Búsqueda y Recuperación de Información en Textos

Víctor Mijangos de la Cruz

I. Introducción a la Recuperación de Información

Información del curso

Nombre	Búsqueda y Recuperación de Información en Textos
Clave	
Créditos	10
Total de horas	112

Objetivo del curso

Este curso tiene como objetivo conocer y aplicar las teorías y algoritmos de los sistemas de recuperación de información en textos. Es un acercamiento a las metodologías, los métodos y algoritmos que permiten obtener información a partir de datos textuales (no estructurados) para así poder representar el conocimiento que estos contienen.

Temario

1. Introducción
   1. Definición
   2. Representación de información en textos
   3. Corpus y lingüística de corpus
2. Métodos formales
   1. Expresiones regulares
   2. Stemming y Lematización
   3. Recuperación booleana
   4. Similitud y métricas entre cadenas
3. Información en textos
   1. Teoría de la información
   2. Extracción y vinculación de términos
   3. Etiquetado NER y POS
4. Representación de significado
   1. Ontologías
   2. Modelos distribucionales
   3. Semántica latente
   4. Representaciones de documentos
5. Clasificación de textos
   1. El problema de la clasificación de textos
   2. Agrupamiento de documentos
   3. Evaluación
6. Aplicaciones específicas
   1. Sistemas de búsqueda
   2. Sistemas de recomendación
   3. Minería de opiniones

Bibliografía

Manning, C., Raghavan, P. y Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Zhai, C. y Massung, S. (2016). *Text Data Management and Anaysis: A Practical Introduction to Information Retrieval and Text Mining*. ACM Books.

Mitra, B. y Craswell, N. (2018). "An Introduction to Neural Information Retrieval". *Foundations and Trends in Information Retrieval*, Vol. 13, No. 1, pp. 1-126.

Evaluación

La evaluación del curso consistirá en la elaboración de un **proyecto final** que retome los temas vistos en el curso.

Se considerarán los siguientes puntos:

- Se desarrollará la programación de una aplicación de recuperación de información en textos.
- Se documentará la colección textual que se utilizó para dicha aplicación.
- El código de la aplicación se documentará cuidadosamente, explicando los métodos y algoritmos utilizados.
- Se podrá trabajar en equipos (máximo 3 personas).

Introducción a la recuperación de información

Información, búsqueda y recuperación

La información es de gran importancia, pues a partir de ella podemos generar **conocimiento**.

Podemos encontrar **información** en muchas fuentes; por ejemplo, en libros.

Organizar esta información nos permite:

- Poder acceder a información de interés más fácilmente.
- Reducir los tiempos de búsqueda.
- Conocer qué documentos son más relevantes para mis intereses.

Información

Una definición informa de información puede ser la siguiente:

Información

Por información podemos entender un conjunto de datos que buscan comunicar un mensaje o significado.

La información se puede transmitir por diferentes **medios**: escritos (libros, periódicos), audiovisuales (televisión, radio), digitales, etc.

Información e internet

Los medios digitales han permitido que la información sea **más accesible** y que ésta pueda **producirse más fácilmente**.

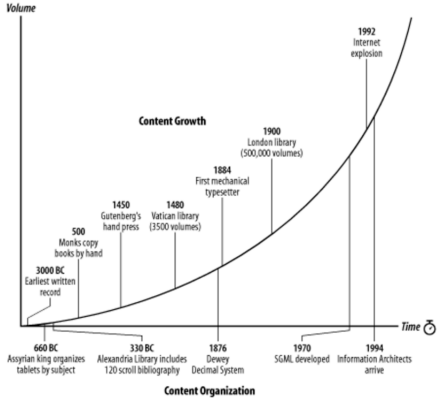
El internet y la World Wide Web son medios que contienen grandes cantidades de información.

Existen claras diferencias entre los medios tradicionales de información y los que encontramos en la web (Rosenfeld y Morville, 2002):

Concepto	**Libros**	**Sitios web**
Componentes	Cubierta, título, autor, capítulos, secciones, índice, etc.	Página principal, links, mapa del sitio, etc.
Dimensión	Páginas 2-dimensionales, secuencia lineal	Información multidimensional, hiper-textos
Límites	Finitos (material)	Intangible

Crecimiento de la información

Crear métodos que nos permitan organizar, buscar y recuperar información es sumamente importante, debido a como ha crecido la cantidad de información (Rosenfeld y Morville, 2002).



