Aprendizaje Automático

Minería de Textos

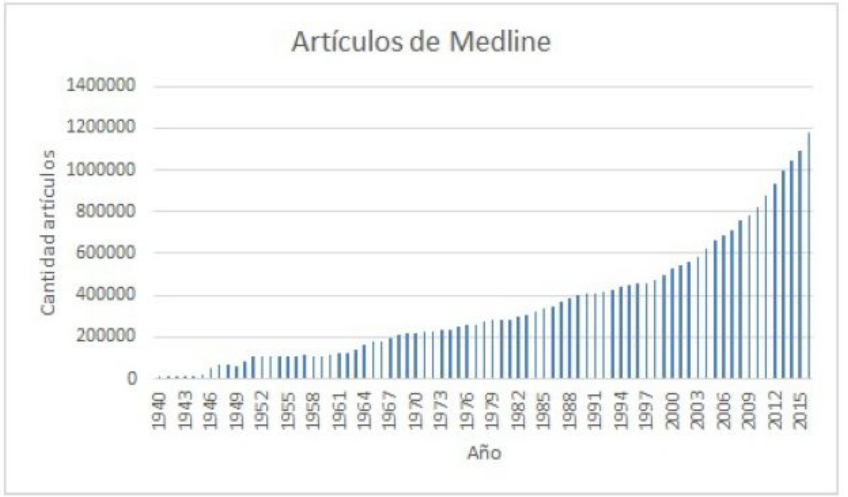
Viviana Cotik

1er cuatrimestre 2019

Motivación

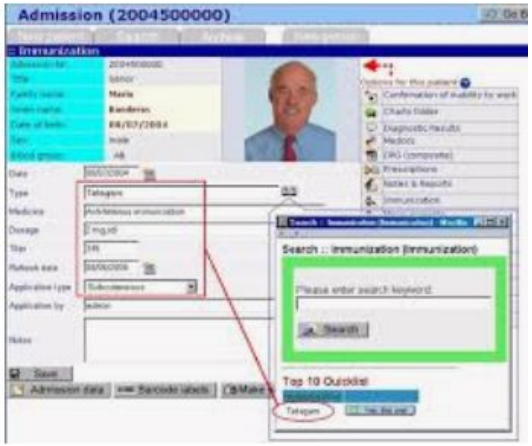
Crecimiento de la cantidad de información digital disponible.

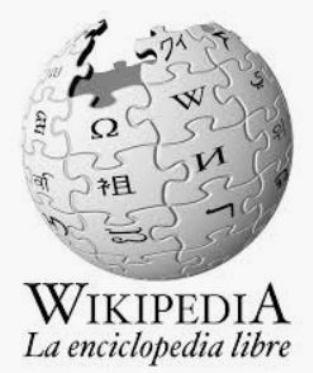
Mayormente de formato textual

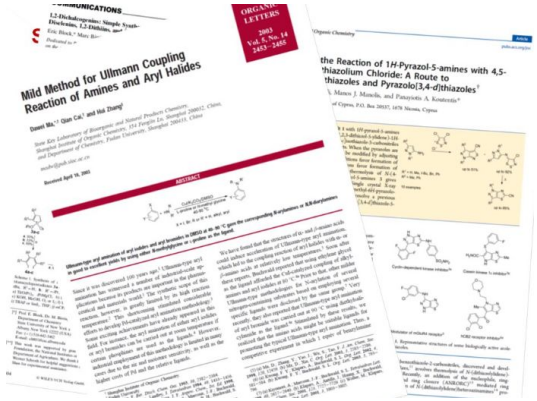


Género


NEWS


Admission (2004500000)
Immunization
Search :: Immunization (Immunization)
Please enter search keyword
Search
Top 10 Quicklist
Admission data
Barcode labels


