Clase 1: Aspectos Generales (Parte A)

Enfoques Clásicos y Neuronales a la Minería de Texto

Marcelo Errecalde[1,2]

[1]Universidad Nacional de San Luis, Argentina

[2]Universidad Nacional de la Patagonia Austral, Argentina

Resumen

1. **Nuestro grupo de investigación**
2. **Datos del Curso**
3. **Introducción**
   - Lenguaje, lingüística y computación
   - Minería de Texto (en contexto)
   - Importancia del AAT
   - Dificultades del PLN
   - Tareas y Aplicaciones

La UNSL en la Argentina

La UNSL en la Argentina

EL LIDIC en la Universidad

Universidad Nacional de San Luis

Facultades:

- Facultad de Química Bioquímica y Farmacia
- Facultad de Cs. Físico Matemáticas y Naturales <–
- Facultad de Ciencias Humanas
- ...

EL LIDIC en la Universidad

Universidad Nacional de San Luis

Facultades:

- Facultad de Química Bioquímica y Farmacia
- Facultad de Cs. Físico Matemáticas y Naturales <–
- Facultad de Ciencias Humanas
- ...

Facultad de Ciencias Físico Matemáticas y Naturales

- Instituto de Matemática Aplicada San Luis
- Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC) <–
- ....

Datos del LIDIC

Creación: Año 2000, Dr. Raúl Héctor Gallard

Dirección Dr. Marcelo Luis Errecalde

PROICO 03-2018, Dra. Marcela Printista

Proyectos: PROICO 03-1018, Dr. Guillermo Leguizmón

PROICO 03-1816, Dr. Marcelo Errecalde

Administrativas

Profesor Responsable:

Marcelo Luis Errecalde

Administrativas

Profesor Responsable:

Marcelo Luis Errecalde

Contacto:

merreca@unsl.edu.ar

o bien

merrecalde@gmail.com

Administrativas

Profesor Responsable:

Marcelo Luis Errecalde

Contacto:

merreca@unsl.edu.ar

o bien

merrecalde@gmail.com

Material del curso:

Disponible en
https://github.com/merrecalde/rio_2020

¿Qué se pretende con este curso?

¿Qué se pretende con este curso?

- Que sea útil para ustedes .... (no fácil, por los distintos perfiles)

¿Qué se pretende con este curso?

- Que sea útil para ustedes .... (no fácil, por los distintos perfiles)
- Introducir los conceptos y herramientas básicas fundamentales vinculados a la minería de textos (MT)/análisis automático de textos.

¿Qué se pretende con este curso?

- Que sea útil para ustedes .... (no fácil, por los distintos perfiles)
- Introducir los conceptos y herramientas básicas fundamentales vinculados a la minería de textos (MT)/análisis automático de textos.
- Analizar y entender las dificultades, posibilidades y principales aplicaciones de este tipo de técnicas.

¿Qué se pretende con este curso?

- Que sea útil para ustedes .... (no fácil, por los distintos perfiles)
- Introducir los conceptos y herramientas básicas fundamentales vinculados a la minería de textos (MT)/análisis automático de textos.
- Analizar y entender las dificultades, posibilidades y principales aplicaciones de este tipo de técnicas.
- Analizar con una mirada crítica los trabajos realizados en el área y/o tener las herramientas para implementar soluciones a problemas específicos.

Dificultad de balancear las áreas

Dificultad de balancear las áreas

Recuperación de la Información

Procesamiento del Lenguaje Natural

Minería de Textos

Aprendizaje Automático

Características del curso

- ¿Teoría?:

Características del curso

- ¿Teoría?: la necesaria para entender los conceptos subyacentes (más intuitiva que definiciones)

Características del curso

- ¿Teoría?: la necesaria para entender los conceptos subyacentes (más intuitiva que definiciones)
- ¿Práctica?:

Características del curso

- ¿Teoría?: la necesaria para entender los conceptos subyacentes (más intuitiva que definiciones)
- ¿Práctica?: haremos algunos ejercicios como parte de la evaluación del curso

Características del curso

- ¿Teoría?: la necesaria para entender los conceptos subyacentes (más intuitiva que definiciones)
- ¿Práctica?: haremos algunos ejercicios como parte de la evaluación del curso
- ¿Implementación/máquina?:

Características del curso

- ¿Teoría?: la necesaria para entender los conceptos subyacentes (más intuitiva que definiciones)
- ¿Práctica?: haremos algunos ejercicios como parte de la evaluación del curso
- ¿Implementación/máquina?: se verán ejemplos en Python. Notebooks disponibles para práctica más intensiva.

