

LLICENCIATURA EN LINGÜÍSTICA

13305 Fonaments de processament del llenguatge natural

1. Dades descriptives de l'assignatura:

Curs acadèmic: 2007-2008

Nom de l'assignatura: Fonaments de Processament del Llenguatge Natural

Codi: 13305

Tipus: obligatòria

Estudis: Lingüística

Nombre de crèdits: 4

Nombre total d'hores de dedicació: 100h.

Temporalització:

- curs: 1r curs
- tipus: trimestral
- període: 1r trimestre

Professora: Núria Bel Rafecas

Llengua de docència: català (grup gran) català i castellà (grups petits)

Edifici: Rambla

Horari: tarda

2. Presentació de l'assignatura

Aquesta assignatura és una introducció als conceptes computacionals que són centrals al processament del llenguatge natural, al mateix temps que exemplifica l'ús dels conceptes més bàsics de l'anàlisi lingüística en aquest domini. L'assignatura està centrada en fer que l'alumne adquireixi les bases de la definició de processos algorítmics, en les característiques de l'ús d'eines informàtiques i en les particularitats del material lingüístic que fa del seu processament un domini específic d'especialització.

3. Prerequisits per al seguiment de l'itinerari formatiu

Per cursar aquesta assignatura és imprescindible poder llegir textos en anglès i familiaritat amb l'ús de programes i eines informàtiques.

4. Competències a assolir en l'assignatura

4.1 Específiques

CE1 - Tècniques de processament. Saber construir: Expressions Regulars, Autòmats i Transductors, Gramàtiques lliures de context basades en unificació i avaluar-ne els resultats.

CE2 - Identificar components de l'**arquitectura** d'un sistema de processament. Saber reconèixer parts de sistemes complexos.

CE3 - **Saber definir formalment** les unitats lingüístiques en diferents nivells de representació

CE4 - Identificar les condicions per a discriminar possibles àrees **d'aplicació** pel que fa al processament del llenguatge natural. Conèixer la importància de les restriccions

4.2 Generals

CG1 – Assolir familiaritat en l'ús d'eines informàtiques de programació, tant pels conceptes com per la pràctica

CG2- Assumir la naturalesa del treball computacional:

- capacitat d'**autocrítica** pel que fa als errors propis en la implementació de sistemes (error humà i poca autonomia del comportament computacional),
- capacitat per **aplicar coneixements teòrics** en un cas pràctic

- concisió en l'expressió del coneixement
- capacitat **d'anàlisi** per resoldre problemes de complexitat creixent
- tenacitat** en el treball pràctic

CG3 – Adquirir autonomia i criteris de rellevància mitjançant el treball amb bibliografia auxiliar.

5. Objectius generals de l'assignatura

El curs pretén que l'alumne sàpiga usar algorismes i tècniques fonamentals per al processament del llenguatge natural, tant mitjançant mètodes basats en dades estadístiques com en informació simbòlica obtinguda amb diferents eines: analitzadors d'estats finits, analitzadors basats en gramàtiques lliures de context i en gramàtiques d'unificació.

Es farà especial èmfasi en que l'alumne aprengui com aplicar els conceptes lingüístics i els mètodes computacionals bàsics per al processament dels diferents nivells de representació lingüística: superficial, morfològica, sintàctica i semàntica. També aprendrà a reconèixer la seva rellevància en els diferents àmbits d'aplicació del PLN.

Les sessions de pràctiques proposaran a l'estudiant la solució d'exercicis amb l'ús dels diferents mètodes, programes i eines per al processament del llenguatge natural que està estudiant. Els exercicis estaran relacionats amb el programa de teoria i tenen com a objectiu l'exemplificació pràctica dels coneixements teòrics.

6. Avaluació

6.1. Criteris generals d'avaluació

La nota final de l'assignatura es farà segons els següents percentatges:

- 40% nota de les pràctiques i
- 60% nota de l'examen final (teoria i pràctica).

Cada pràctica serà avaluada de 1 a 3 tenint com a criteris la dedicació (ús de material de classe i de lectures auxiliars) i la solució de l'exercici: 1 presentat, 2 exercici treballat, 3 exercici resolt).

El càlcul de la nota final es farà d'acord amb la següent fórmula:

$$\text{Nota final} = ((\text{Nota mitja de pràctiques} * 10) / 3) * 0,4 + \text{Nota examen} * 0,6$$

Condicions per l'aplicació:

- Lliurament de totes les pràctiques
- Nota mitja de les pràctiques i Nota de l'examen igual o superior a 4

L'avaluació de les pràctiques té com a objectiu motivar l'esforç constant que permetrà l'alumne anar assimilant correctament l'aplicació dels coneixements que es van presentant a classe i a la bibliografia auxiliar.

L'examen avaluarà:

- la capacitat d'aplicar els coneixements adquirits durant el curs a escenaris diferents dels que s'han vist. Es tracta de copsar si l'alumne pot diferenciar entre un problema en concret i el tipus de problemes que pot resoldre l'aplicació d'una tècnica concreta.
- l'adquisició d'informació clau, que ha de ser entesa i assumida per tal que pugui ser fonaments d'altres coneixements de la llicenciatura.

6.2. Avaluació de les competències

La competència CE1 s'avaluarà principalment en l'examen final, tot i que les sessions de seminari i pràctiques donaran també informació sobre el procés d'adquisició.

