5ª Jornada de Difusión Tecnológica

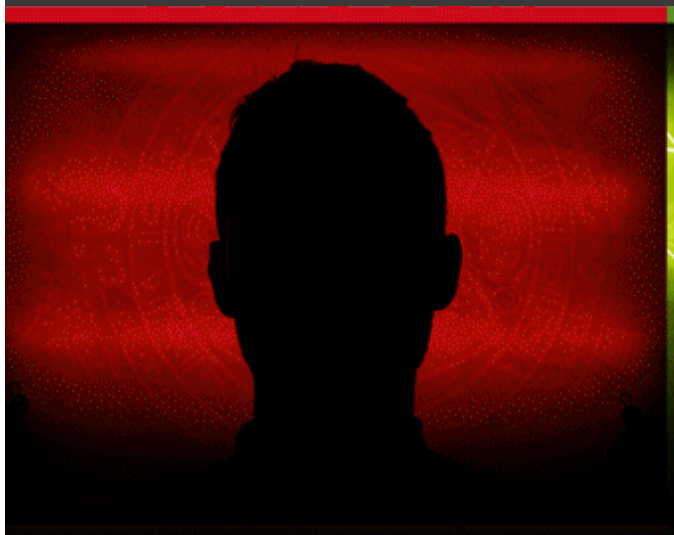
La Minería de Textos y Opinión, oportunidad para la adquisición de información no estructurada.

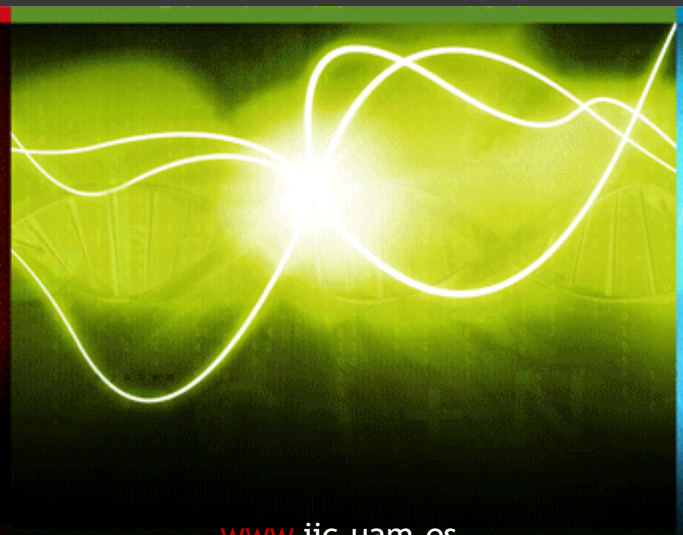
23 noviembre 2010

instituto de ingeniería del conocimiento

D. Antonio Moreno

Profesor Titular e Investigador del área Text & Opinion Mining del IIC





www.iic.uam.es

Esquema de la presentación

1. Estado de la cuestión en minería de datos y de opinión
2. Estado de la cuestión en semántica computacional
3. Algunas aplicaciones

¿Qué es la minería de datos (data mining)?

- *Data mining* es encontrar patrones dentro de documentos que permitan establecer relaciones entre conceptos.