Organización del Curso

5 clases + evaluación

- Clase 1: Aspectos generales (conceptos, más bien teórica)

Organización del Curso

5 clases + evaluación

- Clase 1: Aspectos generales (conceptos, más bien teórica)
- Clase 2: Pre-procesamiento de textos

Organización del Curso

5 clases + evaluación

- Clase 1: Aspectos generales (conceptos, más bien teórica)
- Clase 2: Pre-procesamiento de textos
- Clase 3: Representación de documentos

Organización del Curso

5 clases + evaluación

- Clase 1: Aspectos generales (conceptos, más bien teórica)
- Clase 2: Pre-procesamiento de textos
- Clase 3: Representación de documentos
- Clase 4: Categorización supervisada de textos

Organización del Curso

5 clases + evaluación

- Clase 1: Aspectos generales (conceptos, más bien teórica)
- Clase 2: Pre-procesamiento de textos
- Clase 3: Representación de documentos
- Clase 4: Categorización supervisada de textos
- Clase 5: Enfoques neuronales al análisis de textos
- Evaluación del Curso: cuestionario teórico/práctico básico, a ser resuelto el día sábado.

El Lenguaje

- El lenguaje revela muchos aspectos de una persona ...

El Lenguaje

- El lenguaje revela muchos aspectos de una persona ...

El Lenguaje

- El lenguaje revela muchos aspectos de una persona ...

El Lenguaje

- El lenguaje revela muchos aspectos de una persona:
  - Pensamientos
  - Creencias
  - Sentimientos
  - Comportamientos
  - Personalidad
  - Nivel educativo
  - Orientación política
  - Edad
  - Género
  - .... y muchos más

El Lenguaje

- El lenguaje revela muchos aspectos de una persona:
  - Pensamientos
  - Creencias
  - Sentimientos
  - Comportamientos
  - Personalidad
  - Nivel educativo
  - Orientación política
  - Edad
  - Género
  - .... y muchos más
- Las ciencias sociales tienen una larga tradición en el estudio del uso del lenguaje para entender mejor estos aspectos.

Lingüística

Lingüística

Ciencia que estudia el lenguaje humano.

Lingüística

Lingüística

Ciencia que estudia el lenguaje humano.

La Lingüística tiene intersección con distintos campos:

Lingüística

Lingüística

Ciencia que estudia el lenguaje humano.

La Lingüística tiene intersección con distintos campos:

Lingüística y Computación

La intersección de ambas ciencias, da origen a distintas ramas que se denominan de diversas formas

Lingüística y Computación

La intersección de ambas ciencias, da origen a distintas ramas que se denominan de diversas formas

Lingüística $\bigcap$ Computación

- Lingüística Computacional (LC).
- Procesamiento (Automático) del Lenguaje Natural (PLN).
  - Procesamiento/Análisis Automático de Textos (PAT).
- Tecnologías/Ingeniería del Lenguaje(TL/IL).
- Minería de Textos (MT)
- Otros

Minería de Textos - Análisis Automático de Textos

En esta presentación nos centraremos en el enfoque usualmente denominado mediante alguno de los siguientes nombres:

- minería de textos (MT, *text mining*),
- análisis (automático) de textos (AAT, *text analytics*)
- o aprendizaje automático con textos (machine learning from text)

que podemos (intentar) definir como:

Minería de Textos - Análisis Automático de Textos

En esta presentación nos centraremos en el enfoque usualmente denominado mediante alguno de los siguientes nombres:

- minería de textos (MT, *text mining*),
- análisis (automático) de textos (AAT, *text analytics*)
- o aprendizaje automático con textos (machine learning from text)

que podemos (intentar) definir como:
... el proceso de analizar texto no estructurado mediante técnicas de aprendizaje automático para extraer conocimiento (modelos descriptivos y predictivos) útil para propósitos particulares.

