Extracción de información

Tecnologías de búsqueda en la web

Marcelo Mendoza

Departamento de Informática
Universidad Técnica Federico Santa María

Leyes del texto

George Kingsley Zipf

El principio del mínimo esfuerzo (1949, Zipf)

George Kingsley Zipf (1949), Human behavior and the principle of least effort, Addison-Wesley Press

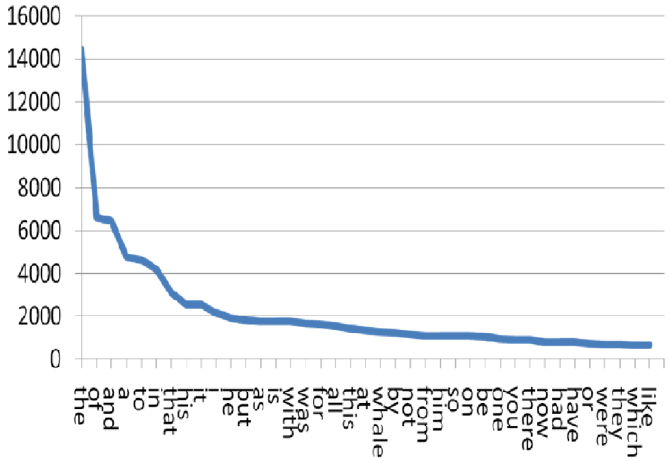
Ley de Zipf

Reuters[1]

[1]Agencia de noticias

Ley de Zipf

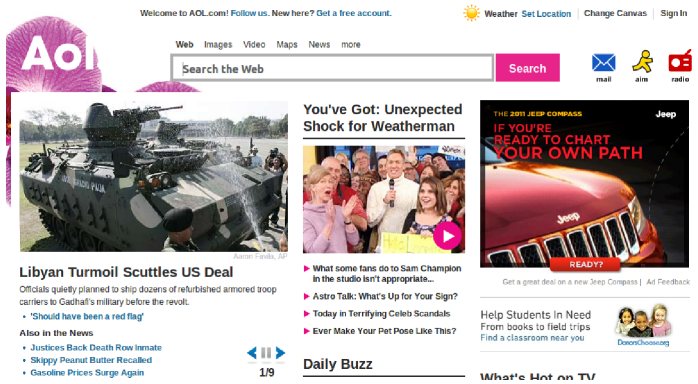
Zipf para Reuters[2]



[2]Dataset de noticias, disponible on-line

Ley de Zipf

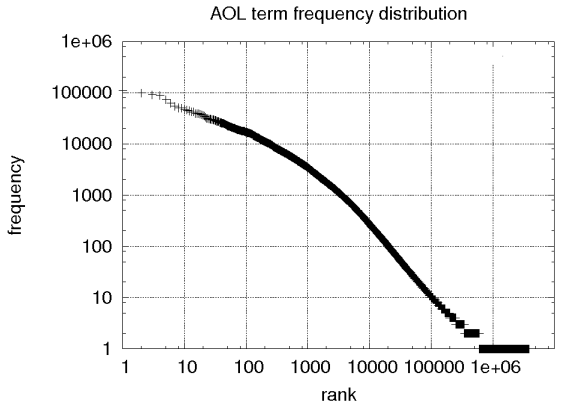
AOL[3]



[3]sitio con autenticacion, America On-Line

Ley de Zipf

Zipf para AOL query log[4]



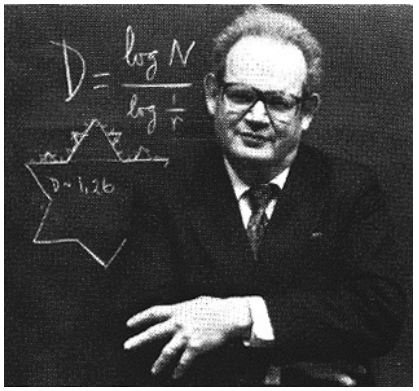
[4]Dataset de consultas formuladas a AOL, disponible on-line

Limitaciones del ajuste Zipf

Word	Freq. (f)	Rank (r)	$f \cdot r$	Word	Freq. (f)	Rank (r)	$f \cdot r$
the	3332	1	3332	turned	51	200	10200
and	2972	2	5944	you'll	30	300	9000
a	1775	3	5235	name	21	400	8400
he	877	10	8770	comes	16	500	8000
but	410	20	8400	group	13	600	7800
be	294	30	8820	lead	11	700	7700
there	222	40	8880	friends	10	800	8000
one	172	50	8600	begin	9	900	8100
about	158	60	9480	family	8	1000	8000
more	138	70	9660	brushed	4	2000	8000
never	124	80	9920	sins	2	3000	6000
Oh	116	90	10440	Could	2	4000	8000
two	104	100	10400	Applausive	1	8000	8000

