UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

"Hipervinculación de documentos con Máquinas de Soporte Vectorial"

Tesis
Para obtener el Grado de
Maestro en Ciencias de la Computación

Que Presenta

Alan Josué Serrano León

Asesor:
Dr. René Arnulfo García Hernández

TIANGUISTENCO, MÉX. ENERO 2018

UAEM | Universidad Autónoma del Estado de México

DICTÁMEN DE AUTORIZACIÓN Y OBTENCIÓN DE GRADO DE MAESTRÍA

Tianguistenco, Méx. , a 15 de enero de 2018

Título del proyecto:

Hipervinculación de Documentos con Máquinas de Soporte Vectorial

Tesista:

Ingeniero en Software Alan Josué Serrano León

Dictamen: **No. de revisión:** 6

- ☐ **Rechazado**
- ☐ **Sujeto a modificaciones**
- ☐ **Aceptado, condicionado**
- ■ **Aceptado**

Observaciones generales:

Aceptado para la impresión
Aceptado para la defensa de grado

Tutor Adjunto	**Tutor Académico**	**Tutor Adjunto**
Dra. Yulia Nikolaevna Ledeneva	**Dr. René Arnulfo García Hernández**	**Dr. José Luis Tapia Fabela**

Declaración de originalidad del trabajo escrito

Mediante esta carta hago constar que el trabajo de tesis presentado en este documento es original porque cita debidamente los contenidos utilizados como soporte a la investigación presentada, por lo que exonero a la Universidad Autónoma del Estado de México de cualquier problema de derechos de propiedad intelectual.

Ingeniero en Software
Alan Josué Serrano León

Dedico este trabajo a:

A mis queridos padres porque me enseñaron desde el día que nací a caer pero que siempre hay que volver a levantarse, les agradezco por mostrarme el camino de ignorar el dolor, a lo largo de mi vida esto me hizo ser más fuerte y comprender el sufrimiento. Doy gracias a esto porque esta etapa de mi vida me enseñó bastante y no habría podido seguir sin su apoyo espiritual ni sus enseñanzas.

Agradecimientos

Quisiera poder agradecer con hechos todo el apoyo y la paciencia que tuvieron conmigo a lo largo de la Maestría la Doctora Yulia, el Doctor René y el Doctor José Luis, no solo como investigadores para el desarrollo de una tesis, más bien como personas por darme la oportunidad de ser alumno de la Maestría en Computación y permitirme permanecer a pesar de las circunstancias.

Resumen

En la actualidad el acceso a la información se da por medio de hipervínculos, los cuales interconectan los textos entre si únicamente si contienen una relación. Varios investigadores han estudiado la forma en que los humanos crean los hipervínculos y han tratado de replicar el modo de trabajo específicamente de la colección de Wikipedia. El uso de hipervínculos se ha pensado como un prometedor recurso para la recuperación de información, que fue inspirado por el análisis de citas de la literatura (Merlino-Santesteban, 2003). Según Dreyfus (Dreyfus, 2003) la hipervinculación no tiene ningún criterio específico, ni tampoco jerarquías. Por ello cuando todo puede vincularse indiscriminadamente y sin obedecer un propósito o significado en particular, el tamaño de la red y la arbitrariedad entre sus hipervínculos, hacen extremadamente difícil para un usuario encontrar exactamente el tipo de información que busca. En las organizaciones, la familiaridad y la confianza durante mucho tiempo han sido identificadas como las dimensiones de credibilidad de la fuente de información en publicidad (Eric Haley, 1996). Un hipervínculo, como una forma de información, puede, por lo tanto, tener un mayor impacto cuando se presenta por un objetivo conocido (Stewart & Zhang, 2003). Mientras tanto, los hipervínculos entre los sitios web pueden generan confianza en el remitente y el receptor del enlace, por lo que estas interacciones tienen efectos positivos de reputación para el destinatario (Stewart, 2006) (Lee, Lee, & Hwang, 2014).

El estudio de documentos por medio de los hipervínculos es un área importante de investigación en minería de datos, en una red social a menudo lleva una gran cantidad de información estructural formada por los hipervínculos creando nodos compartidos dentro de la comunidad. Algunas importantes aplicaciones de los métodos de minería de datos para redes sociales son la recomendación social mediante las experiencias similares de los usuarios (Alhajj & Rokne, 2014). En marketing y publicidad se aprovechan las cascadas en las redes sociales y se obtienen beneficios sobre modelos de propagación de la información (Domingos & Richardson, 2001). Las empresas de publicidad están interesados en cuantificar el valor de un solo nodo en la red, tomando en cuenta que sus acciones pueden desencadenar cascadas a sus nodos vecinos. Los resultados de (Allan, 1997) (Bellot et al., 2013) (Agosti, Crestani, & Melucci, 1997) (Blustein, Webber, & Tague-Sutcliffe, 1997) sugieren que el descubrimiento de hipervínculos automatizado no es un problema resuelto y que cualquier evaluación de los sistemas de descubrimiento de Hipervínculos de Wikipedia debe basarse en la evaluación manual, no en los hipervínculos existentes.

Contenido

Página

Lista de figuras

Lista de tablas

CAPÍTULO 1.

Introducción

El acceso a la información se ha transformado por el uso de nuevas tecnologías de información y comunicación ocasionando modificaciones en la forma de acceder a la información. Por lo que es posible tener nuevas formas de recopilar y analizar datos (López & Gómez, 2006) (Ananiadou, Friedman, & Tsujii, 2004). El presente y futuro de la información con el uso masivo de computadoras da lugar a la aparición del ciberespacio como un rastro de la presencia humana en Wikis, Blogs, Chats, y Redes Sociales. El acceso a la información se da por medio de hipervínculos, los cuales interconectan los textos entre si únicamente si contienen una relación. De esta manera se crean múltiples conexiones para que los usuarios puedan navegar de acuerdo al tema de interés y crear su propia secuencia de exploración. Básicamente un hipervínculo se refiere a la navegación entre fragmentos de información textual, sonora, gráfica y audiovisual (López & Gómez, 2006). Según Dreyfus (Dreyfus, 2003), la hipervinculación no tiene ningún criterio específico ni jerarquías. Por ello, cuando todo puede vincularse indiscriminadamente y sin obedecer un propósito o significado en particular, el tamaño de la red y la arbitrariedad entre sus hipervínculos, hacen extremadamente difícil para un usuario encontrar exactamente el tipo de información que busca. Esta problemática acerca de los hipervínculos se encuentra presente en la Docencia por la gestión de documentos que se pretende incorporar en un modelo basado en hipervínculos para administrar las guías docentes, temarios, normas, etc. Ya que constituye un recurso muy importante en la educación superior, con la puesta en marcha del Espacio Europeo de Educación Superior

(EEES) que resalta la necesidad de fuentes de información docente (Ocaña & García, 2012). En las organizaciones, la familiaridad y la confianza durante mucho tiempo han sido identificadas como las dimensiones de credibilidad de la fuente de información en publicidad (Eric Haley, 1996). Un hipervínculo, como una forma de información, puede tener un mayor impacto cuando se presenta por un objetivo conocido (Stewart & Zhang, 2003). Mientras tanto, los hipervínculos entre los sitios web pueden generan confianza en el remitente y el receptor del enlace, por lo que estas interacciones tienen efectos positivos de reputación para el destinatario (Stewart, 2006) (Lee, Lee, & Hwang, 2014).

Por ejemplo, el motor de búsqueda *Google* tiene un sistema llamado *PageRank* que ayuda a determinar la relevancia de una página web al momento de indexar, se basa en la estructura de los hipervínculos como el indicador del valor de una página web. El recuento de los hipervínculos entrantes a una página determinada, hace una aproximación de la importancia o calidad de dicha página. *PageRank* adopta esta idea y normaliza el número de los hipervínculos en una página. Básicamente el cálculo inicial de *PageRank* interpreta un hipervínculo de una página *A* a una página *B* como un voto, de la página *A*, para la página *B*. Por lo tanto, *PageRank* calcula la relevancia de una página y muestra su importancia en internet. Según Brin y Page (Brin & Page, 1998) mencionan que la probabilidad de que un usuario navegue por medio de hipervínculos en internet es de 0.25 en lugar de que escriba la URL en el navegador directamente.

Una *Wiki* muy reconocida por su colección de documentos en hipertexto es *Wikipedia*, una enciclopedia digital disponible en Internet, con más de 37 millones de artículos escritos por voluntarios en más de 287 idiomas (Wikipedia, 2015a). Estas publicaciones son de libre acceso, y las personas pueden editar los artículos incluyendo lenguas indígenas como náhuatl y maya o lenguas muertas, como el latín, el chino clásico o el anglosajón. Su crecimiento ha sido tan espectacular que al día de hoy es la más grande y popular obra de referencia en internet. La edición en español se inició el 20 de mayo de 2001 y actualmente cuenta con 1, 216, 036 artículos. De acuerdo con los últimos análisis de Wikipedia en español hay 4, 016, 675 usuarios, de los cuales se encuentran activos 16, 730 (Wikipedia, 2015b).

Wikipedia abarca varias temáticas y la cantidad de documentos en hipertexto disponibles es tomada como referencia para el estudio de la tarea de hipervinculación (Huang, Xu, Trotman, & Geva, 2007). Existen otros sitios con hipervínculos (por ejemplo UNO noticias) donde los eventos relevantes que ocurren en el mundo deben producir noticias para informar sobre este tipo de eventos y su proceso es muy dinámico debido a que el número que noticias que se generan diariamente en el mundo. Por lo tanto, la cantidad de hipervínculos que se tienen que crear para poder distribuirse a los usuarios a través de internet incrementa exponencialmente.

El procesamiento del lenguaje natural (PLN) es una rama de la inteligencia artificial que se origina para comprender el lenguaje humano. El PLN pertenece a la intersección de la

lingüística aplicada y las ciencias de la computación, y estudia los métodos para que una computadora realice tareas relacionadas con el lenguaje humano y tenga un cierto grado de entendimiento del contenido (Gelbukh, 2017) (Covington, 1994).

El PLN se encuentra en constante desarrollo y se aplica en múltiples actividades como la traducción automática, resúmenes y en los sistemas de recuperación de información, etc. Aún existen múltiples obstáculos para el PLN pero han tenido un gran avance en aplicaciones informáticas para hacer más flexible el acceso y almacenamiento de la información (Sosa, 1997). Desde un enfoque computacional como lingüístico se utilizan distintas técnicas de inteligencia artificial para cumplir diferentes tareas y poder modelar el conocimiento.

Varios investigadores han estudiado la forma en que los humanos crean los hipervínculos y han tratado de replicar el modo de trabajo específicamente de la colección de Wikipedia, han separado en dos categorías como reconocer palabras que puedan ser útiles para crear hipervínculos como lo son las entidades nombradas que se clasifican en nombres de personas, organizaciones, ubicaciones y expresiones numéricas incluyendo hora, fecha, dinero y porcentajes (Sil, 2013a) (Bunescu, 2006) (Cucerzan, 2007) (Varma et al., 2009) (Rao, McNamee, & Dredze, 2013). Estas palabras proporcionan información general sobre el contenido de un documento (Bordea & Buitelaar, 2010) (Treeratpituk, Teregowda, Huang, & Giles, 2010) (Mihalcea, 2007) (Camacho, 2015). Tales frases clave constituyen una descripción de un documento que pueden ser utilizadas en particular para la recuperación de información, el agrupamiento automático de documentos, clasificación e hipervinculación de documentos (Lopez & Romary, 2010).

Otros autores han reunido esfuerzos para desambiguar entidades y frases clave debido a que existen términos que pueden tener varios significados dependiendo del contexto en el que se encuentran (Morgan & Keulen, 2013) (Hakimov & Oto, 2012) (Navigli, 2009; Navigli, 2012). La desambiguación es una tarea importante en la recuperación de información y PLN (Morgan & Keulen, 2013), por lo que se tiene que identificar con qué sentido una palabra está siendo usada dentro de una oración. Para esto, debemos comprender que la ambigüedad ocurre cuando una palabra suele tener varios significados y presenta confusión al lector (Ramos, 2006). Autores como Mihalcea (Mihalcea, 2007), Bunescu (Bunescu, 2006) y Sil (Sil, 2013b) no hacen una desambiguación como tal para la hipervinculación de documentos de Wikipedia, estos autores calculan la similitud de una ventana que contiene cierta cantidad de palabras y los documentos obtenidos de una búsqueda para crear un hipervínculo. Las áreas de investigación NER, frases clave y desambiguación son muy dependientes y no existen muchas obras que examinen esta dependencia (Morgan & Keulen, 2013), pero en trabajos como (Jin, Kıcıman, Wang, & Loynd, 2014) (Moro & Navigli, 2015) (Sidorov, 2014) mencionan que para analizar y descubrir conocimiento se requiere la unificación de varios enfoques para hacer frente a una tarea determinada.

En 2007, el foro INEX estudió la calidad de hipervínculos generados automáticamente ya que nunca habían sido cuantificados. En el trabajo de Camacho (Camacho, 2015) se hizo un estudio, donde se menciona que existen 49 hipervínculos por página en Wikipedia (2008). Por otro lado, en el foro INEX se señala que cada documento tiene 5.5 hipervínculos en promedio (Wikipedia 2006). La diferencia entre un estudio y otro es debido a que los hipervínculos que se tomaron en el foro INEX para ser cuantificados fueron únicamente los hipervínculos que tenían enlaces dentro del mismo corpus, utilizando 114, 336 documentos de Wikipedia.

El uso de hipervínculos se ha pensado como un prometedor recurso para la recuperación de información, que fue inspirado por el análisis de citas de la literatura (Merlino-Santesteban, 2003). Para comprender que es un hipervínculo la definición del libro HTML5 (Pilgrim, 2010) nos dice que es un enlace asociado a un elemento de un documento que apunta a otro documento con hipertexto. El Hipertexto es un texto electrónico compuesto por bloques de palabras (hipervínculo) que unen a otros textos. Por lo tanto, se manifiesta en múltiples conexiones entre los diferentes documentos, los hipervínculos no tienen un principio y un fin; al mismo tiempo que ninguno puede ser visto como principal y la cantidad no tiene límite porque tiene de base la infinidad del lenguaje.

La gran mayoría de los documentos académicos permanecen ocultos en el caso de los clásicos (textos impresos) y es difícil de seguir físicamente lejos de sus referencias. Mientras que el hipertexto, facilita el seguimiento de las referencias individuales y la navegación entre las lecturas. El lector tiene que ser muy activo ya que tiene total libertad y se encuentra sumergido en un abundante enriquecimiento de conocimiento (Landow, 1995). Lo anterior es la definición de hipervínculo, pero no se ha estudiado ¿Cómo está conformado un hipervínculo? A continuación se muestra una extensión de la definición con el análisis de 16, 033 documentos de Wikipedia 2008 en español. El total de hipervínculos son 1, 196, 686 los cuales están conformados con diferentes longitudes de términos. En la tabla 1.1 están las longitudes y el número total de hipervínculos de cada tamaño.

Congreso[1] de[2] los[3] Diputados[4]

Hipervínculo tamaño 4

Tamaño	Número
1	766,178
2	180,281
3	157,708
4	43,645
5	16,763
6	22,564
7	3,352
8	2,005
9	1,253
10	830
11	599
12	384
13	292
14	173
15	167
16	116
17	81
18	69
19	50
20	37

Tamaño	Número
21	33
22	20
23	20
24	6
25	13
26	13
27	2
28	4
29	4
30	1
31	1
32	1
33	1
34	3
35	3
36	2
37	1
38	1
39	2
40	0

41	0
42	0
43	0
44	0
45	0
46	2
47	1
48	1
49	0
50	0
Σ	1,196,682

Tabla 1.1 Número total de hipervínculos por longitud.

En la figura 1.1 se muestra el total de términos (2, 082, 529) que forman los hipervínculos en los documentos, de los cuales 1, 319, 133 son entidades nombradas; 1, 794, 600 son el resto de los términos (frases clave) y por último 287, 929 palabras vacías (*stopwords*).

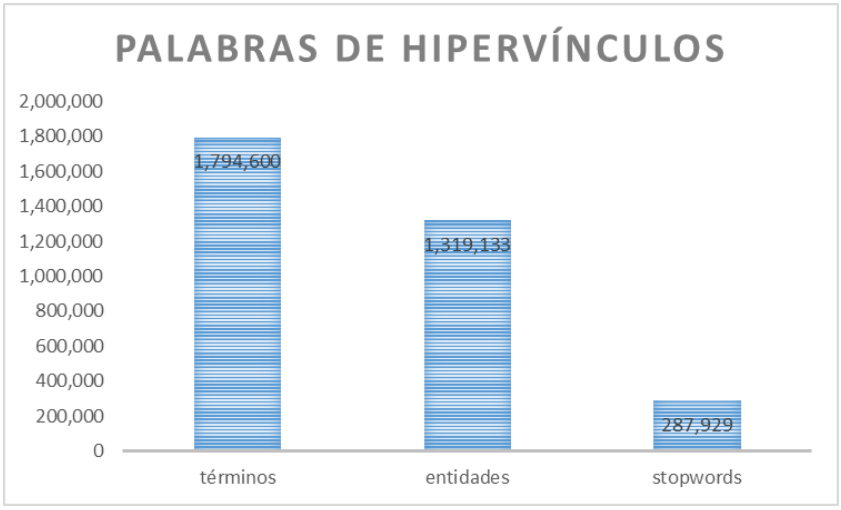

Figura 1.1 Palabras que forman un hipervínculo.

En la figura 1.2 se muestra la clasificación de 1, 319, 133 entidades con el *Reconocedor de entidades nombradas* (por sus siglas en inglés, NER) de Stanford. La clasificación consta de 4 categorías (persona, organización, lugar, otros).

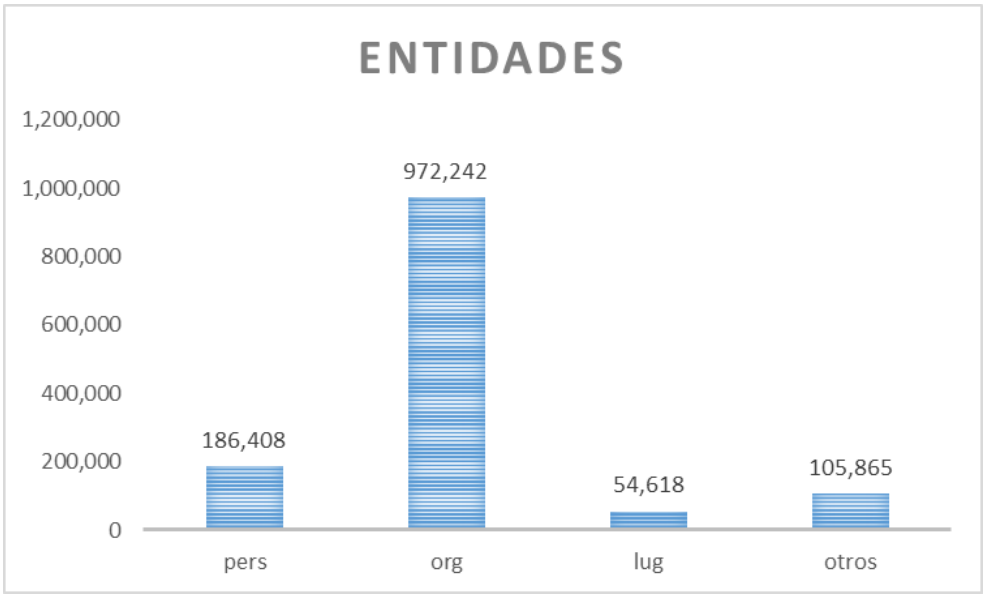

Figura 1.2 Clasificación de entidades.

Por último, y particularmente en la colección de Wikipedia, tenemos que el mayor uso de hipervínculos en un documento se encuentran principalmente en las primeras 27 palabras y disminuye el uso de hipervínculos con el paso de los términos en el resto del documento. En la figura 1.3 se puede observar la gráfica de la distribución de los hipervínculos en los documentos de Wikipedia.

