SAR

SISTEMAS DE ALMACENAMIENTO Y RECUPERACIÓN DE INFORMACIÓN

Créditos: 3 (T) + 1,5 (P)

Emilio Sanchis (R)	esanchis@dsic.upv.es
Encarna Segarra	esegarra@dsic.upv.es
Salvador España	sespana@dsic.upv.es
José Ángel González	jogonba2@inf.upv.es

Indice

- Sobre la asignatura.
  - Descripción general de la asignatura
  - Relación con otras disciplinas
  - Objetivos indispensables
  - Contenidos Teoría y Prácticas
  - Bibliografía
  - Evaluación
- Investigación
- Pensemos...

Descripción general de la asignatura

- Se presentarán algoritmos para el acceso a la información no estructurada compuesta por grandes volúmenes de datos.
- Se estudiarán estructuras de datos eficientes para la implementación de diccionarios y búsqueda de términos.
- Se aplicarán las técnicas estudiadas a tareas de recuperación de información, indexación,...

Relación con otras asignaturas

- (11548) Bases de datos y sistemas de información
- (11551) Estructuras de datos y algoritmos
- (11563) Tecnología de sistemas de información en la red

Objetivos indispensables

- Conocer y desarrollar técnicas de aprendizaje computacional y diseñar e implementar aplicaciones y sistemas que las utilicen sobre grandes volúmenes de datos.
- Aprender de manera autónoma nuevos conocimientos y técnicas adecuados para la concepción, el desarrollo, la evaluación o la explotación de sistemas informáticos.
- Desarrollar habilidades de aprendizaje necesarias para emprender estudios posteriores con un alto grado de autonomía.

Competencia Transversal

Comunicación efectiva

- Oral: transmitir convicción y seguridad, ilustrar el discurso para facilitar su comprensión y adaptarlo a las condiciones formales exigidas en presentaciones orales de duración media (10-15 minutos aproximadamente)
- Escrita: seleccionar la información relevante y ordenarla de forma lógica para elaborar un documento que sea comprensible, utilizando los recursos adecuados.

Contenidos Teoría

1. Introducción a la Recuperación de Información
2. Índices invertidos, términos y consultas
3. Diccionarios y búsqueda con tolerancia
4. Modelo de espacio vectorial
5. Evaluación de Sistemas de Recuperación de Información
6. Búsqueda en la web
7. Construcción y compresión de índices
8. Algoritmos de búsqueda secuencial
9. Nuevas tendencias en Recuperación de Información

Contenidos Prácticas

PL0. Introducción a Python

PL1. Pig Latin

PL2. Cuenta palabras

PL3. Mono infinito

PL4. Proyecto (compartido con asignatura Algorítmica)

Las prácticas empezarán el 19 de Febrero

Bibliografía

***A Introduction to Information Retrieval**:*
Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze.
*Cambridge University Press, **2009**.*

Modern Information Retrieval:
Ricardo Baeza-Yates and Berthier Ribeiro-Neto
*Addison Wesley, First printed **1999***

***Managing Gigabytes**.*
Ian H. Witten, Alistair Moffat, and Timothy C. Bell
*Morgan kaufmann publishers, **1999***

Speech and Language Processing (3rd ed. draft, dec. 2020)
Dan Jurafsky and James H. Martin
https://web.stanford.edu/~jurafsky/slp3/

Evaluación

40% pruebas escritas:

Acto1 (20%)

Acto2 (20%)

Acto3 (Recuperación)

40% trabajos laboratorio (en grupos de 4)

- 10% seguimiento del trabajo a lo largo del curso
- 30% desarrollo de un trabajo de laboratorio (en grupo)

20% estudio y exposición de un tema (en grupos de 4)

*Todo alumno puede presentarse al Acto 3, la nota que contará será la última obtenida

*Alumnos con dispensa de asistencia: misma evaluación

INVESTIGACIÓN

Campos de investigación relacionados

- Recuperación de Información monolingüe

(Information Retrieval).

- Recuperación de Información Multilingüe

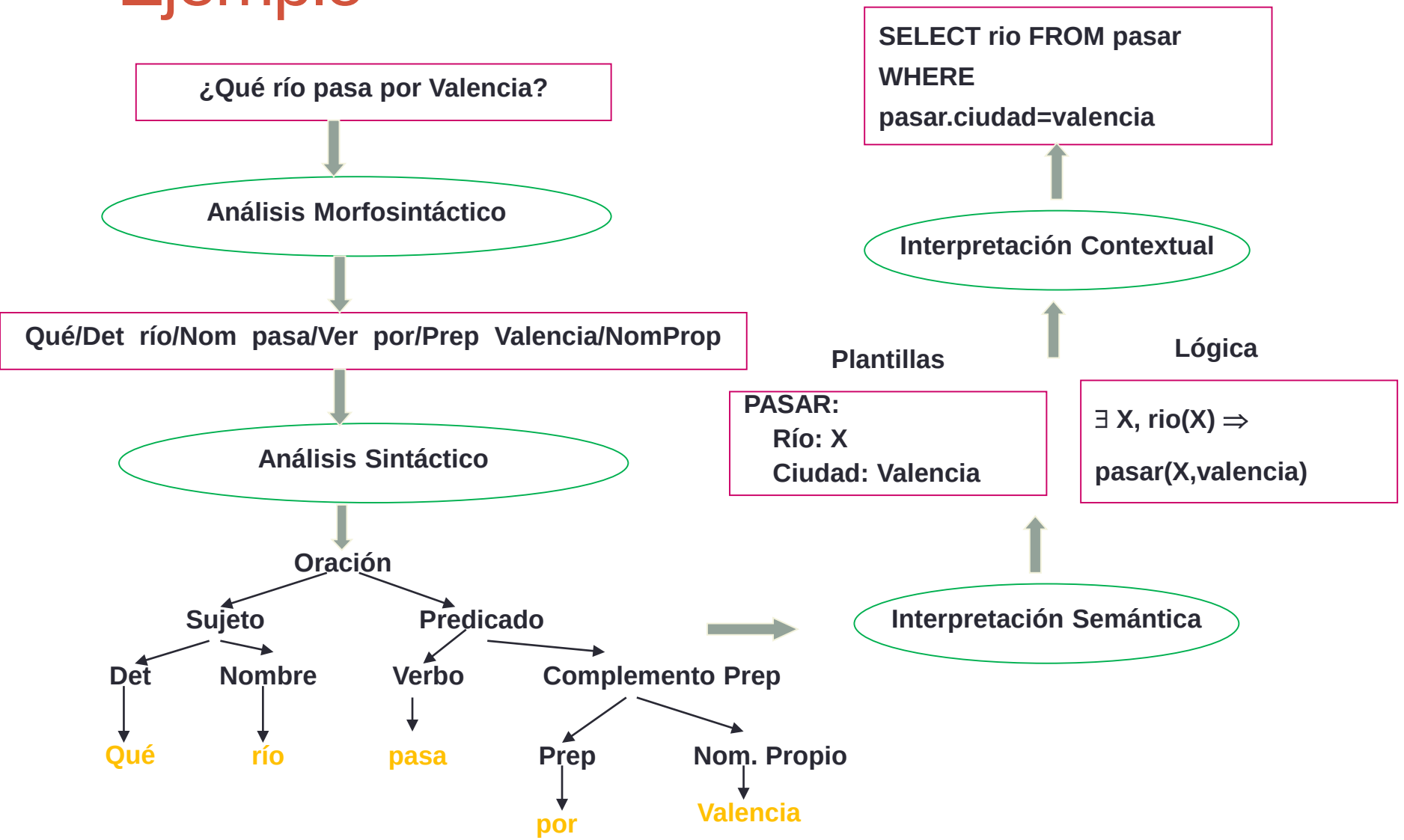
(Cross Language Information Retrieval).

- Lingüística Computacional

(Computational Linguistic):

  - Se ocupa de la aplicación de métodos computacionales en el estudio científico del lenguaje.
  - Aproximaciones de los lingüistas computacionales a los lenguajes se llaman habitualmente *Procesamiento del Lenguaje Natural (Natural Language Processing)*.
  - Aproximaciones de la ingeniería al lenguaje se llaman habitualmente *Language Engineering (LE)* o *Human Language Technologies (HLT)*.

Ejemplo


¿Qué río pasa por Valencia?
Análisis Morfosintáctico
Qué/Det río/Nom pasa/Ver por/Prep Valencia/NomProp
Análisis Sintáctico
Oración
Sujeto
Predicado
Det
Nombre
Verbo
Complemento Prep
Qué
río
pasa
Prep
Nom. Propio
por
Valencia
Interpretación Semántica
Plantillas
PASAR:
Río: X
Ciudad: Valencia
Lógica
∃ X, rio(X) ⇒
pasar(X,valencia)
Interpretación Contextual
SELECT rio FROM pasar
WHERE
pasar.ciudad=valencia

Búsqueda de respuestas

(Question Answering -QA)

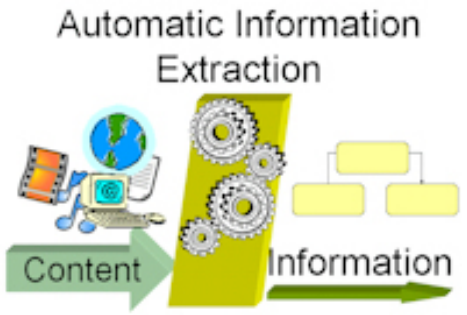
Muchos sistemas de QA tienen como paso previo la tarea de Recuperación de Documentos relevantes a la consulta

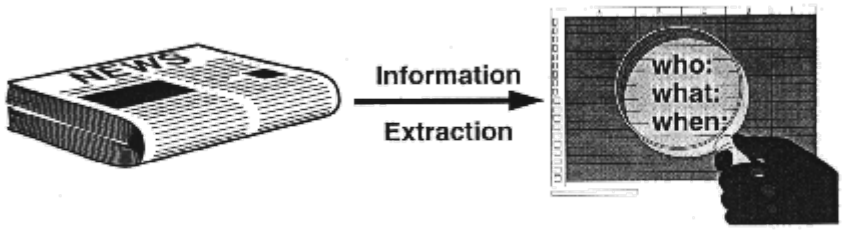
1989 **Remedia** Publications, Comprehension
First Man Walks on the Moon!(THE MOON, July 20,1969)- "That's one small step for man, one giant leap for mankind." Those words were spoken from the moon today. They were said by Neil Armstrong. He will be known for all time as the first man to walk on the moon. This trip had been planned for nine years. The trip itself will last almost two weeks. That's how long it takes to travel to the moon, study it for a few days, and return to Earth.
Armstrong picked up moon rocks to bring back to Earth. He says some are purple.
Armstrong will leave some things behind, too. One is a United States flag. The other is a plaque. It lists the names of Armstrong and his fellow moon-visitors. It says they came in peace for all mankind. More trips are planned to the moon in the near future.

1. **Who** was the first man to walk on the moon?
2. **What** did he say?
3. **When** did this story happen?
4. **Where** are more trips planned in the near future?
5. **Why** did the visitors come to the moon, according to the plaque they left?

Sistemas de extracción de información (*Information Extraction*):

- La información a buscar está predefinida (*plantillas*).
- **Entrada**: texto no estructurado.
- **Salida**: texto estructurado en forma de plantillas.





Ejemplo a partir de documentos no estructurados:

Hadson Corp. said it expects to report **a third quarter net loss** of $ 17 million to $ 19 million because of special reserves and continued low natural gas prices.
The Oklahoma City energy and defense concern said it will record a $ 7. 5 million reserve for its defense group, including a $ 4. 7 million charge related to problems under a fixed price development contract and $ 2. 8 million in overhead costs that won' t be reimbursed.
In addition, Hadson said it will write off about $ 3. 5 million in costs related to international exploration leases where exploration efforts have been unsuccessful.
The company also cited interest costs and amortization of goodwill as factors in **the loss** .
A year earlier, net income was $ 2. 1 million, or six cents a share, on revenue of $ 169. 9 million

---

Company Losses

company name	company description	loss description	amount	link to text
Hadson Corp.	The Oklahoma City energy and defense concern	a third quarter net loss	$ 17 million to $ 19 million	source

Clasificación de textos

(*Automated Text Categorization*):

- Asignar a los documentos categorías previamente establecidas.
- Aplicación: *Content/Text Filtering (TF)*.
  - Sistema que decide qué documentos son relevantes para el usuario en función de su perfil previamente establecido.
- Ejemplos:
  - Filtrado de correo (anti-spam)
  - Filtrado de páginas web con contenidos violentos.

PENSEMOS.....

La época actual de la sociedad de la información

La ventaja:

- La disponibilidad de grandes volúmenes de información.

El problema:

- La localización de la información que interesa a cada persona.

Solución:

- Los sistemas de búsqueda de información.
  - **Entrada:** palabras clave a buscar.
  - **Salida:** relación de documentos ordenada.
  - Ejemplos:
    - Google, Bing, ...

¿Qué inconvenientes encuentras en los buscadores de información disponibles en la actualidad?

Los buscadores de información:

Ventajas: gran rapidez de funcionamiento.

Problemas:

Precisión baja: nivel bajo de comprensión del texto.

No realizan todo el trabajo: buscar la información dentro del documento.

Posible solución:

- Aplicación de técnicas de PLN para mejorar los resultados.
- Sistemas de búsqueda de respuestas (*Question Answering*):

  Entrada: frases completas en lenguaje natural.

  Salida: trozos de texto con la información requerida.

¿Qué información debería almacenar (y de qué manera) un buscador de información para que funcionase de forma óptima?

La solución:

- Se verá en la asignatura