WIKIPEDIA
La enciclopedia libre


Mild Method for Ullmann Coupling Reaction of Amines and Aryl Halides
ORGANIC LETTERS
2003
Vol. 5, No. 14
2453 - 2455


twitter

Dominio

Ficha | Críticas [80] | Tráilers [1] | Imágenes [12] | DVD/VoD [2] | *Las 100 mejores películas del siglo XXI (BBC)*

Año 2003
País Corea del Sur
Director Kim Ki-duk
Reparto Oh Yeong-su, Kim Ki-duk, Kim Jung-yeong, Seo Jae-gyeong, Kim Yeong-min, Ha Yeo-jin, ...
Género Drama | Drama psicológico. Religión

7,6
19.623 votos

Sinopsis
Dos mor
más jove
una rana
monje a
algo que
alguien l

Jo *Bogotá (Colombia)*

10

Un bello cuento simbólico

26 de diciembre de 2006 • **186** de **212** usuarios han encontrado esta crítica útil

Críticas

En un escenario simple y bello se presenta de manera simbólica el desarrollo de la conciencia, desde su nacimiento en forma de conciencia social o colectiva que aparece en la primavera, a partir los primeros días de la infancia, hasta el descubrimiento de la conciencia individual en los días de la madurez humana, representados por

Pepe wrote a review Jul 9
Newcastle, South Africa • **16** contributions • **1** helpful vote

My favourite hotel

"We had a long day travelling with a flight delay and reached Romance Boutique Hotel much to o
pleasant surprise.We were warmly greeted and helped by Hakan and Onder, our room amazed u
spotless and cosy and we had some lovely welcome treats waiting for us, a cute…"

Read more

Date of stay: July 2019

Naturaleza del texto

- **informal:**
  - abundancia de errores ortográficos,
  - abreviaturas no estándar,
  - abreviaturas ambiguas y no referenciadas
  - falta de signos de puntuación
  - errores gramaticales

  Ej: twitter, ciertos informes médicos

- **formal**:
  - oraciones bien formadas

  Ej: notas periodísticas, artículos científicos, ciertos informes médicos

Procesamiento del lenguaje natural (NLP)

Lenguaje natural: el hablado y escrito por personas.
NLP:

- técnicas computacionales para procesar lenguaje natural, de forma tal de **analizarlo** y/o **generarlo**.
- diferencia con otros sistemas de procesamiento de datos: tiene que tener **conocimiento acerca del lenguaje** [Jurafsky y Martin].
- Ej. de **aspectos del lenguaje** a considerar:
  - morfología
  - sintaxis
  - semántica
  - pragmática
  - ...

Procesamiento del lenguaje natural (NLP)

Ejemplos de aplicaciones:

- traducción automática
- resumen automático
- autopredictor (celular, mails)
- autocorrector
- sistemas de respuesta a preguntas (question answering -QA)
- generación de lenguaje natural
- diálogo automático
  - Hombre: cuáles son los horarios de buses a Mar del Plata para mañana?
  - Máquina: 7:30, 8, 8:30, 9, 10, 11, ...

Extracción de información (IE)

- extracción de **información estructurada** a partir de textos
- se puede pensar como tarea de **llenar *plantillas***, con espacios en blanco.
- Ejemplo de tareas:
  - **reconocimiento de entidades nombradas (NER)**
  - **extracción de relaciones (RE)**
- Ejemplos:
  - médico

An example of an ultrasonography report (in English)

27518 —14y 11m—20070103—950051 Normal kidney echostructure implant. Dilation not detected in the Urinary tract. Plenified of normal characteristics. Color Doppler examination: normal characteristics. IR: 0.67. liver preserved homogeneous echostructure. Spleen homogeneous of 7.8 cm. Both kidneys native small echogenic . Evidence in retroperitoneal of solid mass allready known with calcifications and lobular extending left flank suggesting .... It measures approximately 6.3 x 6.8 x 5 cm.

