

Extracción de información

Tecnologías de búsqueda en la web

Marcelo Mendoza



Departamento de Informática
Universidad Técnica Federico Santa María

Leyes del texto

MM

INF-335

1 / 52

MM

INF-335

2 / 52

Extracción

Leyes del texto

Extracción

Leyes del texto

George Kingsley Zipf

El principio del mínimo esfuerzo (1949, Zipf)



HUMAN BEHAVIOR OR THE PRINCIPLE OF LEAST EFFORT

An Introduction to Human Ecology

By
GEORGE KINGSLEY ZIPF, Ph.D.
Harvard University

ADDISON-WESLEY PUBLISHING COMPANY
350 SUMMIT STREET, READING, MASSACHUSETTS 01867
1949

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

MM

INF-335

3 / 52

MM

INF-335

4 / 52

Ley de Zipf

Reuters¹

REUTERS U.S. News & Markets Sectors & Industries Analysis & Opinion Search News & Quotes SEARCH

BREAKING NEWS:
Obama says NATO considering military options against Libya

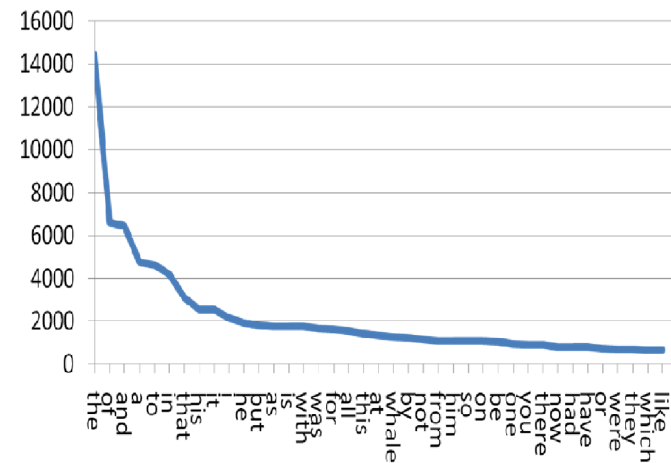
Global central banks point to more acute price risks
BASEL, Switzerland (Reuters) - A spike in food and oil prices has made the threat of inflation more acute, leading central banks said on Monday, but they warned tightening of policy in response will not proceed at the same pace.

CONTINUE READING

ISSUES IN DEPTH
Gaddafi counter-attacks, talks sought with rebels
Government forces seeking to dislodge rebels from Libya's strategically important coast struck at an oil town on Monday amid quickening efforts to prevent more humanitarian suffering and a military offensive against... *Paul G. Anderson, Reuters*

MARKETS
OPEN
US Indices
DOW -47.64 12,122.24 -0.39%
NASDAQ -35.72 2,748.95 -1.29%
S&P 500 -7.12 1,314.03 -0.54%
TRI US INDEX -0.82 1,920.03

Ley de Zipf

Zipf para Reuters²¹Agencia de noticias

MM

INF-335

5 / 52

²Dataset de noticias, disponible on-line

MM

INF-335

6 / 52

Ley de Zipf

AOL³

Welcome to AOL.com! Follow us. New here? Get a free account. Weather Set Location Change Canvas Sign In

Web Images Video Maps News more

Search the Web Search

Libyan Turmoil Scuttles US Deal
Officials quietly planned to ship dozens of refurbished armored troop carriers to Gaddafi's military before the revolt.
• 'Should have been a red flag'

Also in the News
• Justices Back Death Row Inmate
• Skippy Peanut Butter Recalled
• Gasoline Prices Surge Again

You've Got: Unexpected Shock for Weatherman
► What some fans do to Sam Champion in the studio isn't appropriate...
► Astro Talk: What's Up for Your Sign?
► Today in Terrifying Celeb Scandals
► Ever Make Your Pet Pose Like This?

Daily Buzz

What's Hot on TV

The 2011 Jeep Compass
IF YOU'RE READY TO CHART YOUR OWN PATH
READY?

