde se usan estas palabras, esto no ocurre con el rankeo de TextRank. Por otro lado, en el procesamiento con tesis de investigación se mostró que los apartados donde se encuentran estas palabras mejor rankeadas se encuentran en el Marco teórico y Método.

5. Discusión y conclusiones

- ¿Qué podemos recuperar del algoritmo de grado de fractalidad?

La escritura es un acto difícil, debido a que se intenta brindar todas las herramientas para poder transmitir el tema principal del escrito. Para esto, se realizan diferentes procesos, como estructuración de las ideas, la escritura y la reestructuración de las ideas nuevamente. En cada paso, el escritor se coloca en el papel del lector, agregando información necesaria para entender el tema principal. La expresión de las ideas y la comprensión son dos procesos cognitivos distintos pero que se modifican el uno al otro, es decir, al escribir se toma en cuenta el proceso de comprensión del lector, lo cual condiciona el proceso de escritura. La correcta posición de las palabras y oraciones están sometidas a la estructura de las ideas y permite una transmisión con claridad, exactitud y fluidez en el discurso. Es así que la colocación de las palabras en un texto es una de muchas combinaciones posibles siguiendo reglas gramaticales y orden semántico. La ubicación de las palabras entonces resulta ser un estímulo y una estrategia del autor que permite al lector interpretar el significado del texto de acuerdo a su conocimientos y experiencia.

La comparación de los resultados del algoritmo de grado de fractalidad con otros algoritmos de extracción de keywords, tales como TextRank y RAKE, muestra que el algoritmo propuesto por Najafi & Danooreh no obtiene resultados similares a los de estos algoritmos, creados específicamente para la extracción de keywords. El algoritmo de TextRank, además, tiene diferentes resultados al procesar artículos científicos y obtiene resultados que resultan más cercanos al proceso cognitivo del humano, autor del artículo, que escoge las palabras clave. Estas diferencias apuntan a que, a pesar de que en los trabajos previos se hable por igual de “keywords”, ambas medidas están ofreciendo información respecto a diferentes conceptos o tipos de “importancia”, y a esto se une el hecho de que en la bibliografía previa no parece existir una definición o explicación clara de qué quiere decir que una palabra sea “importante” en el texto, es decir, que una palabra sea “clave”. La hipótesis presentada en esta investigación, permite al menos, plantear un primer análisis respecto a qué criterios diferentes y/o compartidos subyacen a la idea de “importancia” en estos métodos de extracción de keywords, y así potencialmente reflexionar en el futuro sobre en qué medida estos están, de hecho, también en el proceso cognitivo humano.

Para comprobar nuestra hipótesis, la clave para ahondar en esta diferencia se encuentra en la observación de la distribución que tienen estas palabras. Al analizar la distribución de las palabras extraídas por TextRank notamos que se encuentran distribuidas uniformemente en todo el texto, característica que en Najafi & Danooreh (2015) se atribuía a las palabras poco importantes.

Estas metodologías de los algoritmos de acuerdo a la relación entre importancia y distribución se contraponen. Como mencionamos anteriormente, para Najafi & Danoreeh (2015) a mayor concentración de la palabra en un apartado del texto es mayor su importancia y que la palabra tenga una distribución uniforme significa que no es importante. En contraste, el algoritmo de TextRank valora las palabras de acuerdo a las asociaciones y ocurrencias que existan entre palabras, lo cual apunta a que tendrán una distribución uniforme en el texto. En ambos algoritmos si las palabras del texto son barajadas, la valoración de las palabras será diferente que, en el original, siendo importante para los dos procesos la ubicación de las palabras. Es así que se concluye que el algoritmo de grado de fractalidad extrae palabras con un tipo de importancia distinta que la de los algoritmos computacionales que intentan emular el proceso cognitivo del humano.

Para el caso de los artículos científicos, las keywords son elegidas por el autor del escrito. Estas palabras se valoran como representativas del contenido del texto, y claro está que el autor conoce el trabajo desarrollado. Al comparar los resultados del grado de fractalidad y la elección de keywords en artículos científicos no existe similitud, por lo que se concluye que el algoritmo no extrae las keywords en los términos de ese tipo de importancia. Esto se pudo ver afectado por la carga subjetiva en la elección de las palabras puesto que se mostró que en el primer artículo existe una palabra clave que solo tiene dos ocurrencias en todo el texto.