Minería de Texto - Análisis Automático de Textos

En esta presentación nos centraremos en el enfoque usualmente denominado mediante alguno de los siguientes nombres:

- minería de textos (MT, *text mining*),
- análisis (automático) de textos (AAT, *text analytics*)
- o aprendizaje automático con textos (machine learning from text)

... y que surge de la interacción de 3 áreas principales:

Minería de Texto - Análisis Automático de Textos

En esta presentación nos centraremos en el enfoque usualmente denominado mediante alguno de los siguientes nombres:

- minería de textos (MT, *text mining*),
- análisis (automático) de textos (AAT, *text analytics*)
- o aprendizaje automático con textos (machine learning from text)

... y que surge de la interacción de 3 áreas principales:

- Procesamiento del Lenguaje Natural (PLN)
- Recuperación de la Información (RI)
- Aprendizaje Automático (AA)

Minería de Texto en contexto

Minería de Texto en contexto

Minería de Texto en contexto

Estas 3 áreas, se relacionan directamente a dos diferentes concepciones respecto a cómo visualizar un texto y como representarlo:

- Textos como **bolsas de palabras**:
  1. Originado en RI.
  2. El orden de las palabras no es relevante.
  3. Efectiva en aplicaciones como clasificación, modelización de tópicos y sistemas recomendadores.
- Textos como **conjuntos de secuencias**
  1. Las sentencias individuales son extraídas como cadenas/secuencias.
  2. Efectiva en aplicaciones que requieren mayor interpretación semántica (respuesta a preguntas (question answering), extracción de información, etc.).
  3. Enfoque cercano a la modelización del lenguaje y el PLN.

Minería de Datos, de Textos y de la Web

Es útil considerar el proceso completo de descubrimiento de conocimiento en Bases de Datos (en inglés Knowledge Discovery in Databases - KDD)

Minería de Datos, de Textos y de la Web

Minería de Datos (MD)

- proceso de descubrir información útil (patrones), de forma automática, en grandes repositorios de datos.
- Incluye técnicas estadísticas y de aprendizaje automático para la extracción de conocimiento.

Minería de Datos, de Textos y de la Web

Minería de Datos (MD)

- proceso de descubrir información útil (patrones), de forma automática, en grandes repositorios de datos.
- Incluye técnicas estadísticas y de aprendizaje automático para la extracción de conocimiento.

Minería de Textos (MT)

- Preparación de los datos: uso intensivo de PLN/PAT/IR.
- Tareas de MD + propias (extracción de información).

Minería de Datos, de Textos y de la Web

Minería de Datos (MD)

- proceso de descubrir información útil (patrones), de forma automática, en grandes repositorios de datos.
- Incluye técnicas estadísticas y de aprendizaje automático para la extracción de conocimiento.

Minería de Textos (MT)

- Preparación de los datos: uso intensivo de PLN/PAT/IR.
- Tareas de MD + propias (extracción de información).

Minería de la Web (MW)

1. Muchas similitudes con la MT.
2. Agrega tareas propias de la Web (Minería de estructura y de uso).

¿Porqué es *importante* el AAT?

1. Información y Conocimiento es el recurso más importante que posee la raza humana.

¿Porqué es *importante* el AAT?

1. Información y Conocimiento es el recurso más importante que posee la raza humana.
2. Este conocimiento se ha comunicado, guardado y manejado en la forma de lenguaje natural -griego, latín, inglés, español, etc

¿Porqué es *importante* el AAT?

1. Información y Conocimiento es el recurso más importante que posee la raza humana.
2. Este conocimiento se ha comunicado, guardado y manejado en la forma de lenguaje natural -griego, latín, inglés, español, etc
3. En la actualidad, además de almacenar este conocimiento en libros y documentos, también se lo hace en forma digital.

¿Porqué es *importante* el AAT?

1. "Inundación" de información textual (principal/ proveniente de la Web)

¿Porqué es *importante* el AAT?

1. "Inundación" de información textual (principal/ proveniente de la Web)

¿Porqué es *importante* el AAT?

1. "Inundación" de información textual (principal/ proveniente de la Web)

¿Porqué es *importante* el AAT?

1. "Inundación" de información textual (principal/ proveniente de la Web)
2. Importancia de inf. textual en la empresa (ley del 80 %).

¿Porqué es *importante* el AAT?