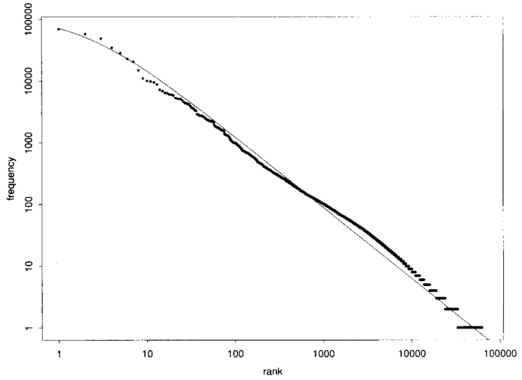
Producto $f \cdot r$ en el libro *Tom Sawyer*, versión en inglés.

Benoit Mandelbrot (1954)



Mandelbrot, B. (1954) Structure formelle des textes et communcation. Word, 10:1-27.

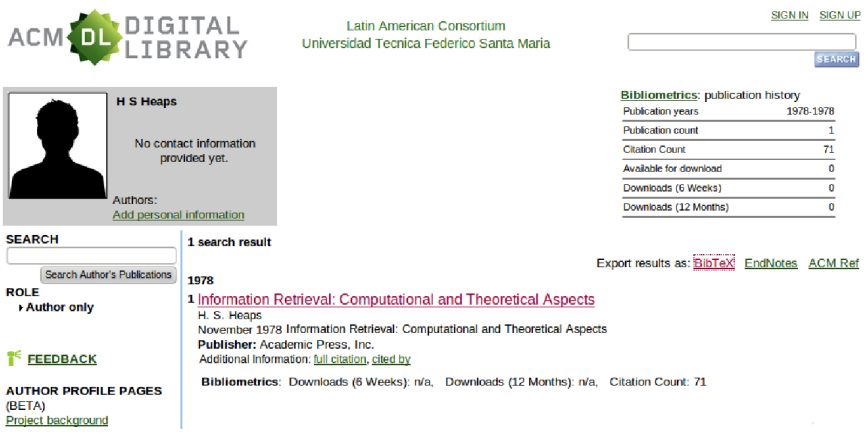
Ajuste de Mandelbrot



Ajuste Mandelbrot en el corpus *Brown* [5].

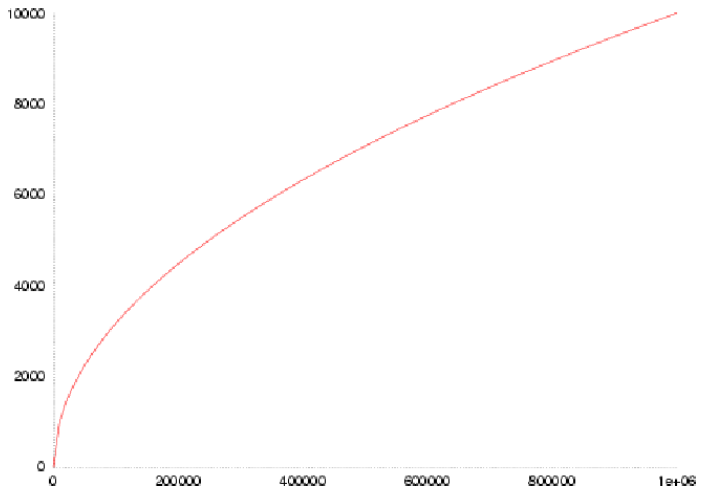
[5] The Brown Corpus was the first million-word electronic corpus of English, created in 1961 at Brown University. This corpus contains text from 500 sources, and the sources have been categorized by genre, such as news, editorial, and so on

Harold Stanley Heaps (1978)



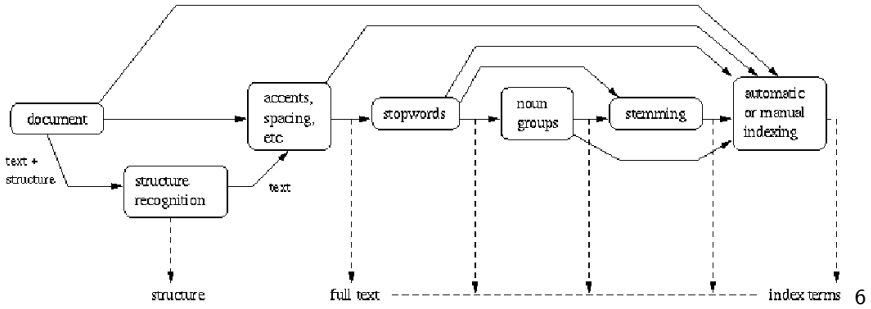
Ley de Heaps

V



Preprocesamiento del texto

Preprocesamiento de texto



[6] Ref.: R. Baeza & B. Ribeiro, Modern Information Retrieval, 1999.

Problemas recurrentes con idiomas

和尚

Chino

ノーベル平和賞を受賞したワンガリ・マータイさんが名誉会長を務めるMOTTAINAIキャンペーンの一環として、毎日新聞社とマガジンハウスは「私の、もったいない」を募集します。皆様が日ごろ「もったいない」と感じて実践していることや、それにまつわるエピソードを８００字以内の文章にまとめ、簡単な写真、イラスト、図などを添えて１０月２０日までにお送りください。大賞受賞者には、５０万円相当の旅行券とエコ製品２点の副賞が贈られます。

Japonés

استقلت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.

← START

'Algeria achieved its independence in 1962 after 132 years of French occupation.'

Árabe

Árabe

Stopwords (inglés)

	Stopwords
A	a, about, again, all, almost, also, although, always, among, an, and, another, any, are, as, at
B	be, because, been, before, being, between, both, but, by
C	can, could
D	did, do, does, done, due, during
E	each, either, enough, especially, etc
F	for, found, from, further
H	had, has, have, having, here, how, however
I	i, if, in, into, is, it, its, itself
J	just
K	kg, km
M	made, mainly, make, may, mg, might, ml, mm, most, mostly, must
N	nearly, neither, no, nor
O	obtained, of, often, on, our, overall
P	perhaps, pmid
Q	quite
R	rather, really, regarding
S	seem, seen, several, should, show, showed, shown, shows, significantly, since, so, some, such
T	than, that, the, their, theirs, them, then, there, therefore, these, they, this, those, through, thus, to
U	upon, use, used, using
V	various, very
W	was, we, were, what, when, which, while, with, within, without, would

[7]

[7] http://www.pubmed.gov

Stopwords (español)

a, acá, ahí, ajena, ajenas, ajeno, ajenos, al, algo, alguna, algunas, alguno, algunos, algún, allá, allí, aquel, aquella, aquellas, aquello, aquellos, aquí, cada, cierta, ciertas, cierto, ciertos, como, cómo, con, conmigo, consigo, contigo, cualquier, cualquiera, cualquieras, cuan, cuanta, cuantas, cuánta, cuántas, cuanto, cuantos, cuán, cuánto, cuántos, de, dejar, del, demasiada, demasiadas, demasiado, demasiados, demás, el, ella, ellas, ellos, él, esa, esas, ese, esos, esta, estar, estas, este, estos, hacer, hasta, jamás, junto, juntos, la, las, lo, los, mas, más, me, menos, mía, mientras, mío, misma, mismas, mismo, mismos, mucha, muchas, muchísima, muchísimas, muchísimo, muchísimos, mucho, muchos, muy, nada, ni, ninguna, ningunas, ninguno, ningunos, no, nos, nosotras, nosotros, nuestra, nuestras, nuestro, nuestros, nunca, o, os, otra, otras, otro, otros, para, parecer, poca, pocas, poco, pocos, por, porque, que, qué, quien, quienes, quienesquiera, quienquiera, quién, si, siempre, sí, sín, Sr, Sra, Sres, Sta, suya, suyas, suyo, suyos, tal, tales, tan, tanta, tantas, tanto, tantos, te, tener, ti, toda, todas, todo, todos, tomar, tuya, tuyo, tú, un, una, unas, unos, usted, ustedes, varias, varios, vosotras, vosotros, vuestra, vuestras, vuestro, vuestros, y, yo.

Tres algoritmos de stemming: comparación

Texto de ejemplo: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Porter: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Lovins: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Paice: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

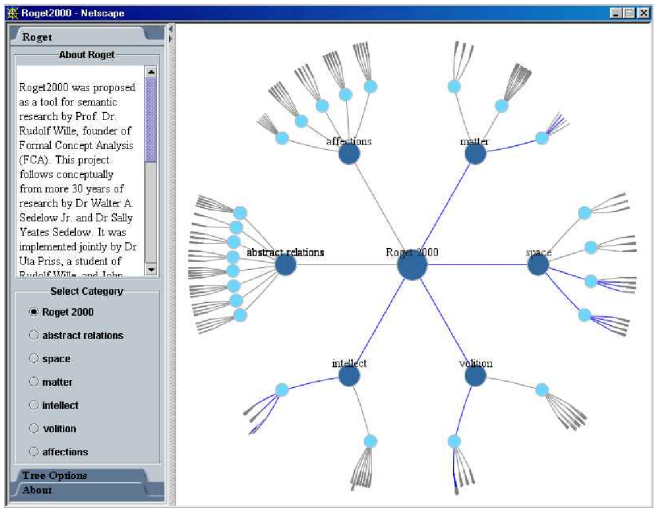
Martin Porter (1980)

http://tartarus.org/~martin/

Corpus

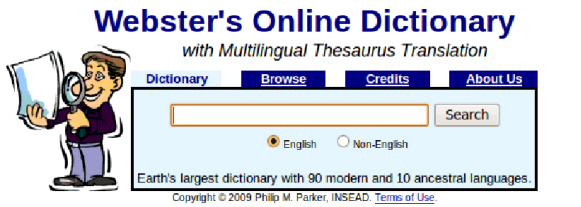
Peter Roget's Thesaurus

Roget's 2000



Ver más en http://www.roget.org

Webster on-line



Ver más en http://www.websters-online.dictionary.org

WordNeT

WordNet Search - 3.0 - WordNet home page - Glossary - Help

Word to search for: car Search WordNet

Display Options: (Select option to change) Change

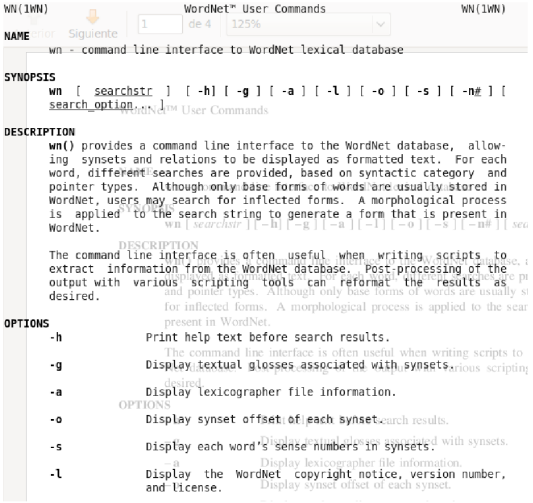
Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Noun

- S: (n) **car**, auto, automobile, machine, motorcar (a motor vehicle with four wheels; usually propelled by an internal combustion engine) *"he needs a car to get to work"*
- S: (n) **car**, railcar, railway car, railroad car (a wheeled vehicle adapted to the rails of railroad) *"three cars had jumped the rails"*
- S: (n) **car**, gondola (the compartment that is suspended from an airship and that carries personnel and the cargo and the power plant)
- S: (n) **car**, elevator car (where passengers ride up and down) *"the car was on the top floor"*
- S: (n) cable car, **car** (a conveyance for passengers or freight on a cable railway) *"they took a cable car to the top of the mountain"*

Ver más en http://www.wordnet.princeton.edu

WordNeT



WN(1WN) WordNet™ User Commands WN(1WN)

NAME

wn - command line interface to WordNet lexical database

SYNOPSIS

wn [searchstr] [-h] [-g] [-a] [-l] [-o] [-s] [-n#] [search_option...]

DESCRIPTION

wn() provides a command line interface to the WordNet database, allowing synsets and relations to be displayed as formatted text. For each word, different searches are provided, based on syntactic category and pointer types. Although only base forms of words are usually stored in WordNet, users may search for inflected forms. A morphological process is applied to the search string to generate a form that is present in WordNet.

The command line interface is often useful when writing scripts to extract information from the WordNet database. Post-processing of the output with various scripting tools can reformat the results as desired.

OPTIONS

-h	Print help text before search results.
-g	Display textual glosses associated with synsets.
-a	Display lexicographer file information.
-o	Display synset offset of each synset.
-s	Display each word's sense numbers in synsets.
-l	Display the WordNet copyright notice, version number, and license.

Hipónimos en WordNet (car)

wn car -hypon

Sense 1:

car, auto, automobile, machine, motorcar

- ambulance
- beach wagon, station wagon, wagon, beach waggon
- bus, jalopy, heap
- cab, hack, taxi, taxicab
- compact, compact car
- convertible
- coupe
- cruiser, police cruiser, patrol car, police car
- electric, electric automobile, electric car
- gas guzzler
- hardtop

...

Hipérnimos en WordNet (car)

wn car -hypen

Sense 1

car, auto, automobile, machine, motorcar

- motor vehicle, automotive vehicle
- self-propelled vehicle
- wheeled vehicle
- vehicle
- conveyance, transport
- instrumentality, instrumentation
- artifact, artefact
- object, physical object
- entity
- whole, whole thing, unit
- object, physical object
- entity

...

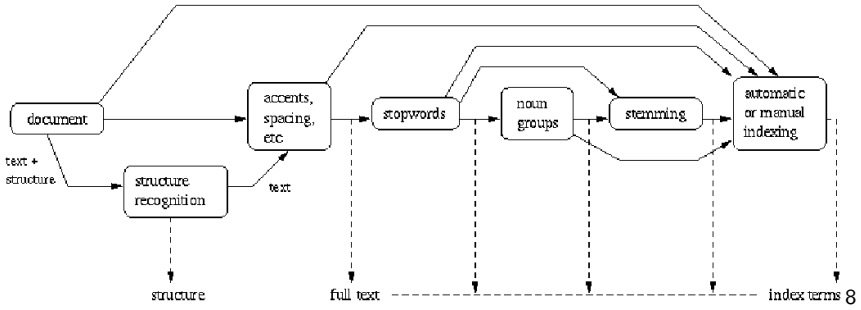
Merónimos en WordNet (car)

wn car -meron
Sense 1
car, auto, automobile, machine, motorcar
- HAS PART: accelerator, accelerator pedal, gas pedal
- HAS PART: air bag
- HAS PART: auto accessory
- HAS PART: automobile engine
- HAS PART: automobile horn, car horn, motor horn, horn, hooter
- HAS PART: buffer, fender
- HAS PART: bumper
- HAS PART: car door
- HAS PART: car mirror
- HAS PART: car seat
- HAS PART: car window

...

Métodos estándar en IR

Preprocesamiento de texto clásico en IR



[8] Ref.: R. Baeza & B. Ribeiro, Modern Information Retrieval, 1999.

Token v/s término

- Token – String delimitado que aparece en el texto.
- Término – token con significado según un corpus (por ejemplo diccionario)

Tokenización

- Input:
  amigos, Romans, habitantes. habia una vez ... Cesar ...
- Output:
  amigo romano habitante cesar ...
- Cada token es candidato a término.
- Cuáles elegimos? Depende del corpus.

Tokenizar es difícil – (sobretodo en Inglés)

Ejemplo 1: *Mr. O'Neill thinks that the boys' stories about Chile's capital aren't amusing.*

Ejemplo 2: *Miss lilly higgins sings shimmy in mississippi's spring.*

Detección de términos

- 3/12/91
- 12/3/91
- Marzo 12, 2008
- B-52
- 100.2.86.144
- (32) 234-2333
- 800.234.2333
- Los primeros IR no indexaban números, pero generalmente son útiles.

Tokenización ambigua en Chino

莎拉波娃现在居住在美国东南部的佛罗里达。今年4月
9日，莎拉波娃在美国第一大城市纽约度过了18岁生
日。生日派对上，莎拉波娃露出了甜美的微笑。

Tokenización ambigua en Chino

Si son usados de forma separada hay dos significados, si se indexan juntos hay un significado.

Tokenización ambigua en otros idiomas

- Composición en Alemán
- Computerlinguistik → Computer + Linguistik
- Lebensversicherungsgesellschaftsangestellter
- → leben + versicherung + gesellschaft + angestellter
- tusaatsiarunnanngittualuujunga (no escucho muy bien!)
- Sueco, Finés, Griego, entre otros.

Normalización

- Es necesario "normalizar" términos en texto indexado así como los términos de las consultas.
- Ejemplo: Queremos 'matching' entre *U.S.A.* y *USA*
- Implícitamente lo que estamos haciendo es definir clases de equivalanecia de términos.
- Alternativa: hacer expansión asimétrica.
  - window → window, windows
  - windows → Windows, windows
  - Windows
- Mas poderosas pero menos eficientes
- Por qué no colocar *window*, *Window*, *windows*, y *Windows* en la misma clase de equivalencias?