Datos estructurados y no-estructurados

Información estructurada

Hablamos de información estructurada cuando contamos con un conjunto de datos que están bien definidos y sujetos a un formato estándar.

El ejemplo más claro de la información estructurada son las **bases de datos**.

Información no-estructurada

La información no-estructurada corresponde a un conjunto de datos que no tiene una estructura clara, y que no es fácil para leer por una computadora.

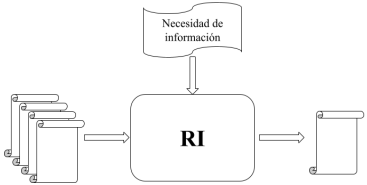
Podemos hablar de **información semi-estructurada**, cuando existe cierta estructura en parte de los datos; por ejemplo, una página web (contiene HTML y datos textuales).

Recuperación de Información

Cuando contamos con datos no-estructurada, se vuelve necesario desarrollar técnicas que permitan extraer información de ellos de manera sencilla y eficiente.

Recuperación de Información

La recuperación de información (RI) busca encontrar material de una naturaleza no-estructurada, que satisfaga una necesidad de información, dentro de una colección generalmente grande.



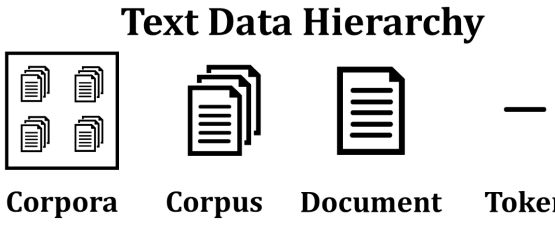
Documentos

Documentos

Entenderemos por documento a toda aquella unidad sobre la que decidamos construir un sistema de recuperación de información. Se conforma de tokens, $d = \{w_1, w_2, ..., w_n\}$.

Colección

Una colección o corpus es un conjunto de documentos, $\mathcal{C} = \{d_1, d_2, ..., d_N\}$, sobre el que aplicaremos la recuperación de información



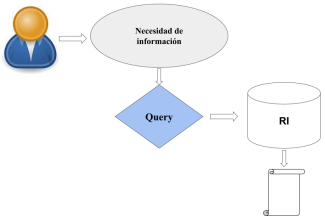
Necesidad de información y consulta

Necesidad de información

Entendemos por necesidad de información al tema o tópico sobre el que un usuario desea conocer y profundizar.

Query

Una query o consulta es aquello que el usuario transmite a la computadora para comunicar una necesidad de información; por tanto, sobre el que se hará la búsqueda.



Relevancia y efectividad

Tarea de recuperación ad hoc

Llamamos tarea de recuperación ad hoc cuando un sistema provee documentos de una colección que son **relevantes** para un usuario, en base a una query.

Relevante

Decimos que un documento es relevante si el usuario percibe que éste contiene información de valor con respecto a la necesidad de información.

Para medir la **efectividad** de un sistema, se usa:

- **Precisión:** Cuántos documentos recuperados son relevantes.
- **Exhaustividad:** (Recall) Cuántos documentos relevantes fueron recuperados.

Escalas de la recuperación de información

Según el tamaño de la colección/corpus de documentos, se puede hablar de 3 escalas a las que se realiza la recuperación de información:

- **Recuperación de información personal:** Las búsquedas se realizan sobre documentos personales, en una computadora personal. Por ejemplo, dentro de correos electrónicos, o sobre archivos almacenados en una PC.
- **Búsqueda en dominios específicos:** En este caso, las búsquedas se realizan sobre datos institucionales o empresariales, enfocadas a dominios específicos. Por ejemplo, en datos gubernamentales.
- **Búsqueda web:** Realiza búsquedas sobre colecciones muy grandes (de has miles de millones de documentos), que se encuentran en la web. Por ejemplo, buscadores web.

Grepping

Una forma común de hacer una búsqueda dentro de documentos es **escanear** el documento **linealmente** hasta encontrar los patrones que coincidad.

Una forma en que se puede hacer esto es con el comando **grep**, que refiere a **g**lobal search, **r**egular **e**xpression, y **p**rint.