Get a great deal on a new Jeep Compass | Ad Feedback

Help Students in Need From books to field trips Find a classroom near you *DonorChoose.org*

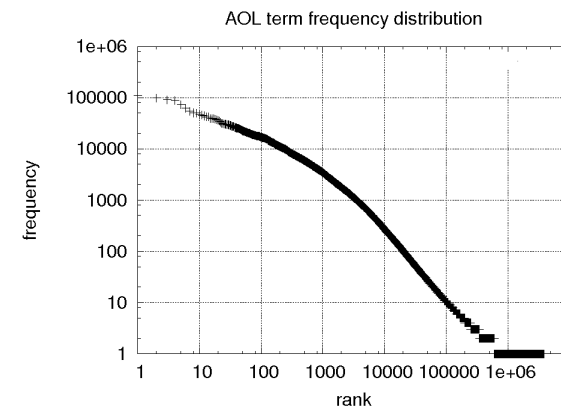
³sitio con autentificacion, America On-Line

MM

INF-335

7 / 52

Ley de Zipf

Zipf para AOL query log⁴⁴Dataset de consultas formuladas a AOL, disponible on-line

MM

INF-335

8 / 52

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.

MM

INF-335

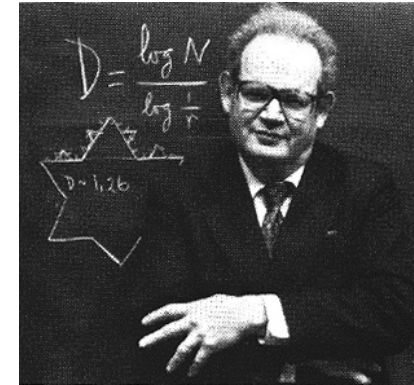
9 / 52

MM

INF-335

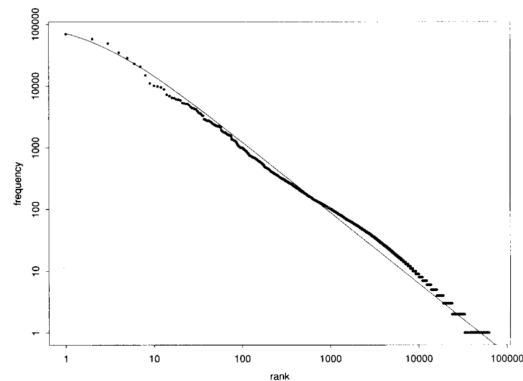
10 / 52

Benoit Mandelbrot (1954)



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

Ajuste de Mandelbrot

Ajuste Mandelbrot en el corpus Brown ⁵.

⁵The 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

MM

INF-335

11 / 52

MM

INF-335

12 / 52

Harold Stanley Heaps (1978)

ACM DL DIGITAL LIBRARY

Latin American Consortium
Universidad Técnica Federico Santa María

SIGN IN SIGN UP

SEARCH

H S Heaps

No contact information provided yet.

Authors:
[Add personal information](#)

SEARCH

Search Author's Publications

ROLE

Author only

FEEDBACK

AUTHOR PROFILE PAGES

(BETA)
[Project background](#)

1 search result

1978

1 Information Retrieval: Computational and Theoretical Aspects

H. S. Heaps

November 1978 Information Retrieval: Computational and Theoretical Aspects

Publisher: Academic Press, Inc.

Additional Information: [Full citation](#), [Cited by](#)

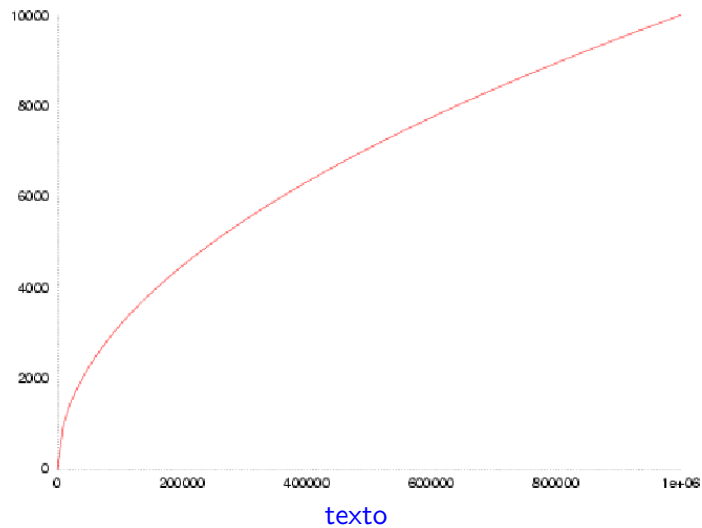
Bibliometrics: Downloads (6 Weeks): n/a, Downloads (12 Months): n/a, Citation Count: 71