A pesar de que el algoritmo se propuso para extracción de keywords, la concepción del procesamiento de lenguaje escrito utilizando el fractal, resulta una buena propuesta, puesto que el algoritmo de grado de fractalidad podría identificar palabras que articulan el contenido global del texto; es decir, al analizar las palabras mejor rankeadas notamos que no se encuentran lejos del campo semántico global y dan estructura al contenido del texto. De la misma forma el grado de fractalidad podría mostrar un patrón característico del texto y/o de la forma de escritura del autor y es ahí donde se asocia el tipo de importancia que captura esta medida, tal como se mostró en los resultados por capítulos del libro de Charles Darwin.

El algoritmo de grado de fractalidad muestra que hay una relación de las palabras rankeadas y su distribución en el texto, que se confirma al procesar varios libros de Charles Darwin. Esta relación podría estar asociada a textos extensos como los libros, pero en textos que son más pequeños no se cumple. Esto no existe entre las palabras obtenidas al procesar artículos científicos y más aún las keywords establecidas por autores tampoco cumplen la relación distribución-importancia, tal como se plantea en la hipótesis.

El número de ocurrencias de las palabras en un texto no depende directamente de la longitud del texto. Es decir, un texto corto puede contener muchas ocurrencias de alguna palabra que haga referencia a un concepto esencial para este, mientras que un texto extenso puede solo tener pocas ocurrencias de la misma, o viceversa; todo depende de la estructura. En promedio, los textos extensos contienen varias ocurrencias de las palabras que hacen referencia a un concepto esencial en comparación con los textos cortos (Katz, 1996). Los algoritmos previos han tomado en cuenta la frecuencia de las palabras en un texto como uno de los elementos al computar su importancia, pero no se ha combinado esta información con información relativa a cómo se distribuyen esas palabras en un texto. Un gran número de ocurrencias pueden estar distribuidas de forma más o menos uniforme en el texto y añadir la información respecto a su distribución permite ofrecer un mejor recuento de la estructura de los contenidos del texto.

El proceso de extracción de palabras clave, parece asumir que los textos tratan de un tema en específico y a la vez arrojan una serie de palabras a través de las cuales se articula el contenido semántico del texto. En la práctica, los textos pueden tratar sobre más de tema, de los cuales unos pueden ser tema global o temas locales, importantes en un determinado fragmento del texto (Reynar, 1999). A partir de los análisis realizados en este trabajo, se considera que la combinación de herramientas más eficaces en la extracción de keywords y medidas que aporten información sobre la distribución de las palabras en el texto, como el algoritmo de fractalidad, se puede avanzar para obtener información sobre la distribución de los temas abordados en un texto. Es así que hacemos la siguiente conjetura: Si una palabra tiene una distribución uniforme en el texto, indica que es un tema o tópico principal que vincula otros. En caso contrario si existe una menor dispersión (es decir, se encuentran más "concentrada") significa que la palabra toma relevancia dentro de una sección (o varias) del texto (puede ser en uno o varios párrafos contiguos) y dan soporte al tema principal puesto que son

subtemas o subtópicos. Esto abre una posibilidad en la investigación sobre el modelado de la estructura jerárquica de contenidos en un texto en PLN.

Tomando como unidad textual los capítulos, los resultados de las tesis de investigación confirman la hipótesis, ya que se identificaron palabras que se ubican en apartados específicos como el Marco Teórico, Método y Resultados. Estos apartados son relevantes pues sustentan la investigación y el método empleado con sus resultados obtenidos de su aplicación. Como se mencionaba respecto al tipo de importancia de esas palabras extraídas, al concentrarse en estos, indican la importancia del apartado.