  - Estructura de eventos a partir de mails (gmail)

Minería de textos (TM)

- análisis de información para **descubrir patrones o conocimiento no mencionado explícitamente en el texto**.
- puede **incluir un sistema que hace extracción de información**
- Ejemplo de tareas:
  - opinion mining (minado de opiniones)
  - interacción medicina-medicina (drug-drug interaction)
  - análisis de sentimientos
  - detección de spam
  - determinación de autoría

NLP, IE, TM

Algunas tareas:

- segmentación de texto
- desambiguación de palabras (word-sense disambiguation)
- resolución de anáforas y co-referencia
- desambiguación de sentidos:
  - mañana: próximo día, primera parte del día
  - banco: banco del río, institución financiera, soportar a una persona
  - Dr: doctor, drive
- expansión de abreviaturas y acrónimos

Soluciones

Implementación de soluciones:

- **Por reglas.**
  - Necesidad de expertos del dominio. Desarrollo trabajoso. Mantenimiento lento.
- **Aprendizaje automático.**
  - Mejores resultados. Necesidad de grandes volúmenes de datos.
- **Híbridas**

Recursos:

- diccionarios, lexicones, tesauros, word-embeddings
- tokenizadores, analizadores sintácticos, asignadores de etiquetas (PoS taggers), etc

Difieren para distintos **idiomas** y para distintos **dominios**.

Algunas definiciones

Abreviatura: secuencia de letras utilizadas para abreviar representación de una palabra. Ej. Av., cm., Dr.

Acrónimo: palabras formadas por una o pocas letras iniciales de cada palabra de una expresión compuesta. Ej, OVNI, AFIP, IBM.

Ambas pueden ser:

- **ambiguas**.
  - **RA:** rheumatoid arthritis, renal artery, right atrium, refractory anemia, radioactive, ...
  - **10 m**: ¿minutos o metros?
- **no estándares:** R.D., RD, RDER para riñón derecho

Algunas definiciones

Stop words:

- Palabras comunes como determinantes y preposiciones (de, la, ..).
- Consideradas irrelevantes para ciertas tareas (**en cuyo caso se eliminan**).
- Algunas tareas, como atribución de autoría, las usan.
- No existen listas universales de stop words.

Algunas definiciones

Morfología (en lingüística): Estudia las palabras, su estructura y los mecanismos de formación de las mismas.

La **palabra** está formada de **morfemas** (unidad mínima con significado de una palabra). Existen dos clases:

- **raíz:** morfema **principal** de la palabra (niñ-)
- **afijo:** **agregan significado** (a: femenino, s: plural)
  - **prefijos**: **a**norexia
  - **sufijos**: traslad**able**
  - **infijos** (en el medio de la palabra)
  - **circunfijos**: **ge**spiel**t** (participio en alemán)

Algunas definiciones

- Se pueden **combinar morfemas** de distintas formas para **crear palabras**:
  - **inflexión:** raíz + morfema que expresa función gramática o atributo (tiempo, persona, número, género) niñ+a+s
  - **derivación**: se cambia el significado o se cambia la etiqueta gramatical: ***anti***depresivo
  - **composición:** combinación de muchas raíces. Ej. pelirrojo, sordomudo. Muy comunes en alemán. Ej: **Bezirks**schornstein**feger**meister

Algunas definiciones

n-gramas: secuencias contiguas de n objetos (ej: palabras o letras).

- unigrama,
- bigrama,
- trigrama..

Modelo de lenguaje (LM): modelos que asignan probabilidades a secuencias de palabras. Sirven por ejemplo para la tarea speech to text.

En **LM** se usan **n-gramas** para:

- **estimar la probabilidad de la última palabra de un n-grama dadas las palabras anteriores** y
- para asignar probabilidades a las secuencias enteras.

Algunos problemas:

- texto no gramatical
- falta de estructura
- presencia de abreviaturas y acrónimos
- textos con palabras de distintos idiomas
- ambigüedad
- errores ortográficos
- polisemia
- términos (Nueva York es un sólo término, no dos distintos)

Representación de las palabras

- **one hot vector (localist representation)** cada palabra constituye una posición de un vector (el valor corresponde con el nº de veces que ha aparecido o con un 1 si aparece). Ej: (0, 0, 0, 0, 1, 0, 0) representa hotel.
  - mucho vocabulario: vector muy grande. En dicc. aprox 250.000. Con morfología derivacional, más composición muchas más.
  - no representa similaridad semántica entre palabras.
- **word-embeddings** (distributed representation).
  - cada palabra está representada por un vector denso (sin ceros). Ej. 50-300 coordenadas.
  - basado en representar palabras por su contexto. si se sabe el contexto en el que se usa se sabe el significado de la palabra
  - palabras relacionadas semánticamente están cerca en el espacio
  - distintas formas de aprenderlas: **Word2vec** (red neuronal shallow)**, GloVe, ...**

Word embeddings

Proyección en
2 dimensiones de vectores
de aprox 100 dimensiones:

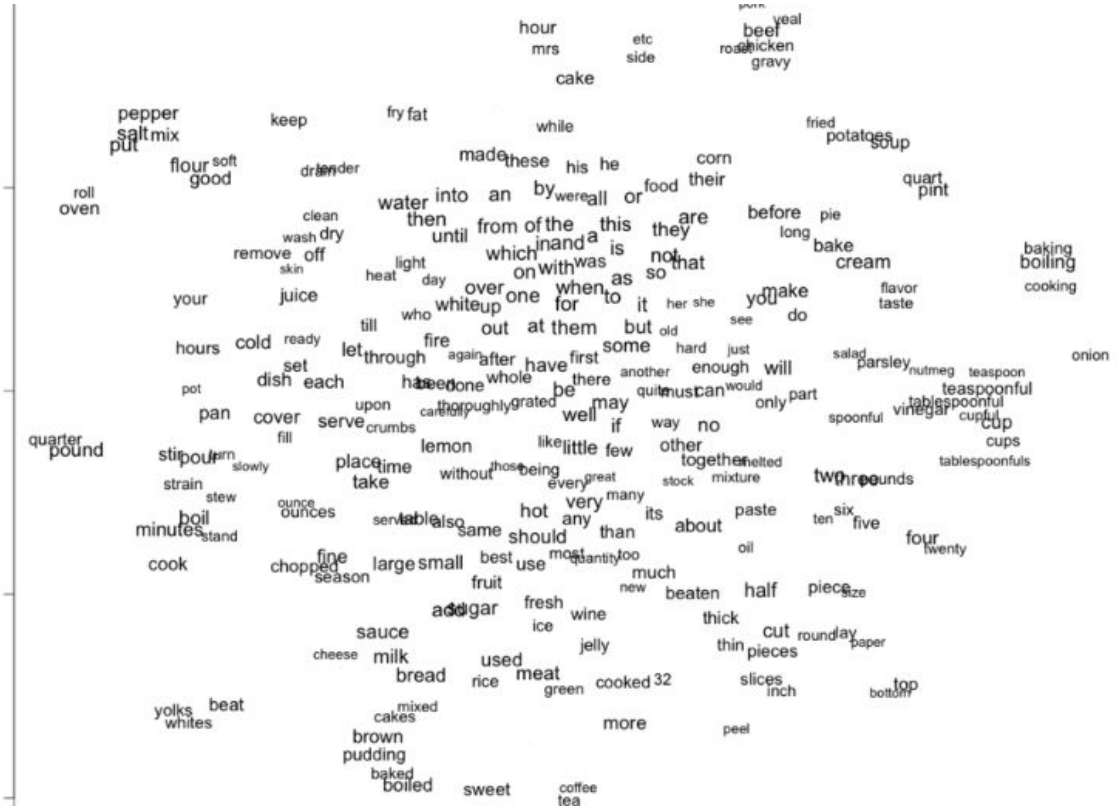


Gráfico: TowardsDataScience

Stemming y lematización

Objetivo: reducir formas inflectivas y derivacionales a una forma básica común.

