Curso UNED

Procesamiento del Lenguaje Natural y aplicaciones para poesía.

Núria Bel (nuria.bel@upf.edu)

Universidad Pompeu Fabra

Julio - 2016

Objetivos

El procesamiento del lenguaje natural proporciona analizadores lingüísticos que aportan información sobre textos y permiten la automatización de tareas, aumentando la capacidad de análisis de los investigadores en humanidades digitales. El objetivo de este curso es que los estudiantes conozcan los analizadores básicos, como los utilizan algunas herramientas específicas y los apliquen al análisis de textos de poesía.

Organización del tiempo

El curso es de 120 minutos repartidos de la siguiente manera:

Actividad	**Duración en min.**
Introducción	5
Ejercicio 1: palabras y listas	5
Procesamiento del Lenguaje Natural	10
Codificación de caracteres y formatos de documentos. Ejercicio 2.	15
Programas de análisis lingüístico	15
Ejercicio 3: Etiquetas de anotación lingüística	10
Una herramienta de análisis: Contawords	15
Ejercicio 4: usando Contawords	15
Otras herramientas de análisis	15
Resumen y conclusiones	15

Evaluación

Para poder evaluar la correcta adquisición de conocimientos y capacidades, se propondrán los siguientes ejercicios.

1) Ejercicio 1: planteamiento general de solución de tareas de análisis en términos de “si … entonces”.

2) Ejercicio 2: Textos procesables: codificación de caracteres, encontrando UTF8 en los procesadores y editores de texto.
3) Ejercicio 3: Etiquetarios de categorías gramaticales.
4) Ejercicio 4: usar ContaWords funciones básicas de procesamiento

Referencias bibliográficas y materiales web

EAGLES (1996a). Recommendations for the morphosyntactic annotation of corpora, *Eag-tcwg-mac/r*, ILC-CNR, Pisa. En http://www.ilc.cnr.it/EAGLES96 (último acceso junio 2016).

EAGLES (1996b). Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora, *Eag-clwg-morphsyn/r*, ILC-CNR, Pisa. A common proposal and applications to european languages. En http://www.ilc.cnr.it/EAGLES96 (último acceso junio 2016).

Gustafson-Capková, Sofia, Britt Hartmann, 2008. Manual of the Stockholm Umeå Corpus version 2.0. Stockholm University. https://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf, (último acceso junio 2016).

Lancaster BNC tagset: http://www.natcorp.ox.ac.uk/docs/c7spec.html, (último acceso junio 2016).

STTS tagset: http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf, (último acceso junio 2016).

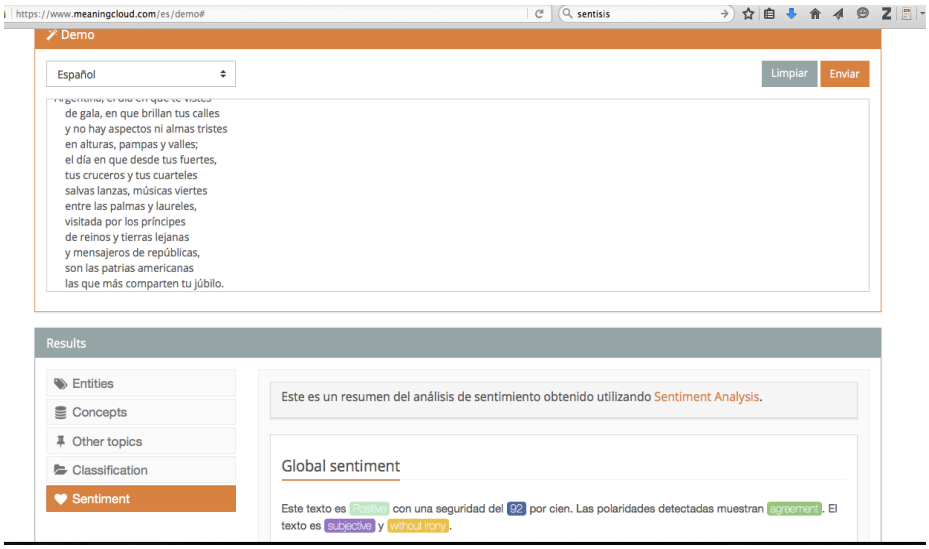
Taylor, Ann; Marcus, Mitchell and Santorini, Beatrice, 2003, "The Penn Treebank: An Overview", en Abeillé, Anne ed. *Treebanks: Building and Using Parsed Corpora*, Springer Netherlands, Dordrecht", 5—22, http://dx.doi.org/10.1007/978-94-010-0201-1_1.

EJERCICIOS

Ejercicio 1:

•Mira este programa. https://www.meaningcloud.com/demo

• Busca "Sentiment" en las opciones de análisis. Nos dice si un texto puede ser interpretado como una opinión positiva o negativa.



Ejercicio 1.1.

Piensa primero, ¿cómo crees que puede hacerlo? Se parece mucho al corrector ortográfico que acabamos de explicar. ¿Puedes deducir su funcionamiento como una tarea "si … entonces …"?

Prueba ahora estos dos textos por separado ¿Hace el programa lo que esperabas?
"Este poema es el más bonito que he leído jamás"
"Este poema es el peor que he leído jamás"

Ejercicio 1.2.

• De este extracto del poema de Rubén Darío, *Canto a la Argentina*, qué palabras pondrías en la lista de palabras positivas y cuáles en la de negativas.

Argentina, el día en que te vistes
de gala, en que brillan tus calles
y no hay aspectos ni almas tristes
en alturas, pampas y valles;
el día en que desde tus fuertes,
tus cruceros y tus cuarteles
salvas lanzas, músicas viertes

entre las palmas y laureles,
visitada por los príncipes
de reinos y tierras lejanas
y mensajeros de repúblicas,
son las patrias americanas
las que más comparten tu júbilo.

Ejercicio 2

• Escoge tu obra favorita en castellano y descárgala
• Por favor, ¡que sea un texto larguito! (unos 1000 versos o más)
• Si no está en .txt codificación utf8, ábrelo en un procesador (Word, OpenOffice) y guarda una versión .txt, codificación utf8.

Ejercicio 3

Etiquetas EAGLES para verbos en español usadas por Freeling.

Pos.	Atributo	Valor	Código
1	Categoría	Verbo	V
2	Tipo	Principal	M
		Auxiliar	A
		Semiauxiliar	S
3	Modo	Indicativo	I
		Subjuntivo	S
		Imperativo	M
		Infinitivo	N
		Gerundio	G
		Participio	P
4	Tiempo	Presente	P
		Imperfecto	I
		Futuro	F
		Pasado	S
		Condicional	C
		-	0
5	Persona	Primera	1
		Segunda	2
		Tercera	3
6	Número	Singular	S
		Plural	P
7	Género	Masculino	M
		Femenino	F

Codifica las etiquetas para las siguientes formas verbales, de acuerdo con los códigos de etiquetas EAGLES:

cantaría

abriríamos
maté
leyera
correrás
oídas

Ejercicio 4

1. Utiliza ContaWords en http://contawords.iula.upf.edu
2. Sigue las instrucciones para subir el archivo que has guardado antes como .txt, utf8 y seleccionar la lengua.
3. Encuentra y acciona el comando "Ejecutar".
4. Descarga el resultado cuando aparezca en pantalla. Puede tardar un poco.
5. Cambia el nombre al documento Excel cuando lo guardes en tu ordenador.
6. Inspecciona los resultados del análisis de ContaWords y responde a estas preguntas:
    1. En la pestaña de nombres, ¿por qué están todos en singular?
    2. Encuentra cuáles son las 10 palabras más frecuentes del texto ¿Tienen alguna característica en común?

7. Crea un Filtro en la pestaña "G-freq-bigram" siguiendo las instrucciones de las transparencias del curso.
8. Con el filtro creado, selecciona las opciones para que se muestren únicamente las combinaciones de Adjetivo-Nombre, y las de Nombre-Adjetivo.

SOLUCIONES A LOS EJERCICIOS

Solución Ejercicio 1.1

• Efectivamente, como en el caso del corrector ortográfico tiene una lista de palabras que considera "positivas" y otras "negativas". El programa recorre el texto y por cada palabra que encuentra que está en la lista suma un punto para positivo o para negativo, según la palabra encontrada.

Solución Ejercicio 1.2.

¿Se parecen tus listas a estas?

Palabras positivas	Palabras negativas
gala	tristes
brillan	
laureles	
júbilo	

Para esta tarea se tienen en cuenta nombres, verbos, adjetivos y adverbios, pero no artículos o pronombres, y de estas las que aportan "polaridad" positiva o negativa. Las demás, se ignoran.

Solución Ejercicio 2.

En las transparencias de la sesión están las instrucciones. Para verificar que se ha hecho correctamente, cerrar el documento y volverlo a abrir con el procesador. Este pedirá confirmación sobre la codificación en UNICODE o UTF8.
Si no te ha salido bien, siempre puedes escoger un poema ya en .txt utf8 en esta dirección.
https://www.gutenberg.org/browse/languages/es

Solución Ejercicio 3

cantaría	VMIC3S0
abriríamos	VMIC1P0
maté	VMIS1S0
leyera	VMSI3S0
correrás	VMIF2S0
oídas	VMP00PF

Información sobre etiquetas EAGLES en FreeLing en:
http://nlp.lsi.upc.edu/freeling-old/doc/tagsets/tagset-es.html

Solución ejercicio 4

Si hay algún problema con tu texto, prueba con el texto Canto a la Argentina, que está en el fichero "Texto-utf8-RubenDario.txt" que está en los Materiales del curso.

4.6.1. Porque se trata de "lemas", representantes del paradigma flexivo. Las palabras tal como aparecen en el texto están en la pestaña "E-freq-word".

4.6.2. Deberías haberlo hecho en la pestaña "E-freq-word". Las palabras más frecuentes son, normalmente, palabras gramaticales que se suelen considerar "stop words" y que se pueden eliminar en muchos programas de análisis lingüístico.

4.7. Recuerda que tienes que insertar una fila en blanco y con el cursor sobre ella ir al menú Datos-Filtro. También se puede encontrar en algunas versiones de Excel en el menú: