Googlastic search

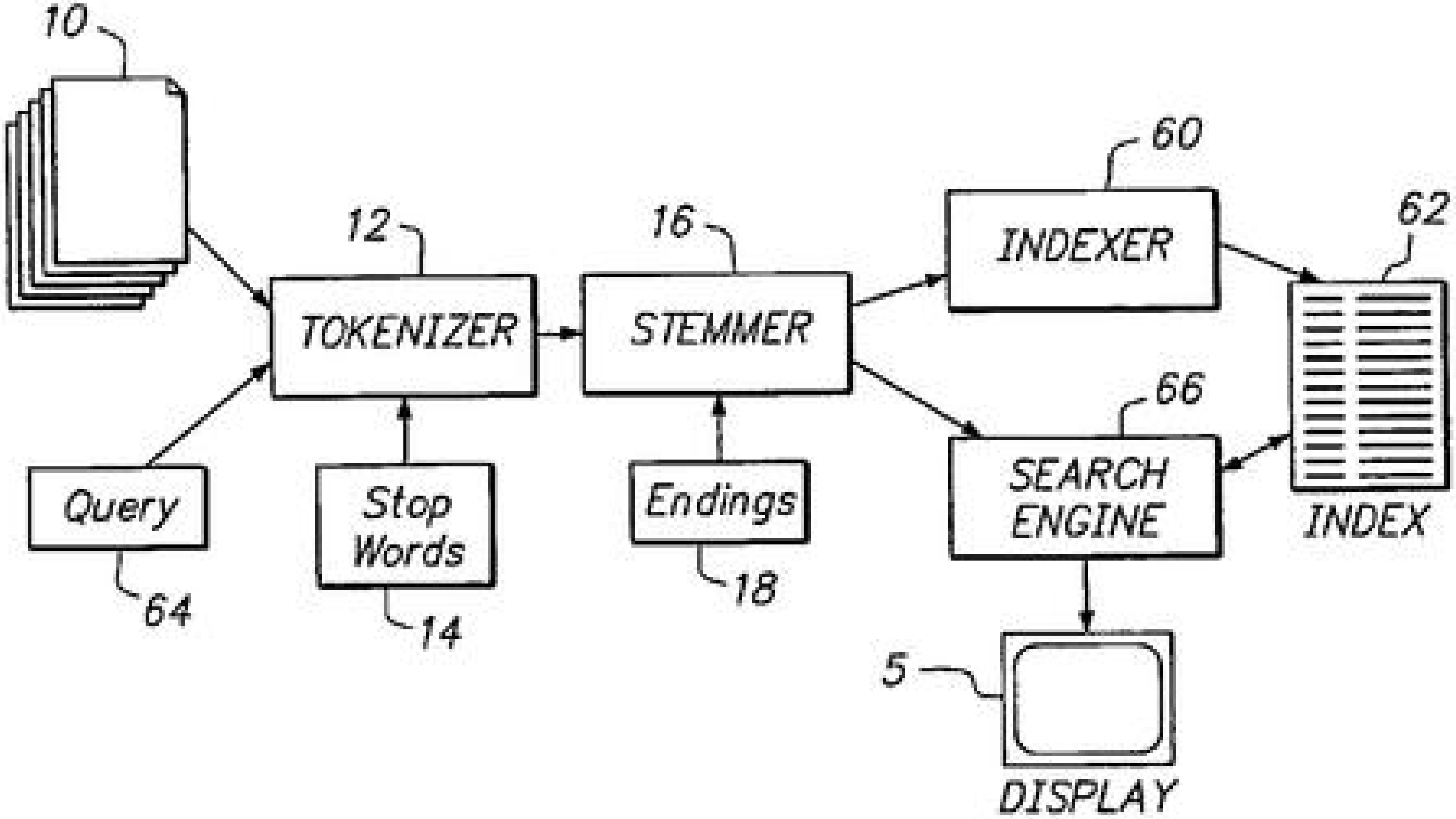
Information Retrieval

Antonio Sánchez | @plutec_net
asanchez@plutec.net

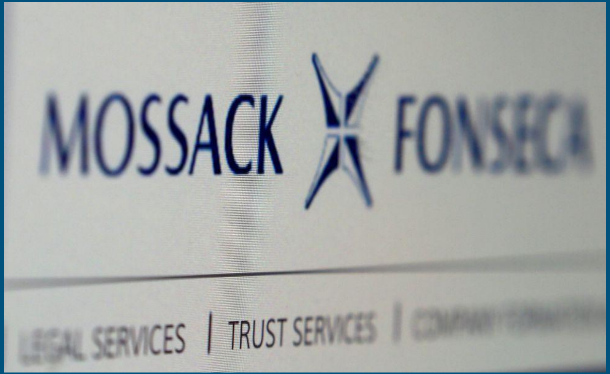
https://github.com/plutec/betabeers_IR

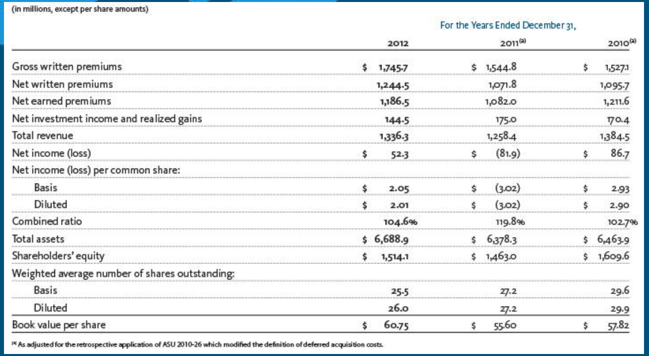
Introducción

- ¿Googlastic search?
- Recuperación de información
- ¿Cómo funciona?
- NLTK para Python


10
12
TOKENIZER
16
STEMMER
60
INDEXER
62
INDEX
66
SEARCH ENGINE
Query
64
Stop Words
14
Endings
18
5
DISPLAY

Documentos




(in millions, except per share amounts)

	For the Years Ended December 31,		
	2012	2011[a]	2010[a]
Gross written premiums	$ 1,745.7	$ 1,544.8	$ 1,527.1
Net written premiums	1,244.5	1,071.8	1,095.7
Net earned premiums	1,186.5	1,082.0	1,211.6
Net investment income and realized gains	144.5	175.0	170.4
Total revenue	1,336.3	1,258.4	1,384.5
Net income (loss)	$ 52.3	$ (81.9)	$ 86.7
Net income (loss) per common share:			
Basis	$ 2.05	$ (3.02)	$ 2.93
Diluted	$ 2.01	$ (3.02)	$ 2.90
Combined ratio	104.6%	119.8%	102.7%
Total assets	$ 6,688.9	$ 6,378.3	$ 6,463.9
Shareholders' equity	$ 1,514.1	$ 1,463.0	$ 1,609.6
Weighted average number of shares outstanding:			
Basis	25.5	27.2	29.6
Diluted	26.0	27.2	29.9
Book value per share	$ 60.75	$ 55.60	$ 57.82

[a] As adjusted for the retrospective application of ASU 2010-26 which modified the definition of deferred acquisition costs.

Tokenization

“You aren’t” -> [“You”, “aren”, “t”]

“You aren’t” -> [“You”, “arent”] **¿rent?**

“You aren’t” -> [“You”, “are” “not”] **Perfecto!**

Tokenization

“Tú no eres” -> [“Tú”, “no”, “eres”]

“Dame el abrelatas” -> [“Dame”, “el”, “abrelatas”]

