TextRank

Bringing order into texts

Índice

- PageRank
- TextRank
- Etiquetado: tareas y herramientas
- Una pequeña aportación

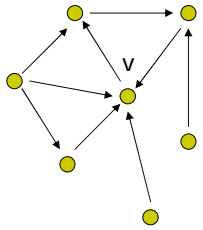
PageRank

Las dos ideas de Google en el 98

- Los artículos:
  - *The PageRank Citation Ranking: Bringing Order to the Web.* L. Page, S. Brin, R. Motwani, T. Winograd
  - *The Anatomy of a Large-Scale Hypertextual Web Search Engine.* S. Brin, L. Page
- Las ideas:
  - PageRank: Método para calcular la relevancia de las páginas independientemente de la consulta
  - Índices inversos: Método para encontrar rápidamente los documentos asociados a una palabra

PageRank

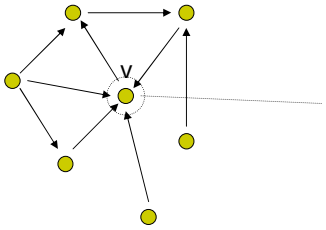
Todos los nodos son iguales, pero …



- In(v) = 4
- Out(v) = 1

PageRank

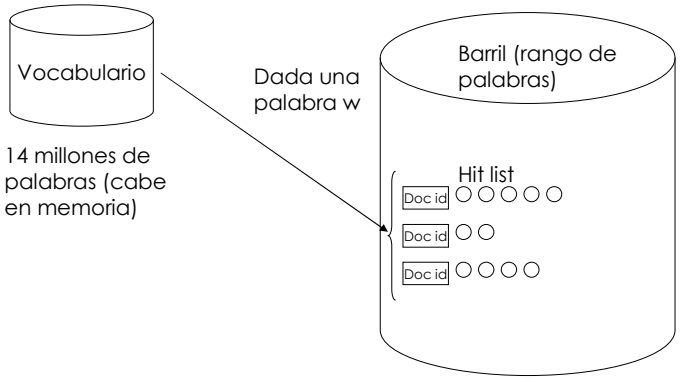
… algunos son más iguales que otros



$$PR(V_i) = (1-d) + d * \sum_{j\ In(Vi)} \frac{1}{|Out(V_j)|} PR(V_j)$$

PageRank

Indexar por palabras



TextRank

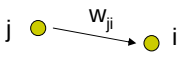
PageRank aplicado a textos

- *TextRank: Bringing Order into Texts.* R. Mihalcea, P. Tarau
  - Buscar conexiones entre unidades de texto
  - Construir un grafo
  - Aplicar PageRank
  - Usar el valor resultante para decidir algo sobre la unidad textual

TextRank

Ponderado

- En internet no tiene mucho sentido tener enlaces múltiples o parciales
- En los grafos de texto sí puede ser útil

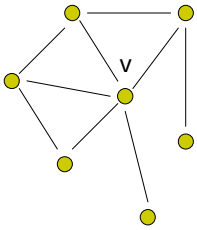


$$PR(V_i) = (1-d) + d * \sum_{j\ In(Vi)} \frac{w_{ji}}{\sum_{j\ Out\ (Vi)} w_{jk}} PR(V_j)$$

TextRank

Grafos sin dirección

- En internet tampoco tiene sentido
- Pero en textos puede que sí
  - In(v) = Out(v) = Número de arcos ligados a v

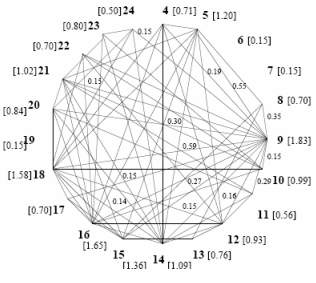


- In(v) = 5
- Out(v) = 5

TextRank

Generación de resúmenes

- Las unidades textuales son frases
- Arcos sin dirección
- El peso es una medida de distancia

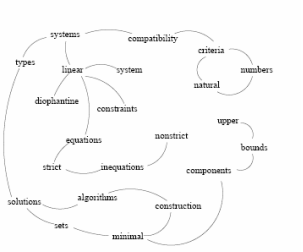


TextRank

Extracción de palabras clave

- Las unidades textuales son palabras
- Dos palabras están conectadas si están a menos de N palabras de distancia
- Se pueden filtrar por categorías sintácticas

Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types systems and systems of mixed types.



TextRank

No supervisado pero potente

- En ninguna de las aplicaciones anteriores se usa material de entrenamiento
  - Corpus con palabras clave
  - Corpus con resúmenes
- Sin embargo, consigue resultados similares a otras propuestas que sí lo usan
  - Supervisado: Ejemplos de train + test
  - No supervisado: Sólo ejemplos de test
- ¿Porqué?