Stemming: reducción de afijos para obtener la raíz. Ej: running: run, considered: consider, satisfaciendo: satisfac, went: went

Lematización: obtención de la forma básica o entrada de diccionario de una palabra (llamada lema). Ej: did: do, satisfaciendo: satisfacer, went: go.

Arquitectura básica de un sistema de IE o NLP

1. **Reconocimiento de idioma**
2. **Segmentación de oraciones y segmentación de palabras (tokenización)**
3. **Análisis morfológico** (stemming, lematización). **Normalización.**
4. **Asignación de etiquetas morfosintácticas** (part of speech tagging - **PoS tagging**) (Incluye desambiguación de etiquetas)
5. **Análisis sintáctico** (shallow parsing, dependency tree parsing)
6. **Constituyentes básicos o chunks**
7. **NLP: Análisis semántico (léxico, proposicional)**
7. **IE: Reconocimiento de entidades nombradas (NER) y extracción de relaciones (RE)**

Módulos de la arquitectura. Segmentación de oraciones

2. Segmentación de oraciones

Identificar oraciones. Las tareas se suelen hacer de oración a oración. Se consideran: ., ?, !, entre otros.

El Dr**.** se dirigió a la Av**.** 9 de Julio al 2**.**600. Allí atendió al hijo del diputado**.**
La charla *¿cómo escribir un artículo científico**?*** será dada el próximo viernes**.**

El Dr. se dirigió a la Av. 9 de Julio al 2.600.
Allí atendió al hijo del diputado.

La charla *¿cómo escribir un artículo científico?* será dada el próximo viernes.

Módulos de la arquitectura. Tokenización

2. Tokenización

Identificación de unidades llamadas *tokens*.

El Dr. Blanco denunció el robo de $3000.
Tokens: El, Dr., Blanco, denunció, el, robo, de, $, 3000, .
O'Connor, l'enfant, african-american

Específico del idioma. Ej.

- En **alemán** se usan módulos que separan las **palabras compuestas**
- En **japonés** y en **chino** las palabras no siempre están delimitadas por ideogramas. Ej: 冰箱: heladera, sin embargo 冰: hielo 箱: caja
- **inglés** y **español**: won't (will not) , del (de el)

Módulos de la arquitectura. PoS tagging

4. Asignación de etiquetas morfosintácticas (PoS tagging)

Objetivo: Identificar clases de palabras que aparecen en contextos parecidos y sufren transformaciones similares. Ej:

- **sustantivo**: perro, gigante, zar, bajo
- **verbo**: ser, estar, parecer, simular, bajo
- **adjetivo** colorado, grande, absurdo, bajo
- **adverbio** desafortunadamente, bajo
- **preposición**: a, de, por, desde, sobre
- **pronombre**: yo, mi, tuyo, el
- **determinante**: la, una, esos, este, el
- **conjunción** y, o, ni, sino
- ....

Módulos de la arquitectura. PoS tagging

PoS Tagset: Conjunto de rótulos o etiquetas.

Ejemplos:

- **Brown Corpus**: 1M palabras, 87 tags
- **Penn Treebank**: corpus del Wall Street Journal, 1M palabras, 46 tags. El más usado en la actualidad para el inglés.
- **Español**: EAGLE tagset

POS Tagging: Proceso de asignar una etiqueta de clase de palabra (PoS tag) a cada palabra de un texto.

Módulos de la arquitectura. PoS tagging

4. Asignación de etiquetas morfosintácticas (PoS tagging)

Allí	allí	*RG*
atendió	atender	*VMIS3S0*
a	a	*SP*
el	el	*DA0MS0*
hijo	hijo	*NCMS000*
de	de	*SP*
el	el	*DA0MS0*
diputado	diputado	*NCMS000*
.		*Fp*

No hay un único PoS tag por palabra. Hay que desambiguar.

cuenta	contar	*VMIP3S0*	El arquitecto cuenta el dinero.
cuenta	cuenta	*NCFS000*	La cuenta le dio bien

Módulos de la arquitectura. PoS tagging

¿Cómo se desarrolla un PoS Tagger?