“Dame el abrelatas” -> [“Dame”, “el”, “abre”, “latas”]

Tokenization

Regalo de navidad

“das Weihnachtsgeschenk” -> [“das”, “Weihnachtsgeschenk”] Meeec

“das Weihnachtsgeschenk” -> [“das”, “Weihnacht”, “Geschenk”] Gooood!

Tokenization

Fuente + idioma

Twitter: k, q, qe –> que | RT –> Retweet

Doc. médicos: AAA –> abdominal aortic aneurysm | CHO –> carbohydrate

Doc. financieros: CEO –> Chief executive officer | ETA -> Estimated Time of Arrival

Stemmer

Sacar la raíz de las palabras eliminando sufijos

“Yo tengo un libro” -> “yo teng un libr”

“Mi casa es muy grande” -> “mi cas es muy grand”

“María tiene una casa” -> “maria tien una cas”

Stemmer

- Porter (1980, Inglés)
  - http://9ol.es/porter_js_demo.html
- SnowBall (~2001, Multi-idioma)
  - http://snowballstem.org/demo.html

Google incorporó stemmer en 2003

Stop/empty words

Palabras que no aportan valor al documento.

“El libro es amarillo” -> “Libro es amarillo”

“Mi casa tiene un sótano” -> “Mi casa tiene sótano” (El mi no se quita porque indica posesión)

he	drink	ink	likes	pink	thing	wink	
2	1	0	2	0	0	1	He likes to wink, he likes to drink.
1	3	0	1	0	0	0	He likes to drink, and drink, and drink.
1	1	1	1	0	1	0	The thing he likes to drink is ink.
1	1	1	1	1	0	0	The ink he likes to drink is pink.
1	1	1	1	1	0	1	He likes to wink and drink pink ink.

NLTK

Natural **L**anguage **T**ool**K**it es una librería para Python que facilita la implementación de algoritmos de IR (Information Retrieval)

```
$ pip install nltk
```

Code for me!

Software

¡Gracias!

@plutec_net | asanchez@plutec.net
https://github.com/plutec