Etiquetado: tareas y herramientas

Algunas definiciones

- Tratamiento secuencial de textos
- Tratamiento basado en análisis sintáctico
  - Parcial
  - Completo
- Etiquetado POS (*Part Of Speech*)
  - La tarea secuencial por excelencia
  - Existen muchos recursos y herramientas

Etiquetado: tareas y herramientas

Etiquetado POS

His	APPG
face	NN
took	VVD
on	RP
a	AT
sudden	JJ
pallor	NN
,	YC
became	VVD
beaded	VVN
with	IW
sweat	NN
,	YC
and	CC
he	PPHS
seemed	VVD
…	

Corpus Susanne

Etiquetado: tareas y herramientas

Reconocimiento de entidades

El	O
presidente	O
del	O
Consejo	B-ORG
por	I-ORG
la	I-ORG
Paz	I-ORG
,	O
organismo	O
observador	O
de	O
Perú	B-LOC
,	O
Francisco	B-PER
Díez	I-PER
Canseco	I-PER
,	O
consideró	O
…	

Corpus CoNLL-2002

Etiquetado: tareas y herramientas

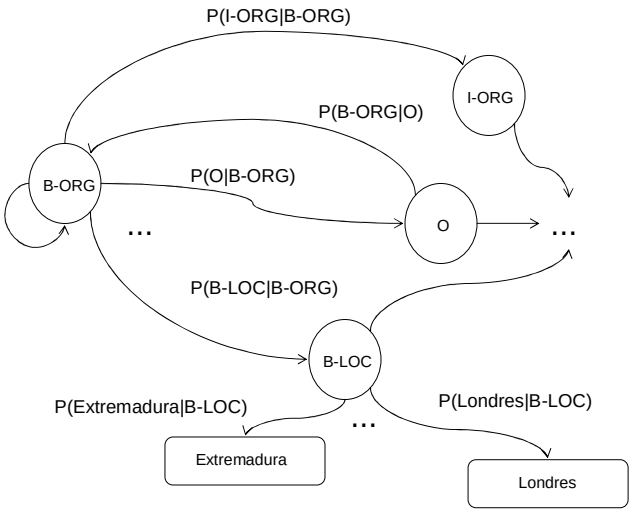
Análisis sintáctico superficial

But	CC	O
analysts	NNS	B-NP
reckon	VBP	B-VP
underlying	VBG	B-NP
support	NN	I-NP
for	IN	B-PP
sterling	NN	B-NP
has	VBZ	B-VP
been	VBN	I-VP
eroded	VBN	I-VP
by	IN	B-PP
the	DT	B-NP
chancellor	NN	I-NP
's	POS	B-NP
failure	NN	I-NP
to	TO	B-VP
announce	VB	I-VP
any	DT	B-NP
…		

Corpus CoNLL-2000

Etiquetado: tareas y herramientas

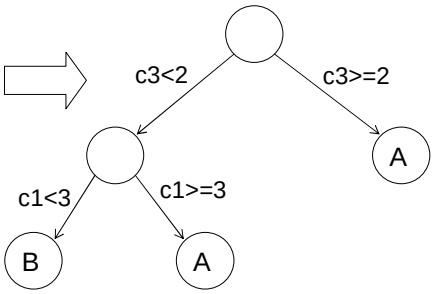
TnT: Modelos de Markov



Etiquetado: tareas y herramientas

TreeTagger: Árboles de decisión

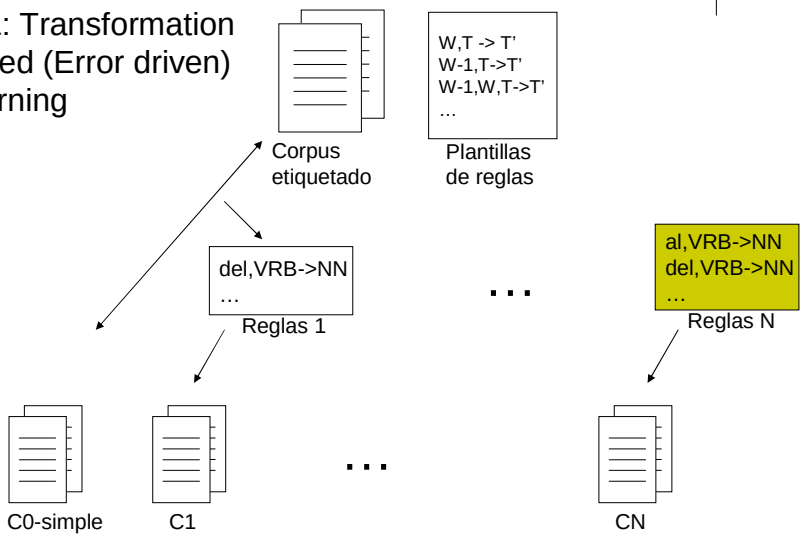
c1	c2	c3	c4	clase
2	6	4	11	A
3	5	3	12	A
1	6	1	1	B
...	...	...	...	...
10	6	1	5	A