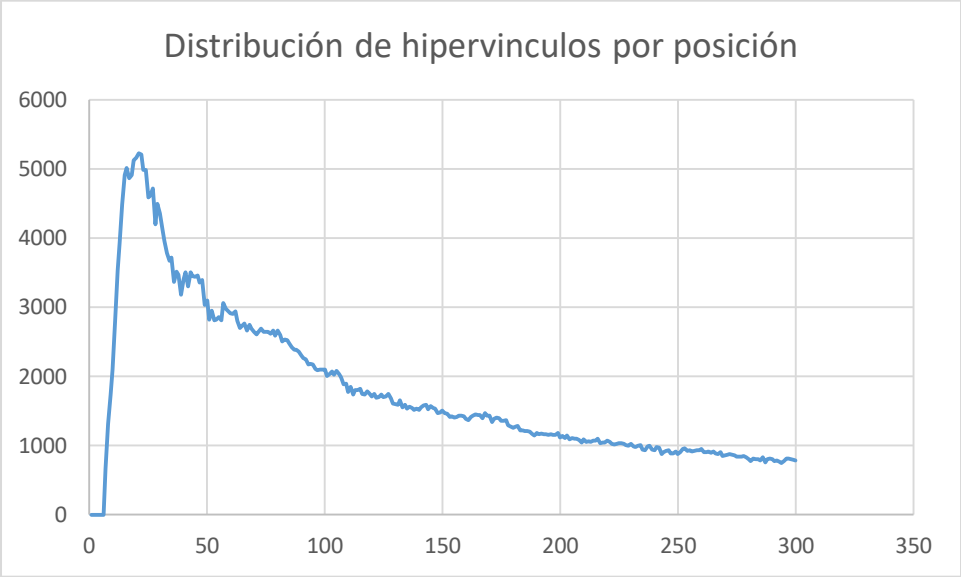

Figura 1.3 Distribución de hipervínculos en los documentos HTML.

1.1 Planteamiento del problema

La hipervinculación de documentos consta de las etapas de selección de términos para crear un hipervínculo, la búsqueda de los documentos, y por último el redireccionamiento a una página. Se sabe que los hipervínculos están conformados de entidades nombradas de manera conjunta con otros términos que pueden ser frases clave. Sin embargo, en el estado del arte se han tomado por separado las técnicas. En esta investigación se aborda el problema de hipervinculación unificando las dos tareas para determinar los términos correctos que representen un texto de un hipervínculo. Por lo tanto, el problema queda planteado de la siguiente manera:

¿Cómo clasificar las frases de un texto que representen hipervínculos y seleccionar los documentos destinos sin emplear diccionarios de desambiguación?

1.2 Justificación o motivación

El estudio de documentos por medio de los hipervínculos es un área importante de investigación en minería de datos. En una red social a menudo lleva una gran cantidad de información estructural formada por los hipervínculos creando nodos compartidos dentro de la comunidad. Algunas importantes aplicaciones de los métodos de minería de datos para redes sociales son la recomendación social mediante las experiencias similares de los usuarios (Alhajj & Rokne, 2014). En *marketing* y publicidad se aprovechan las cascadas en las redes

sociales y se obtienen beneficios sobre modelos de propagación de la información (Domingos & Richardson, 2001). Las empresas de publicidad están interesados en cuantificar el valor de un solo nodo en la red, tomando en cuenta que sus acciones pueden desencadenar cascadas a sus nodos vecinos. Los resultados de (Allan, 1997) (Bellot et al., 2013) (Agosti, Crestani, & Melucci, 1997) (Blustein, Webber, & Tague-Sutcliffe, 1997) sugieren que el descubrimiento de hipervínculos automatizado no es un problema resuelto y que cualquier evaluación de los sistemas de descubrimiento de Hipervínculos de Wikipedia debe basarse en la evaluación manual, no en los hipervínculos existentes.

1.3 Objetivos

General:

Construcción de un modelo de clasificación para determinar los términos que pueden ser delimitados en su contexto para formar una frase que represente un hipervínculo.

Particulares:

- Extraer las características en los documentos de Wikipedia.
- Construir un modelo de clasificación.
- Preparar los datos para el aprendizaje automático supervisado.
- Seleccionar la cantidad mínima necesaria de atributos.
- Evaluar la clasificación.
- Seleccionar el contexto cercano de cada hipervínculo en los documentos.
- Probar medidas de similitud para desambiguar el documento destino de un hipervínculo.
- Hipervincular las frases con los documentos de Wikipedia.
- Realizar la evaluación de los hipertextos resultantes con humanos.

1.4 Hipótesis

Dado que la desambiguación de un término es a través de su contexto, entonces un clasificador puede aprender el contexto que define un hipervínculo sin emplear diccionarios lingüísticos.

1.5 Metodología

El reconocimiento de patrones es la disciplina científica cuyo objetivo es la clasificación de objetos en una serie de categorías o clases. Dependiendo de la aplicación, estos objetos pueden ser cualquier tipo de mediciones que necesitan para ser clasificados. El reconocimiento de patrones antes de la década de 1960 era sobre todo el resultado de la investigación teórica en el ámbito de las estadísticas. Sin embargo, a medida de que la sociedad evoluciona, la automatización y la recuperación de información es cada vez más importante (Theodoridis & Koutroumbas, 2003). El reconocimiento de patrones estadístico se utiliza para cubrir todas las etapas de una investigación a partir de la formulación del problema y recolección de datos a través de la extracción de características, discriminación, clasificación, evaluación de resultados y la interpretación (Webb, 2004).

Etapas de reconocimiento de patrones:

1. Formulación del problema. La obtención de una comprensión clara de los objetivos de la investigación y la planificación de las etapas restantes.

2. Recolección de datos. Realizar mediciones sobre las variables apropiadas y registrar los detalles del procedimiento de recolección de datos.

3. Examen inicial de los datos. Comprobación de los datos, el resumen del cálculo de las estadísticas con el fin de tener una idea de la estructura.

4. Selección o extracción de características. Es el proceso de análisis de datos exploratorio para elegir características que puedan ser usadas en la clasificación de los datos, estas características deben ser apropiadas para la tarea.

5. Aplicar procedimientos de discriminación. El clasificador está diseñado utilizando un conjunto de entrenamiento de los patrones ejemplares. Por lo tanto, hay que aplicar una medida de discriminación y seleccionar los atributos más relevantes.

6. Clasificación supervisada. Basándose en las características extraídas, se aplican métodos de aprendizaje automático según corresponda.

7. Evaluación de los resultados. Esto implica aplicar el clasificador entrenado a una prueba independiente del conjunto de patrones marcados.

8. Interpretación.- Proporcionar una interpretación de la salida del diseño de solución.

1.6 Estructura de la tesis

En el capítulo 2 se describen los conceptos teóricos sobre el objeto de estudio que van a ser administrados durante el desarrollo y las pruebas de la experimentación. También, se listan las diferentes técnicas utilizadas comúnmente para el reconocimiento de patrones, además se describen las técnicas y algoritmos de clasificación, así como el análisis y evaluación de hipervínculos.

El capítulo 3 corresponde al estado del arte, donde se describen los sistemas de otras investigaciones sobre el tema de investigación, las tareas reconocimiento de entidades nombradas y frases clave fueron tomados como base para el desarrollo de la propuesta.

En el capítulo 4 se aborda el problema de hipervinculación unificando las tareas de NER y frases clave. Se describe el método propuesto desde el entrenamiento, la creación de un modelo y la validación de la clasificación de los candidatos para crear un hipervínculo. También se menciona el proceso para la búsqueda de documentos y la hipervinculación.

En el capítulo 5 se presentan los experimentos para comprobar la hipótesis planteada de esta investigación y la evaluación del usuario con el hipertexto.

En el capítulo 6 se presentan las conclusiones de esta investigación y el trabajo futuro.

CAPÍTULO 2.

Marco Teórico

En este capítulo analizamos los conceptos sobre el objeto de estudio. Se describe brevemente el soporte teórico de la problemática de investigación, sin embargo, se analizaron las teorías globales del tema de estudio. El primer punto es el modelo básico de reconocimiento de patrones, luego se describen las técnicas y algoritmos de clasificación, así como el análisis y evaluación de hipervínculos.

2.1 Reconocimiento de patrones

2.1.1 Modelo básico

Se utiliza el término patrón para denotar la *p*-dimensionalidad del vector de datos $x = (x_1, \dots, x_p)^T$ de las mediciones (T denota transposición del vector), cuyos componentes son x_i mediciones de las características de un objeto. Así, las características son las variables especificadas por el investigador y que se consideran importantes para la clasificación. En la discriminación, suponemos que existen grupos o clases C, denotado $w_1, \dots, w_c$, y se asocia con cada patrón de x a una categórica z que indica la pertenencia a una clase o grupo; es decir, si $z = i$, entonces el patrón pertenece a $w_i, i \in \{1, \dots, C\}$.

Ejemplos de patrones son mediciones de una forma de onda acústica en un reconocimiento de voz; mediciones en un paciente hecho con el fin de identificar una enfermedad

(diagnóstico); mediciones en pacientes con el fin de predecir el resultado probable (pronóstico); mediciones sobre las variables meteorológicas (para el pronóstico o predicción); y una imagen digitalizada de reconocimiento de caracteres. Por lo tanto, vemos que el término "patrón", en su sentido técnico, no se refiere necesariamente a la estructura dentro de las imágenes.

Suponemos que tenemos un conjunto de patrones de la clase conocida; $\{(x_i, z_i), i = 1, \dots, n\}$ (el entrenamiento o conjunto de diseño) que utilizamos para diseñar el clasificador (para configurar sus parámetros internos). Una vez que esto se ha hecho. Se puede estimar la pertenencia a una clase de un patrón desconocido x.

La forma derivada para el clasificador de patrones depende de diferentes factores, como la distribución de los datos de entrenamiento o las hipótesis formuladas en relación con su distribución. Otro factor importante es el coste de clasificación errónea. En muchas aplicaciones, el coste de clasificación errónea son difíciles de cuantificar, ya que son combinaciones de varias contribuciones, tales como los costes monetarios, tiempo y otros costes subjetivos (Webb, 2004).

En la clasificación, los datos pueden someterse a varias etapas de transformación. Estas transformaciones las conocemos mejor como pre-procesamiento que va desde la selección o extracción de características. La principal función de realizar pre-procesamiento en general es para reducir la dimensión, elimina información redundante o irrelevante. Este procedimiento recibe el nombre de dimensionalidad intrínseca donde se utiliza únicamente el número mínimo de variables requeridas para obtener una estructura en los datos. En la figura 2.1 se simplifica el procedimiento de la clasificación de patrones (Theodoridis & Koutroumbas, 2003).

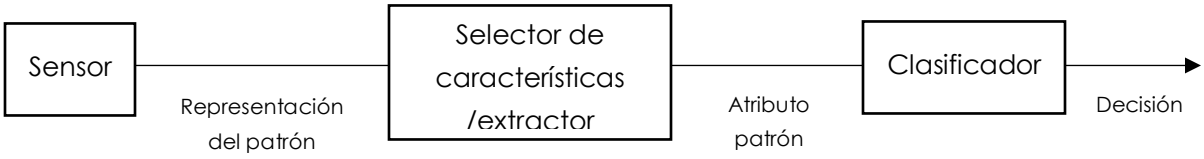

Figura 2.1 Clasificador de patrones.

Al igual que la selección de características, cuando se busca el clasificador a implementar, nos enfrentamos a una serie de algoritmos y resulta difícil decidir cuál es mejor de manera a priori. Por supuesto, algunos algoritmos pueden ser preferidos debido a su menor complejidad computacional y su fácil implementación. Sin embargo, existen problemas de clasificación en donde al implementar distintos algoritmos tienen resultados equivalentes dadas las clasificaciones con el conjunto de entrenamiento.

2.1.2 Método supervisado

En la clasificación no supervisada, simplemente referida a veces como clasificación o agrupamiento, un experto proporciona una etiqueta de categoría o el coste para cada patrón en un conjunto de entrenamiento, y se busca reducir la suma de los costos de estos patrones (Webb, 2004) (Duda, Hart, & Stork, 2000). ¿Cómo podemos estar seguros de que un algoritmo de aprendizaje en particular es lo suficientemente potente como para aprender la solución a un problema dado y que sea estable a las variaciones de parámetros?
Un conjunto de datos de entrenamiento están disponibles y el clasificador fue diseñado por la exploración de la información a priori que se conocía. A esta técnica se le conoce como reconocimiento de patrones supervisado. Sin embargo, hay otros tipos de tareas de reconocimiento de patrones donde los datos de entrenamiento y etiquetas de clase conocidas no están disponibles (Theodoridis & Koutroumbas, 2003).
En la clasificación supervisada tenemos un conjunto de muestras de datos (cada uno consistente en mediciones en un conjunto de variables) con etiquetas asociadas, los tipos de clase. Estos se utilizan como ejemplos en el diseño del clasificador.

2.1.3 Método no supervisado

En la clasificación no supervisada, los datos no están etiquetados y buscamos grupos en los datos y las características que distinguen a un grupo de otro. Se puede aplicar un esquema de agrupamiento a los datos para cada clase por separado y tener muestras representativas para cada grupo dentro de la clase, utilizado como prototipos para esa clase (Webb, 2004). Los beneficios que se pretenden obtener del aprendizaje no supervisado es por el deseo de que la máquina puede mejorar su rendimiento sin ningún tipo de supervisión exterior después de realizar el aprendizaje inicial (Fukunaga, 2009).

2.1.4 Aprendizaje automático

El aprendizaje se obtiene en general mediante la observación y la experimentación. Los procesos de aprendizaje incluyen la adquisición de nuevos conocimientos a través de instrucciones o la práctica. El aprendizaje tiene un fenómeno de múltiples facetas y desde el inicio de la era de la informática se ha hecho un esfuerzo para implantar dicha capacidad en un ordenador. La solución a este problema desafiante sigue siendo estudiada por la inteligencia artificial (Carbonell, Michalski, & Mitchell, 1983).
El uso de máquinas de aprendizaje en la última década se ha extendido en informática. El aprendizaje automático se utiliza en búsquedas en Internet, los filtros de *spam*, los sistemas de recomendación, la colocación de anuncios, la puntuación de crédito, detección de fraudes,

el comercio de acciones, diseño de fármacos y muchas otras aplicaciones. Sin embargo, gran parte del conocimiento que se necesita para desarrollar con éxito las máquinas de aprendizaje automático no están fácilmente disponibles, como resultado proyectos toman mucho más tiempo de lo necesario (Domingos, 2012).

Existen muchos tipos de máquinas de aprendizaje, pero para fines de simplificación se centrará en el más ampliamente utilizado que es la clasificación en el tema de aprendizaje automático. Un clasificador es un sistema que introduce (típicamente) un vector de valores discretos o de funciones continuas y da salida a un único valor discreto llamado clase (Domingos, 2012). Por ejemplo, un filtro de *spam* clasifica los mensajes de correo electrónico en "*spam*" o "correo deseado" y su entrada puede ser un vector booleano $x = (x_1, \ldots, x_j, \ldots, x_d)$, donde $x_j = 1$ si la palabra de orden j aparece en el diccionario del correo electrónico y $x_j = 0$ en caso contrario. Un alumno introduce un conjunto de ejemplos de entrenamiento (x_i, y_i), donde $x_i = (x_i, 1, \ldots, x_i, d)$ es una entrada y y_i es la salida correspondiente, y da salida a un clasificador. La prueba de que el aprendizaje en este clasificador produce la salida correcta y_t para futuros ejemplos x_t (por ejemplo, si el filtro de *spam* clasifica correctamente mensajes de correo electrónico nunca antes vistas como correo deseado o *spam*) (Domingos, 2012).

APRENDIZAJE = REPRESENTEACIÓN + EVALUACIÓN + OPTIMIZACIÓN

Existe una variedad de algoritmos de aprendizaje disponibles y es desconcertante cual se debe usar. Por lo tanto, para no perderse en este espacio y tomar esta decisión es seguir los siguientes tres pasos para no perderse:

Representación:

El clasificador se debe representar en algún lenguaje formal que la computadora pueda manejar. El conjunto de clasificadores que posiblemente pueden aprender se le llama espacio de hipótesis de aprendizaje. Otra cuestión relacionada es la forma de representar la entrada, es decir, cuáles son las características que se van a usar.

Evaluación:

Se necesita tener una función objetivo o función de puntuación para distinguir el resultado de la clasificación.

Optimización:

Se necesita un método para encontrar el mejor clasificador con la mayor puntuación. La elección de la técnica de optimización es muy importante debido a que la eficiencia del aprendizaje determina el resultado final del clasificador. Para tener un mejor aprendizaje y ser optimizado se utilizan diseños más personalizados.

2.2 Clasificación

La clasificación es muy general y tiene muchas aplicaciones, por ejemplo para separar imágenes en clases como paisaje, retrato y ninguna de las anteriores. Sin embargo, en esta tesis nos enfocamos en la Recuperación de Información (Manning, Raghavan, & Schütze, 2008).

Las clases generalmente se denominan temas, y la tarea de clasificación se denomina clasificación de texto o categorización de texto. Para entender la generalidad y el alcance del espacio del problema, presentamos un ejemplo sencillo de un problema clasificación. Se pretende determinar a qué clase(s) pertenece un objeto, dado un conjunto de clases. Tenemos un conjunto de artículos, los cuales tenemos que separar en: documentos donde el tema primordial sean computadoras y documentos donde no se escriba acerca de computadoras. Nos referimos a esto como una clasificación de dos clases, en algunos casos las clases pueden estar estrechamente una de otra. Por lo tanto, a continuación se describen algoritmos utilizados para la clasificación de texto.

2.2.1Algoritmos de clasificación

En los siguientes puntos se describe brevemente algunos algoritmos de clasificación como: Naive bayes bayesian, K-vecinos más cercanos, Máquinas de soporte vectorial, Árboles de decisión Id3 y C4.5.

2.2.1.1 Naive bayes bayesian

La utilización de reglas de Bayes, puede extender la idea para combinar varios clasificadores (Webb, 2004). El enfoque Bayesiano es utilizado para los sistemas de información y se han aplicado desde hace tiempo, se basan sobre las teorías del razonamiento probatorio (sacando conclusiones de la evidencia) (Kowalski, 1997). El enfoque bayesiano podría aplicarse como parte de la ponderación de los términos de un índice, sin embargo, generalmente se aplica como parte del proceso de recuperación de información al calcular la similitud entre un artículo y una consulta específica.

Un ejemplo de aprendizaje Bayesiano donde se busca estimar alguna cantidad como la densidad en x

$$P(x|D)$$

Donde $D = \{x_1, \ldots, x_n\}$ es el conjunto de patrones de entrenamiento de la distribución. La dependencia de la densidad de x en D es a través de los parámetros del modelo asumido para la densidad. Si suponemos un modelo en particular, $p(x|\theta)$, entonces el enfoque bayesiano no basa la estimación de la densidad en una sola estimación de los parámetros, θ, de la probabilidad función de densidad $p(x|\theta)$, pero admite que no sabemos el verdadero valor de θ y escribimos (Webb, 2004).

$$p(x|D = \int p(x|\theta)\, p(\theta|D)\, d\theta$$

Donde por el teorema de Bayes la densidad posterior de θ se puede expresar como:

$$p(\theta|D) = \frac{p(D|\theta)\, p(\theta)}{\int p(D|\theta)\; p(\theta)\, d\theta}$$

El teorema de Bayes permite combinar ante cualquier $p(\theta)$ con alguna probabilidad, $p\,(D|\theta)$ para dar la parte posterior. Para un determinado modelo $p(x|\theta)$ la familia de distribuciones previas para los cuales la densidad posterior $p(\theta|D)$ es de la misma forma funcional y se denomina conjugado con respecto a $p(x|\theta)$.

2.2.1.2 K-vecinos más cercanos

La regla k-vecino más cercano se clasifica con x asignándole la etiqueta con más frecuencia entre las muestras k más cercanas; en otras palabras, se toma una decisión al examinar las etiquetas de los k vecinos más próximos y tomar una votación. Sin embargo, teniendo en cuenta el caso de dos clases con k impar (para evitar lazos), se puede obtener alguna penetración adicional en estos procedimientos (Duda et al., 2000).