Ejemplo para recuperar los documentos que contienen una query:

```
$ grep -l regex *
```

Este tipo de búsqueda puede ser problemática cuando:

1. Se tiene una colección grande de documentos que se quiere procesar rápidamente.
2. Cuando se quieren opciones más flexibles de búsqueda.

Indexación

Una alternativa a la búsqueda lineal es **indexar** los documentos.

Los **términos** pueden funcionar como índices de los documentos. Podemos crear una **matriz de incidencias**:

	$\mathbf{doc_1}$	$\mathbf{doc_2}$	...	$\mathbf{doc_N}$
$term_1$	1	0	...	0
$term_2$	1	1	...	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$term_n$	0	0	...	1

En general, la matriz de incidencia se puede definir como:

$$IncidenceMatrix_{i,j} = incidence(w_i, d_j)$$

Matriz de incidencia

La función de incidencia puede entenderse como una función binaria definida como:

$$incidence(w_i, d_j) = \begin{cases} 1 & \text{si } w_i \in d_j \\ 0 & \text{en otro caso} \end{cases}$$

Este tipo de representación tiene ventajas:

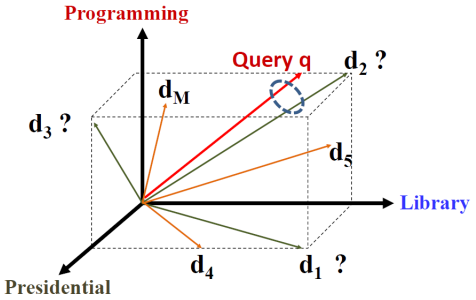
- Se crea una representación para cada documento (columna) y para cada término (renglón).
- Permite búsqueda booleanas (AND, OR, NOT).

Su desventaja es la cantidad de memoria que puede llegar a utilizar en colecciones muy grandes.

Modelo de espacio vectorial

Otra forma común es representar los términos y los documentos como vectores en un espacio d-dimensional, $\mathbb{R}^d$. Para esto, se busca una mapeo:

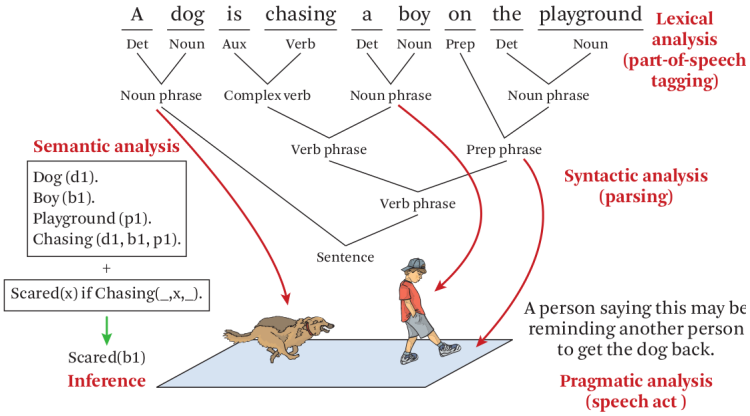
$$d_i \mapsto v_i \in \mathbb{R}^d$$



Lingüística textual y corpus

Codificación de la información

El lenguaje **codifica información** del mundo. Por eso resulta importante entender el lenguaje: su funcionamiento y las estructuras que permiten esta codificación.



Lingüística textual

La **lingüística** estudia el lenguaje humano. En particular, el lenguaje se manifiesta en forma **oral** y **textual**.

Lenguaje textual

El lenguaje textual hace referencia a la expresión del lenguaje por medio de símbolos escritos. Estos símbolos responden a un **sistema de escritura**.

El lenguaje escrito es **secundario**, en el sentido de que es una forma de codificar al lenguaje oral. El lenguaje oral es primario.

Lenguaje escrito

A la lingüística le interesa en principio el **lenguaje hablado u oral**, pues se considera como primario.
El **lenguaje escrito** busca representar al lenguaje hablado a partir de un sistema de escritura.

Sistema de escritura

Un sistema de escritura puede considerarse como un conjuntos de símbolos convencionales que representan un lenguaje.

Sistemas de escritura

Los sistemas de escritura no reflejan ni los sonidos ni las palabras que representan, y pueden ser muy variados:

Tipo	Representación	Ejemplo
Alfabeto (latino)	Busca que cada símbolo represente un sonido	La cebra comía
Fonético	Busca representar sonidos de forma exacta	/la 'se.bra ko.'mja/
Silabario	Representa sílabas o sonidos pronunciables	Ka ta ka na カタカナ
Ideográfico	Representa conceptos	

Escritura

- Muchas veces, los sistemas de escritura no representan de manera precisa los sonidos del habla (en inglés "'read"' puede leerse de dos formas distintas: como pasado o como presente).
- La **ortografía** juega un rol en la escritura. Sin embargo, un sistema de PLN debe lidiar con todo tipo de escritura.
- Ciertas manifestaciones escritas buscan denotar usos del lenguaje ("'Holaaa!!!"', "'Hola :)"',...)
- Ciertos sistemas de escritura pueden causar conflictos de codificación, o usar símbolos que en otras lenguas no son sonidos (en algunas lenguas, el apostrofe /'/ representa un sonido glotal).
- No todas las lenguas cuentan con un sistema de escritura.

Corpus

Para tratar el lenguaje escrito se toman datos empíricos para crear un modelo computacional que represente la lengua. Estos datos empíricos se presentan dentro de lo que llamamos **corpus** o colección:

Corpus

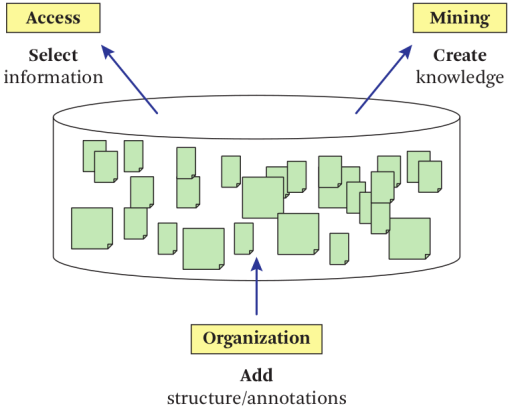
Un corpus es una recopilación bien organizada de muestras del lenguaje (documentos) a partir de materiales escritos o hablados, agrupados bajo criterios mínimos.

Un documento es una entidad d_i, $i \in \mathbb{N}$ compuesto por lenguaje escrito; es decir, determinado por un sistema de escritura y una lengua. Un corpus o colección se puede definir como:

$$\mathcal{C} = \{d_1, d_2, ..., d_N\}$$

Importancia de los corpus

Las colecciones y corpus deben ser **organizados** para poder **acceder a información** y **crear conocimiento**.



Importancia de corpus

Algunas tareas relevantes de la **extracción de información** son:

- Filtrado de información.
- Búsqueda de información en colecciones.
- Agrupamiento de documentos.

Por su parte, tareas que implican **creación de conocimiento** son:

- Análisis de tópicos.
- Visualización de la información de los textos.
- Clasificación de documentos en categorías.

Corpus digitales

Cuando hablamos de colecciones o corpus, estos pueden comprender diferentes formatos:

- Formato físico: colecciones de archivos físicos, como bibliotecas.
- Formato digital: colecciones informatizadas de archivos.

En RI nos enfocamos en documentos en formato digital. Tenemos 3 distinciones:

1. **Archivo digital**: Repertorio de textos que no tienen relación entre ellos.
2. **Biblioteca digital**: Repertorio que sigue normas de contenido y en formato estándar.
3. **Corpus digital**: Recopilación de documentos codificado en modo estándar y homogéneo, orientado a ser procesados.

Tipología de corpus textuales

Podemos distinguir varios tipos de corpus textuales según sus características:

1. Según su especificidad pueden ser **específicos** o **generales**.
2. Según su propósito pueden ser **de propósito específico** o **multipropósito**.
3. Según su tamaño, se habla de:
   1. Grande: Contienen una cantidad considerable de texto ($<$ 10 millones de palabras).
   2. Pequeño: Contienen una cantidad menor de textos ($\sim$ 1 millón de palabras).
   3. Monitor: Contiene un volumen fijo de textos que se actualiza constantemente.