Etiquetado: tareas y herramientas

fnTBL: Basado en transformaciones

TBL: Transformation Based (Error driven) Learning



Etiquetado: tareas y herramientas

MaxEnt: Máxima Entropía

- *P(e,c):* probabilidad de que la etiqueta *e* se corresponda con el contexto *c*
- El Modelado de máxima entropía plantea calcular P de manera que:
  - Se ajuste lo mejor posible a los casos conocidos
  - Distribuya el resto de la probabilidad uniformemente entre los casos desconocidos
- Las características expresan una relación de co-ocurrencia entre una predicción y algo en el contexto

$$f_j(e,c) = \begin{cases} 1 \text{ si } e = \text{DET y palabra}(c)=\text{“that”} \\ 0 \text{ en otro caso} \end{cases}$$

Etiquetado: tareas y herramientas

MBT: Basado en memoria

- MBT: Memory Based Tagger
- MBL: Memory Based Learning
- Optimización del método de los k-vecinos más cercanos
- Paquete adaptado para tareas PLN

Una pequeña aportación

Objetivo

- Implementar un método de etiquetado basado en TextRank:
  - Secuencial
  - Supervisado

Una pequeña aportación

La idea

- Vértices:
  - Extraídos del texto
  - Pareja palabra-etiqueta

<abstract, NOM>
<abstract, ADJ>
<abstract, VER>

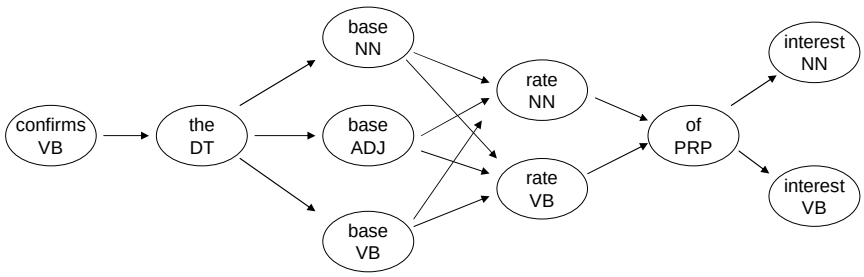
- Arcos:
  - Sacados del corpus
  - P(t|t-1) * P(w|t)

P(t|t-1) = C(t-1,t)/C(t-1)
P(w|t) = C(w,t)/C(t)

Una pequeña aportación

Un ejemplo

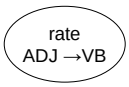
"The Ministry of Finance confirms the base rate of interest for half a year."



Una pequeña aportación

Un par de variantes

- Con trigramas y bigramas



- Inverso: de izquierda a derecha
  - Se utiliza la probabilidad P(t-1|t)
- Combinado con *stacking*
  - Se genera una base de datos con las propuestas del método original y del inverso
  - Se aprende de esa base de datos

Una pequeña aportación

Resultados

	Susanne	Penn
Línea base	79.15 %	80.01 %
TnT	93.61 %	95.48 %
TreeTagger	91.27 %	94.28 %
fnTBL	93.01 %	95.04 %
MBT	91.16 %	94.40 %
MaxEnt	93.09 %	95.47 %
TextRank	90.32 %	92.14 %
TextRankI	89.84 %	91.70 %
TextRankC	91.51 %	93.28 %

	NER-E	NER-B	Chunk
Línea base	71.90 %	72.64 %	63.08 %
TnT	94.78 %	88.97 %	89.62 %
TreeTagger	90.58 %	84.79 %	84.40 %
fnTBL	94.30 %	90.49 %	89.54 %
MBT	94.38 %	88.71 %	90.61 %
MaxEnt	95.03 %	87.52 %	92.83 %
TextRank	92.72 %	86.75 %	87.34 %
TextRankI	90.85 %	87.78 %	78.84 %
TextRankC	92.93 %	89.71 %	89.24 %

Una pequeña aportación

Posibles ampliaciones

- Heurísticas para palabras desconocidas
  - Ya incluidas por la mayoría de las herramientas comparadas
- Aplicar la idea a otro tipo de problemas
  - No secuenciales