- A partir de **reglas**
  - Asignar a cada palabra todos sus tags posibles.
  - Usar reglas predefinidas para eliminar tags (cuando hay más de uno).
  - Repetir hasta que cada palabra tenga sólo un exactamente un tag.
- Con **aprendizaje automático:**
  - Support Vector Machines (SVM)
  - Modelos Ocultos de Markov (HMM)
  - Redes neuronales profundas

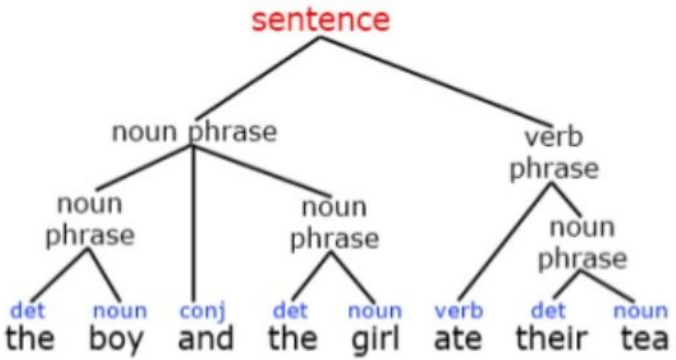
Tiene muy buena performance para determinados **dominios**

Módulos de la arquitectura. Análisis sintáctico

5. Análisis sintáctico

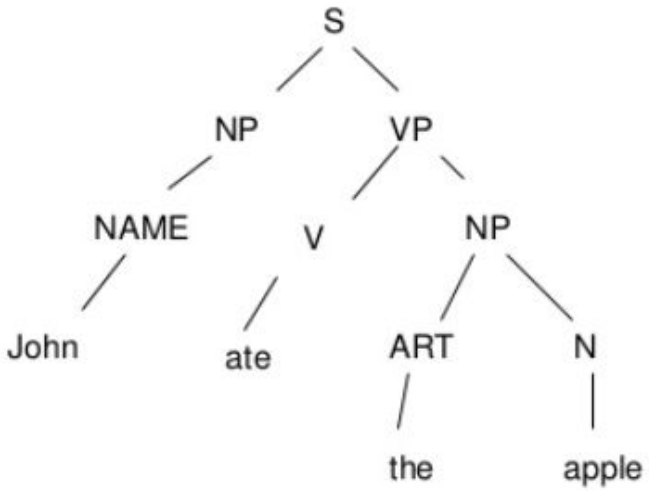
Se puede contar con **gramática**: reglas para construcción de frases

Se arma un árbol sintáctico (por ej: bottom up)



John ate the apple.

1. S -> NP VP
2. VP -> V NP
3. NP -> NAME
4. NP -> ART N
5. NAME -> John
6. V -> ate
7. ART-> the
8. N -> apple



Módulos de la arquitectura. Análisis sintáctico

Alternativas al parsing completo (muy costoso computacionalmente):

- shallow parsing y chunking

También existen árboles de dependencias

Módulos de la arquitectura. Análisis semántico

7. Análisis semántico

Ej: El gato come palomas

léxico

El gato: entidad –> ser vivo –> animal -> felino doméstico determinado
come: acción -> voluntaria -> ...
palomas: entidad->ser vivo-> animal -> ave

proposicional

Existe gato (x) and existe comida(y) and come (x,y)

Módulos de la arquitectura. NER

7. Reconocimiento de Entidades nombradas (NER)

Objetivo: Identificar instancias de una clase de información específica en el texto y asignarles una clase. Puede incluir normalización.

