Clase N°21

Análisis de Sentimiento

Motivación

Vimos que existen herramientas para abordar el estudio cuantitativo de textos.

Motivación

Vimos que existen herramientas para abordar el estudio cuantitativo de textos.

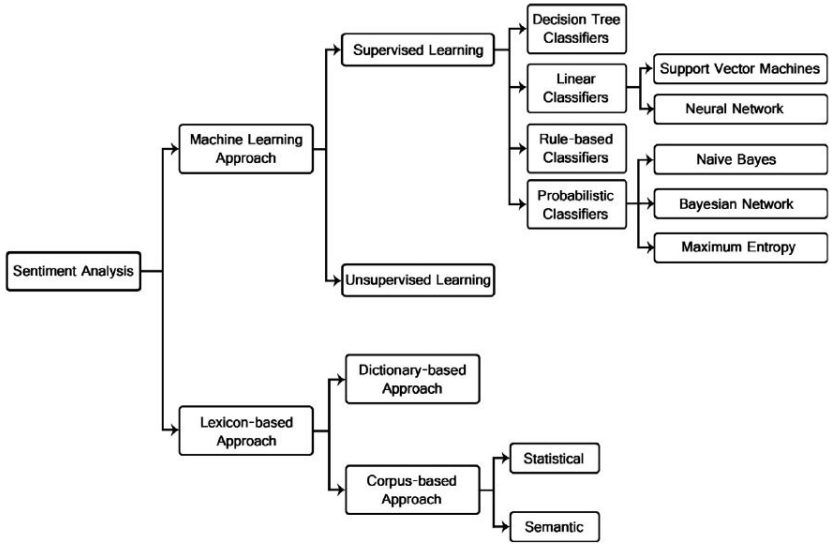
Con varias metodologías, se busca clasificar textos según el tema del que hablan; o bien, vincularlos según métricas de similaridad.

Motivación

Vimos que existen herramientas para abordar el estudio cuantitativo de textos.

Con varias metodologías, se busca clasificar textos según el tema del que hablan; o bien, vincularlos según métricas de similaridad.

No obstante, existe un universo gigante de técnicas que buscan atacar el problema de la connotación sentimental de los textos, y eso es lo que veremos hoy.



Generalidades

El Análisis de Sentimiento (*Sentimental Analysis - Opinion Mining*) es un área dentro del PLN que busca estudiar el sentimiento expresado en un texto, a partir de las palabras y expresiones contenidas en el mismo.

Generalidades

El Análisis de Sentimiento (*Sentimental Analysis - Opinion Mining*) es un área dentro del PLN que busca estudiar el sentimiento expresado en un texto, a partir de las palabras y expresiones contenidas en el mismo.

Suele utilizarse en (y pareciera haber nacido a partir de) contextos de opiniones de consumidores sobre artículos, o de clientes sobre servicios.

Generalidades

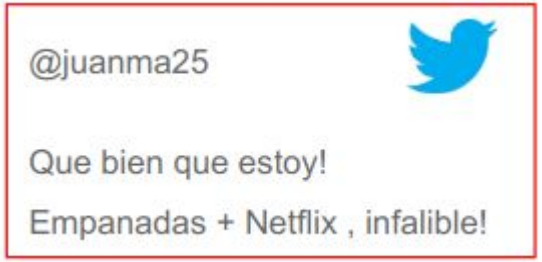
El Análisis de Sentimiento (*Sentimental Analysis - Opinion Mining*) es un área dentro del PLN que busca estudiar el sentimiento expresado en un texto, a partir de las palabras y expresiones contenidas en el mismo.

Suele utilizarse en (y pareciera haber nacido a partir de) contextos de opiniones de consumidores sobre artículos, o de clientes sobre servicios.

Pero también puede aplicarse, entre otras cosas, a estudiar la cobertura mediática sobre determinados temas y posiciones valorativas en torno a las mismas.