Recapitulando la investigación, los resultados de los métodos automatizados para extracción de keywords, el análisis de la elección de los keywords por parte del autor en artículos científicos y tesis de investigación, la comparación de los resultados de los procesos anteriores y los resultados del procesamiento por capítulos de los textos por el algoritmo de fractalidad, nos dan cuenta de la gran cantidad de matices que tiene el concepto de importancia. Por un lado, la metodología de los métodos automatizados sustenta esta idea del tipo de importancia, puesto que, al intentar emular el proceso cognitivo de extracción de información, su procesamiento es distinto, justo como se formuló en la hipótesis. Por otro lado, estas metodologías se apoyan en la distribución de las palabras en el texto y su frecuencia para elegir keywords, esto es, en el algoritmo de fractalidad mientras más concentrado se encuentre una palabra es más importante, mientras que en el caso de TextRank, a mayor distribución mayor importancia. Aunado a esto, del análisis de los artículos científicos y tesis de investigación, la distribución de las keywords elegidas por el autor nos muestra que el patrón que promueven los algoritmos no los sigue el humano; en algunos casos las palabras elegidas tienen una frecuencia pequeña, en otros más grande o su distribución es uniforme o no. Sin embargo, en algunos casos, el TextRank obtiene keywords cercanas a las que el humano elegiría con un tipo de importancia, pero el grado de fractalidad rankea palabras que tienen una relación con el campo semántico del texto, es decir, no son palabras que el humano podría indicar que no representan nada del escrito. Por último, la concentración de estas palabras obtenidas por el grado de fractalidad en ciertos capítulos, promueve un tipo de importancia del apartado, tal como se mostró en las tesis de investigación y libros, sustentando así nuestra hipótesis.

Es así que en el presente estudio se analizó la relación de distribución de las palabras con una noción de tipo de importancia, desde diferentes metodologías, y se determinó en qué medida el concepto de fractal aplicado a lenguaje escrito constituye una unidad de medida útil para caracterizar y modelar la distribución de las palabras que captura en una estructura del texto.

En relación con esta investigación, a través de la lingüística cuantitativa y computacional se estudia la incertidumbre de las palabras, a través del término *Word Entrop*y, un concepto teórico de la información, que mide el nivel de incertidumbre en un texto, es decir, qué tan organizado o desorganizado es. Mehri & Darooneh (2011) utilizan esta métrica para extraer keywords en textos escritos, obteniendo "mejores resultados" que TextRank. *Word Entropy* también se ha utilizado en oraciones para identificar cómo se distribuye el contenido de la información en las diferentes posiciones de una oración (Yu et al., 2016).

Por otro lado, Serrano, Flammini & Mencze (2009) afirman que "si un término ocurre con más frecuencia que otro en las primeras líneas de un documento sería suficiente para detectar *burstiness* de la palabra y, en consecuencia, el tema del texto". El termino *burstiness* es la característica de algunas palabras a aparecer agrupadas al inicio del documento, implicando que reaparezca nuevamente en el texto.

Estos últimos conceptos están alineados al análisis de las propiedades distributivas de las palabras en un texto escrito, por lo que la relación entre distribución y algún tipo de importancia existe con sus diferentes matices.

Por último, como hipótesis para el futuro, el algoritmo de fractalidad nos podría proporcionar información útil sobre otras piezas lingüísticas llamadas fillers o pausas llenas (mmm, este, ajá, …) cuya distribución en discursos orales no está claramente explicada, y que no contribuyen al significado pero que podrían asemejarse más a "fenómenos naturales"; no regidos por reglas gramaticales propias del lenguaje, pero si a su ubicación en el discurso.

6. Referencias

Barber HA, Kutas M. (2007). Interplay between computational models and cognitive elec trophysiology in visual word recognition. *Brain Res*, 53(1), 98-123.

Barron-Cedeño, A., Sierra, G., Drouin, P. & Ananiadou. S. (2009). An improved automatic term recognition method for spanish. In Proc of the 10th Int. *CNF on Computational Linguistics and Intelligent Text Processing*, 125–136

Bowker, L. & Pearson, J. (2003) *Working with Specialized Languaje*. Taylor & Rancis e-Library.

Calles, F. (2001) La distribución del texto escrito en la página: una técnica de comunicación visual. *Investigación Universitaria Multidisciplinaria Revista de Investigación de la Universidad Simón Bolívar, (3)*

Contreras, M. (2018) Aplicación del algoritmo RAKE en la indización de documentos digi tales. *Investigación Bibliotecológica: archivonomía, bibliotecología e información*, 32 (75), 109-123.