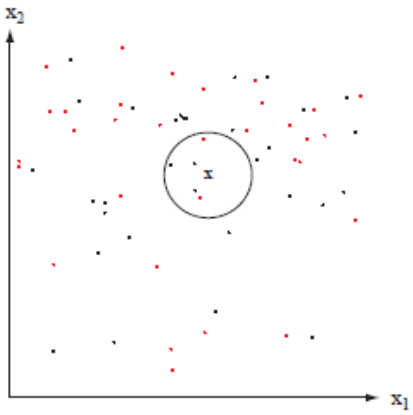

Figura 2.2 Región k-vecino más cercano (Duda et al., 2000).

En la figura 2.2 la consulta para k-vecino más cercano se inicia en el punto de prueba y crece una región esférica hasta que encierre muestras de entrenamiento k, se etiqueta el punto de prueba por un voto de la mayoría de estas muestras. En este caso $k = 5$, y el punto de prueba es x.

La fase de entrenamiento del algoritmo consiste en almacenar los vectores característicos y las etiquetas de clase de los ejemplos de entrenamiento. En la fase de clasificación, la evaluación del ejemplo (cuya clase es desconocida) está representada por un vector en el espacio característico. La distancia entre los vectores almacenados y el nuevo vector se calculan y se seleccionan los k ejemplos más cercanos; el nuevo ejemplo se clasifica con la clase más repetida en los vectores seleccionados.

2.2.1.3 Máquinas de soporte vectorial

Una máquina de soporte vectorial (SVM en inglés) calcula directamente el hiperplano de separación que maximiza el margen o la distancia al punto de ejemplo más cercano. Varios puntos estarán a la misma distancia; estos puntos se conocen como los vectores de soporte y el clasificador resultante es una combinación lineal de estos vectores (Büttcher, Clarke, & Cormack, 2010).

SVM depende de procesamiento previo de los datos para representar patrones en una dimensión alta, típicamente mucho más alta que el espacio de características originales (Duda et al., 2000). Con un mapeo no lineal apropiado a una dimensión suficientemente alta, los datos de dos categorías siempre se pueden separar por un hiperplano. Se asume que cada patrón x_k se ha transformado en $y_k = \varphi(x_k)$.

El objetivo en la formación de una máquina de soporte vectorial es encontrar el hiperplano de separación con el margen más grande. Se espera que cuanto mayor sea el margen, mejor es la generalización del clasificador. Como se ilustra en la Fig. 2.2 la distancia desde un hiperplano a un patrón (transformado) es $|g(\boldsymbol{y})|/||\boldsymbol{a}||$, y suponiendo que un margen positivo b existe, entonces la ecuación implica:

$$\frac{z_k g(\boldsymbol{y}_k)}{||\boldsymbol{a}||} \geq b \quad k = 1, \ldots, n;$$

El objetivo es encontrar el vector de pesos de a que maximiza a b. Por supuesto, el vector solución puede escalar de forma arbitraria y aún conservar el hiperplano, por lo tanto para asegurar la singularidad imponemos la restricción $b\,||\boldsymbol{a}|| = 1$; es decir, exigimos la solución a las ecuaciones 1 y 2 también minimizar $||\boldsymbol{a}||^2$.

Los vectores de soporte son las muestras de entrenamiento que definen la separación del hiperplano óptimo como se muestra en la figura 2.3 (Duda et al., 2000). Es decir, son los patrones con más información para la tarea de clasificación. Dado un conjunto de ejemplos (muestra de entrenamiento) podemos etiquetar las clases para construir un modelo. Este modelo puede predecir la clase de una nueva muestra (muestra de validación).

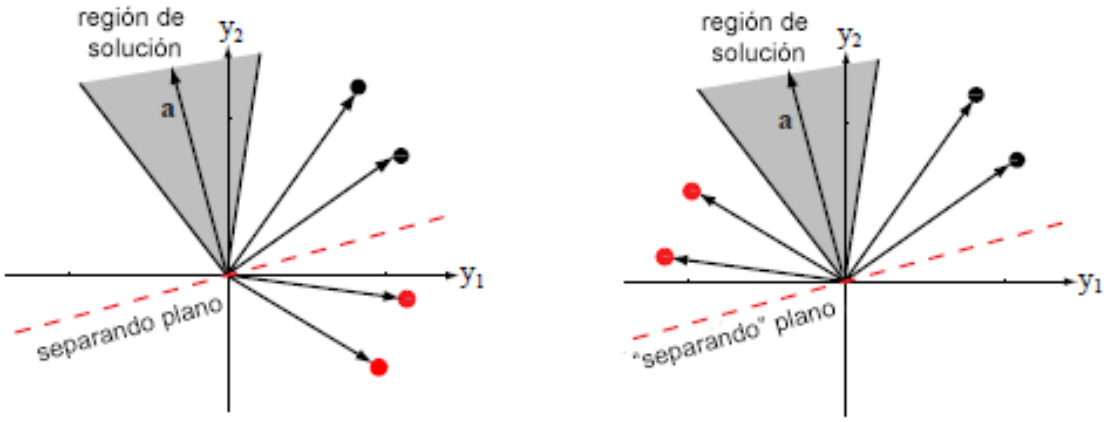

Figura 2.3 Separación del plano (Duda et al., 2000).

2.2.1.4 Árboles de decisión

La metodología de árbol de decisión es un método de minería de datos comúnmente utilizado para establecer sistemas de clasificación basados en múltiples variables o para desarrollar algoritmos de predicción, segmentación y estratificación (IBM, 2017). Este método clasifica una población en segmentos tipo rama que construyen un árbol invertido con un nodo raíz, nodos internos y nodos hoja. El algoritmo no es paramétrico y puede tratar de manera eficiente conjuntos de datos grandes y complicados sin imponer una estructura paramétrica complicada (Song & Lu, 2015).

Los métodos basados en árboles (o árboles de decisión) son bastante populares en minería de datos, pudiéndose usar para clasificación y regresión. Estos métodos se derivan de una metodología previa denominada detección de interacción automática (Diazaraque, 1998). Son útiles para la exploración inicial de datos y apropiados cuando hay un número elevado de datos, y existe incertidumbre sobre la manera en que las variables explicativas deberían introducirse en el modelo. Sin embargo, los árboles de decisión no constituyen una herramienta demasiado precisa de análisis.

En conjuntos pequeños de datos es poco probable que revelen los árboles de decisión la estructura de ellos, de modo que su mejor aplicación se encuentra en grandes masas de datos donde pueden revelar formas complejas en la estructura que no se pueden detectar con los métodos convencionales de regresión. Cuando el tamaño de la muestra es lo suficientemente grande, los datos del estudio se pueden dividir en conjuntos de datos de capacitación y validación. Usar el conjunto de datos de capacitación para construir un modelo de árbol de decisión y un conjunto de datos de validación para decidir sobre el tamaño de árbol apropiado necesario para lograr el modelo final óptimo.

2.2.1.5 Árbol de decisión Id3

El ID3 se considera un algoritmo de árbol de decisión muy simple (Quinlan, 1986). Se utiliza la técnica de ganancia de información como criterio de división El árbol ID3 deja de crecer cuando todos los casos pertenecen a un solo valor de una entidad de destino o cuando la mejor ganancia de información no es mayor que cero. ID3 no aplica ningún procedimiento de poda ni maneja atributos numéricos o valores faltantes.

ID3 recibió su nombre debido a que fue el tercero de una serie de procedimientos de identificación o "ID". ID3 está diseñado para entradas nominales sin ordenar (Duda et al., 2000). Si el problema implica variables con valores reales, que se han agrupado por primera vez en intervalos, cada intervalo debe ser tratado como un atributo nominal desordenado. Cada división tiene un factor de ramificación B_j, donde B_j es el número de contenedores de atributos discretos de la variable j elegido para la división. Tales árboles tienen su número de niveles igual al número de entrada variables. El algoritmo continúa hasta que todos los nodos son puros o no hay más variables que se dividirán.

2.2.1.6 Árbol de decisión C4.5

El algoritmo C4.5 es una evolución de ID3 el cual utiliza la relación de ganancia como criterio de división (Quinlan, 1993). La división cesa cuando el número de instancias para ser dividido está por debajo de un cierto umbral. La poda basada en error se realiza después de la fase de crecimiento. C4.5 puede manejar atributos numéricos. También puede inducir la formación de un conjunto que incorpora los valores que faltan mediante el uso de criterios del cociente de ganancia corregidos. El algoritmo C4.5 (Duda et al., 2000) utiliza la heurística para la poda basados en la significación estadística de las divisiones.

2.2.2 Extracción de características

Unos de los principales problemas asociados con el reconocimiento de patrones es la dimensionalidad de los datos. En general el número de características disponibles del diseño de clasificación es grande. La necesidad de reducir el número de características a un mínimo óptimo es la complejidad computacional, en algunos casos también puede proporcionar una mejor precisión de clasificación debido a el tamaño finito de la muestra (Zongker & Jain, 1996).

En resumen el procedimiento para la selección de características es dado un número n de características, seleccionar las características más importantes con el fin de reducir su número y al mismo tiempo conservar tanto como sea posible la información de su clase discriminatoria. Si seleccionamos características con poco poder de discriminación, como consecuencia el diseño de un clasificador creado daría lugar a un rendimiento deficiente. Por otro lado, si se seleccionan características ricas en información, el diseño del clasificador se puede simplificar en gran medida. En una descripción más cuantitativa, se debe tener como objetivo mantener gran distancia entre clases y pequeña distancia dentro de la misma clase en el espacio de características del vector. Las formas más básicas de selección es examinar las características de forma individual y descartar las que tienen poca capacidad discriminatoria, otra mejor alternativa consiste en examinarlas en combinaciones (Fukunaga, 2009).

2.2.2.1 Ganancia de información

El aumento de la información se emplea con frecuencia como un criterio de calidad de término en el campo del aprendizaje automático. La ganancia de información mide la cantidad de bits de información obtenida para la predicción de categoría al conocer la presencia o ausencia de un término en un documento (Yang & Pedersen, 1997). Considerando $\{c_i\}\ {}^{m}_{i}= 1$ denota el conjunto de categorías en el espacio objetivo. La ganancia de información del término t se define como:

$$\begin{aligned} G(t) = \quad & \sum {}^{m}_{i}= 1 \Pr(c_i) \log \Pr(c_i) \\ & + \Pr(t) \sum {}^{m}_{i}= 1 \Pr(c_i|t) \log \Pr(c_i|t) \\ & + \Pr(t) \sum {}^{m}_{i}= 1 \Pr(c_i|t) \log \Pr(c_i|t) \end{aligned}$$

Esta definición es más general que la empleada en los modelos de clasificación binarios. Usamos la forma más general porque los problemas de categorización de texto normalmente tienen un espacio de categoría $m - ary$ (donde m puede ser de hasta decenas de miles), y

necesitamos medir la calidad de un término globalmente con respecto a todas las categorías en promedio.

En el corpus de entrenamiento, para cada término calculamos la ganancia de información y eliminamos del espacio de características aquellos términos cuya ganancia de información sea menor que un umbral predeterminado. El cálculo incluye la estimación de las probabilidades condicionales de una categoría dada a un término, y los cálculos de entropía en la definición. La estimación de probabilidad tiene una complejidad temporal de $O\ (N)$ y la complejidad del espacio de $O\ (V\ N)$ donde N es el número de documentos de entrenamiento y V es el tamaño del vocabulario. Los cálculos de entropía tienen una complejidad temporal de $O\ (V\ m)$.

2.2.2.2 Binario

Los enfoques principales para generar consultas en la recuperación de información son el enfoque booleano y en lenguaje natural. Las consultas de lenguaje natural se representan fácilmente dentro de modelos estadísticos y son utilizables por las medidas de similitud. Los problemas surgen cuando las consultas Booleanas están asociadas con sistemas de índices ponderados. Salton (Salton, Fox, & Wu, 1983) demostró que usar la tecnología de recuperación booleana convencional tiene varias desventajas:

1. El tamaño de la salida como respuesta a una consulta dada es difícil de controlar; dependiendo de la frecuencia de los términos en la consulta y las combinaciones de términos reales, una gran cantidad de resultados se puede obtener o no se puede recuperar ningún resultado en absoluto.
2. La salida obtenida como respuesta a una consulta no está clasificada en orden de importancia para el usuario. Cada elemento recuperado es tan importante como cualquier otro artículo recuperado.
3. No hay forma alguna de asignar factores de importancia o pesos a los términos adjuntos a los documentos o la consulta. Por lo tanto, todos los términos incluidos en los documentos y la consulta tienen la misma importancia.
4. Las formulaciones de consulta booleanas pueden producir resultados intuitivos: por ejemplo, en respuesta a una o consulta ("$A\ o\ B\ o\ \ \ldots\ Z$"), un elemento que contiene solo un término en la consulta se considera tan importante como un elemento que contiene todos los términos.

2.2.2.3 Frecuencia del término en una colección de documentos

La frecuencia del documento es la cantidad de documentos en los que aparece un término. Se calcula la frecuencia del documento para cada término en el corpus de entrenamiento y se elimina del espacio de características aquellos términos cuya frecuencia de documento sea menor que un umbral predeterminado. Los términos raros no son informativos para la predicción de categoría o no influyen en el rendimiento global. En cualquier caso, la eliminación de términos raros reduce la dimensionalidad del espacio de características y es posible que mejore la precisión de categorización (Yang & Pedersen, 1997).

2.2.3 Extracción de información

Hay dos procesos asociados con la extracción de información: determinación de hechos para entrar en campos estructurados en una base de datos y extracción de texto que se puede utilizar para resumir un artículo. En el primer caso, sólo un subconjunto de hechos importantes en un artículo pueden ser identificados y son extraídos. En el segundo caso, todos los conceptos principales en el ítem deben estar representados en el resumen.

El proceso de extracción de datos sirve para la construcción automática de archivos para crear la entrada en los índices. El objetivo es procesar los elementos entrantes para extraer los términos del índice que entrarán en una base de datos estructurada. Un sistema de extracción de información sólo analiza aquellas partes de un documento que potencialmente contienen información relevante para los criterios de extracción. El objetivo de la extracción de datos es en la mayoría de los casos para actualizar una base de datos estructurada con datos adicionales. Las actualizaciones pueden ser de un vocabulario controlado que se define por las reglas de extracción (Kowalski, 1997).

2.2.3.1 N-gramas

Los n-gramas también son utilizados para crear las entradas a los índices, la dificultad para extraer las secuencias de caracteres depende de los detalles de cada lenguaje. Los n-gramas de caracteres representan una tokenización alternativa a la complejidad de un lenguaje (Büttcher et al., 2010).

Los n-gramas se pueden ver como una estructura de datos única en los sistemas de información, tienen una longitud fija y son una serie consecutiva de "n" caracteres. A diferencia del *stemming* que generalmente trata de determinar la raíz de una palabra que representa el

significado semántico, en cambio con los n-gramas la semántica no es importante (Kowalski, 1997).
Algunos ejemplos son bigramas, trigramas y pentagramas como la frase en inglés "sea colony", ver la siguiente tabla.

se ea co ol lo on ny	Bigramas (sin símbolos de interconexión)
sea col olo lon ony	Trigramas (sin símbolos de interconexión)
#se sea ea# #co col olo lon ony ny#	Trigramas (con símbolo de interconexión #)
#sea# #colt colon olony lony#	Pentagramas (con símbolo de interconexión #)

Tabla 2.1 Bigramas, trigramas y pentagramas.

El símbolo # se utiliza para representar el símbolo de interconexión que es cualquiera de un conjunto de símbolos (por ejemplo, espacio en blanco, punto, punto y coma, dos puntos, etc.).

2.2.4 Evaluación de la clasificación

Para evaluar la clasificación se utiliza la evaluación de los sistemas de recuperación de información, esto ha colocado una referencia estándar. Sin embargo, todavía hay debate sobre la precisión y la utilidad de los resultados del uso de un corpus de prueba (Kowalski, 1997). El propósito de la evaluación de la clasificación es para obtener rápidamente resultados aceptables. Se utilizan ejemplos del mundo real, estos conjuntos de datos son suministrados por los usuarios y son procesados para crear un modelo. El número de modelos depende de los parámetros que son elegidos y la escalabilidad. La selección del modelo para cada problema considera la precisión de entrenamiento, prueba y la escalabilidad.

2.3 Análisis de hipervínculos

Los antecedentes del análisis de hipervínculos para la búsqueda web son el análisis de citas bibliográficas. La bibliometría busca cuantificar los patrones de citas que se encuentran entre los artículos académicos. Las citas representan la atribución de autoría de un artículo académico a otros, el análisis de hipervínculos trata los enlaces de una página web a otra como un otorgamiento de autoría. Sin embargo, la métrica de calidad por la cantidad de

hipervínculos a otras páginas web no es suficientemente sólida, debido a que se puede aumentar artificialmente el conteo, este fenómeno se conoce como *spam* de hipervínculo (Manning et al., 2008).

2.4 Evaluación de hipervínculos

las investigaciones de recuperación de información sirvieron como base para formar el laboratorio INEX (Bellot et al., 2013). INEX evalúa los resultados de los sistemas de recuperación de información con el fin de mejorar su rendimiento. Sin embargo, surgió una comunidad de investigación basada en la interacción de los sistemas (Borlund & Ingwersen, 1998) (Xie, 2008). El objetivo es evaluar los sistemas de búsqueda utilizando métodos de evaluación centrados en el usuario.

2.4.1 PageRank

El algoritmo *PageRank* se convirtió en un elemento clave del motor de búsqueda *Backrub*, que maduró rápidamente en *Google* (Büttcher et al., 2010) (Manning et al., 2008). La intuición clásica detrás del algoritmo de *PageRank* imagina a una persona que navega por la Web al azar siguiendo los siguientes pasos:

1. Seguir un hipervínculo de la página actual haciendo clic en él ó
2. Seleccione una página al azar y escriba su URL en la barra de direcciones.

En cualquier momento, la probabilidad de que siga un enlace se corrige como d. Por lo tanto, la probabilidad de un salto es $1 - d$.

6. Los valores razonables para d pueden variar de 0.75 a 0.90, 0.85 utilizado por la literatura.

El algoritmo inicial del *PageRank*

$$PR(A) = (1 - d) + d \sum_{i=1}^{n} \frac{PR(i)}{C(i)}$$

Donde:

$PR(A)$ es el *PageRank* de la página A.
d es un factor de amortiguación que tiene un valor entre 0 y 1.
$PR(i)$ son los valores de *PageRank* que tienen cada una de las páginas **i** que enlazan a A.
$C(i)$ es el número total de enlaces salientes de la página i (sean o no hacia A).

Una página que está vinculada por muchas páginas con un *PageRank* alto también obtiene un *PageRank* alto. Es decir, el *PageRank* de una página se define recursivamente y depende del número y el *PageRank* de todas las páginas que lo vinculan.

2.4.2 Ranking

Al estructurar datos existen muchos métodos de recuperación, el más simple y utilizado es el ordenamiento de documentos en una colección según su relevancia dada una consulta. Cuando se escribe una consulta en un sistema de recuperación de información, los términos se expresan como vectores (Büttcher et al., 2010).
La notación para escribir los vectores de términos es la siguiente:

$$(t1, t2, t3, \dots, tn)$$

La representación como vector es útil cuando los términos se reflejan en una consulta y cuando el orden de los términos es significativo. En el ranking se usa la notación qt para indicar el número de veces que el término t aparece en la consulta. En la tabla 2.2 se explica cómo se construye básicamente el índice invertido.

Método básico de índice invertido	
Primer (término)	Devuelve la primera posición en la que se produce el término.
Último (término)	Devuelve la última posición en la que se produce el término.
Siguiente (término, actual $\boldsymbol{t}$)	Devuelve la siguiente posición en la que aparece el término después de la posición actual.
Anterior (término, actual $\boldsymbol{t}$)	Devuelve la posición anterior en la que el término aparece antes de la posición actual.

Tabla 2.2 Índice invertido.

2.4.3 Evaluación manual o humana

La recuperación de información requiere una evaluación cuidadosa para demostrar su rendimiento en los sistemas de recuperación de información. Existen medidas de evaluación que se utilizan de forma estándar como precisión y recuerdo para la recuperación de documentos y tareas relacionadas. Sin embargo, no existe una métrica para medir la utilidad de un hipervínculo según un usuario de acuerdo a la relevancia del documento.

