



El Tokenizador

Partiendo el documento en trozos

La Agencia Tributaria ha decretado el embargo preventivo de los ingresos televisivos y publicitarios del Atlético de Madrid y 7 equipos de Segunda, entre ellos el recién ascendido Albacete. La cantidad que se embargaría asciende a 43,5 millones de euros debido al "riesgo de impago" de la deuda que tienen contraídos dichos clubes con el Ministerio de Hacienda.

Todas las palabras

La	Agencia	tributaria	ha	decretado	el	...
----	---------	------------	----	-----------	----	-----

Minúsculas

la	agencia	tributaria	ha	decretado	el	...
----	---------	------------	----	-----------	----	-----

Sin palabras huecas

agencia	tributaria	decretado	embargo	preventivo	...
---------	------------	-----------	---------	------------	-----

Extrayendo raíces

agenc	tribut	decret	embarg	prevent	...
-------	--------	--------	--------	---------	-----

Agrupando raíces

agenc	tribut / haciend	decret/ Dictam	embarg	prevent/ provis	deud	...
-------	---------------------	-------------------	--------	--------------------	------	-----

Palabras huecas

él	apenas	cualquier	durante
ésta	aproximadamente	cuando	e
ésta	aquí	cuanto	ejemplo
éste	así	cuatro	el
éstos	aseguró	cuenta	ella
última	aunque	da	ellas
últimas	ayer	dado	ello
último	bajo	dan	ellos
últimos	bien	dar	embargo
a	buen	de	en
añadió	buna	debe	encuentra
aún	buenas	deben	entonces
actualmente	bueno	debido	entre
adelante	buenos	decir	era
además	cómo	dejó	eran
afirmó	cada	del	es
agregó	casi	demás	esa
ahí	cerca	dentro	esas
ahora	cierto	desde	ese
al	cinco	después	eso
algún	comentó	dice	esos
algo	como	dicen	está
alguna	con	dicho	están
algunas	conocer	dieron	esta
alguno	consideró	diferente	estaba
algunos	considera	diferentes	estaban
alrededor	contra	dijeron	estamos
ambos	cosas	dijo	estar
ante	creo	dio	estará
anterior	cual	donde	estas
antes	cuales	dos	...

Extracción de raíces

Lematizador:

- casa => casa, casero => casa
- Usa conocimiento lingüístico para extraer el lema
- Se puede realizar junto al *P.O.S. tagging*
- Son necesarios recursos adicionales (corpus etiquetado, generador de etiquetadores, reglas morfológicas)

Stemmer:

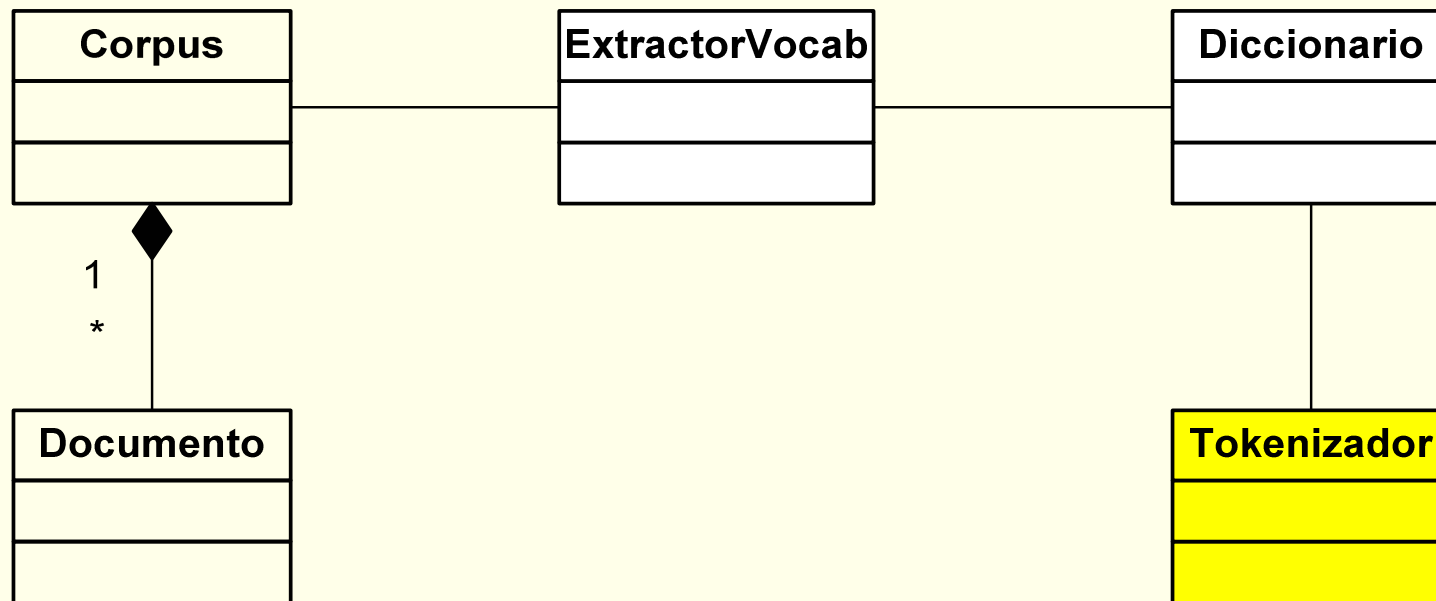
- casa => cas, casero => cas
- No produce palabras sino raíces
- No está fundamentado lingüísticamente
- Se apoyan en reglas heurísticas
- No requieren recursos adicionales
- Algoritmo de Porter

Corpus para el lematizador

El	DET
casero	NOM
me	PRON
dio	VER
la	DET
factura	NOM
pero	CONJ
no	ADV
la	PRON
encuentro	VER
...	

El	DET	el
casero	NOM	casa
me	PRON	me
dio	VER	dar
la	DET	el
factura	NOM	factura
pero	CONJ	pero
no	ADV	no
la	PRON	la
encuentro	VER	encontrar
...		

Clases a desarrollar



La clase Tokenizador

```
import java.io.BufferedReader;
import java.io.FileReader;
import java.io.IOException;
import java.util.*;
import java.util.regex.*;
/*
 * Número aproximado de líneas de código: 80
 */
public class Tokenizador {
    ...
    public Tokenizador() {...}
    public Tokenizador(Diccionario dic){...}
    public void normalizaMinusculas() {...}
    public void defineExpresionRegular(String er) {...}
    public void definePalabrasHuecas(String nombreFichero) {...}
    public void tokeniza(String texto){...}
    public boolean hayMasTokens(){...}
    public String lexema(){...}
    public int indice() {...} // Sólo para tokenizadores
                               // con diccionario
    ...
}
```

Más información en: [Tokenizador.html](#)