1. "Inundación" de información textual (principal/ proveniente de la Web)
2. Importancia de inf. textual en la empresa (ley del 80 %).
3. Texto: vehículo principal para el intercambio de información formal.

¿Porqué es *importante* el AAT?

1. "Inundación" de información textual (principal/ proveniente de la Web)
2. Importancia de inf. textual en la empresa (ley del 80 %).
3. Texto: vehículo principal para el intercambio de información formal.
4. Texto: vehículo principal para el intercambio de inf. informal.

¿Porqué es *importante* el AAT?

1. "Inundación" de información textual (principal/ proveniente de la Web)
2. Importancia de inf. textual en la empresa (ley del 80 %).
3. Texto: vehículo principal para el intercambio de información formal.
4. Texto: vehículo principal para el intercambio de inf. informal.
5. Aplicaciones de alto impacto: detección de spams, filtrado de noticias, detección de plagios, determinación del perfil del autor, análisis de opinión, análisis de tendencias, organización de patentes en categorías, clasificación y organización de páginas Web, etc.

Dificultades del PLN

1. Principal dificultad del PLN: resolución de ambiguedades (acústicas, semánticas, etc)

Dificultades del PLN

1. Principal dificultad del PLN: resolución de ambiguedades (acústicas, semánticas, etc)
2. Ambigüedad: una unidad de lenguaje se puede interpretar de más de una manera:

Dificultades del PLN

1. Principal dificultad del PLN: resolución de ambiguedades (acústicas, semánticas, etc)
2. Ambigüedad: una unidad de lenguaje se puede interpretar de más de una manera:
   - a nivel de palabra: ejemplo, fuera:
     - “como si fuera esta noche la última vez”, ser
     - “como si se fuera a la escuela”, ir
     - “está fuera de la ciudad”, adverbio

Dificultades del PLN

1. Principal dificultad del PLN: resolución de ambiguedades (acústicas, semánticas, etc)
2. Ambigüedad: una unidad de lenguaje se puede interpretar de más de una manera:
   - a nivel de palabra: ejemplo, fuera:
     - “como si fuera esta noche la última vez”, ser
     - “como si se fuera a la escuela”, ir
     - “está fuera de la ciudad”, adverbio
   - a nivel de oración: “veo al gato con el telescopio”

Dificultades del PLN

1. Principal dificultad del PLN: resolución de ambiguedades (acústicas, semánticas, etc)
2. Ambigüedad: una unidad de lenguaje se puede interpretar de más de una manera:
   - a nivel de palabra: ejemplo, fuera:
     - “como si fuera esta noche la última vez”, ser
     - “como si se fuera a la escuela”, ir
     - “está fuera de la ciudad”, adverbio
   - a nivel de oración: “veo al gato con el telescopio”
3. Las ambigüedades se originan al considerar estas unidades fuera del contexto local y global.

Dificultades del PLN

1. Principal dificultad del PLN: resolución de ambiguedades (acústicas, semánticas, etc)
2. Ambigüedad: una unidad de lenguaje se puede interpretar de más de una manera:
   - a nivel de palabra: ejemplo, fuera:
     - "como si fuera esta noche la última vez", ser
     - "como si se fuera a la escuela", ir
     - "está fuera de la ciudad", adverbio
   - a nivel de oración: "veo al gato con el telescopio"
3. Las ambigüedades se originan al considerar estas unidades fuera del contexto local y global.
4. Su resolución implica un análisis a un nivel mayor que considere 3 tipos de conocimiento:
   - Conocimiento lingüístico
   - Conocimiento extra-lingüístico
   - Conocimiento obtenido del mismo texto

Algunas tareas “típicas” del AAT

- Análisis de sentimientos
- Modelización de tópicos
- Extracción de información
- Análisis de autoría
  1. Atribución de autoría (authorship attribution)
  2. Detección de plagio (plagiarism detection)
  3. Determinación del perfil del autor (author profiling)
  4. Detección de inconsistencias estilísticas
- Categorización temática
- Agrupamiento de documentos