La medida de utilidad para un hipervínculo de acuerdo a Manning (Manning et al., 2008) es la felicidad del usuario al navegar, algunos factores considerados son la velocidad de respuesta, el tamaño del índice y lo más importante es la relevancia de los resultados. Pero no todos los usuarios tienen la misma percepción y no coinciden con las nociones de calidad.

Otro ejemplo donde no se mide en el paradigma de relevancia básica es la metodología de Blustein (Blustein et al., 1997), en ella presenta un método de evaluación para medir el rendimiento del usuario usando el hipertexto, cada usuario debe leer dos veces el mismo documento en texto plano y con hipertexto. Se debe evitar leer los dos documentos para poder controlar la confusión, así como el estado mental del lector. El usuario debe escribir un resumen del documento que había leído, con la finalidad de saber si entendieron el texto. En el caso de dar clic en hipervínculos en el documento, al final se mostrará una lista y se pide que califiquen en una escala de cinco puntos.

2.5 Resumen del capítulo

En este capítulo se vieron los distintos métodos de reconocimiento de patrones; método supervisado, no supervisado y semisupervisado. En la clasificación se mencionan los distintos algoritmos como: Naive bayes, K-vecinos, Máquinas de soporte vectorial, Árboles de decisión Id3 y C4.5. La extracción de características es imprescindible, por lo tanto se describe; Ganancia de información, Enfoque binario y la Frecuencia de documentos. Por último en el punto 2.4 se explica el conocimiento que se obtiene de la evaluación de hipervínculos.

CAPÍTULO 3.

Estado del Arte

En este capítulo se analiza la compilación de resultados de otras investigaciones sobre el tema de investigación, se expresa la forma de abordar el tema y como establecieron la metodología los investigadores.

3.1 Reconocimiento de entidades

En la Tabla 3.1 se resume la extracción de entidades nombradas de distintos autores y la forma en que se desambiguan.

Tabla 3.1 Sistemas de reconocimiento de entidades.

Autor	**Desambiguador**
(Bunescu, 2006)	SVM Rank
(Cucerzan, 2007)	Producto escalar en vector categoría/término
(Varma et al., 2009)	Coseno entre en candidato y el contexto de la mención

(Rao et al., 2013)	Emparejamiento con una Base de conocimiento
(Martínez, 2005)	Etiquetado BIO y etiquetado PoS

3.2 Extracción de frases clave

El sistema *GROBID* (Lopez & Romary, 2010) analiza la estructura de los artículos científicos, formando un conjunto de características estructurales. Un segundo conjunto de características contiene las propiedades de fraseología, informatividad y las medidas de palabras clave. La experimentación fue realizada en la tarea 5 de *Semeval*, donde tan solo 144 artículos de ACM fueron utilizados para el entrenamiento. La colección de documentos científicos de este corpus fue creada por autores y lectores, en donde en cada documento ellos han asignado frases clave.

El método *MFSRank* (R. E. López, Barreda, Tejada, & Cuadros, 2011) está basado en grafos para extraer frases clave usando información semántica. La primera etapa consta de extraer secuencias frecuentes maximales (MFS), para construir los nodos de un grafo. En la segunda etapa, se calculan los valores de las MFS con el algoritmo de *PageRank*. Para la evaluación se utilizó el corpus de la tarea 5 de *Semeval*.

En el método de Lahiri (Lahiri, Choudhury, & Caragea, 2014) se experimenta con distintas medidas de centralidad en palabras y sintagmas nominales. Menciona Lahiri que existen medidas de centralidad que funcionan igual o mejor que *PageRank*, incluso que son mucho más simples. Para la evaluación se utiliza el corpus de la tarea 5 de *Semeval*.

En el trabajo de Camacho (Camacho, 2015) se hace la detección de patrones léxicos que tiene el contexto izquierdo y derecho de los hipervínculos. Este conocimiento se descubre de la construcción de hipervínculos creados por los anotadores humanos siguiendo las directrices de Wikipedia. Los patrones léxicos se identificaron para localizar las frases de texto candidatas a hipervínculo, y estos patrones se transformaron en patrones de búsqueda. El proceso de minería de datos se realiza con secuencias frecuentes maximales, después de las 5 fases del proceso de "*Knowledge Discovery in Text*" (KDT). Para la evaluación se utiliza un corpus externo de Wikipedia.

En la tabla 3.2 se muestran los sistemas de extracción de palabras clave y los modelos que se implementaron para realizar la tarea.

Tabla 3.2 Sistemas de extracción de palabras clave.

Autor	**Modelo**
(Lopez & Romary, 2010)	Modelo Machine learning SVM
(R. E. López et al., 2011)	Maximal Frequent Sequences MFS
(Lahiri et al., 2014)	Algoritmo Naïve y medidas de centralidad
(Camacho, 2015)	Patrones léxicos y MFS

3.3 Wikification

Wikification es el proceso de anotar las menciones de conceptos en un documento con la URL de la página de Wikipedia sobre ese concepto. Algunas herramientas como *DBPedia*, *SpotLight* o *Freebase* se usan en lugar de Wikipedia, pero la idea básica es la misma. La forma estándar de la tarea es encontrar conceptos candidatos en el artículo, después verificar si esos candidatos pueden coincidir con los títulos de las páginas de Wikipedia.

La mayoría de investigadores han apoyado la tarea general *Wikification*. Sin embargo, el concepto de *Milne* (Milne & Witten, 2008) en la evaluación es diferente, explica cómo se puede utilizar SVM para identificar términos significativos dentro del texto no estructurado, y crear los hipervínculos hacia los artículos apropiados de Wikipedia. La precisión para detectar candidatos a hipervínculo es del 75%. Para la experimentación seleccionaron un subconjunto de 50 documentos del corpus AQUAINT. Es una colección de historias de noticias del Servicio de Noticias *Xinhua*, *New York Times* y la *Associated Press*. Seleccionaron al azar los documentos, restringiendo la selección a documentos cortos con 250-300 palabras para evitar saturar la capacidad de atención de los evaluadores humanos.

Para evaluar la calidad de los hipervínculos que produjo un sistema donde el evaluador recibió el texto del artículo de noticias como hipertexto. El hipervínculo es mostrado con un cuadro

emergente que contiene el primer párrafo del artículo relevante de Wikipedia. Esto permite que tanto el contexto del hipervínculo como su destino sean vistos al mismo tiempo, como se muestra en la figura 3.1 del ejemplo "Baghdad".

Figura 3.1 Hipertexto con destinos hacia Wikipedia 2007.

El evaluador recibió las siguientes opciones para especificar si el ejemplo del hipervínculo "Baghdad" es válido:

- No, Bagdad no es una ubicación plausible para un hipervínculo.
- No, Bagdad es una ubicación plausible, pero el enlace no va al artículo correcto de Wikipedia.
- Tipo de - Bagdad es un hipervínculo plausible al artículo correcto de Wikipedia, pero el artículo no es útil o lo suficientemente relevante como para que valga la pena vincularlo.
- Sí - Bagdad es un enlace plausible al artículo correcto de Wikipedia, y este artículo es útil y relevante.

La contribución de Milne es un método de extracción de frases clave en texto plano que ha sido evaluado con un rendimiento humano. En la tabla 3.3 se muestran los trabajos que se basaron en *Wikification* para la detección de frases clave y los distintos métodos que utilizaron para intentar superar los resultados de la *baseline*.

Tabla 3.3 Métodos que implementaron para mejorar los resultados de *Wikification*.

DM	mention detection	SVM	Support Vector Machine
CL	Local compatibility	LM	Lexical Match
C	Coreference		
RS	Semantic Relatedness		
RG	Relational Graph		
RI	Relational Inference		
RE	Relation Extraction		

Autor	**Dominio**	**DM**	**CL**	**C**	**RS**	**RG**	**supervisado**	**SVM**	**Naïve Bayes**	**Lexical Match**	**RI**
(Huang et al., 2014)	Wikipedia	✓	✓	✓	✓	✓	semi				
(Milne & Witten, 2008)	Wikipedia				✓	✓		✓	✓		
(Csomai & Mihalcea , 2007)	Wikipedia noticias			✓							
(Cheng & Roth, 2013)	Wikipedia									✓	✓
(Kim, Banchs, & Li, 2015)	Wikipedia, noticias						✓	✓			

3.4 Desambiguación

La siguiente tabla muestra las partes de un documento de Wikipedia que se utilizaron para la búsqueda de documentos utilizando una entidad nombrada o frase clave como consulta.

Tabla 3.4 Partes de un documento de Wikipedia utilizado para la búsqueda.

Autor	Extracción	Título	Redirección	Negritas
(Bunescu, 2006)	NER	✓	✓	
(Cucerzan, 2007)	NER	✓	✓	
(Varma et al., 2009)	NER	✓	✓	✓
(Rao et al., 2013)	NER	✓	✓	
(Csomai & Mihalcea, 2007)	Frases clave	✓	✓	

La tarea *Entity-linking* presenta el problema de clasificar las entidades nombradas como lo pueden ser principalmente personas, organizaciones o localizaciones. Cada entidad puede tener ambigüedad en su significado, por lo tanto afecta el rendimiento de los sistemas de recuperación de información. Los siguientes autores presentan métodos robustos para la desambiguación aprovechando el contexto de las bases de conocimiento.

Tabla 3.5 métodos de desambiguación para reconocimiento de entidades nombradas.

EE	extracción de entidad	MS	método supervisado
P	persona	BC	base de conocimiento
O	organización	AD	arboles de decisión
L	localización		
M	miscelánea		

Autor	**dominio**	**EE**					**MS**	**BC**	**NER**	**SVM**	**AD**	**Grafo**	**Page Rank**
			P	O	L	M							
(Fernández et al., 2012)	noticias	si	✓	✓	✓		si						
(Hoffart et al., 2011)	Wikipedia		✓	✓	✓			*DBpedia, Freebase*, o YAGO	Stanford NER, Web graph	✓		✓	
(Saha & Ekbal, 2013)	Noticias		✓	✓	✓	✓	si, no		ME, CRF, MBL, HMM	✓	✓		
(Giuliano, Lavelli, & Romano, 2007)	Noticias		✓	✓	✓			lista de palabras	ME, KMS, ML, HMMs, CRF	✓			
(Blanco et al., 2015)	ANCORA							*Freebase*, *Wordnet*				✓	✓
(Pérez, 2008)	Noticias		✓	✓	✓	✓	si, no		HMM, freeling	✓	✓		

3.5 Resumen del capítulo

En este capítulo se presentaron los métodos enfocados a la tarea *Entity-linking*, NER, extracción de frases clave y *Wikification*. La tarea NER únicamente identifica las entidades nombradas a diferencia de *Entity-linking*, el cual abarca desde el reconocimiento hasta la vinculación de las entidades. Con los enfoques presentados de los sistemas es posible realizar un método que unifique las tareas de extracción de frases clave y NER, usando de técnicas de aprendizaje automático.

CAPÍTULO 4.

Método Propuesto

La hipervinculación de documentos consta de 3 etapas: (1) de selección de términos para hipervinculación, (2) búsqueda de los documentos y, por último, (3) el redireccionamiento a una página. Se sabe que los hipervínculos están conformados de entidades nombradas de manera conjunta con otros términos que pueden ser frases clave. Por lo tanto, se aborda el problema de hipervinculación unificando las tareas de NER y frases clave para seleccionar los términos correctos que representen un texto de un hipervínculo. En el método propuesto haremos uso de técnicas de aprendizaje automático, específicamente máquinas de soporte vectorial (SVM). Con SVM clasificaremos los mejores candidatos para crear un hipervínculo. Para la búsqueda de los documentos y el redireccionamiento a una página se realizará a partir de un ranking con una métrica de similitud.

4.1 Características

La particularidad más importante de la extracción de características para clasificar hipervínculos validos es que éstas se identifican y se seleccionan sin utilizar ningún conocimiento dependiente del dominio o recursos específicos del idioma. Para ello se utiliza una medida de discriminación conocida como ganancia de información, donde tenemos n números de atributos y la ganancia se obtiene para cada característica. La característica con la ganancia mayor de información normalizada se selecciona para la toma de decisiones.

En la identificación de características asociamos cada hipervínculo con atributos de frases clave o entidades nombradas de los artículos de Wikipedia, tomamos en cuenta las etiquetas HTML tales como:
Negrita
Cursiva
Hipervínculo <a>
Título <h1></h1>
Párrafo
El texto dentro de estas etiquetas, principalmente de la etiqueta <a> observamos que las palabras tienen los siguientes atributos:
1.- Empieza con mayúsculas.
2.- Todas las letras son mayúsculas.
3.- Todas las letras son minúsculas.
4.- Es un acrónimo.
5.- Es un acrónimo con definición.
6.- Contiene signos.
7.- Título de la página.

Para el resto de las etiquetas y palabras tenemos los siguientes atributos:
8.- Signos de puntuación.
9.- Es una *stopword*.
10.- Solo dígitos.
11.- Largo de cadena.
12.- Puede tener *stem*.
13.- Términos en negritas.
14.- Términos en cursivas.
15.- Frecuencia del término en el documento.
16.- Frecuencia del término por párrafo en cada documento.
17.- Posición de un párrafo dentro de un documento.
18.- Posición del término en el documento por párrafo.
19.- Posición de una oración dentro de un documento.
20.- Relación con el título.
21.- Número total de tokens en cada documento.
22.- Términos con acentos.
23.- Solo palabras (sin dígitos o símbolos).
24.- Bigramas.
25.- Trigramas.

Tenemos como ejemplo la siguiente oración "A cappella La música a cappella es música vocal sin acompañamiento instrumental.". Aplicando las 25 características con el texto obtenemos la siguiente bolsa de palabras.

A(1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 175, 2, 1, 1, 1, 0, 285, 0, 1, A, A), cappella (0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 8, 0, 0, 0, 11, 2, 1, 2, 1, 0, 285, 0, 1, A cappella, A cappella), La (1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 2, 0, 0, 0, 32, 1, 1, 3, 1, 0, 285, 0, 1, cappella La, A cappella La), música (0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 6, 0, 0, 0, 9, 2, 1, 4, 1, 0, 285, 1, 1, La música, cappella La música), a (0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 175, 1, 1, 5, 1, 1, 285, 0, 1, música a, La música a), cappella (0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 8, 0, 1, 1, 11, 2, 1, 6, 1, 1, 285, 0, 1, a cappella, música a cappella), es (0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 2, 0, 0, 0, 23, 1, 1, 7, 1, 0, 285, 0, 1, cappella es, a cappella es), música (0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 6, 0, 0, 0, 9, 2, 1, 8, 1, 0, 285, 1, 1, es música, cappella es música), vocal (0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 3, 1, 1, 9, 1, 0, 285, 0, 1, música vocal, es música vocal), sin (0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 3, 0, 0, 0, 2, 1, 1, 10, 1, 0, 285, 0, 1, vocal sin, música vocal sin), acompañamiento (0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 14, 0, 0, 0, 2, 1, 1, 11, 1, 0, 285, 0, 1, sin acompañamiento, vocal sin acompañamiento), instrumental (0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 12, 1, 0, 0, 1, 1, 1, 12, 1, 0, 285, 0, 1, acompañamiento instrumental, sin acompañamiento instrumental), . (0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 10, 1, 1, 13, 1, 0, 285, 0, 0, instrumental ., acompañamiento instrumental .).

Describiremos las características más relevantes utilizadas por Bunescu (Bunescu, 2006), Varma (Varma et al., 2009), Cucerzan (Cucerzan, 2007), Lopez (Lopez & Romary, 2010) y Rao (Rao et al., 2013), ellos utilizaron como base las siguientes características:

Título , bigramas, trigramas, negritas, acrónimo, empieza con mayúsculas.

Las etiquetas para las partes del discurso (en inglés POS) dependen de bases de conocimiento para su posterior análisis de estructuras sintácticas (ver Anexo 1), aprovechamos el conocimiento obtenido de estos autores con las partes del discurso para identificar la posición de cada una de estas y además obtener las características estructurales html como el trabajo de Lopez y Romary (Lopez & Romary, 2010):

Título, resumen, introducción, título de sección, conclusión, título de referencia.

Solo nos centramos en las siguientes características ocupadas en el estado del arte debido a que el método propuesto es independiente del lenguaje.

Características de NER:

- Título de la página.
- Bigramas.
- Trigramas.
- Términos en Negritas.
- Es un acrónimo.
- Es un acrónimo con definición.
- Empieza con mayúsculas.

Características de estructura:

- Todas las letras son mayúsculas.
- Términos en Cursivas.
- Frecuencia del término en el documento.
- Frecuencia del término por párrafo en cada documento.
- Posición de un párrafo dentro de un documento.
- Posición del término en el documento por párrafo.
- Posición de una oración dentro de un documento.

Características añadidas:

- Todas las letras son minúsculas.
- Contiene signos.
- Relación con el título.
- Número total de tokens en cada documento.
- Términos con acentos.
- Solo palabras (sin dígitos o símbolos).
- Signos de puntuación.
- Es una *stopword*.
- Solo dígitos.
- Largo de cadena.
- Puede tener *stem*.

Utilizamos ganancia de información para discriminar y seleccionar los atributos más relevantes, teniendo las alternativas para poder hacer el entrenamiento y la clasificación procedemos a realizar la tokenización. Las características finalmente aprobadas son:

1.- Empieza con mayúsculas.
2.- Todas las letras son mayúsculas.
3.- Todas las letras son minúsculas.
4.- Es un acrónimo.
5.- Contiene signos.
6.- Título de la página.
7.- Signos de puntuación.
8.- Es una *stopword*.
9.- Solo dígitos.
10.- Términos en negritas.
11.- Términos en cursivas.
12.- Frecuencia del término por párrafo en cada documento.
13.- Posición de un párrafo dentro de un documento.
14.- Posición del término en el documento por párrafo.
15.- Posición de una oración dentro de un documento.
16.- Relación con el título.
17.- Número total de términos en cada documento.

4.2 Tokenización

La tokenización posibilita el entrenamiento a partir de un conjunto que sea suficiente representativo, los artículos de Wikipedia tienen formato HTML. Por lo tanto, se necesita hacer un preproceso para limpiar todas las etiquetas que no se utilizan para la etapa de características, las cuales son utilizadas para definir el hiperplano en la clasificación. Para la limpieza de los artículos de Wikipedia se usan expresiones regulares que forman un patrón de búsqueda para usarlo en el componente FORMATEXT (ver anexo 2). Como resultado de Tokenización obtenemos un archivo de texto, como ejemplo tenemos la siguiente oración "A cappella La música a cappella es música vocal sin acompañamiento instrumental." y la salida es:

1 1 0 1 0 1 0 0 1 0 A
1 2 0 2 0 2 0 0 1 0 cappella
1 3 1 1 1 1 0 0 0 0 La
1 4 1 2 1 2 0 0 0 0 música
1 5 1 3 1 3 1 1 0 0 a
1 6 1 4 1 4 1 1 0 0 cappella
1 7 1 5 1 5 0 0 0 0 es

1 8 1 6 1 6 0 0 0 0 música
1 9 1 7 1 7 0 0 0 0 vocal
1 10 1 8 1 8 0 0 0 0 sin
1 11 1 9 1 9 0 0 0 0 acompañamiento
1 12 1 10 1 10 0 0 0 0 instrumental
1 13 1 11 1 11 0 0 0 0 .

La posición de cada columna en el ejemplo anterior son algunas características estructurales y se describen a continuación:

0 Nombre o identificador del documento.
1 No. Palabra en el documento.
2 No. Párrafo.
3 No. Palabra en el párrafo.
4 No. Oración.
5 No. Palabra en la oración.
6 Negrita.
7 Cursiva.
8 Relación con título.
9 Hipervínculo.
10 Palabra en crudo.

4.3 Clasificación

En la hipervinculación de documentos inicialmente se eligen los términos para crear un hipervínculo, en la etapa de selección de características los atributos de las entidades con nombre de manera conjunta con las frases clave determinarán los términos correctos que representen un texto de un hipervínculo por medio de una clasificación. El algoritmo que se ubica entre los mejores para la tarea de clasificación de texto es la Máquina de Soporte Vectorial (SVM), debido a la alta dimensionalidad del espacio de características.