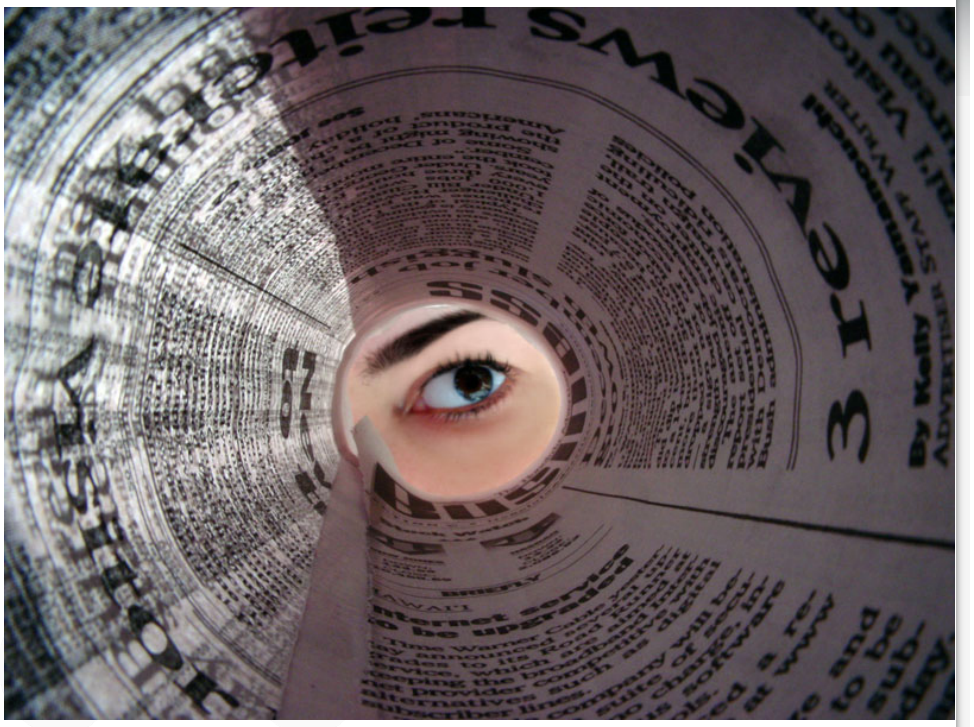
  - "por ejemplo" indica una relación entre *concepto* y *ejemplo*.

  - "regalos de Navidad, por ejemplo, una corbata, un frasco de perfume"

- Ciertos conceptos sirven para clasificar el contenido de un documento.

  "nómina" "préstamo" "cuenta corriente" "tarjeta de crédito"

  Dominio financiero



Los tres problemas

1. Reconocer los conceptos clave de cada dominio
2. Descubrir sus relaciones entre ellos, especialmente las relaciones jerárquicas
3. Descubrir los sinónimos de cada concepto en el dominio

¿Qué es un concepto clave?

- **Términos**: *hipoteca concedida*
- **Sintagmas**: *concesión de hipoteca*
- **Expresiones**: *me concedieron la hipoteca*

→ Todos ellos pueden ser sinónimos (patrones sintácticos que se refieren al mismo significado) y reconocerlos es esencial para interpretar el contenido de un documento

¿Qué es la minería de opiniones?

- Extracción de opiniones de fuentes de información
- Presentación de las opiniones en un formato estructurado

Una opinión es una **expresión subjetiva** y personal acerca de cosas, personas, situaciones…

Está muy relacionada con:

- Extracción de información (IE)
- Recuperación de información (IR)
- Procesamiento del lenguaje natural (NLP)
- Aprendizaje automático (ML)
- Minería de datos (DM)
- Clasificación de textos (TC)