Algunas tareas "típicas" del AAT

- Análisis de sentimientos ⇐
- Modelización de tópicos
- Extracción de información
- Análisis de autoría
  1. Atribución de autoría (authorship attribution)
  2. Detección de plagio (plagiarism detection)
  3. Determinación del perfil del autor (author profiling)
  4. Detección de inconsistencias estilísticas
- Categorización temática
- Agrupamiento de documentos

Análisis de Sentimientos

Análisis de Sentimientos (basado en aspectos)

Aspect	Polarity
Staff	Positive
Food	Positive
Cleanliness	Positive
Beds	Negative

Algunas tareas “típicas” del AAT

- Análisis de sentimientos
- Modelización de tópicos
- Extracción de información
- Análisis de autoría
  1. Atribución de autoría (authorship attribution)
  2. Detección de plagio (plagiarism detection)
  3. Determinación del perfil del autor (author profiling)
  4. Detección de inconsistencias estilísticas
- Categorización temática
- Agrupamiento de documentos

Algunas tareas "típicas" del AAT

- Análisis de sentimientos
- Modelización de tópicos ⇐
- Extracción de información
- Análisis de autoría
  1. Atribución de autoría (authorship attribution)
  2. Detección de plagio (plagiarism detection)
  3. Determinación del perfil del autor (author profiling)
  4. Detección de inconsistencias estilísticas
- Categorización temática
- Agrupamiento de documentos

Modelización de tópicos

Algunas tareas “típicas” del AAT

- Análisis de sentimientos
- Modelización de tópicos
- Extracción de información
- Análisis de autoría
  1. Atribución de autoría (authorship attribution)
  2. Detección de plagio (plagiarism detection)
  3. Determinación del perfil del autor (author profiling)
  4. Detección de inconsistencias estilísticas
- Categorización temática
- Agrupamiento de documentos

Algunas tareas "típicas" del AAT

- Análisis de sentimientos
- Modelización de tópicos
- Extracción de información ⇐
- Análisis de autoría
  1. Atribución de autoría (authorship attribution)
  2. Detección de plagio (plagiarism detection)
  3. Determinación del perfil del autor (author profiling)
  4. Detección de inconsistencias estilísticas
- Categorización temática
- Agrupamiento de documentos

Extracción de Información (EI)

- Tarea que convierte texto no (o semi) estructurado en representaciones estructuradas (ej., estructuras tabulares para bases de datos).

Extracción de Información (EI)

- Tarea que convierte texto no (o semi) estructurado en representaciones estructuradas (ej., estructuras tabulares para bases de datos).
- Las estructuras obtenidas contienen información útil y etiquetada referida a nombres/entidades, relaciones y eventos identificados en el texto original.

Ejemplo de EI

EI de actividad terrorista a partir de un cable de noticia.

Dado el siguiente texto:

```
19 March - A bomb went off this morning near a
power tower in San Salvador leaving a large
part of the city without energy, but no
casualties have been reported. According to
unofficial sources, the bomb - allegedly
detonated by urban guerrilla commandos - blew
up a power tower in the northwestern part of
San Salvador at 0650 (1250 GMT).
```

Ejemplo de EI (II)

... se genera ("rellena") el siguiente "template":

Incident type	bombing
Date	March 19
Location	El Salvador: San Salvador (city)
Perpetrator	urban guerilla commandos
Physical target	power tower
Human target	-
Effect on physical target	destroyed
Effect on human target	no injury or death
Instrument	bomb

EI en el dominio médico (reportes de mamografías)

Input: Mammography report		
There is an area of architectural distortion in the left breast 9 o'clock in the anterior depth with associated punctate calcifications which represents a post surgical scar. Associated with this area of architectural distortion is skin thickening.		
Output: Information Frames		
Imaging Observation	**Frame1**	**Relation** = there is **Observation** = architectural distortion **Laterality** = left **Clock face** = 9 **Depth** = anterior
	Frame 2	**Relation** = with associated **Observation** = calcification **Type** = punctuate **Relation** = associated with **Observation** = skin thickening
	Frame 3	**Relation** = represents **Observation** = post surgical scar

Algunas tareas “típicas” del AAT

- Análisis de sentimientos
- Modelización de tópicos
- Extracción de información
- Análisis de autoría
  1. Atribución de autoría (authorship attribution)
  2. Detección de plagio (plagiarism detection)
  3. Determinación del perfil del autor (author profiling)
  4. Detección de inconsistencias estilísticas
- Categorización temática
- Agrupamiento de documentos