4.3.1 Entrenamiento

Para el entrenamiento tomamos un conjunto aleatorio de artículos en español de la colección de Wikipedia 2008. Al introducir a la Máquina de Soporte Vectorial los artículos de Wikipedia, define un modelo que representa el conjunto de entrenamiento. El espacio de características

Debe ser concatenado y los datos de formación pueden tener un aspecto similar a la tabla siguiente.

Tabla 4.1 Datos de entrenamiento.

Token	Atributo 1	Atributo 2	Atributo 3	Atributo 4	Atributo 5	Atributo 6	…	N
A	1	0	0	0	0	0		
cappella	0	0	1	0	0	0		
La	1	0	0	0	0	0		
música	0	0	1	0	0	0		
a	0	0	1	0	0	0		
cappella	0	0	1	0	0	0		
es	0	0	1	0	0	0		
música	0	0	1	0	0	0		
vocal	0	0	1	0	0	0		
sin	0	0	1	0	0	0		
acompañamiento	0	0	1	0	0	0		
.	0	0	1	0	0	0		
Token N	…	…	…	…	…	…		

Durante el proceso de formación, la Máquina Soporte Vectorial creará un modelo en un espacio de n dimensiones, donde n depende de la cantidad total que se utiliza en el espacio de características. El modelo distribuye cada uno de los ejemplos de formación en las categorías de verdadero o falso.

4.3.2 Validación (Predicción de hipervínculos)

Una vez que la formación del modelo está completa, se puede usar para clasificar nuevos datos como verdaderos o falsos. Al utilizar el modelo para predecir cuales son los términos que pueden ser hipervínculos, la muestra de validación debe mantener el mismo formato que la muestra de entrenamiento, así como el mismo tamaño de dimensiones. Los datos booleanos de salida tienen un formato como la tabla siguiente:

Tabla 4.2 Valores booleanos de clasificación.

Token	Valor booleano
1	1
2	0
3	1
4	1
5	0
6	0
7	0
8	0
9	0
10	1

Los valores booleanos de la tabla son verdadero (número 1) y falso (número 0). Para la clasificación de los términos que pueden ser hipervínculos, aquellos tokens que tengan como valor "verdadero" son tomados como candidatos y se podrán utilizar para la búsqueda de documentos y posteriormente crear su hipervínculo.

4.4 Búsqueda de documentos

En el trabajo de la tesis "Sistema de Construcción Automática de Hipervínculos Independiente del Lenguaje" (León, 2015) se diseñó un sistema que resuelve el problema de la vinculación de documentos electrónicos. Este sistema consta de 4 fases principales como se muestra en la figura 4.1.

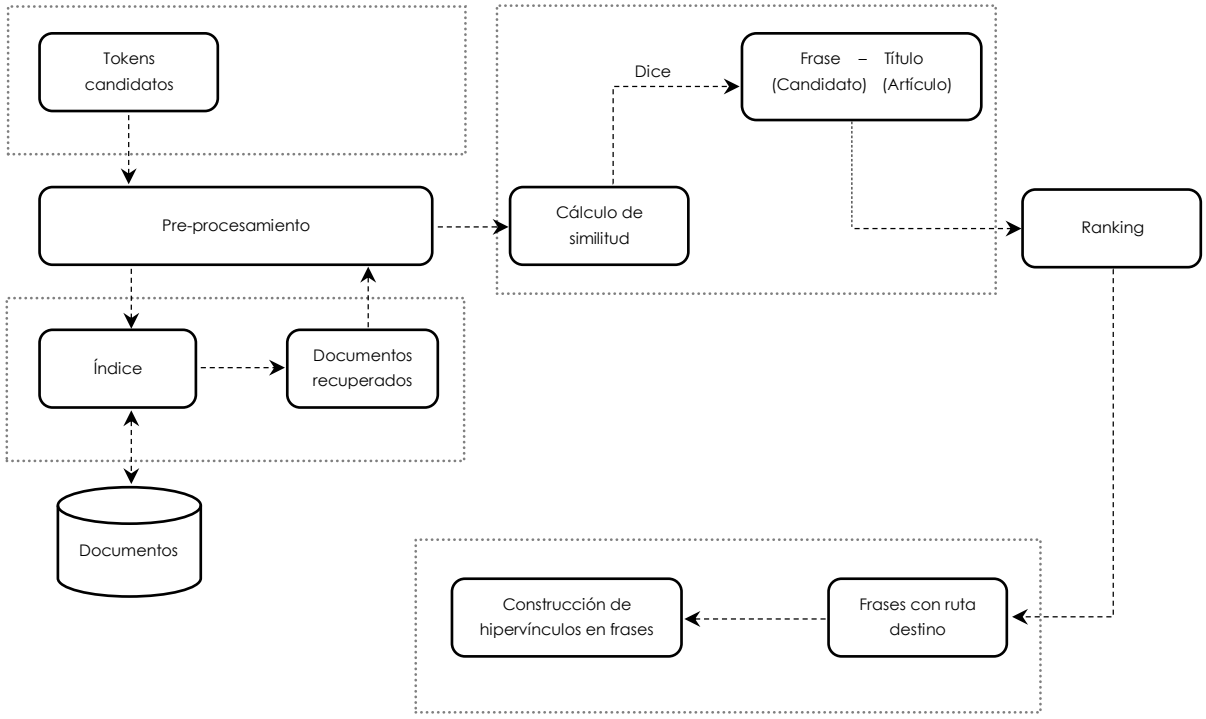

Figura 4.1 Arquitectura del sistema.

La primera fase (Consulta), está conformada por las frases candidatas, las cuales vamos a sustituir por los tokens candidatos resultantes de la clasificación con SVM.

La segunda fase (Glosario de términos y estructuración de archivos), después del pre-procesamiento de los tokens candidatos se realiza la búsqueda del documento destino.
Para ello, se cuenta con un glosario de términos gestionado en una bases de datos MySQL con 7, 198, 504 registros, de ahí se recupera un conjunto de documentos que contiene uno o varios términos del título del artículo.

La tercera fase (Búsqueda del documento destino), para determinar cuál es el documento destino pertinente de acuerdo al título de Wikipedia y token candidato, se utiliza la similitud de Dice, se hace el cálculo del token candidato y los títulos de los documentos de Wikipedia recuperados. De tal manera que, con las puntuaciones de la similitud de Dice se determina el mejor documento destino para cada frase candidata en el *ranking*.

La cuarta fase (Hipertexto), se tienen todas los tokens candidatos con sus respectivas rutas para crear los hipervínculos con destino a los documentos de Wikipedia correspondientes.

4.5 Hipervinculación

La etapa de hipervinculación consiste en la creación de hipervínculos en los tokens candidatos con los destinos de los documentos obtenidos del ranking. La estructura del hipertexto es creada con *Bootstrap*. El hipertexto se visualiza en el navegador como se aprecia en la figura 4.2.

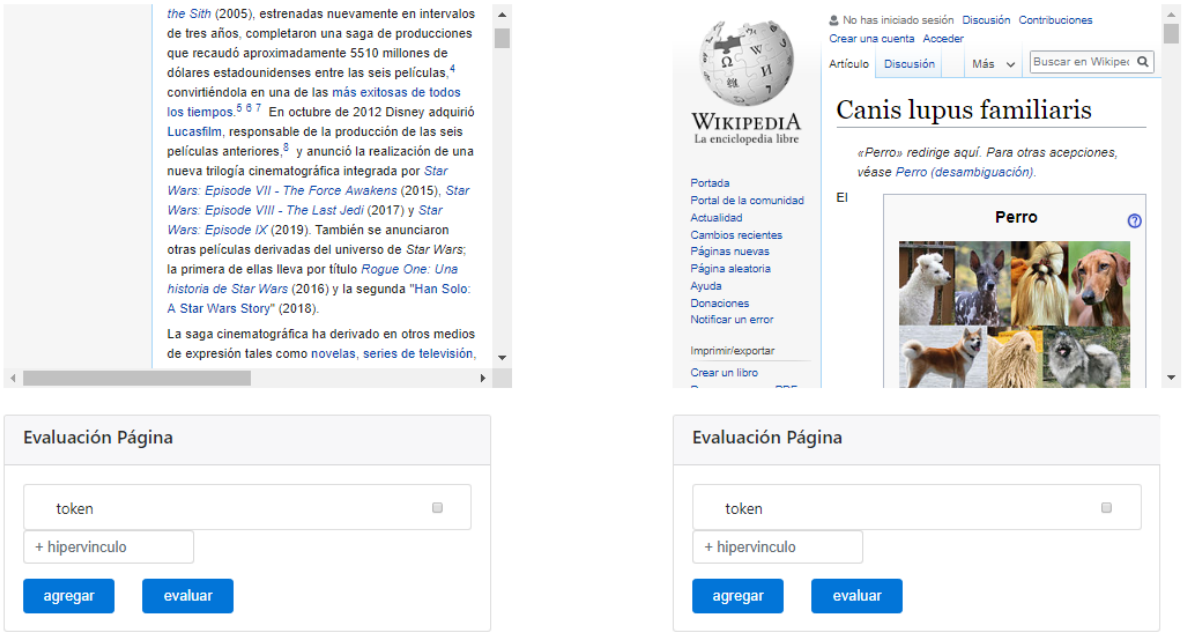

Figura 4.2 Interfaz de evaluación.

4.6 Evaluación del usuario

La recuperación de información requiere de una evaluación cuidadosa para demostrar su rendimiento. La medición de la efectividad de un método de recuperación depende en evaluaciones de relevancia. Para poder hacer una evaluación se necesita: (1) el conjunto de documentos devueltos por la consulta y (2) el conjunto de documentos relevantes del contexto en la colección. A partir de estos dos conjuntos podemos entonces calcular dos medidas estándar: recuerdo y precisión. Sin embargo, no todos los usuarios tienen la misma percepción de calidad, por lo tanto, en este trabajo nos enfocamos en el rendimiento del usuario usando el hipertexto, cada usuario debe leer y navegar en los hipervínculos para determinar cuáles fueron creados por humanos y cuales fueron construidos de manera automática por la máquina.

4.7 Resumen del capítulo

En este capítulo se describió la arquitectura del método propuesto que resuelve el problema de la hipervinculación de documentos electrónicos. Las etapas constan de la selección de términos para crear un hipervínculo, la búsqueda de los documentos, y por último de la hipervinculación del documento en un *ranking*. Contar con un glosario de términos bastante amplio es importante porque aporta un conjunto de documentos en cualquier consulta, y la medida de similitud determina los destinos de los hipervínculos. Finalmente con el hipertexto creado, los usuarios de prueba se sumergen en los hipervínculos y prueban su habilidad para evaluar cuales hipervínculos son creados por humanos y cuales son creados por la máquina.

CAPÍTULO 5.

Experimentación

En este capítulo se comprueba la hipótesis planteada de esta investigación. En el método propuesto presentamos un modelo general para la hipervinculación, y tenemos conocimiento de las tareas principales de la pregunta de investigación ¿Cómo clasificar las frases de un texto que representen hipervínculos y seleccionar los documentos destinos sin emplear diccionarios de desambiguación? Nuestra hipótesis es que un clasificador como SVM es capaz de aprender el contexto que define un hipervínculo, de este modo el modelo puede predecir cuales términos son candidatos a hipervínculo. Implementamos la librería LIBSVM, ya que es de libre distribución y es un paquete software pensado para resolver problemas de clasificación mediante SVM. Puesto que no necesitamos desambiguar los candidatos para la recuperación de documentos, el proceso de búsqueda es un *ranking* de similitud con los candidatos y los títulos de Wikipedia. Por último, la evaluación del usuario con el hipertexto se realiza por medio de una interfaz.

5.1 Experimento con 17 características

Las características que utilizamos como primera instancia para el experimento inicial son las siguientes:

1.- Empieza con mayúsculas.
2.- Todas las letras son mayúsculas.

3.- Todas las letras son minúsculas.
4.- Es un acrónimo.
5.- Contiene signos.
6.- Título de la página.
7.- Signos de puntuación.
8.- Es una *stopword*.
9.- Solo dígitos.
10.- Términos en negritas.
11.- Términos en cursivas.
12.- Frecuencia del término por párrafo en cada documento.
13.- Posición de un párrafo dentro de un documento.
14.- Posición del término en el documento por párrafo.
15.- Posición de una oración dentro de un documento.
16.- Relación con el título.
17.- Número total de tokens en cada documento.

Para el entrenamiento tomamos un conjunto aleatorio de 681 artículos en español de la colección de Wikipedia 2008 y con 17 atributos para crear el modelo. El ejemplo de la concatenación con 17 atributos se muestra en el anexo 3.

Para la validación del modelo creado con 17 atributos del espacio de características, se utilizaron aleatoriamente 5,400 tokens. La exactitud para este modelo es de 72.7721% con una ventana de tamaño 3 para el contexto.

5.2 Experimento con Ganancia de Información

Utilizamos la medida de ganancia de información para calcular qué atributos poseen mayor relevancia, de tal manera que ocupemos aquellos atributos que son necesarios para aumentar la exactitud del modelo. Únicamente se tokenizaron 681 artículos de Wikipedia para obtener la siguiente tabla 5.1.

Tabla 5.1 Valores de Ganancia de Información para cada característica.

N	GI	Característica
1	0.14334	Empieza con mayúsculas.
2	0.12665	Todas las letras son minúsculas.
3	0.08666	Es una *stopword*.
4	0.04659	Signos de puntuación.
5	0.04401	Solo dígitos.
6	0.04307	Posición del término en el documento por párrafo.
7	0.03938	Posición de una oración dentro de un documento.
8	0.03085	Posición de un párrafo dentro de un documento.
9	0.02078	Contiene signos.
10	0.01715	Número total de términos en cada documento.
11	0.00927	Todas las letras son mayúsculas.
12	0.00632	Es un acrónimo.
13	0.00465	Términos en Negritas.
14	0.00334	Título de la página.
15	0.00199	Relación con el título.
16	0	Frecuencia del término por párrafo en cada documento.
17	0	Términos en Cursivas.

Con respecto a los valores de ganancia de información, se creó el modelo. De esta manera podemos comprobar la importancia del uso de la medida de discriminación. Para el entrenamiento tomamos el mismo conjunto de 681 artículos y de acuerdo a la tabla 5.1 se utilizaron solo los primeros 10 atributos.

Para la validación del modelo también se utilizaron 5,400 tokens con una ventana de tamaño 3 para el contexto, y obtuvimos un aumento de la exactitud con 77.2761 % de clasificación, por lo tanto comprobamos la importancia de utilizar ganancia de información.

5.3 Experimento con validación cruzada

Creamos un tercer modelo con los 10 primeros atributos de la tabla 5.1 y esta vez en la tokenización utilizamos 13 mil artículos de Wikipedia, para tener la mayor cantidad de ejemplos posible y comprobar la eficiencia. Utilizamos la técnica de validación cruzada para garantizar que los datos de la participación de los resultados son independientes de los datos de prueba y los datos de entrenamiento. La validación consiste en repetir y calcular la media aritmética

de n medidas de evaluación sobre diferentes particiones. Para este experimento n es igual a 10 y la exactitud del modelo es de 80.05538 % con una ventana de tamaño 3 para el contexto. El valor de la exactitud superó los experimentos anteriores, sin embargo no comparamos la exactitud con el estado del arte, debido a que los datos de entrenamiento no tienen sobreajuste (como en un corpus), ya que se trata de la colección de Wikipedia. Como el caso de estudio son los hipervínculos no se cuenta con un corpus con el que se pueda comparar, tampoco existen investigaciones similar al método propuesto. Por esta razón, la validación cruzada es una forma de predecir el ajuste de un modelo hipotético.

5.4 Búsqueda de documentos

Para realizar las búsquedas se generó previamente un glosario de términos con la colección de Wikipedia 2017, el cual cuenta con 7, 198, 504 registros. En esta etapa se preparan los documentos con los candidatos de manera anticipada con un preprocesamiento para realizar las consultas. Con los tokens candidatos a hipervínculo obtenidos de la clasificación SVM se realiza la búsqueda del documento destino para cada consulta. Considerando los tokens candidatos "Linux" y "sistema informático" vamos a mostrar el ejemplo del proceso de búsqueda. En la consulta se recuperaron 299 títulos de Wikipedia que contienen el Token "Linux", en el anexo 6 se muestra la recuperación completa de la consulta y en el anexo 4 solo se muestran 15 títulos. Para el token candidato "sistema informático", se recuperaron 3, 101 títulos y por la cantidad decidimos solo mostrar 15 títulos en el anexo 5.

5.5 Hipervinculación

Para determinar el documento destino de cada token candidato, se utiliza la similitud de Dice, calculando la similitud entre el token candidato y los títulos de los documentos de Wikipedia recuperados. Con las puntuaciones del ranking se determina el mejor documento destino para cada frase candidata con el mayor puntaje. Los valores de similitud del token "Linux" se muestran en la tabla 5.2 y la tabla 5.3 muestra los puntajes del token "sistema informático".
La recuperación completa de las url por la consulta del token "Linux" se encuentra en el anexo 6 Búsqueda de documentos.

Tabla 5.2 Valores de similitud del token "Linux".

	Título	URL	Ranking
1	Linux HA	https://es.wikipedia.org/wiki/Linux_HA	0.727272727
2	GNU Linux	https://es.wikipedia.org/wiki/GNU_Linux	0.666666667
3	Arch Linux	https://es.wikipedia.org/wiki/Arch_Linux	0.615384615
4	Linux Libre	https://es.wikipedia.org/wiki/Linux_Libre	0.571428571
5	Portabilidad del núcleo Linux y arquitecturas soportadas	https://es.wikipedia.org/wiki/Portabilidad_del_núcleo_Linux_y_arquitecturas_soportadas	0.133333333
6	Proceso de arranque en linux	https://es.wikipedia.org/wiki/Proceso_de_arranque_en_linux	0.193548387
7	Disputas de sco sobre linux	https://es.wikipedia.org/wiki/Disputas_de_sco_sobre_linux	0.2
8	PXES Universal Linux Thin Client	https://es.wikipedia.org/wiki/PXES_Universal_Linux_Thin_Client	0.228571429
9	Softlanding Linux System (SLS)	https://es.wikipedia.org/wiki/Softlanding_Linux_System_(SLS)	0.242424242
10	Alt Linux	https://es.wikipedia.org/wiki/Alt_Linux	0.666666667
11	Amber Linux	https://es.wikipedia.org/wiki/Amber_Linux	0.571428571
12	ArchBang Linux	https://es.wikipedia.org/wiki/ArchBang_Linux	0.470588235
13	Arquitectura de Sonido Avanzada para Linux	https://es.wikipedia.org/wiki/Arquitectura_de_Sonido_Avanzada_para_Linux	0.177777778
14	Arranque remoto sin disco en linux	https://es.wikipedia.org/wiki/Arranque_remoto_sin_disco_en_linux	0.162162162
15	Linux	https://es.wikipedia.org/wiki/Linux	**1.0**

Tabla 5.3 Valores de similitud del token "sistema informático".

	Título	URL	Ranking
1	Sistema informático	https://es.wikipedia.org/wiki/Sistema_informático	**0.944444444**
2	Sistema operativo Windows	https://es.wikipedia.org/wiki/Sistema_operativo_Windows	0.380952381
3	Sistema operativo web	https://es.wikipedia.org/wiki/Sistema_operativo_web	0.368421053
4	Sistema Operativo Robótico	https://es.wikipedia.org/wiki/Sistema_Operativo_Robótico	0.428571429
5	Sistema (informatica)	https://es.wikipedia.org/wiki/Sistema_(informatica)	0.756756757
6	Sistema operativo	https://es.wikipedia.org/wiki/Sistema_operativo	0.470588235
7	Sistema (informática)	https://es.wikipedia.org/wiki/Sistema_(informática)	0.789473684
8	Sistema de informacion	https://es.wikipedia.org/wiki/Sistema_de_informacion	0.631578947
9	Sistema operativo multiusuario	https://es.wikipedia.org/wiki/Sistema_operativo_multiusuario	0.304347826
10	Sistema Xbox One	https://es.wikipedia.org/wiki/Sistema_Xbox_One	0.363636364
11	Sistema X Window	https://es.wikipedia.org/wiki/Sistema_X_Window	0.424242424
12	Sistema abierto (informatica)	https://es.wikipedia.org/wiki/Sistema_abierto_(informatica)	0.577777778
13	Sistema de información	https://es.wikipedia.org/wiki/Sistema_de_información	0.631578947
14	Sistema abierto (informática)	https://es.wikipedia.org/wiki/Sistema_abierto_(informática)	0.652173913

15	Sistema informatico	https://es.wikipedia.org/wiki/Sistema_informatico	0.857142857

Finalmente para crear el Hipertexto de todos los tokens candidatos y sus respectivas url, se selecciona el mayor puntaje de cada ranking en la consulta. El valor de mayor puntuación para el token "Linux" es la posición 15 de la tabla 5.2 con una puntuación de similitud completa.