Algunos datasets textuales

Nombre	Contenido	Tamaño	url
DBpedia	Artículos de Wikipedia	10 mil	http://downloads.dbpedia.org/wiki-archive/Downloads2015-10.html
TREC	Documentos para diferentes tareas de RI	28 tracks	https://trec.nist.gov/data.html
GOV2	Documentos gubernamentales	$\sim$ 25 mil docs	http://ir.dcs.gla.ac.uk/test_collections/gov2-summary.htm
Amazon Reviews	Opiniones de usuarios de Amazon	493 MB	https://www.kaggle.com/bittlingmayer/amazonreviews
IMDB Dataset	Opiniones de películas	50 mil	ai.stanford.edu/~amaas/data/sentiment/

Algunos tipos de corpus

Tipo de corpus	Subtipo	Descripción	Ejemplos
Textual	Monolingüe	Corpus de textos en una sola lengua.	CREA http://corpus.rae.es/creanet.html
	Paralelo	Corpus de textos en dos lenguas, alineados oración-oración, párrafo-párrafo.	Tsunkua https://tsunkua.elotl.mx/
	Anotado	Corpus con etiquetas que indican propiedad lingüísticas (POS, Semánticas, Clases)	CESS http://lod.iula.upf.edu/resources/project_CESS-ECE
Oral		Corpus con grabaciones o transcirpciones	Transcriber http://trans.sourceforge.net/en/presentation.php

Corpus anotados

Un corpus puede constar del texto únicamente. Pero también puede ser que un corpus tenga información metatextual. Diremos que se trata de un **corpus etiquetado**.
Generalmente, el etiquetado se corresponde con los niveles del lenguaje:

- Etiquetado fonológico. Información sobre sonidos del habla.
- Etiquetado morfológico. Muestra información sobre la estructura de las palabras.
- Etiquetado morfosintáctico. Contiene información sobre las categorías gramaticales de palabras (etiquetado POS).
- Etiquetado sintáctico. Información sobre estructura de frase y oraciones.
- Etiquetado semántico. Nivel de etiquetado con información sobre el significado de palaras, oraciones o documentos.

Etiquetado morfosintáctico y sintáctico

El **etiquetado POS** (Parts Of Speech) asigna a cada palabra su categoría gramatical:

(El, da0ms0), (grupo, ncms000), (estatal, aq0cs0), (Electricité de France, np00000), (anunció, vmis3s0), (hoy, rg), (la, da0fs0), (compra, ncfs000),..

Por su parte, el **etiquetado sintáctico** es un etiquetado jerárquico que determina las funciones dentro de las oraciones:

(O (FN (Sust Juan)) (FV (V come) (FN (Det unos) (Sust tacos))))

Etiquetado NER

El **etiquetado NER** (Named Entity Recognition) es un tipo de etiquetado semántico que busca identificar entidades: lugares, fechas, personas, instituciones, etc.

Ya que estas entidades suelen ser términos compuestos, el etiquetado NER se suele acompañar por **etiquetado BIO** (Beginning-Inside-Outside), que indica los elementos que componen un término:

(Juan, B-PER), (vive, O), (en, O), (Ciudad, B-LOC), (de, I-LOC), (México, I-LOC)

Otro tipo de etiquetados semánticos pueden ser el etiquetado de opiniones, emociones, temas, etc.

Palabras

El concepto de palabra es esencial, una definición práctica dentro de PLN es la siguiente:

Palabra

Una palabra es una cadena de caracteres entre dos espacios en blancos.

Surgen diversos **problemas** de esta definición:

- No todas las lenguas usan espacios en su escritura (chino, japonés).
- Sólo se aplica a lenguaje escrito. ¿Qué pasa con grabaciones acústicas?
- Casos conflictivos como "'dormirse"' contra "'se duerme"'.

Palabras

Otra definición de palabra puede darse a partir de los elementos que buscamos en éstas:

Palabra

Una palabra es una unidad lingüística con estructura, significado y función.

Forma	Significado	Función
gato	"Mamífero carnívoro de la familia de los félidos, digitígrado, doméstico, de unos 50 cm de largo desde la cabeza hasta el arranque de la cola, que por sí sola mide unos 20 cm, de cabeza redonda, lengua muy áspera, patas cortas y pelaje espeso, suave, de color blanco, gris, pardo, rojizo o negro, que se empleaba en algunos lugares para cazar ratones"	Sustantivo

Tipos y tokens

El concepto de palabra no es fácil de tratar por una computadora, esta puede variar según la lengua y puede encontrar problemas en los estándares ortográficos y los sistemas de escritura.