De Granda, JI, García, F & Callol, L. (2003) Importancia de las palabras clave en las bús quedas bibliográficas. *Rev Esp Salud Publica*, 77(6), 765-767.

De Granda, JI, García, F, Roig, F, Escobar J, Gutiérrez - Jiménez T & Callol, L (2005). Las palabras clave como herramientas imprescindibles en las búsquedas bibliográficas. Análisis de las áreas del sistema respiratorio a través de Archivos de Bronconeumo logía. *Arch Bronconeumol*, 41, 78-83.

Dode, A. & Hasani, S. (2017). PageRank Algorithm. *IOSR Journal of Computer Engineering (IOSR-JCE)* 19 (1), 01-07

Flach, Peter A. (2012). *Machine learning : the art and science of algorithms that make sense of data*. Cambridge ; New York :Cambridge University Press

Fedorenko E, Ivanova A, Dhamala R, Bers MU. (2019). The Language of Programming: A Cognitive Perspective. *Trends Cogn Sci*, 23(7), 525-528.

Ingwersen, P. (1996) Cognitive Perspectives Of Information Retrieval Interaction: Ele ments Of A Cognitive Ir Theory *Journal of Documentation*, 52(1), 3-50.

Katz, S. (1996) Distribution of content words and phrases in text and language modelling. *Natural Language Engineering 2* (1), 15-59

Khurana, D., Koli, A., Khatter, K. & Singh, S. (2017). Natural language processing: State of the art, current trends and challenges. *arXiv*, 1-25.

Mandelbrot B. (1982). *The Fractal Geometry in Nature*. W.H. Freeman and Company.

Mehri A & Darooneh A. (2011) The role of entropy in word ranking. *Physica A, 390, 3157–3163*

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Text. *EMNLP*.

Muñoz-Martin, B. (2016) Descriptores y palabras clave. *ORL,* 7(3), 179-183.

Najafi, E. & Darooneh A. (2015) The fractal patterns of words in a text: A method for auto matic keyword extraction. *PlosONE* 10(6)

Pareyon G. (2007) Fractal theory and language: the form of macrolinguistics. *Form and Symmetry: Art and Science.* Buenos Aires Congress

Schunk, D. (2012) *Teorías del aprendizaje. Una perspectiva educativa.* México: PEARSON EDUCACIÓN.

Shannon B. (1993) Fractal patterns in language. *Newideas in Psychol*, 11(1), 105-109.

Tolosa, G y Bordignon, F. (2007) *Introducción a la Recuperación de Información Conceptos, modelos y algoritmos básicos*. Argentina: Universidad Nacional de Luján.

Urbizagástegui, R. & Restrepo, C. (2011). La ley de Zipf y el punto de transición de Goffman en la indización automática. Investigación bibliotecológica. *Scielo,* 25(54), 71-92.

Ursino M., Cuppini C. & Magosso E. (2010). A Computational Model of the Lexical-Semantic System Based on a Grounded Cognition Approach. *Frontiers in Psychology*. (1)2010, 221

Vallez, M. & Pedraza-Jiménez, R. (2007). El Procesamiento del Lenguaje Natural en la Re cuperación de Información Textual y áreas afines. *Hipertext.net*, 5.

Walker GM, Hickok G. (2016). Bridging computational approaches to speech production: The semantic-lexical-auditory-motor model (SLAM). *Psychon Bull*, 23(2), 339-352.

Xuebo Song, Pradip K. Srimani, and James Z. Wang. (2019). HWE: Hybrid Word Embed dings For Text Classification. In Proceedings of the 2019 3rd International Confer ence on Natural Language Processing and Information Retrieval (NLPIR 2019). *As sociation for Computing Machinery*, New York, NY, USA, 25–29.

Zhang, Y., Rahman, M.M., Braylan, A., Dang, B., Chang, H., Kim, H., McNamara, Q., An gert, A., Banner, E., Khetan, V., McDonnell, T., Nguyen, A.T., Xu, D., Wallace, B.C., & Lease, M. (2016). Neural Information Retrieval: A Literature Review.