Les competències CE2, CE3, CE4 i totes les generals s'adquireixen bàsicament en les pràctiques que aniran fent al llarg del trimestre i el seu seguiment i avaluació anirà conjuntament amb les assistències al desenvolupament de la pràctica i la seva avaluació. Les pràctiques estan dissenyades per tal d'augmentar la complexitat de diferents factors, cada un d'ells relacionat amb les diferents competències. D'aquesta manera es fa una avaluació gradual

7. Continguts de l'assignatura

7.1. Continguts teòrics

1. Introducció.

- Breu història del Processament del Llenguatge Natural.
- Objectius i àmbits d'aplicació del PLN

BLOC 1: Les paraules

2. Les paraules

- Identificació d'unitats: els mots. Autòmats i Expressions Regulars.
- Categories lèxiques i categories funcionals. Propietats estadístiques, morfològiques i semàntica lèxica
- Compostos i col·locacions. Mesures d'associació lèxica.
- Morfemes i analitzadors morfològics. Transductors d'estats finits i morfologia de doble nivell

3. Cadenes de paraules i Models de llenguatge

- Seqüències i models de n-grames.
- Etiquetatge i Models de llenguatge. Categories, etiquetes, desambiguació basada en regles, estocàstica i Cadenes de Markov.

BLOC 2. Les oracions

4. Les oracions

4.1. Sintaxi: L'estructura de l'oració

- Categories, constituents i funcions.
- Gramàtiques, regles lliures de context i arbres.
- Analitzadors: algorismes i tècniques de cerca.

4.2. Processament d'informació sintàctica.

- Concordança. Coordinació. Subcategorització.
- Trets i unificació. Formalismes d'Unificació.
- Gramàtiques lliures de context augmentades amb unificació.

4.3. Anàlisi sintàctica probabilística amb gramàtiques lliures de context. Anàlisi probabilística lexicalitzada.

4.4. Semàntica: El significat de les oracions

- Representació de la informació semàntica.
- Informació semàntica a les gramàtiques d'unificació.

7.2 Pràctiques

1. Fonaments de processament: expressions regulars i autòmats
2. Dades estadístiques del llenguatge: anàlisi de corpus

3. Anàlisi sintàctica i analitzadors: gramàtiques lliures de context i formalismes d'unificació (DCG en PROLOG, PATR)

Bibliografia bàsica

Allen, J. 1995, Natural Language Understanding (second edition), Benjamin Cummins Publishing.

Dale, R., H. Moisl, H. Somers, 2000, Handbook of Natural Language Processing, Marcel Dekker, Inc., New York.

Gazdar G. y Ch. Mellish, 1989, Natural Language Processing in Prolog, Adison Wesley

Jurafsky, D. & J. Martin, 2000, Speech and Language Processing, Prentice Hall [enguany encara treballarem amb la primera edició, tot i que hi ha una segona del 2008 amb canvis substancials]

8. Metodologia

Atesa la distribució de feines que té l'alumne, el professor donarà suport i assistirà l'estudiant en el seu treball personal. L'assignatura estarà dirigida a que l'alumne assimili que ha de doblar com a mínim les classes presencials amb treball personal.

En l'organització de l'assignatura, tindrem en compte les següents dedicacions setmanals:

- Sessions llargues, 1:30
- Sessions curtes: 45 m. en grup petit
- 1 pràctica setmanal de 2 h. de treball personal¹
- Estudi complementari a les sessions presencials, 2h de treball personal

A més a més, l'estudiant pot recórrer al professor en el seu temps de tutories. El professor animarà als estudiants a formular les seves preguntes per correu electrònic atès que aquesta forma fa el professor accessible quan s'està treballant, i perquè obliga l'estudiant a formular el seu problema, esforç que en la majoria dels casos equival a resoldre el problema.

7. Recursos didàctics

7.1. Web de l'assignatura

The screenshot shows a web browser window with the following content:

Fonaments de Processament del Llenguatge Natural
13305 – UPF - Lingüística 2004-2005

Prof. [Núria Bel Rafecas](#)
Despatx 226, Institut Universitari de Lingüística Aplicada, IULA, Edifici Rambla
Telf. 93542 – 2307
correu_e: nuria.bel@upf.edu

Podeu consultar més informació sobre:

- [Objectius](#)
- [Temari i bibliografia bàsica](#)
- [Recursos](#)
- [Avaluació](#)

Horari de l'assignatura		
Classe	divendres 17:05 a 18:35	Aula: 202
Seminari	dilluns 18:45 a 19:30	Aula: 102
	dilluns 19:40 a 20:25	Aula: 416
Tutoria	dilluns 17:30 a 18:30 i/o hores convingudes	Despatx 226

Programació

DIA	Seminari	lectures auxiliars	DIA	Classe
27-09	1. Introducció a PLN Lectura a classe	Jurafsky, D. & J. Martin, <i>Speech and Natural Language Processing</i> , Prentice Hall, 2000. (cap. 1)	1-10	1. Introducció a PLN Breu història. Objectius i

La pàgina web conté informació sobre el pla docent de l'assignatura: Objectius, temari i bibliografia, Recursos i detalls de l'Avaluació. A més consta la programació detallada (sessió a sessió) del curs, juntament amb la bibliografia obligatòria i accés a la pràctica i tota la informació necessària.

7.2. Recursos per a les classes teòriques

Les classes teòriques tindran de base presentacions projectades que estaran disponibles abans de la classe corresponent en format pdf. En la web l'estudiant també trobarà referències a les lectures complementàries a les explicacions donades en classe.

7.3. Recursos per a les pràctiques

Per a les pràctiques, el web de l'assignatura té l'enunciat de la pràctica amb una explicació completa per a la realització de la pràctica i material addicional de suport per a la seva realització.

Les classes pràctiques també usaran materials disponibles a internet i demostracions d'aplicacions que son interessants per visualitzar els objectius del PLN.

En les que ha estat possible, les pràctiques inclouen fitxer d'autoavaluació, per tal que l'estudiant sigui el més autònom possible en la realització i compleció de la pràctica.