An example of an ultrasonography report (in English)

27518 —14y 11m—20070103—950051 Normal kidney echostructure implant. Dilation not detected in the Urinary tract. Plenified of normal characteristics. Color Doppler examination: normal characteristics. IR: 0.67. liver preserved homogeneous echostructure. Spleen homogeneous of 7.8 cm. Both kidneys native small echogenic . Evidence in retroperitoneal of solid mass allready known with calcifications and lobular extending left flank suggesting .... It measures approximately 6.3 x 6.8 x 5 cm.

Módulos de la arquitectura. NER

7. Reconocimiento de entidades nombradas

Métodos:

- **Reglas**. Usando terminologías. Ej.
  - PoS taggear,
  - usar terminologías para detectar qué términos podrían ser entidades de interés
  - derivar reglas a partir de árboles de dependencias.
- **Aprendizaje automático**:
  - Conditional Random Fields (CRF) y Modelos ocultos de Markov (HMM)
  - SVM, NB
  - redes neuronales profundas
- Híbridos

Módulos de la arquitectura. NER

7. Reconocimiento de entidades nombradas

Evaluación:

- Precision, recall, F1 o Fβ con $\beta \neq 1$

Criterios de coincidencia:

- **total (exact match):** mismo comienzo, mismo fin, mismo tipo de entidad
- **parcial (partial match):** (mismo comienzo, o mismo fin, o incluido, o..)

Dependiendo del dominio conviene una u otra. Siempre informar con algún criterio estándar.

Módulos de la arquitectura. Extracción de relaciones

7: Extracción de relaciones

Objetivo: detectar un tipo de relación específico entre entidades nombradas. Ej:

- **interacción medicamento-medicamento** (drug drug interaction **-DDI-**)
- **casamiento** entre dos personas (binaria), agregando locación del mismo (ternaria).
- **ubicación** en el cuerpo de un **hallazgo médico**

Uso de técnicas de aprendizaje automático

Posibles pasos:

1. **Segmentar** oraciones, tokenizar, (expandir abreviaturas, (normalizar)
2. realizar **PoS tagging**
3. **extraer features** (ej: PoS tag completo, PoS tag reducido, prefijos de 3 caracteres, sufijos de 4 caracteres, son todas mayúsculas? son todos números?, etc.)
4. **dividir los datos** (entrenamiento, validación, test)
5. **entrenar el modelo** con la técnica seleccionada. elegir parámetros óptimos con conjunto de validación
6. **entrenar** modelo con **dataset de desarrollo**, **reportar** resultados con **dataset de test**.

Otras Tareas

Clasificación de documentos (ej: spam-no spam, tendencia política de un artículo)

Detección de negaciones. Por ej. en informes médicos más del 50% de los hallazgos aparecen como negados.

Métodos basados en:

- reglas (ej. distancia entre término de negación y hallazgo)
- reglas a partir de análisis sintáctico
- aprendizaje automático

Algunos Recursos

Freeling: http://nlp.lsi.upc.edu/freeling/demo/demo.php. Para español
NLTK: http://www.nltk.org/. Swiss-army knife de NLP.

Principales conferencias de estas temáticas: ACL, COLING, EMNLP, NAACL (se pueden buscar los proceedings)

Bibliografía

Libros:

- Speech and Language Processing , Prentice Hall. Daniel Jurafsky and James H. Martin (2008)
- Foundations of Statistical Natural Language Processing. The MIT Press, 1 edition. Manning, C. D. and Schütze, H. (1999).
- Natural Language Processing with Python. O'Reilly 1 edition. Bird, S., Klein, E., and Loper, E. (2009).

Cursos online:

Natural language processing with Deep Learning. CS224n. Stanford. Chris Manning

Natural language processing Dan Jurafsky Chris Manning. Coursera.

Otros:

Procesamiento del Habla. Text to Speech. Agustín Gravano. Campus, FCEyN,UBA

Fin de la cursada

Temas tesis para más adelante, NLP, IE, Machine Learning, BioNLP

Viviana Cotik
vcotik@dc.uba.ar

Encuesta de la cátedra

Fechas