Bibliometrics: publication history	
Publication years	1978-1978
Publication count	1
Citation Count	71
Available for download	0
Downloads (6 Weeks)	0
Downloads (12 Months)	0

Export results as: [BibTeX](#) [EndNotes](#) [ACM Ref](#)

Ley de Heaps

V



Preprocesamiento del texto

MM

INF-335

13 / 52

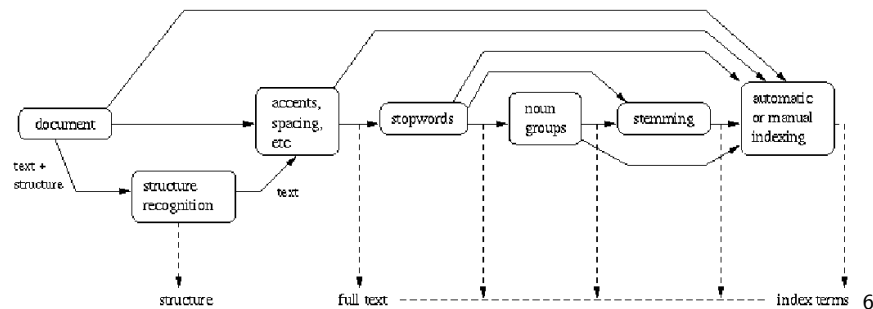
MM

INF-335

14 / 52

Preprocesamiento de texto

Problemas recurrentes con idiomas



和尚

Chino

ノーベル平和賞を受賞したワシントン・マーティン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
← START
← START
'Algeria achieved its independence in 1962 after 132 years of French occupation.'

Árabe

Árabe

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

MM

INF-335

15 / 52

MM

INF-335

16 / 52

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>

MM

INF-335

17 / 52

MM

INF-335

18 / 52

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

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, cualesquiera, 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, tuyas, 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.

Martin Porter (1980)



<http://tartarus.org/~martin/>

MM

INF-335

19 / 52

MM

INF-335

20 / 52

Peter Roget's Thesaurus

Corpus



MM

INF-335

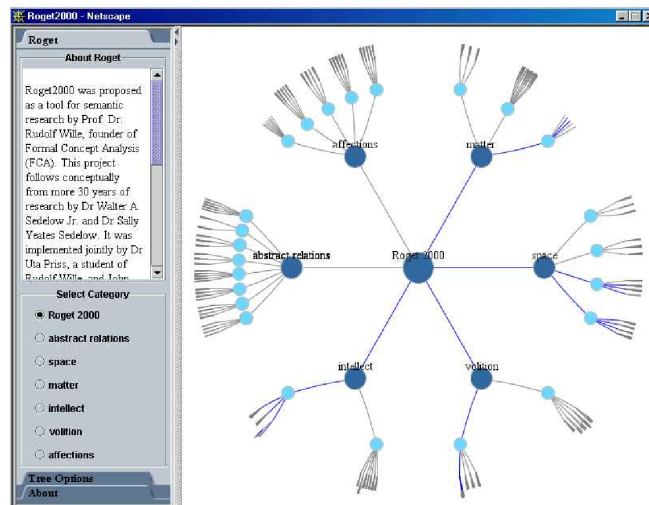
21 / 52

MM

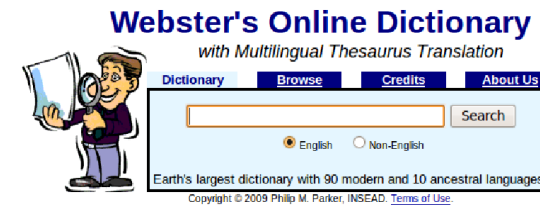
INF-335

22 / 52

Roget's 2000



Webster on-line



Coming in 2009: timelines, translations, sound effects, and a big surprise!