Algunas tareas "típicas" del AAT

- Análisis de sentimientos
- Modelización de tópicos
- Extracción de información
- Análisis de autoría ⇐
  1. Atribución de autoría (authorship attribution)
  2. Detección de plagio (plagiarism detection)
  3. Determinación del perfil del autor (author profiling)
  4. Detección de inconsistencias estilísticas
- Categorización temática
- Agrupamiento de documentos

Algunas tareas "típicas" del AAT

- Análisis de sentimientos
- Modelización de tópicos
- Extracción de información
- Análisis de autoría ⇐
  1. Atribución de autoría (authorship attribution)
  2. Detección de plagio (plagiarism detection)
  3. Determinación del perfil del autor (author profiling) ⇐
  4. Detección de inconsistencias estilísticas
- Categorización temática
- Agrupamiento de documentos

Deerminación del perfil de autor

Tareas: determinar

1. edad
2. género
3. personalidad
4. orientación política
5. ...

Algunas tareas “típicas” del AAT

- Análisis de sentimientos
- Modelización de tópicos
- Extracción de información
- Análisis de autoría
  1. Atribución de autoría (authorship attribution)
  2. Detección de plagio (plagiarism detection)
  3. Determinación del perfil del autor (author profiling)
  4. Detección de inconsistencias estilísticas
- Categorización temática
- Agrupamiento de documentos

Algunas tareas "típicas" del AAT

- Análisis de sentimientos
- Modelización de tópicos
- Extracción de información
- Análisis de autoría
  1. Atribución de autoría (authorship attribution)
  2. Detección de plagio (plagiarism detection)
  3. Determinación del perfil del autor (author profiling)
  4. Detección de inconsistencias estilísticas
- Categorización temática ⇐
- Agrupamiento de documentos

Categorización de Textos (por temas/tópicos)

Algunas tareas “típicas” del AAT

- Análisis de sentimientos
- Modelización de tópicos
- Extracción de información
- Análisis de autoría
  1. Atribución de autoría (authorship attribution)
  2. Detección de plagio (plagiarism detection)
  3. Determinación del perfil del autor (author profiling)
  4. Detección de inconsistencias estilísticas
- Categorización temática
- Agrupamiento de documentos

Algunas tareas "típicas" del AAT

- Análisis de sentimientos
- Modelización de tópicos
- Extracción de información
- Análisis de autoría
  1. Atribución de autoría (authorship attribution)
  2. Detección de plagio (plagiarism detection)
  3. Determinación del perfil del autor (author profiling)
  4. Detección de inconsistencias estilísticas
- Categorización temática
- Agrupamiento de documentos $\Leftarrow$

Agrupamiento de documentos

Agrupamiento de documentos

Agrupamiento de documentos

Agrupamiento de documentos

Agrupamiento de documentos

Agrupamiento de documentos

Tareas “avanzadas” del AAT

1. Generación automática de respuestas (Question Answering).
2. Generación automática de resúmenes (Summarization).
3. Traducción automática
4. Interfaces en Lenguaje Natural.

Aplicaciones

1. Propaganda dirigida
2. Evaluación de campañas, políticos y productos
3. Detección de plagio (intrínseco y extrínseco) - plagio de patentes
4. Detección de “bullying” y pedófilos en la Web
5. Detección anticipada de signos de depresión, anorexia, rumores y tendencias suicidas
6. Atribución de autoría en casos forenses
7. Determinación del perfil del autor (sexo, grupo etario, nacionalidad, rasgos de personalidad, orientación política, sexual y religiosa)
8. Identificación de información de calidad
9. Predicción del comportamiento del mercado, precios de productos y servicios, ganancias obtenidas en un rubro, difusión de enfermedades y ocurrencia de catástrofes.

Aplicaciones recientes: descripción de escenas

Aplicaciones recientes: responder preguntas sobre imágenes

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Aplicaciones recientes: calidad de información en Wikipedia

Artículos Destacados

Detección de Fallas

Visualización

Métricas de Calidad

Detección de pedófilos en la Web

- Datos de entrenamiento en www.perverted-justice.com
- Competencias recientes pan.webis.de/clef12/pan12-web

Ejemplo: