



PROGRAMA DE ESTUDIO

PROCESAMIENTO DE CORPUS TEXTUALES Y ORALES

Asignatura

Clave

8 °, 9°

Semestre

06

Créditos

Ingeniería Eléctrica

División

Ingeniería en Computación

Departamento

Ingeniería en Computación

Carrera en que se imparte

Asignatura:

Obligatoria
de elección

Optativa

Horas:

Teóricas

Prácticas

Total (horas):

Semana

16 Semanas

Modalidad: Curso

Asignatura obligatoria antecedente: Ninguna.

Asignatura obligatoria consecuyente: Ninguna.

Objetivo(s) del curso:

El alumno conocerá los métodos, problemas y aplicaciones del Procesamiento del Lenguaje Natural basado en el procesamiento de corpus textuales y será capaz de imaginar soluciones a problemas concretos en este campo.

Temario

NÚM.	NOMBRE	HORAS
1.	Fundamentos teóricos del procesamiento de corpus textuales	12.0
2.	Constitución y compilación de corpus	8.0
3.	Anotación de corpus textuales	12.0
4.	Herramientas y técnicas de análisis	10.0
5.	Aplicaciones	6.0
		48.0
	Prácticas de laboratorio	0.0
	Total	48.0

PROCESAMIENTO DE CORPUS TEXTUALES Y ORALES

(2 / 4)



1 Fundamentos teóricos del procesamiento de corpus textuales

Objetivo: El alumno conocerá las bases teóricas para el procesamiento de corpus textuales.

Contenido:

- 1.1 Antecedentes
- 1.2 Teoría de la información
- 1.3 Teoría de la probabilidad
- 1.4 Técnicas de *clustering*

2 Constitución y compilación de corpus

Objetivo: El alumno se familiarizará con las generalidades de los corpus textuales, así como de su compilación y constitución.

Contenido:

- 2.1 Tipología de corpus lingüísticos
- 2.2 Corpus textuales
- 2.3 Corpus orales
- 2.4 Representatividad y balance del contenido de los corpus
- 2.5 Interfaces electrónicas para los corpus

3 Anotación de corpus textuales

Objetivo: El alumno conocerá los esquemas de aplicación de etiquetas o anotaciones a los corpus textuales y dominará el lenguaje XML para estructurar documentos.

Contenido:

- 3.1 Etiquetado de partes de la oración (POS)
- 3.2 Etiquetado sintáctico
- 3.3 Etiquetado semántico
- 3.4 Etiquetado de fenómenos pragmático-discursivos
- 3.5 Otros niveles de anotación
- 3.6 Estándares y lineamientos para la anotación de corpus: XML
 - 3.6.1 Analizadores y motores
 - 3.6.2 Edición y composición
 - 3.6.3 Definición del tipo de documento
 - 3.6.4 Entidades
 - 3.6.5 Lenguajes de enlace y estilo extensible
 - 3.6.6 Características avanzadas

4 Herramientas y técnicas de análisis

Objetivo: El alumno dominará las técnicas de corpus textuales y conocerá las herramientas de análisis disponibles.

Contenido:

- 4.1 Problemas de *tokenización* de corpus textuales



- 4.2 Frecuencias de n-gramas de palabras gráficas, tipos de palabra y otras estructuras
- 4.3 Generación de concordancias de las estructuras
- 4.4 Colocaciones y medidas de asociación
- 4.5 Análisis fraseológico y oracional
- 4.6 Herramientas de análisis disponibles

5 Aplicaciones

Objetivo: El alumno se familiarizará con las aplicaciones de los corpus textuales tanto en la ingeniería como en las investigaciones de lengua natural.

Contenido:

- 5.1 Ingeniería lingüística
 - 5.1.1 Lexicografía y terminología
 - 5.1.2 Traducción automática
 - 5.1.3 Procesamiento de voz/habla (análisis y síntesis)
- 5.2 Lingüística
 - 5.2.1 Investigación fonológica, morfológica, morfosintáctica, sintáctica, gramatical, entre otros tipos.
 - 5.2.2 Investigación diacrónica.

Bibliografía básica:

GARSDIE, R., LEECH, G., MCENERY, A
Annotation. Linguistic Information from Computer Text Corpora
 Addison Wesley Longman, 1997

MANNING, Christopher, FRIDRICH Schütze
Foundations in Statistical Natural Language Processing
 Cambridge
 The MIT Press, 1999

MASON, Oliver
Programming for Corpus Linguistics. How to do Text Analysis with Java
 Edinburgh
 Edinburgh University Press, 2000

MCENERY, T., WILSON, A.
Corpus Linguistics
 2nd edition
 Edinburgh
 Edinburgh University Press, 2001

OAKES, Michael
Statistics for Corpus Linguistics
 Edinburgh
 Edinburgh University Press, 1998



Bibliografía complementaria:

GOLDFARB, CH. y PRESCOD, P.
Manual de XML
 Prentice Hall, 1999

MCLAUGHLIN, B.
Java and XML
 O'Reilly, 2000

Sugerencias didácticas:

Exposición oral	<input checked="" type="checkbox"/>	Lecturas obligatorias	<input type="checkbox"/>
Exposición audiovisual	<input checked="" type="checkbox"/>	Trabajos de investigación	<input type="checkbox"/>
Ejercicios dentro de clase	<input checked="" type="checkbox"/>	Prácticas de taller o laboratorio	<input checked="" type="checkbox"/>
Ejercicios fuera del aula	<input checked="" type="checkbox"/>	Prácticas de campo	<input type="checkbox"/>
Seminarios	<input type="checkbox"/>	Otras	<input type="checkbox"/>

Forma de evaluar:

Exámenes parciales	<input type="checkbox"/>	Participación en clase	<input checked="" type="checkbox"/>
Exámenes finales	<input type="checkbox"/>	Asistencias a prácticas	<input type="checkbox"/>
Trabajos y tareas fuera del aula	<input checked="" type="checkbox"/>	Otras	<input type="checkbox"/>

Perfil profesiográfico de quienes pueden impartir la asignatura

Ingeniero en computación familiarizado con el procesamiento y análisis automático de corpus textuales.