15	**Linux**	**https://es.wikipedia.org/wiki/Linux**	1.0

En la tabla 5.3 el valor con mayor puntuación para el token "sistema informático" es la posición 1.

1	**Sistema informático**	**https://es.wikipedia.org/wiki/Sistema_informático**	0.944444444

5.6 Evaluación del usuario

La interacción humana con el hipertexto esta mediada por una interfaz, en donde los usuarios después de leer el texto, eligen los hipervínculos para navegar y agregan aquellos que utilizaron. La finalidad de agregar solo aquellos que utilizaron es debido a que el mismo usuario valora la calidad del hipervínculo, es decir, el usuario decide si el hipervínculo en el que navegó en internet, fue creado por un humano, o en su defecto por una máquina. En la tabla 5.4 podemos observar cómo valoran cada hipervínculo, el símbolo "✓" determina que el hipervínculo fue creado por un humano.

Tabla 5.4 Valoración de hipervínculos del Método Propuesto.

	Usuarios									
Hipervínculo	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
Congreso de los Diputados de España	✓	✓	✓	✓	✓	✓		✓		✓
Chaos Computer Club	✓	✓		✓	✓	✓	✓	✓		
Bajos	✓		✓			✓	✓		✓	
Rhythm & Blues	✓	✓		✓	✓		✓		✓	
The Amazing Race	✓	✓	✓	✓		✓		✓		✓

Primera guerra mundial		✓	✓	✓		✓	✓			✓
Lista de asteroides					✓					
el futuro .	✓	✓	✓				✓			

La tabla 5.4 tiene hipervínculos creados por el método propuesto, los usuarios no estaban informados de que el hipertexto que leyeron, era creado de forma automática por la máquina, con la intención de no alterar el discernimiento del usuario. De igual forma para la lectura del hipertexto de Wikipedia, los usuarios no estaban informados y los hipervínculos que los usuarios eligieron se muestran en la tabla 5.5.

Tabla 5.5 Valoración de hipervínculos de Wikipedia.

	Usuarios									
Hipervínculo	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
datos	✓	✓	✓	✓	✓	✓			✓	✓
análisis de datos	✓	✓	✓	✓		✓	✓		✓	✓
inteligencia empresarial	✓	✓	✓	✓		✓	✓			✓
Sistemas de Soporte a Decisiones	✓	✓	✓		✓		✓		✓	✓
repositorio de datos	✓		✓	✓	✓	✓		✓	✓	
top-down	✓	✓	✓		✓	✓		✓	✓	✓
bases de datos	✓		✓	✓	✓	✓	✓	✓	✓	
ETL		✓	✓	✓		✓		✓	✓	
Extract, transform and load		✓		✓		✓	✓		✓	
sistema operativo	✓	✓	✓	✓	✓		✓	✓	✓	

La interfaz muestra ambos hipertextos con el mismo formato para controlar los efectos de discernimiento del hipertexto que se utiliza para presentar texto a los lectores. Cada usuario experimental debe leer dos documentos, el contenido y la longitud de los dos documentos son similares. En la tabla 5.6 se muestra el total de puntos que los usuarios dieron a cada hipervínculo para el hipertexto creado por el Método Propuesto.

Tabla 5.6 Valoración total del Método Propuesto.

Hipervínculo	Evaluación
Congreso de los Diputados de España	8
Chaos Computer Club	7
Bajos	5
Rhythm & Blues	6
The Amazing Race	7
Primera guerra mundial	6
Lista de asteroides	1
el futuro .	4
∑	44

En la tabla 5.7 los puntos para cada hipervínculo del hipertexto con Wikipedia tienen valores altos, notamos que los usuarios al utilizar Wikipedia, se encuentran familiarizados con el formato y la forma en que crean los hipervínculos para sus artículos.

Tabla 5.7 Valoración total de Wikipedia.

Hipervínculo	Evaluación
datos	8
análisis de datos	8
inteligencia empresarial	7
Sistemas de Soporte a Decisiones	7
repositorio de datos	7
top-down	8
bases de datos	8
ETL	6
Extract, transform and load	5
sistema operativo	8
∑	72

5.7 Resumen del capítulo

En este capítulo se explicó cada etapa del método propuesto, para el conjunto de entrenamiento explicamos como utilizamos ganancia de información para tener mejores resultados en el conjunto de validación, de igual forma la importancia del glosario de términos en la búsqueda de documentos permite hacer el cálculo de similitud para obtener el documento pertinente para cada hipervínculo con el ranking. Por último se genera el hipertexto, que permite a los usuarios visualizar los hipervínculos y navegar entre los documentos, la valoración de los hipervínculos genera resultados que pueden ser útiles para sacar conclusiones de cómo podría mejorar el método propuesto.

CAPÍTULO 6

Conclusiones y Trabajo Futuro

En este capítulo se presentan las conclusiones de esta investigación. Finalmente, en el punto de trabajo futuro se mencionan las posibles líneas de investigación que surgen a partir de esta tesis.

6.1 Conclusiones

Presentamos un método general para la hipervinculación de documentos en las etapas de selección de términos (reconocimiento de entidades nombradas y frases clave), búsqueda de documentos con un glosario de términos gestionado en una bases de datos MySQL, y finalmente el redireccionamiento a una página por medio de la mayor puntuación de similitud con el coeficiente de Dice.

En la selección de términos se utilizaron los algoritmos de aprendizaje SVM, clasificamos los mejores candidatos para crear un hipervínculo. Se entrenó una SVM para construir un modelo con 17 características y una ventana de tamaño 3 para contexto alcanzando una precisión del 72.77 %. Sin embargo, pudimos mejorar estos resultados utilizando solo las características con Ganancia de Información y una combinación ponderada de documentos. Con estos ajustes y la técnica de validación cruzada donde n =10 la exactitud lograda es 80.05 %. Además, el modelo de búsqueda permitió hacer las consultas de todos los candidatos puesto

que se tiene un glosario de términos con 7, 198, 504 registros de todos los títulos de las páginas de Wikipedia 2017.

En la evaluación de los hipervínculos descubrimos que la falta de solidez de los enlaces de hipertexto o la incompletitud de la estructura del hipervínculo, hace que el usuario desconfié y tiene el juicio de que fue error de la máquina. Proponemos que las investigaciones adicionales se centren en los métodos de evaluación basados en el ser humano, porque al diseñar un experimento de evaluación basada en humanos, nos encontramos con varios problemas, para iniciar el hipertexto tiene que ser útil para las personas o deben saber que contenido les interesa para que se beneficien al usarlo. La interacción de un usuario con una interfaz afecta la forma en que percibe el hipertexto que lee, por esta razón la evaluación al hipertexto lo deben hacer las personas y tiene que ser según el criterio de si ayuda o no a completar sus tareas.

6.2 Trabajo futuro

Realizar pruebas con otros kernel de clasificación, por ejemplo:
-kernel linear (x, x').
-kernel polynomial.
-kernel radial.
-Métodos de evaluación basados en el ser humano.
-Evaluar estructuras de hipertexto.

Referencias

Agosti, M., Crestani, F., & Melucci, M. (1997). On the use of information retrieval techniques for the automatic construction of hypertext. *Information Processing & Management*, *33*(2), 133-144.

Alhajj, R., & Rokne, J. (Eds.). (2014). Data Mining. En *Encyclopedia of Social Network Analysis and Mining*. New York, NY: Springer New York. Recuperado a partir de http://link.springer.com/10.1007/978-1-4614-6170-8

Allan, J. (1997). Building Hypertext Using Information Retrieval. *Inf. Process. Manage.*, *33*(2), 145-159.

Ananiadou, S., Friedman, C., & Tsujii, J. (2004). Introduction: named entity recognition in biomedicine. *Journal of Biomedical Informatics*, *37*(6), 393-395. https://doi.org/10.1016/j.jbi.2004.08.011

Bellot, P., Doucet, A., Geva, S., Gurajada, S., Kamps, J., Kazai, G., … Mothe, J. (2013). Overview of INEX 2013. En *Information Access Evaluation Multilinguality, Multimodality, and Visualization* (Vol. 8138, pp. 269-281). Berlin, Heidelberg: Springer Berlin Heidelberg. Recuperado a partir de http://dx.doi.org/10.1007/978-3-642-40802-1_27

Blanco, R., Boldi, P., & Marino, A. (2015). Using graph distances for named-entity linking. *Science of Computer Programming*. https://doi.org/10.1016/j.scico.2015.10.013

Blustein, J., Webber, R. E., & Tague-Sutcliffe, J. (1997). Methods for evaluating the quality of hypertext links 1997 Information Processing Management. *Information Processing & Management*, *33*(2), 255-271.

Bordea, G., & Buitelaar, P. (2010). DERIUNLP: A context based approach to automatic keyphrase extraction. En *Proceedings of the 5th international workshop on semantic evaluation* (pp.

146–149). Association for Computational Linguistics. Recuperado a partir de http://dl.acm.org/citation.cfm?id=1859694

Borlund, P., & Ingwersen, P. (1998). Measures of relative relevance and ranked half-life: performance indicators for interactive IR. En *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 324-331). Melbourne, Australia: ACM.

Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. En *Seventh International World-Wide Web Conference*. Brisbane, Australia.

Bunescu, R. C. (2006). Using Encyclopedic Knowledge for Named Entity Disambiguation (pp. 9-16). Presentado en 11th Conference of the European Chapter of the Association for Computational Linguistics, Italy.

Büttcher, S., Clarke, C. L. A., & Cormack, G. V. (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press.

Camacho, marcela. (2015). *Detección de Fragmentos de Texto como Candidato a Hipervínculo* (Tesis de Maestría). UAEM, Tianguistenco, Edo de México.

Carbonell, J. G., Michalski, R. S., & Mitchell, T. M. (1983). An overview of machine learning. Berlin, Heidelberg: Springer-Verlag.

Cheng, X., & Roth, D. (2013). Relational inference for wikification. *Urbana*, *51*, 61801.

Covington, M. A. (1994). What is NLP? En *Natural Language Processing for Prolog Programmers* (pp. 1-2). New Yersey: Prentice Hall.

Csomai, A., & Mihalcea, R. (2007). Linking educational materials to encyclopedic knowledge. Recuperado a partir de

http://digital.library.unt.edu/ark:/67531/metadc30992/m2/1/high_res_d/Mihalcea-2007-Linking_Educational_Materials_to_Encyclopedic.pdf

Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, *7*, 708–716.

Diazaraque, J. M. M. (1998). *Análisis de Cluster y Arboles de Clasificación*. Universidad Carlos III de Madrid: Universidad Complutense de Madrid.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87.

Domingos, P., & Richardson, M. (2001). Mining the network value of customers. In: Proceedings of the seventh ACM SIGKDD international (pp. 57-66). Presentado en conference on knowledge discovery and data mining (KDD), San Francisco, California. Recuperado a partir de http://doi.acm.org/10.1145/502512.502525

Dreyfus, H. (2003). *Acerca de Internet* (Colección Nuevas Tecnologías y Sociedad). Barcelona: Editorial UOC.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification* (2nd edition). John Wiley & Sons, Inc. Recuperado a partir de http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.114.8622

Eric Haley. (1996). Exploring the Construct of Organization as Source: Consumers' Understandings of Organizational Sponsorship of Advocacy Advertising. *Journal of Advertising*, *25*(2), 19-35.

Fernández, N., Arias Fisteus, J., Sánchez, L., & López, G. (2012). IdentityRank: Named entity disambiguation in the news domain. *Expert Systems with Applications*, *39*(10), 9207-9221. https://doi.org/10.1016/j.eswa.2012.02.084

Fukunaga, K. (2009). *Introduction to statistical pattern recognition* (2. ed., [reprint]). San Diego: Academic Press.

Gelbukh, A. (2017, junio 13). Procesamiento del lenguaje natural: estado de la investigación. Centro de Investigación en Computación, Instituto Politécnico Nacional. Recuperado a partir de http://iibi.unam.mx/publicaciones/217/organizacion_del_conocimiento_24_gelbukh_alexander.html

Giuliano, C., Lavelli, A., & Romano, L. (2007). Relation extraction and the influence of automatic named-entity recognition. *ACM Transactions on Speech and Language Processing*, *5*(1), 1-26. https://doi.org/10.1145/1322391.1322393

Gómez-Adorno, H., Sidorov, G., Pinto, D., & Vilarino, D. (2014). Automatic Linguistic Pattern Identification Based on Graph Text Representation. *Research in Computing Science*, *71*. Recuperado a partir de https://www.researchgate.net/profile/Helena_Gomez_Adorno/publication/267211521_Automatic_Linguistic_Pattern_Identification_Based_on_Graph_Text_Representation/links/5447dd440cf22b3c14e28273.pdf

Hakimov, S., & Oto, S. A. (2012). Named Entity Recognition and Disambiguation using Linked Data and Graph-based Centrality Scoring. *In Proceedings of the 4th International Workshop on Semantic Web Information Management*, 1-7.

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., … Weikum, G. (2011). Robust disambiguation of named entities in text. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 782–792). Association for Computational Linguistics. Recuperado a partir de http://dl.acm.org/citation.cfm?id=2145521

Huang, D. W. C., Xu, Y., Trotman, A., & Geva, S. (2007). Overview of INEX 2007 Link the Wiki Track. *Faculty of Information Technology, Queensland University of Technology, Brisbane Queensland Australia. Springer-Verlag Berlin Heidelberg*.

Huang, H., Cao, Y., Huang, X., Ji, H., & Lin, C.-Y. (2014). Collective Tweet Wikification based on Semi-supervised Graph Regularization. En *ACL (1)* (pp. 380–390). Baltimore, Maryland, USA: Association for Computational Linguistics. Recuperado a partir de http://www.aclweb.org/website/old_anthology/P/P14/P14-1036.pdf

IBM. (2017). Modelos de árboles de decisión. Recuperado a partir de https://www.ibm.com/support/knowledgecenter/es/SS3RA7_18.1.0/modeler_mainhelp_client_ddita/clementine/nodes_treebuilding.html

Jin, Y., Kıcıman, E., Wang, K., & Loynd, R. (2014). Entity linking at the tail: sparse signals, unknown entities, and phrase models (pp. 453-462). ACM Press. https://doi.org/10.1145/2556195.2556230

Kim, S., Banchs, R. E., & Li, H. (2015). Wikification of Concept Mentions within Spoken Dialogues Using Domain Constraints from Wikipedia. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, *17*(21), 2225–2229.

Kowalski, G. (1997). *Information retrieval systems: theory and implementation*. Boston: Kluwer Academic.

Lahiri, S., Choudhury, S. R., & Caragea, C. (2014). Keyword and keyphrase extraction using centrality measures on collocation networks. *arXiv preprint arXiv:1401.6571.*

Landow, G. P. (1995). Definición del hipertexto y su historia como concepto. En *Hipertexto: la convergencia de la teoría crítica, contemporánea y la tecnología* (pp. 13-49). Barcelona: Paidós Ibérica.

Lee, K. C., Lee, S., & Hwang, Y. (2014). The impact of hyperlink affordance, psychological reactance, and perceived business tie on trust transfer. *Computers in Human Behavior*, *30*, 110-120.

León, A. J. S. (2015). *Sistema de Construcción Automática de Hipervínculos Independiente del Lenguaje* (Tesis de Licenciatura). Universidad Autónoma del Estado de México, Tianguistenco, Edo de México.

López, dania M. O., & Gómez, M. C. S. (2006). Técnicas de recolección de datos en entornos virtuales más usadas en la investigación cualitativa. *Revista de Investigación Educativa*, *24*, 205-222.

Lopez, P., & Romary, L. (2010). HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID. En *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL* (pp. 248–251). Uppsala, Sweden.

López, R. E., Barreda, D., Tejada, J., & Cuadros, E. (2011). MFSRank: An unsupervised method to extract keyphrases using semantic information. *Advances in artificial intelligence*, (Springer Berlin Heidelberg), 338-344.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval* (Cambridge University Press). New York, NY, USA. Recuperado a partir de https://nlp.stanford.edu/IR-book/

Martínez, T. I. S. (2005). *TAKING ADVANTAGE OF EXISTING NAMED ENTITY TAGGERS BY MACHINE LEARNING*. The National Institute of Astrophysics, Optics and Electronics, INAOE.

Merlino-Santesteban, C. (2003). Análisis de conectividad en la recuperación de información web. *Ciência dóna Informação*, *32*(3), 113–119.

Mihalcea, R. (2007). Wikify!: linking documents to encyclopedic knowledge. En *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 233–242). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=1321475

Milne, D., & Witten, I. H. (2008). Learning to link with wikipedia. En *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 509–518). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=1458150

Morgan, M. B. H., & Keulen, M. van. (2013). Named Entity Extraction and Disambiguation: The Missing Link. *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval*.

Moro, A., & Navigli, R. (2015). SemEval-2015 task 13: multilingual all-words sense disambiguation and entity linking. *Proc. of SemEval*, 288–297.

Ocaña, F. A., & García, A. del M. (2012). Modelo de gestión de documentos docentes en un centro universitario, basado en hipervínculos. *RUSC. Universities and Knowledge Society Journal*. Recuperado a partir de http://www.redalyc.org/articulo.oa?id=78023425011

Pérez, C. R. S. (2008). *Clasificación de Entidades Nombradas utilizando Información Global* (Maestría). INAOE, Tonantzintla, Puebla. Recuperado a partir de https://ccc.inaoep.mx/~mmontesg/tesis%20estudiantes/TesisMaestria-CarolinaSanchez.pdf

Pilgrim, M. (2010). *HTML5: up and running* (First Edition). Sebastopol, CA: O'Reilly.

Quinlan, J. R. (1986). La inducción de árboles de decisión. *Machine Learning*, *1*, 81-106.

Quinlan, J. R. (1993). C4.5: Programas para el aprendizaje automático. En *Morgan Kaufmann*. Los Altos.

Ramos, S. T. (2006). *Aprendizaje Supervisado de Colocaciones para la Resolución de la Ambigüedad Sintáctica* (Tesis de Maestría). Instituto Politécnico Nacional, México, D.F.

Rao, D., McNamee, P., & Dredze, M. (2013). Entity linking: Finding extracted entities in a knowledge base. En *Multi-source, Multilingual Information Extraction and Summarization* (pp. 93–115).

Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-642-28569-1_5

Saha, S., & Ekbal, A. (2013). Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. *Data & Knowledge Engineering*, *85*, 15-39. https://doi.org/10.1016/j.datak.2012.06.003

Salton, G., Fox, E. A., & Wu, H. (1983). Extended Boolean Information Retrieval. *Communications of the ACM*, *26*(12), 1022-1036.

Sil, A. (2013a). Exploring re-ranking approaches for joint named-entity recognition and linking (pp. 11-18). ACM Press. https://doi.org/10.1145/2513166.2513177

Sil, A. (2013b). Re-ranking for Joint Named-Entity Recognition and Linking. *Association for Computing Machinery*. Recuperado a partir de http://dx.doi.org/10.1145/2505515.2505601.

Song, Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

Sosa, E. (1997). Procesamiento del lenguaje natural: revisión del estado actual, bases teóricas y aplicaciones (Parte I). *Revista nacional científica y profesional*. Recuperado a partir de http://www.elprofesionaldelainformacion.com/contenidos/1997/enero/procesamiento_del_lenguaje_natural_revisin_del_estado_actual_bases_tericas_y_aplicaciones_parte_i.html

Stewart, K. J. (2006). How Hypertext Links Influence Consumer Perceptions to Build and Degrade Trust Online. *Journal of Management Information Systems*, *23*(1), 183–210.

Stewart, K. J., & Zhang, Y. (2003). Effects Of Hypertext Links On Trust Transfer. *Proceedings of the 5th international conference on Electronic commerce (ICEC '03)*, 235-239.

Theodoridis, S., & Koutroumbas, K. N. (2003). *Pattern Recognition* (Second Edition). USA: Elsevier.

Treeratpituk, P., Teregowda, P., Huang, J., & Giles, C. L. (2010). Seerlab: A system for extracting key phrases from scholarly documents. En *Proceedings of the 5th international workshop on semantic evaluation* (pp. 182–185). Association for Computational Linguistics. Recuperado a partir de http://dl.acm.org/citation.cfm?id=1859703

Varma, V., Praveen Bysani, & Kranthi Reddy. (2009). IIIT Hyderabad at TAC 2009. *Proceedings of the Text Analysis Conference*.

Webb, A. R. (2004). *Statistical pattern recognition* (2. ed., reprint). Chichester: Wiley.

Wikipedia. (2015a, abril 8). Wikipedia. En *Wikipedia, la enciclopedia libre*. Recuperado a partir de http://es.wikipedia.org/w/index.php?title=Wikipedia&oldid=81269788

Wikipedia. (2015b, mayo 24). Wikipedia:Usuarios muy activos. En *Wikipedia, la enciclopedia libre*. Recuperado a partir de http://es.wikipedia.org/w/index.php?title=Wikipedia:Usuarios_muy_activos&oldid=82695306

Xie, I. (2008). Interactive IR Models. *Interactive Information Retrieval in Digital Environments*, 183-214. https://doi.org/10.4018/978-1-59904-240-4.ch007

Yang, Y., & Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. *Proceedings of the 14th International Conference on Machine Learning*.

Zongker, D., & Jain, A. (1996). Algorithms for feature selection: An evaluation (pp. 18-22 vol.2). IEEE. https://doi.org/10.1109/ICPR.1996.546716

Anexo 1.

Etiquetado POS (Part of Speech)

Etiqueta	Descripción	Ejemplos
ao0000	Adjetivo (ordinal)	*primera*, *segundo*, *últimos*
aq0000	Adjetivo (descriptivo)	*populares*, *elegido*, *emocionada*, *andaluz*
conjunciones		
cc	Conjunción (coordinativa)	*y*, *o*, *pero*
cs	Conjunción (subordinante)	*que*, *como*, *mientras*
determinadores		
da0000	Artículo (definido)	*el*, *la*, *los*, *las*
dd0000	Demostrativo	*este*, *esta*, *esos*
de0000	"Exclamativo" (TODO)	*qué* (*¡Qué pobre!*)
di0000	Artículo (indefinido)	*un*, *muchos*, *todos*, *otros*
dn0000	Número	*tres*, *doscientas*
dp0000	Posesivo	*sus*, *mi*
dt0000	Interrogativo	*cuántos*, *qué*, *cuál*
Puntuación		
f0	Otro	*&*, *@*
faa	Signo de exclamación invertida	*¡*
fat	Signo de exclamación	*!*
fc	Coma	*,*
fd	Colon	*:*
fe	Comilla doble	*"*
fg	Guión	*-*
fh	Barra inclinada	*/*
fia	Signo de interrogación invertido	*¿*
fit	Signo de interrogación	*?*
fp	Período / parada completa	*.*
fpa	Paréntesis izquierdo	*(*
fpt	Paréntesis derecho	*)*
fs	Elipsis	*..., etcétera*

ft	Signo de porcentaje	*%*
fx	Punto y coma	*;*
fz	Comilla simple	'
interjecciones		
i	Interjección	*ay*, *ojalá*, *hola*
Sustantivos		
nc00000	Sustantivo común desconocido (neologismo, palabra de préstamo)	*minidisc*, *hooligans*, *re-flotamiento*
nc0n000	Sustantivo común (número invariante)	*hipótesis*, *campus*, *golf*
nc0p000	Sustantivo común (plural)	*años*, *elecciones*
nc0s000	Sustantivo común (singular)	*lista*, *hotel*, *partido*
np00000	Nombre propio	*Málaga*, *Parlamento*, *UFINSA*
pronombres		
p0000000	Impersonal se	*se*
pd000000	Pronombre demostrativo	*éste*, *eso*, *aquellas*
pe000000	"Exclamativo" Pronombre	*qué*
pi000000	Indefinido Pronombre	*muchos*, *uno*, *tanto*, *nadie*
pn000000	Numeral Pronombre	*dos miles*, *ambos*
pp000000	Personal Pronombre	*ellos*, *lo*, *la*, *nos*
pr000000	Relativo Pronombre	*que*, *quien*, *donde*, *cuales*
pt000000	Interrogativo Pronombre	*cómo*, *cuánto*, *qué*
px000000	Posesivo Pronombre	*tuyo*, *nuestra*
Adverbios		
rg	Adverbio (general)	*siempre*, *más*, *personalmente*
rn	Adverbio (negación)	*no*
Preposiciones		
sp000	Preposición	*en*, *de*, *entre*
verboos		
vag0000	Verbo (Auxiliar, gerundio)	*habiendo*
vaic000	Verbo (Auxiliar, Indicativo, condicional)	*habría*, *habríamos*
vaif000	Verbo (Auxiliar, Indicativo, futuro)	*habrá*, *habremos*
vaii000	Verbo (Auxiliar, Indicativo, imperfecto)	*había*, *habíamos*

vaip000	Verbo (Auxiliar, Indicativo, presente)	*ha*, *hemos*
vais000	Verbo (Auxiliar, Indicativo, preterito)	*hubo*, *hubimos*
vam0000	Verbo (Auxiliar, imperativo)	*haya*
van0000	Verbo (Auxiliar, infinitivo)	*haber*
vap0000	Verbo (Auxiliar, participio)	*habido*
vasi000	Verbo (Auxiliar, subjuntivo, imperfecto)	*hubiera*, *hubiéramos*, *hubiese*
vasp000	Verbo (Auxiliar, subjuntivo, presente)	*haya*, *hayamos*
vmg0000	Verbo (principal, gerundio)	*dando*, *trabajando*
vmic000	Verbo (principal, Indicativo, condicional)	*daría*, *trabajaríamos*
vmif000	Verbo (principal, Indicativo, futuro)	*dará*, *trabajaremos*
vmii000	Verbo (principal, Indicativo, imperfecto)	*daba*, *trabajábamos*
vmip000	Verbo (principal, Indicativo, presente)	*da*, *trabajamos*
vmis000	Verbo (principal, Indicativo, preterito)	*dio*, *trabajamos*
vmm0000	Verbo (principal, imperativo)	*da*, *dé*, *trabaja*, *trabajes*, *trabajemos*
vmn0000	Verbo (principal, infinitivo)	*dar*, *trabjar*
vmp0000	Verbo (principal, participio)	*dado*, *trabajado*
vmsi000	Verbo (principal, subjuntivo, imperfecto)	*diera*, *diese*, *trabajáramos*, *trabajésemos*
vmsp000	Verbo (principal, subjuntivo, presente)	*dé*, *trabajemos*
vsg0000	Verbo (semiAuxiliar, gerundio)	*siendo*
vsic000	Verbo (semiAuxiliar, Indicativo, condicional)	*sería*, *serían*
vsif000	Verbo (semiAuxiliar, Indicativo, futuro)	*será*, *seremos*
vsii000	Verbo (semiAuxiliar, Indicativo, imperfecto)	*era*, *éramos*
vsip000	Verbo (semiAuxiliar, Indicativo, presente)	*es*, *son*

vsis000	Verbo (semiAuxiliar, Indicativo, preterito)	*fue*, *fuiste*
vsm0000	Verbo (semiAuxiliar, imperativo)	*sea*, *sé*
vsn0000	Verbo (semiAuxiliar, infinitivo)	*ser*
vsp0000	Verbo (semiAuxiliar, participio)	*sido*
vssf000	Verbo (semiAuxiliar, subjuntivo, futuro)	*fuere*
vssi000	Verbo (semiAuxiliar, subjuntivo, imperfecto)	*fuera*, *fuese*, *fuéramos*
vssp000	Verbo (semiAuxiliar, subjuntivo, presente)	*sea*, *seamos*
fechas		
w	Fecha	*octubre*, *jueves*, *2002*
numerales		
z0	Número	*547.000*, *04*, *52,52*
zm	Número calificador (moneda)	*dólares*, *euros*
zu	Número calificador (Otras unidades)	*km*, *cc*

Anexo 2.

Regex FORMATEXT

(?m)(.*).html
^(.*)<h1 (.*)>(.*)</h1>
¬<h1> $3 </h1>¬
^(.*) (.*)
$2
^ (.+) $
¬ <o> $1 </o> ¬
^[< /@/"/&\t\w (].*

¬

\&amp\;
\&
\<
@MENQ
\>
@MAYQ
@MENQ
<
@MAYQ
>
\,
@COMA
\@COMA
\,
\.
@PUNTO
\@PUNTO
\.
\;
@PUNTOCOMA
\@PUNTOCOMA
\;
\:
@DOSPUNTOS
\@DOSPUNTOS
\:
\(
@PBR
\@PBR
\(
\)
@PCI
\@PCI
\)
\[
@CBR
\@CBR
\[
\]
@CCI
\@CCI
\]
\"
@COMILL
\"
@COMILL
\@COMILL
\"
\?
@INTFIN
\@INTFIN
\?
\¿
@INTINI
\@INTINI
\¿
\!
@ADMFIN
\@ADMFIN
\!
\¡
@ADMINI
\@ADMINI
\¡

Ã³
o
Ã¡
a
Ã-
i
Ãº
u
Ã©
e
.C3.B3

o Ã±	' Â«	Ã\?
ñ	"	a
|	Â»	(.*)Redirecting to(.*)
	"	
â€œ	«	
"	"	
â€˜	»	
'	"	
â€™	ÅŒ	\<sup(.*)\>
'	o	
â€?		\<\/td\>
"	a. C.	
Ã§	Â¿	<a name=(.*)
c	¿	
Ã¨	Ã‰	a\. C\.\;a\.a\. C\.\;C
e	e	a\.C
Ã¨	(.*)class="image"(.*)	
e		\s+
Â¨	<span class(.*)	
"		
Ã¨	^<p><a name(.*)	
e		
Â¨	<br style(.*)	
"		
Âº	<small>	
º		
Ã¨	</small>	
e		
Â¨		
"		
Âº	</sup>	
º		
Â´		

Anexo 3.

Atributos sin ganancia de información

Token	a1	a2	a3	a4	a5	a5	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16	a17
inusitada	0	0	0	0	108	3257	0	32	4474	0	0	1	0	0	0	1	1
las	0	0	0	0	108	3258	0	32	4474	0	0	1	0	0	0	1	0
listas	0	0	0	0	108	3259	0	32	4474	0	0	1	0	0	0	1	1
de	0	0	0	0	108	3260	0	32	4474	0	0	1	0	0	0	3	0
senadores	0	0	0	0	108	3261	0	32	4474	0	0	0	0	1	1	1	0
y	0	0	0	0	108	3262	0	32	4474	0	0	1	0	0	0	3	0
caballeros	0	0	0	0	108	3263	0	32	4474	0	0	1	0	0	0	1	1
,	0	0	0	0	108	3264	0	32	4474	0	0	1	0	0	0	3	1
expulsando	0	0	0	0	108	3265	0	32	4474	0	0	1	0	0	0	1	0
de	0	0	0	0	108	3266	0	32	4474	0	0	1	0	0	0	1	0
Su	0	0	0	0	108	3267	0	32	4474	0	0	1	0	0	0	1	1
orden	0	0	0	0	108	3268	0	32	4474	0	0	1	0	0	0	2	0
social	0	0	0	0	108	3269	0	32	4474	0	0	1	0	0	0	1	1
a	0	0	0	0	108	3270	0	32	4474	0	0	1	0	0	0	1	1
aquellos	0	0	0	0	108	3271	0	32	4474	0	0	1	0	0	0	3	1
a	0	0	0	0	108	3272	0	32	4474	0	0	1	0	0	0	1	0
los	0	0	0	0	108	3273	0	32	4474	0	0	1	0	0	0	1	1
que	0	0	0	0	108	3274	0	32	4474	0	0	1	0	0	0	1	1
consideraba	0	0	0	0	108	3275	0	32	4474	0	0	1	0	0	0	1	1
que	0	0	0	0	108	3276	0	32	4474	0	0	1	0	0	0	1	0
no	0	0	0	0	108	3277	0	32	4474	0	0	1	0	0	0	1	1
eran	0	0	0	0	108	3278	0	32	4474	0	0	1	0	0	0	1	0

Anexo 4.

Títulos y Urls de "Linux"

	Título	URL
1	Linux HA	https://es.wikipedia.org/wiki/Linux_HA
2	GNU Linux	https://es.wikipedia.org/wiki/GNU_Linux
3	Arch Linux	https://es.wikipedia.org/wiki/Arch_Linux
4	Linux Libre	https://es.wikipedia.org/wiki/Linux_Libre
5	Portabilidad del núcleo Linux y arquitecturas soportadas	https://es.wikipedia.org/wiki/Portabilidad_del_núcleo_Linux_y_arquitecturas_soportadas
6	Proceso de arranque en linux	https://es.wikipedia.org/wiki/Proceso_de_arranque_en_linux
7	Disputas de sco sobre linux	https://es.wikipedia.org/wiki/Disputas_de_sco_sobre_linux
8	PXES Universal Linux Thin Client	https://es.wikipedia.org/wiki/PXES_Universal_Linux_Thin_Client
9	Softlanding Linux System (SLS)	https://es.wikipedia.org/wiki/Softlanding_Linux_System_(SLS)
10	Alt Linux	https://es.wikipedia.org/wiki/Alt_Linux
11	Amber Linux	https://es.wikipedia.org/wiki/Amber_Linux
12	ArchBang Linux	https://es.wikipedia.org/wiki/ArchBang_Linux
13	Arquitectura de Sonido	https://es.wikipedia.org/wiki/Arquitectura_de_Sonido_Avanzada_para_Linux

	Avanzada para Linux	
14	Arranque remoto sin disco en linux	https://es.wikipedia.org/wiki/Arranque_remoto_sin_disco_en_linux
15	Linux	https://es.wikipedia.org/wiki/Linux

Anexo 5.

Títulos y Urls de "sistema informático"

	Título	URL
1	Sistema informático	https://es.wikipedia.org/wiki/Sistema_informático
2	Sistema operativo Windows	https://es.wikipedia.org/wiki/Sistema_operativo_Windows
3	Sistema operativo web	https://es.wikipedia.org/wiki/Sistema_operativo_web
4	Sistema Operativo Robótico	https://es.wikipedia.org/wiki/Sistema_Operativo_Robótico
5	Sistema (informatica)	https://es.wikipedia.org/wiki/Sistema_(informatica)
6	Sistema operativo	https://es.wikipedia.org/wiki/Sistema_operativo
7	Sistema (informática)	https://es.wikipedia.org/wiki/Sistema_(informática)
8	Sistema de informacion	https://es.wikipedia.org/wiki/Sistema_de_informacion
9	Sistema operativo multiusuario	https://es.wikipedia.org/wiki/Sistema_operativo_multiusuario
10	Sistema Xbox One	https://es.wikipedia.org/wiki/Sistema_Xbox_One
11	Sistema X Window	https://es.wikipedia.org/wiki/Sistema_X_Window
12	Sistema abierto (informatica)	https://es.wikipedia.org/wiki/Sistema_abierto_(informatica)

13	Sistema de información	https://es.wikipedia.org/wiki/Sistema_de_información
14	Sistema abierto (informática)	https://es.wikipedia.org/wiki/Sistema_abierto_(informática)
15	Sistema informatico	https://es.wikipedia.org/wiki/Sistema_informatico

Anexo 6.

Búsqueda de documentos

https://es.wikipedia.org/wiki/ABC_GNU_Linux	0.5
https://es.wikipedia.org/wiki/Adopción_de_Linux	0.38095238
https://es.wikipedia.org/wiki/Adopcion_de_Linux	0.4
https://es.wikipedia.org/wiki/Advanced_Linux_Sound_Architecture	0.22222222
https://es.wikipedia.org/wiki/Alpine_Linux	0.53333333
https://es.wikipedia.org/wiki/Alt_linux	0.5
https://es.wikipedia.org/wiki/ALT_Linux	0.66666667
https://es.wikipedia.org/wiki/Alt_Linux	0.66666667
https://es.wikipedia.org/wiki/Amber_Linux	0.57142857
https://es.wikipedia.org/wiki/Anexo:Comparación_de_distribuciones_GNU_Linux	0.16326531
https://es.wikipedia.org/wiki/Anexo:Comparación_de_distribuciones_Linux	0.17777778
https://es.wikipedia.org/wiki/Anexo:Comparación_de_LiveDistros_de_Linux	0.17777778
https://es.wikipedia.org/wiki/Anexo:Comparativa_de_distribuciones_Linux	0.18181818
https://es.wikipedia.org/wiki/Anexo:Comparativa_de_secuenciadores_para_Linux	0.16326531
https://es.wikipedia.org/wiki/Anexo:Distribuciones_GNU_Linux	0.24242424
https://es.wikipedia.org/wiki/Anexo:Distribuciones_GNU_Linux_de_España	0.18181818
https://es.wikipedia.org/wiki/Anexo:Distribuciones_GNU_Linux_españolas	0.18181818
https://es.wikipedia.org/wiki/Anexo:Distribuciones_Linux	0.27586207
https://es.wikipedia.org/wiki/Anexo:Lanzamientos_de_Linux_Mint	0.22857143
https://es.wikipedia.org/wiki/Anexo:Lista_de_lanzamientos_de_Linux_Mint	0.18181818
https://es.wikipedia.org/wiki/Arch_linux	0.46153846
https://es.wikipedia.org/wiki/Arch_Linux	0.61538462
https://es.wikipedia.org/wiki/ArchBang_Linux	0.47058824
https://es.wikipedia.org/wiki/Arquitectura_de_Sonido_Avanzada_para_Linux	0.17777778
https://es.wikipedia.org/wiki/Arranque_remoto_sin_disco_en_linux	0.16216216
https://es.wikipedia.org/wiki/Arranque_remoto_sin_disco_en_Linux	0.21621622
https://es.wikipedia.org/wiki/Automotive_Grade_Linux	0.32
https://es.wikipedia.org/wiki/Bayanihan_linux	0.33333333
https://es.wikipedia.org/wiki/Bayanihan_Linux	0.44444444
https://es.wikipedia.org/wiki/Beakos_GNU_Linux	0.42105263
https://es.wikipedia.org/wiki/Big_linux	0.5
https://es.wikipedia.org/wiki/Big_Linux	0.66666667