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

Ver más en <http://www.roget.org>

MM

INF-335

23 / 52

MM

INF-335

24 / 52

WordNeT

WordNet Search - 3.0 - [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

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>

MM

INF-335

25 / 52

WordNeT

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

NAME 1 de 4 125%

wn -> command line interface to WordNet lexical database

SYNOPSIS

```
wn [ searchstr ] [ -h ] [ -g ] [ -a ] [ -l ] [ -o ] [ -s ] [ -ng ] [ 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.

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.

MM

INF-335

26 / 52

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

...

MM

INF-335

27 / 52

Hipérnimos en WordNet (car)

[wn car -hypon](#)

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

...

MM

INF-335

28 / 52

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

MM

INF-335

29 / 52

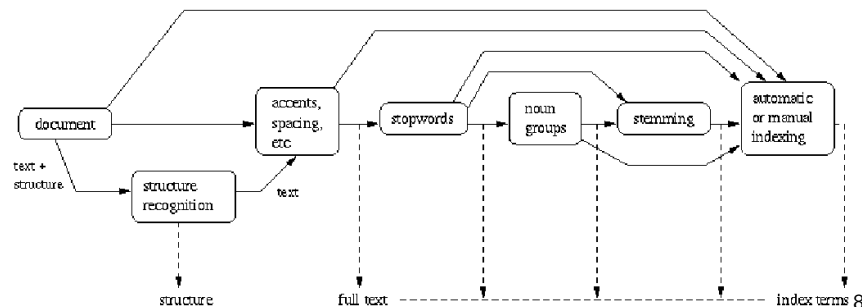
MM

INF-335

30 / 52

Preprocesamiento de texto clásico en IR

Token v/s término



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

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

MM

INF-335

31 / 52

MM

INF-335

32 / 52

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 – (sobre todo 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
- ▶ tusaatsiarunnannngittualuujunga (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 equivalencia** 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 infleccionales a su raíz
- ▶ Ejemplo: *am, are, is* → *be*
- ▶ Ejemplo: *autos, auto, automoviles* → *auto*
- ▶ Ejemplo: *Los autos de los jóvenes son de colores* → *auto joven es color*
- ▶ Lematización implica realizar una reducción hacia la raíz (lema). (*destruccion* → *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
- ▶ Infleccional y derivacional
- ▶ Ejemplo de derivacional: *automata*, *automatico*, *automatizado* se reduce a *automata*

MM

INF-335

45 / 52

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.

MM

INF-335

46 / 52

Algoritmo de Porter: Unas pocas reglas

Regla

SS → SS

IES → I

SS → SS

S →

Ejemplo

caresses → caress

ponies → poni

caress → caress

cats → cat

MM

INF-335

47 / 52

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

MM

INF-335

48 / 52

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

MM

INF-335

49 / 52

MM

INF-335

50 / 52

Ejemplos simples en NLTK

Procesamiento básico

```

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

```

Ejemplos simples en NLTK

▶ Distribución de frecuencia

```

1 > fdist = FreqDist(tokens)
2 > vocabulary = fdist.keys()
3 > fdist.plot(30)

```

▶ WordNet

```

1 > from nltk.corpus import wordnet as wn
2 > wn.synsets('motorcar')
3 > wn.synset('car.n.01').lemma_names
4 > wn.synset('car.n.01').definition
5 > wn.synset('car.n.01').hyponyms()
6 > wn.synset('car.n.01').hypernyms()
7 > wn.synset('car.n.01').part_meronyms()

```

▶ Collocations

```

1 > text3.collocations()

```

MM

INF-335

51 / 52

MM

INF-335

52 / 52