Por tanto, se suele hablar de **tipos** y **tokens**:

Token

Un token es la ocurrencia individual de una palabra dentro de un texto.

Tipo

Los tipos son los diferentes elementos lingüísticos que existen en un corpus.

El **tamaño de corpus** se suele medir en número de tokens.

Tipos y tokens

Considérese el siguiente texto:

> *El perro negro le quitó el hueso al perro amarillo.*

En este texto encontramos tipos y tokens distribuidos de la siguiente forma:

- Número de tokens $= 10$ (palabras en el texto, sin importar repetición).
- Número de tipos $= 8$ (palabras únicas en el texto).

Stopwords

En los textos suelen encontrarse palabras que aportan poca información y que muchas veces son irrelevantes para las aplicaciones de PLN. A estas palabras se les conoce como **stopwords**. Las stopwords suelen ser preposiciones, conjunciones o artículos. Estas se eliminan a partir de una lista de paro.

- Texto con stopwords:

  Homo sapiens es una especie del orden de los primates perteneciente a la familia de los homínidos. También son conocidos bajo la denominación genérica de "'humanos"'.

- Texto sin stopwords:

  Homo sapiens es especie orden primates perteneciente familia homínidos. También son conocidos denominación genérica "'humanos"'.

Concordancias

Las **concordancias** o KWIC (Key Word In Context) presentan una palabra en su contexto dentro del corpus.

El **contexto** suele ser una ventana de *n* palabras (o caracteres) a la derecha y *n* a la izquierda.

CONCORDANCIA			AÑO	AUTOR
levantando a Spencer Tracy como quien levanta un	gatito-	, aunque al final el actor lanza un "vive la d **	1994	PRENSA
énicos en la superficie de los glóbulos rojos del	gatito	los anticuerpos provocan hemólisis intra o ext **	2004	PRENSA
del donante es un paso muy importante. Aunque el	gatito	sea A o AB los únicos anticuerpos que posee so **	2004	PRENSA
ión es una fuente adecuada de anticuerpos para un	gatito	privado de calostro (Casal et al., 1996; Giger **	2004	PRENSA
nicio de la producción de IgG y IgA por parte del	gatito	es tardío se recomienda un plan de vacunación **	2004	PRENSA
.F. 1996 1996 10 505 R Tigre o	gatito	Francisco Ponce Ya hombre, ya. Ya estuvo suave **	1996	PRENSA
ivas, pero en esta ocasión no lucen como el dulce	gatito	al que todos abrirán sus puertas y verlo con m **	1997	PRENSA
ela, The actual. Según dice, el observador "es el	gatito	de la esquina". ¿Bellow un gatito? Es más prob **	1997	PRENSA
servador "es el gatito de la esquina". ¿Bellow un	gatito	? Es más probable que los lectores de este gana **	1997	PRENSA
obedo habla en su carta de su "colegui Nico", "un	gatito	precioso, muy simpático, con quien jugaba por **	1988	PRENSA
dosis de paciencia. Un perro adulto aceptará a un	gatito	sin mucha dificultad. En el caso inverso, la a **	1985	PRENSA
ón, aunque completamente inmóvil, observa cómo un	gatito	se prepara el nido en el corpachón peludo clav **	1985	PRENSA
asa? Seguro que usted también adorará a Nerón, mi	gatito	preferido. Hasta pronto. Miranda pasó de la ri **	1996	PRENSA
uchar los maullidos digitalizados de "Shocks", el	gatito	de los Clinton. Bob Dole, el candidato con más **	1995	PRENSA
cuatro patas a la calle. Yo estaba solita con mi	gatito	y los dos nos hemos salvado. Gracias a Dios mi **	2000	PRENSA
ecológico ¿Por qué? Es que Silvestre se llama el	gatito	animado que persigue al canario Piolín. "Yo ta **	1997	PRENSA
cazar su cola? Es así: Un gato grande vio cómo un	gatito	pequeño trataba de pescarse la cola y le pregu **	2003	PRENSA
qué tratas de pescarte la cola en esa forma?". El	gatito	dijo: "He aprendido que lo mejor para un gato **	2003	PRENSA
y madre extraordinaria. El logotipo y distintivo "gatito	negro con moño rojo" que figura en todo lo con **	1992	PRENSA
ditado por el sello de Eurotrack Estudios, "Lindo	gatito	". En Alicante, la sala Desafinado del Casco An **	2003	PRENSA
ó venir a Chicago y lo vacunó con un cabezazo del	gatito	Leeb en un contraataque iniciado por Cáceres d **	1997	PRENSA
lessio, de 11 años, y escultora debutante, con un	gatito	acurrucado. Cuando el sol empezó a esconderse, **	1997	PRENSA
idas secretas, pero que en aquel tiempo él era un	gatito	ciego, cegado por el amor, y no veía agujeros **	1996	PRENSA
males. A ver, Tamara. Yo yo en casa mía tengo un	gatito	y en casa de mi abuela tengo gallinas. Bueno. **	---	ORAL
, vale. Entonces me estaba diciendo, que tenía un	gatito	, un gatito marrón, tremendo. Apenas me senté y **	---	ORAL

Elaboración de corpus

El proceso de selección y organización de los textos garantiza que los métodos de RI tengan un mejor rendimiento.

Cuando se elabora un corpus, se busca que este cumpla con lo siguiente (McEnenry y Wilson, 2001):

- **Bien seleccionado:** Los documentos deben estar acorde al problema que busquemos tratar.
- **Representativo:** Debe garantizarse la variedad de documentos, de tal forma que se represente adecuadamente el problema.
- **Referencia estandarizada:** Las anotaciones y otro tipo de estructura debe hacerse bajo estándares que garanticen la reproducibilidad.
- **Principios éticos:** La curaduría de los documentos debe garantizar que no se violente o se minorice a ciertos sectores.

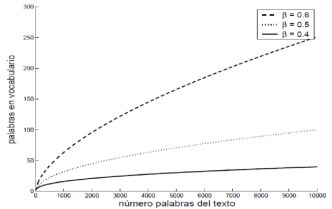
Representatividad léxica

La **representatividad léxica** responde a que se contenga un número de palabras que abarque la mayor parte del léxico utilizado en la lengua (o lenguas) del corpus.

Ley de Herdan

La ley de Herdan es una ley empírica del lenguaje que señala que existe una relación entre el número de tipos (t) y el número de tokens (N) dada por:

$$t \propto N^{1/\beta}$$



Estadísticas de corpus

Algunas estadísticas que nos dan información de la **representatividad** del corpus dependen de los **géneros** o tópicos que se manejen y cómo se distribuyen lo términos en estos.

Frecuencia relativa de término

Dado un término w_i con frecuencia absoluta t_i, su frecuencia relativa está determinada por:

$$f_i = \frac{t_i}{\sum_{k=1}^{n} t_k}$$

Tamaño relativo de género

El tamaño relativo del género g_j determina la cantidad de términos en este género y se calcula como:

$$r_j = \frac{\sum_{i=1}^{n} f_{ij}}{\sum_{k=1}^{n} t_k}$$

donde f_{ij} denota las frecuencia absoluta del término w_i en el género g_j.

Frecuencia corregida

Algunas estadísticas sobre los términos de nuestra colección son:

Frecuencia corregida

La frecuencia corregida de un término w_i pondera la frecuencia en base a a los géneros en que aparece:

$$KF_i = \left(\sum_{j=1}^{N} \sqrt{r_j f_{ij}} \right)^2$$

Dispersión

La dispersión de un término w_i a través de los géneros se puede determinar como:

$$S_i = \frac{KF_i}{t_i}$$

Lecturas recomendadas

Manning, C., Raghavan, P. y Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Manning, C., Raghavan, P. y Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

McEnery, A. y Wilson, A. (2006). *Corpus Linguistics: An introduction*. Edinburgh University Press.

Rosenfeld, L. y Morville, P. (2002). *Information Arquitechture for the World Wide Web*. Sebastopol, CA: O'Really.

Zhai, C. y Massung, S. (2015). ''Introduction". *Text Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. ACM Books, pp. 3-19.