https://es.wikipedia.org/wiki/Black_Cat_Linux	0.44444444
https://es.wikipedia.org/wiki/Black_Lab_Linux	0.44444444
https://es.wikipedia.org/wiki/Blag_linux_and_gnu	0.28571429
https://es.wikipedia.org/wiki/BLAG_Linux_and_GNU	0.38095238
https://es.wikipedia.org/wiki/Bodhi_Linux	0.57142857
https://es.wikipedia.org/wiki/Bonzai_Linux	0.53333333
https://es.wikipedia.org/wiki/CAINE_Linux	0.57142857
https://es.wikipedia.org/wiki/Calculate_Linux	0.44444444
https://es.wikipedia.org/wiki/Canaima_GNU_Linux	0.4
https://es.wikipedia.org/wiki/Canaima_GNU_Linux	0.4
https://es.wikipedia.org/wiki/Chakra_GNU_Linux	0.42105263
https://es.wikipedia.org/wiki/Chakra_Linux	0.53333333
https://es.wikipedia.org/wiki/Clones_de_Red_Hat_Enterprise_Linux	0.21621622
https://es.wikipedia.org/wiki/College_Linux	0.5
https://es.wikipedia.org/wiki/Controversía_por_el_nombre_de_Linux	0.20512821
https://es.wikipedia.org/wiki/Controversia_por_el_nombre_de_Linux	0.21052632
https://es.wikipedia.org/wiki/Controversia_por_la_denominación_GNU_Linux	0.17391304
https://es.wikipedia.org/wiki/Controversia_por_la_denominación_GNU_Linux	0.17391304
https://es.wikipedia.org/wiki/Controversia_por_la_denominacion_GNU_Linux	0.17777778
https://es.wikipedia.org/wiki/Controversia_por_la_denominacion_GNU_Linux	0.17777778
https://es.wikipedia.org/wiki/Cooperative_Linux	0.4
https://es.wikipedia.org/wiki/Corel_Linux	0.57142857
https://es.wikipedia.org/wiki/Coyote_linux	0.4
https://es.wikipedia.org/wiki/Coyote_Linux	0.53333333
https://es.wikipedia.org/wiki/Cray_Linux_Environment	0.32
https://es.wikipedia.org/wiki/CrunchBang_Linux	0.42105263
https://es.wikipedia.org/wiki/Damn_Small_Linux	0.42105263
https://es.wikipedia.org/wiki/Debian_GNU_Linux	0.42105263
https://es.wikipedia.org/wiki/Debian_GNU_Linux	0.42105263
https://es.wikipedia.org/wiki/DeLi_Linux	0.61538462
https://es.wikipedia.org/wiki/Desktop_Light_Linux	0.36363636
https://es.wikipedia.org/wiki/Desktop_Linux_Consortium	0.2962963
https://es.wikipedia.org/wiki/Diskless_Remote_Boot_in_Linux	0.25
https://es.wikipedia.org/wiki/Disputa_sobre_la_autoría_de_Linux	0.21621622
https://es.wikipedia.org/wiki/Disputa_sobre_la_autoria_de_Linux	0.22222222
https://es.wikipedia.org/wiki/Disputas_de_sco_sobre_linux	0.2
https://es.wikipedia.org/wiki/Disputas_de_SCO_sobre_Linux	0.26666667
https://es.wikipedia.org/wiki/Distribción_GNU_Linux	0.32
https://es.wikipedia.org/wiki/Distribución_de_GNU_Linux	0.27586207
https://es.wikipedia.org/wiki/Distribución_de_GNU_Linux	0.27586207

https://es.wikipedia.org/wiki/Distribución_de_linux	0.24
https://es.wikipedia.org/wiki/Distribución_de_Linux	0.32
https://es.wikipedia.org/wiki/Distribución_GNU_Linux	0.30769231
https://es.wikipedia.org/wiki/Distribución_GNU_Linux	0.30769231
https://es.wikipedia.org/wiki/Distribución_linux	0.27272727
https://es.wikipedia.org/wiki/Distribución_Linux	0.36363636
https://es.wikipedia.org/wiki/Distribucion_de_GNU_Linux	0.28571429
https://es.wikipedia.org/wiki/Distribucion_de_GNU_Linux	0.28571429
https://es.wikipedia.org/wiki/Distribucion_de_linux	0.25
https://es.wikipedia.org/wiki/Distribucion_de_Linux	0.33333333
https://es.wikipedia.org/wiki/Distribucion_GNU_Linux	0.32
https://es.wikipedia.org/wiki/Distribucion_GNU_Linux	0.32
https://es.wikipedia.org/wiki/Distribucion_linux	0.28571429
https://es.wikipedia.org/wiki/Distribucion_Linux	0.38095238
https://es.wikipedia.org/wiki/Distribuciones_de_Linux	0.30769231
https://es.wikipedia.org/wiki/Distribuciones_GNU_Linux_españolas	0.21052632
https://es.wikipedia.org/wiki/Distribuciones_GNU_Linux_españolas	0.21052632
https://es.wikipedia.org/wiki/Distribuciones_GNU_Linux_espanolas	0.21621622
https://es.wikipedia.org/wiki/Distribuciones_GNU_Linux_espanolas	0.21621622
https://es.wikipedia.org/wiki/Distribuciones_Linux	0.34782609
https://es.wikipedia.org/wiki/Dizinha_Linux	0.5
https://es.wikipedia.org/wiki/Enoch_Linux	0.57142857
https://es.wikipedia.org/wiki/Escritorio_linux	0.31578947
https://es.wikipedia.org/wiki/Escritorio_Linux	0.42105263
https://es.wikipedia.org/wiki/EterTICs_GNU_Linux	0.38095238
https://es.wikipedia.org/wiki/Eurielec_linux	0.35294118
https://es.wikipedia.org/wiki/Eurielec_Linux	0.47058824
https://es.wikipedia.org/wiki/Familiar_Linux	0.47058824
https://es.wikipedia.org/wiki/Feather_Linux	0.5
https://es.wikipedia.org/wiki/Fedora_linux	0.4
https://es.wikipedia.org/wiki/Fedora_Linux	0.53333333
https://es.wikipedia.org/wiki/Flash_Linux	0.57142857
https://es.wikipedia.org/wiki/Foresight_Linux	0.44444444
https://es.wikipedia.org/wiki/Formatos_de_paquetes_en_GNU_Linux	0.22222222
https://es.wikipedia.org/wiki/Formatos_de_paquetes_en_GNU_Linux	0.22222222
https://es.wikipedia.org/wiki/Formatos_de_paquetes_en_Linux	0.25
https://es.wikipedia.org/wiki/Framebuffer_de_Linux	0.34782609
https://es.wikipedia.org/wiki/Fundación_Linux	0.42105263
https://es.wikipedia.org/wiki/Fundacion_Linux	0.44444444
https://es.wikipedia.org/wiki/Funtoo_Linux	0.53333333

https://es.wikipedia.org/wiki/Galsoft_Linux	0.5
https://es.wikipedia.org/wiki/Gentoo_Linux	0.53333333
https://es.wikipedia.org/wiki/GNU_linux	0.5
https://es.wikipedia.org/wiki/GNU_Linux	0.66666667
https://es.wikipedia.org/wiki/GNU_Linux	0.66666667
https://es.wikipedia.org/wiki/Greenie_Linux	0.5
https://es.wikipedia.org/wiki/High_Availability_Linux	0.30769231
https://es.wikipedia.org/wiki/High-Availability_Linux	0.30769231
https://es.wikipedia.org/wiki/Historia_de_linux	0.3
https://es.wikipedia.org/wiki/Historia_de_Linux	0.4
https://es.wikipedia.org/wiki/HP_Linux_Imaging_and_Printing	0.25
https://es.wikipedia.org/wiki/Huayra_GNU_Linux	0.42105263
https://es.wikipedia.org/wiki/Huayra_Linux	0.53333333
https://es.wikipedia.org/wiki/HuayraMedios_GNU_Linux	0.32
https://es.wikipedia.org/wiki/Hybryde_Linux	0.5
https://es.wikipedia.org/wiki/Impi_Linux	0.61538462
https://es.wikipedia.org/wiki/Instituto_Profesional_Linux	0.26666667
https://es.wikipedia.org/wiki/Jazz_Linux	0.61538462
https://es.wikipedia.org/wiki/Jerarquía_de_directorios_en_Linux	0.21621622
https://es.wikipedia.org/wiki/Jerarquia_de_directorios_en_Linux	0.22222222
https://es.wikipedia.org/wiki/Kalango_Linux	0.5
https://es.wikipedia.org/wiki/Kali_linux	0.46153846
https://es.wikipedia.org/wiki/Kali_Linux	0.61538462
https://es.wikipedia.org/wiki/KaOS_Gnu_Linux	0.47058824
https://es.wikipedia.org/wiki/Kernel_de_linux	0.33333333
https://es.wikipedia.org/wiki/Kernel_de_Linux	0.44444444
https://es.wikipedia.org/wiki/Kernel_estándar_de_Linux	0.28571429
https://es.wikipedia.org/wiki/Kernel_estandar_de_Linux	0.2962963
https://es.wikipedia.org/wiki/Kernel_Linux	0.53333333
https://es.wikipedia.org/wiki/Kurumin_Linux	0.5
https://es.wikipedia.org/wiki/Lihuen_GNU_Linux	0.42105263
https://es.wikipedia.org/wiki/Lihuen_GNU_Linux	0.42105263
https://es.wikipedia.org/wiki/Lilo_linux	0.61538462
https://es.wikipedia.org/wiki/Linpus_Linux	0.53333333
https://es.wikipedia.org/wiki/Linpus_Linux_Lite	0.4
https://es.wikipedia.org/wiki/LINUX	0
https://es.wikipedia.org/wiki/Linux	1
https://es.wikipedia.org/wiki/Linux_(desambiguación)	0.30769231
https://es.wikipedia.org/wiki/Linux_(desambiguacion)	0.32
https://es.wikipedia.org/wiki/Linux_(detergente)	0.38095238

https://es.wikipedia.org/wiki/Linux_(kernel)	0.47058824
https://es.wikipedia.org/wiki/Linux_(núcleo)	0.44444444
https://es.wikipedia.org/wiki/Linux_(nucleo)	0.47058824
https://es.wikipedia.org/wiki/Linux_(programas)	0.4
https://es.wikipedia.org/wiki/Linux_Bangalore	0.44444444
https://es.wikipedia.org/wiki/Linux_Cooperative	0.4
https://es.wikipedia.org/wiki/Linux_Counter	0.5
https://es.wikipedia.org/wiki/Linux_Embebido	0.47058824
https://es.wikipedia.org/wiki/Linux_embebido	0.47058824
https://es.wikipedia.org/wiki/Linux_Empotrado	0.44444444
https://es.wikipedia.org/wiki/Linux_empotrado	0.44444444
https://es.wikipedia.org/wiki/Linux_en_PlayStation_3	0.32
https://es.wikipedia.org/wiki/Linux_failsafe	0.47058824
https://es.wikipedia.org/wiki/Linux_Format	0.53333333
https://es.wikipedia.org/wiki/Linux_Foundation	0.42105263
https://es.wikipedia.org/wiki/Linux_FrameBuffer	0.4
https://es.wikipedia.org/wiki/Linux_From_Scratch	0.38095238
https://es.wikipedia.org/wiki/Linux_from_Scratch	0.38095238
https://es.wikipedia.org/wiki/Linux_Game_Publishing	0.33333333
https://es.wikipedia.org/wiki/Linux_HA	0.72727273
https://es.wikipedia.org/wiki/Linux_ha	0.72727273
https://es.wikipedia.org/wiki/Linux_International	0.36363636
https://es.wikipedia.org/wiki/Linux_IPLE	0.61538462
https://es.wikipedia.org/wiki/Linux_Journal	0.5
https://es.wikipedia.org/wiki/Linux_kernel	0.53333333
https://es.wikipedia.org/wiki/Linux_Libertine	0.44444444
https://es.wikipedia.org/wiki/Linux_Libre	0.57142857
https://es.wikipedia.org/wiki/Linux_libre	0.57142857
https://es.wikipedia.org/wiki/Linux_Lite	0.61538462
https://es.wikipedia.org/wiki/Linux_Magazine	0.47058824
https://es.wikipedia.org/wiki/Linux_Mandriva	0.47058824
https://es.wikipedia.org/wiki/Linux_Mint	0.61538462
https://es.wikipedia.org/wiki/Linux_MultiMedia_Studio	0.30769231
https://es.wikipedia.org/wiki/Linux_nonfb	0.57142857
https://es.wikipedia.org/wiki/Linux_para_la_PlayStation_3	0.26666667
https://es.wikipedia.org/wiki/Linux_para_PlayStation_2	0.2962963
https://es.wikipedia.org/wiki/Linux_Professional_Institute	0.25806452
https://es.wikipedia.org/wiki/Linux_RIP	0.66666667
https://es.wikipedia.org/wiki/Linux_Software_Map	0.38095238
https://es.wikipedia.org/wiki/Linux_Standard_Base	0.36363636

https://es.wikipedia.org/wiki/Linux_Terminal_Server_Project	0.25
https://es.wikipedia.org/wiki/Linux_Ubuntu	0.53333333
https://es.wikipedia.org/wiki/Linux_Unified_Kernel	0.34782609
https://es.wikipedia.org/wiki/Linux_Virtual_Server	0.34782609
https://es.wikipedia.org/wiki/Linux_VServer	0.5
https://es.wikipedia.org/wiki/Lista_de_Distribución_Linux	0.25806452
https://es.wikipedia.org/wiki/Lista_de_Distribucion_Linux	0.26666667
https://es.wikipedia.org/wiki/Lista_de_videojuegos_en_Linux	0.25
https://es.wikipedia.org/wiki/Lunar_Linux	0.57142857
https://es.wikipedia.org/wiki/LXLE_Linux	0.61538462
https://es.wikipedia.org/wiki/Mac_on_linux	0.4
https://es.wikipedia.org/wiki/Mac_on_Linux	0.53333333
https://es.wikipedia.org/wiki/Malware_en_linux	0.31578947
https://es.wikipedia.org/wiki/Malware_en_Linux	0.42105263
https://es.wikipedia.org/wiki/Mandrake_Linux	0.47058824
https://es.wikipedia.org/wiki/Mandriva_Linux	0.47058824
https://es.wikipedia.org/wiki/Mandriva_Linux_One_Gnome	0.2962963
https://es.wikipedia.org/wiki/Mangaka_Linux	0.5
https://es.wikipedia.org/wiki/Manjaro_Linux	0.5
https://es.wikipedia.org/wiki/Maryan_Linux	0.53333333
https://es.wikipedia.org/wiki/MAX_Madrid_Linux	0.42105263
https://es.wikipedia.org/wiki/MCC_Interim_Linux	0.4
https://es.wikipedia.org/wiki/MEPIS_Linux	0.57142857
https://es.wikipedia.org/wiki/Metadistros_herramienta_GNU_Linux	0.22222222
https://es.wikipedia.org/wiki/Minidistribución_de_Linux	0.27586207
https://es.wikipedia.org/wiki/Minidistribucion_de_Linux	0.28571429
https://es.wikipedia.org/wiki/Minidistribuciones_de_GNU_Linux	0.23529412
https://es.wikipedia.org/wiki/Minidistribuciones_de_GNU_Linux	0.23529412
https://es.wikipedia.org/wiki/Minino_linux	0.4
https://es.wikipedia.org/wiki/Núcleo_de_Linux	0.42105263
https://es.wikipedia.org/wiki/Núcleo_Linux	0.5
https://es.wikipedia.org/wiki/Nero_Linux	0.61538462
https://es.wikipedia.org/wiki/Novell_Linux_Desktop	0.34782609
https://es.wikipedia.org/wiki/Nucleo_de_Linux	0.44444444
https://es.wikipedia.org/wiki/Nucleo_Linux	0.53333333
https://es.wikipedia.org/wiki/Oracle_Linux	0.53333333
https://es.wikipedia.org/wiki/Parabola_GNU_Linux	0.38095238
https://es.wikipedia.org/wiki/Portabilidad_de_Linux	0.33333333
https://es.wikipedia.org/wiki/Portabilidad_del_núcleo_Linux_y_arquitecturas_soportadas	0.13333333

https://es.wikipedia.org/wiki/Portabilidad_del_nucleo_Linux_y_arquitecturas_soportadas	0.13559322
https://es.wikipedia.org/wiki/Poseidon_Linux	0.47058824
https://es.wikipedia.org/wiki/Proceso_de_arranque_en_linux	0.19354839
https://es.wikipedia.org/wiki/Proceso_de_arranque_en_Linux	0.25806452
https://es.wikipedia.org/wiki/Progeny_Componentized_Linux	0.26666667
https://es.wikipedia.org/wiki/PS2_Linux	0.66666667
https://es.wikipedia.org/wiki/Puppy_Linux	0.57142857
https://es.wikipedia.org/wiki/PXES_Universal_Linux_Thin_Client	0.22857143
https://es.wikipedia.org/wiki/Real_time_linux	0.33333333
https://es.wikipedia.org/wiki/Reconocimiento_del_habla_en_Linux	0.22222222
https://es.wikipedia.org/wiki/Red_Flag_Linux	0.47058824
https://es.wikipedia.org/wiki/Red_Hat_Enterprise_Linux	0.2962963
https://es.wikipedia.org/wiki/Red_Hat_Linux	0.5
https://es.wikipedia.org/wiki/Red_Star_Linux	0.47058824
https://es.wikipedia.org/wiki/RedOS_Linux	0.57142857
https://es.wikipedia.org/wiki/ROCK_Linux	0.61538462
https://es.wikipedia.org/wiki/RPath_Linux	0.57142857
https://es.wikipedia.org/wiki/Rxart_Linux	0.57142857
https://es.wikipedia.org/wiki/Sabayon_linux	0.375
https://es.wikipedia.org/wiki/Sabayon_Linux	0.5
https://es.wikipedia.org/wiki/SAM_Linux	0.66666667
https://es.wikipedia.org/wiki/Scientific_Linux	0.42105263
https://es.wikipedia.org/wiki/Security_Enhanced_Linux	0.30769231
https://es.wikipedia.org/wiki/Security-Enhanced_Linux	0.30769231
https://es.wikipedia.org/wiki/Semplice_Linux	0.47058824
https://es.wikipedia.org/wiki/Shadow_Linux	0.53333333
https://es.wikipedia.org/wiki/Simple_Linux_Utility_for_Resource_Management	0.17021277
https://es.wikipedia.org/wiki/Slackware_Linux	0.44444444
https://es.wikipedia.org/wiki/SliTaz_GNU_Linux	0.42105263
https://es.wikipedia.org/wiki/SliTaz_GNU_Linux	0.42105263
https://es.wikipedia.org/wiki/SLS_Linux_(Softlanding_Linux_System_)	0.2
https://es.wikipedia.org/wiki/SLS_Linux_(Softlanding_Linux_System_)	0.2
https://es.wikipedia.org/wiki/SLS_Linux_(Softlanding_Linux_System)	0.20512821
https://es.wikipedia.org/wiki/SLS_Linux_(Softlanding_Linux_System)	0.20512821
https://es.wikipedia.org/wiki/Softlanding_Linux_System	0.2962963
https://es.wikipedia.org/wiki/Softlanding_Linux_System_(SLS)	0.24242424
https://es.wikipedia.org/wiki/Sorcerer_Linux	0.47058824
https://es.wikipedia.org/wiki/Storm_Linux	0.57142857
https://es.wikipedia.org/wiki/SUSE_Linux	0.61538462

https://es.wikipedia.org/wiki/SuSE_Linux	0.61538462
https://es.wikipedia.org/wiki/Suse_Linux	0.61538462
https://es.wikipedia.org/wiki/SUSE_Linux_Enterprise_Desktop	0.25
https://es.wikipedia.org/wiki/Symphony_Linux	0.47058824
https://es.wikipedia.org/wiki/The_Linux_Game_Tome	0.36363636
https://es.wikipedia.org/wiki/Tiny_Core_linux	0.33333333
https://es.wikipedia.org/wiki/Tiny_Core_Linux	0.44444444
https://es.wikipedia.org/wiki/TOMOYO_Linux	0.53333333
https://es.wikipedia.org/wiki/Trisquel_distribucion_linux	0.2
https://es.wikipedia.org/wiki/Trisquel_GNU_Linux	0.38095238
https://es.wikipedia.org/wiki/Trisquel_GNU_Linux	0.38095238
https://es.wikipedia.org/wiki/Tuquito_GNU_Linux	0.4
https://es.wikipedia.org/wiki/Tuquito_GNU_Linux	0.4
https://es.wikipedia.org/wiki/Tuquito_linux	0.375
https://es.wikipedia.org/wiki/Vector_Linux	0.53333333
https://es.wikipedia.org/wiki/Videojuegos_en_Linux	0.34782609
https://es.wikipedia.org/wiki/Void_Linux	0.61538462
https://es.wikipedia.org/wiki/Wando_linux	0.42857143
https://es.wikipedia.org/wiki/White_Box_Enterprise_Linux	0.27586207
https://es.wikipedia.org/wiki/Wikipedia:EterTICs_GNU_Linux	0.25806452
https://es.wikipedia.org/wiki/Wireless_tools_for_Linux	0.2962963
https://es.wikipedia.org/wiki/Yellow_Dog_Linux	0.42105263
https://es.wikipedia.org/wiki/Yggdrasil_Linux	0.44444444
https://es.wikipedia.org/wiki/Zenwalk_Linux	0.5
https://es.wikipedia.org/wiki/Linux 1.0	