Normalización: Otros lenguajes

- Acentos: résumé vs. resume
- Cremillas: Universität vs. Universitaet
- Más importante aún: Cómo los usuarios formulan estas consultas frecuentemente?
- En lenguajes donde usualmente se usan acentos, las consultas se formulan sin ellos.
- *PETER WILL NICHT MIT.* → MIT = mit
- *El obtuvo su PhD en el MIT.* → MIT ≠ mit

Mayúsculas

- Reducir todo a minúsculas
- Excepciones posibles: marcas
- MIT vs. mit
- Fed vs. fed

Stop words

- stop words = palabras usadas comúnmente que tienen un bajo valor descriptivo
- Ejemplos: *e*, *y*, *o*, *en*, *de*, *pero*, *para*, *por*, *el*, *la*, *los las*, *desde*, *hasta*, ...
- Los primeros sistemas IR eliminaban las stop words del vocabulario.
- Para consultas de frases son necesarias las stop words, e.g. "Rey de España"
- Actualmente varios motores de búsqueda indexan stop words.

Más clases de equivalencia

- Sonido: (equivalencia fonética, Tchebyshev = Chebysheff)
- Thesauri: (equivalencia semántica, auto = carro)

Lematización

- Reducir formas inflecciones a su raíz
- Ejemplo: *am*, *are*, *is* $\rightarrow$ *be*
- Ejemplo: *autos*, *auto*, *automoviles* $\rightarrow$ *auto*
- Ejemplo: *Los autos de los jóvenes son de colores* $\rightarrow$ *auto joven es color*
- Lematización implica realizar una reducción hacia la raíz (lema). (*destruccion* $\rightarrow$ *destruir*)

Stemming

- Definición de stemming: Proceso heurístico que corta la derivación de las palabras para encontrar la raíz (es un tipo de lematización).
- Es dependiente del lenguaje
- Inflecccional y derivacional
- Ejemplo de derivacional: *automata*, *automatico*, *automatizado* se reduce a *automata*

Algoritmo de Porter

- Algoritmo de stemming más comúnmente usado en Inglés
- Los resultados sugieren de que es al menos tan bueno como otros algoritmos de stemming
- Convenciones + 5 fases de reducción
- Las fases son aplicadas secuencialmente
- Cada fase consiste de un conjunto de reglas.
  - Regla de ejemplo: Eliminar la derivación *ement* si el largo del prefijo es mayor que 1
  - replacement → replac
  - cement → cement
- Convención de ejemplo: Si hay varias reglas que se pueden aplicar en un mismo caso, use aquella que se aplica a un sufijo más largo.

Algoritmo de Porter: Unas pocas reglas

Regla			**Ejemplo**		
SSES	→	SS	caresses	→	caress
IES	→	I	ponies	→	poni
SS	→	SS	caress	→	caress
S	→		cats	→	cat

Tres algoritmos de stemming: comparación

Texto de ejemplo: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Porter: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Lovins: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Paice: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Mejora el stemming la efectividad de IR?

- En general, stemming mejora la efectividad en algunas consultas, y la desmejora en otras.
- Mientras más regular es la gramática, mejor.
- En castellano es difícil (gramática muy irregular).

Programación en NLTK

Ejemplos simples en NLTK

Procesamiento básico

```
> sentence = """Such an analysis can reveal features that are not easily visible
from the variations in the individual genes and can lead to a picture of expression
that is more biologically transparent and accessible to interpretation"""
> len(sentence)
> tokens = nltk.word_tokenize(sentence)
> sorted(set(tokens))
> tagged = nltk.pos_tag(tokens)
> nltk.chunk.ne_chunk(tagged)
```


Ejemplos simples en NLTK

- Distribución de frecuencia

```
> fdist = FreqDist(tokens)
> vocabulary = fdist.keys()
> fdist.plot(30)
```

- WordNet

```
> from nltk.corpus import wordnet as wn
> wn.synsets('motorcar')
> wn.synset('car.n.01').lemma_names
> wn.synset('car.n.01').definition
> wn.synset('car.n.01').hyponyms()
> wn.synset('car.n.01').hypernyms()
> wn.synset('car.n.01').part_meronyms()
```

- Collocations

```
> text3.collocations()
```