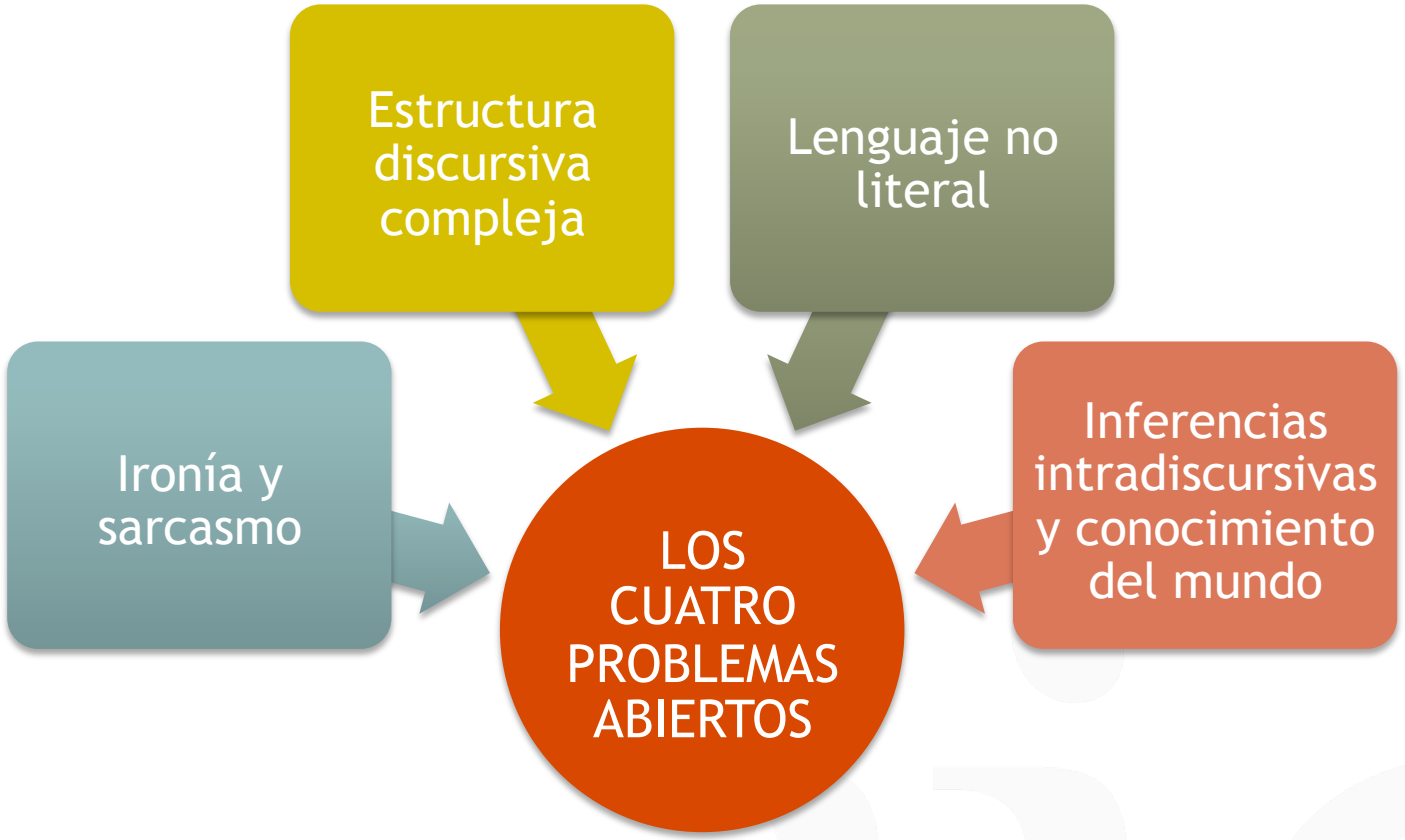
Aplicaciones actuales del análisis de opiniones y de sentimientos

- Ayuda en la toma de decisiones del consumidor
- Medición de popularidad
- Guía para I+D de nuevos productos/mejora de existentes
- Marketing
- Recomendaciones en redes sociales

Problemas con el análisis de opiniones y sentimientos

- La clave: interpretar las palabras dentro de un contexto, no hay léxico intrínsecamente positivo o negativo
  - GRANDE: déficit grande (-) / grandes beneficios (+)
  - MAL: esto está mal (-) / esto no está nada mal (+)
- Distinguir entre sentidos subjetivos y sentidos objetivos de la misma palabra (ambigüedad de sentidos):
  - INTERÉS: **Interés** por algo (subjetivo)
    Tasa de **interés** (objetivo)

Los cuatro problemas abiertos



→ Por dominios, los opiniones sobre productos son más fáciles de analizar que las opiniones políticas o ideológicas.

Semántica Computacional

- Tipos de “Semánticas”
  - Semántica **léxica**: el significado de las palabras (lexicones, ontologías)
  - Semántica **oracional**: significado de las oraciones (bancos de eventos)
  - Semántica **discursiva** o pragmática: significado de textos, fragmentos del discurso (modelos conceptuales de un dominio)
- La mayoría de los sistemas actuales tratan el significado como:
  - Detección de coocurrencias entre palabras que aparecen frecuentemente (Google)
  - Empleo de ontologías y lexicones de dominio (web semántica)

→ Para entender el significado de un texto hay que abordar el significado de las oraciones dentro de un dominio.

Nuestra aportación

- En el IIC combinamos:

1. La experiencia en data mining (procesado de grandes catidades de datos)
2. La experiencia en semántica computacional (modelado de situaciones y escenarios)

- Nuestra metodología:

1. Procesamos el texto a un nivel semántico más profundo que el léxico
2. Aplicamos técnicas de clasificación de *machine learning*

Aplicaciones posibles

- A gran cantidad de datos que no se puedan analizar manualmente (requisito de data mining)
- En un dominio restringido modelizable (requisito de semántica oracional)
- Beneficios: una herramienta con una efectividad y precisión superior al 50% que ahorraría tiempo al analista humano en el análisis de grandes cantidades datos rápidamente.