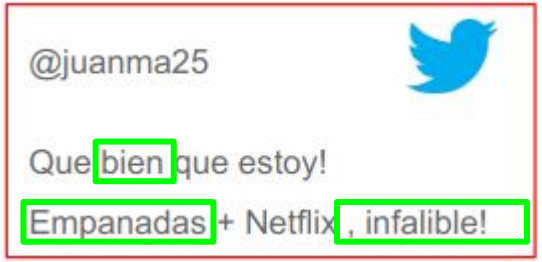
El ejemplo

Sea un conjunto de palabras que conocemos con valoración positiva y negativa, podemos determinar de forma frecuentista la valoración de una frase:



El ejemplo

Sea un conjunto de palabras que conocemos con valoración positiva y negativa, podemos determinar de forma frecuentista la valoración de una frase:



Generalidades

En la literatura, hay consenso sobre que existen tres grandes enfoques:

- Basados en léxico -*lexicon based*-

- Basados en Aprendizaje Automático -*automatic learning*-

- Enfoques híbridos -*hybrid*-

Método Basado en Léxico

Utiliza listados de palabras/expresiones cuya connotación positiva/negativa es conocida, para clasificar sentencias y textos.

Existen varios métodos para la construcción de estos léxicos:

- **a mano**

a favor	en contra
- preciso - controlado	- demanda demasiado tiempo

- diccionarios

- corpus

Método Basado en Léxico

Utiliza listados de palabras/expresiones cuya connotación positiva/negativa es conocida, para clasificar sentencias y textos.

Existen varios métodos para la construcción de estos léxicos:

- a mano

- **diccionarios**

- corpus

a favor	en contra
- automático - accesible en términos de facilidad de encontrar diccionarios, etc	- no puede distinguir connotaciones que varíen de un contexto a otro

Método Basado en Léxico

Utiliza listados de palabras/expresiones cuya connotación positiva/negativa es conocida, para clasificar sentencias y textos.

Existen varios métodos para la construcción de estos léxicos:

- a mano
- diccionarios
- **corpus**

a favor	en contra
- automático - flexible a los cambios de contexto según los textos que le pasemos	- ausencia de determinadas palabras en los pasos de entrenamiento

Método Basado en Aprendizaje Automático

Se usan técnicas de ML (algunas ya vistas durante la cursada) para clasificar textos.

En definitiva, el problema de clasificación termina siendo:

- Existe un conjunto de textos, cada uno asignado a una clase (positivo, negativo, estrellas, etc)
- Se entrena el modelo de forma tal de relacionar las features relevantes a las clases
- Se utiliza el modelo ya entrenado para predecir el sentimiento de nuevos textos

Método Basado en Aprendizaje Automático - NB

Dentro del universo de métodos de clasificación, uno de los utilizados es el de Naïve-Bayes

La idea es computar la probabilidad de que, dadas ciertas features, se obtenga cierta clase, basándonos en las probabilidades inversas obtenidas del entrenamiento.

$$P(label|features) = \frac{P(label) \cdot P(features|label)}{P(features)}$$ ***Bayes***

Naïve
independencia de las f

$$= \frac{P(label) \cdot P(f_1|label) \cdot P(f_2|label) \cdot ... \cdot P(f_m|label)}{P(features)}$$

¿Hay más?

No sólo de sentimientos valorativos se trata, sino que aparecen campos como el análisis de emociones.

Además, sobre la construcción de listados de palabras/sentencias suele haber mucho trabajo.

Así también como lo que refiere a la utilización de ciertos contextos bastante exhaustivos para trabajar en otros, ej: Wikipedia ---> Twitter

Problemas abiertos

Todavía quedan misterios sin resolver:

- El de las bases de datos

- El del idioma

- El del procesamiento de lenguaje natural como preprocesamiento para encarar el análisis de sentimiento

Métodos de la notebook

Si bien existen herramientas gratuitas e incluso online para análisis de sentimiento en inglés, para el español no parece haber tanta suerte.

Acá tres librerías que vienen a aportar en este sentido:

- https://github.com/FernanOrtega/SentiLeak (lexicon based)
- https://github.com/sentiment-analysis-spanish/sentiment-spanish (ml based)
- https://github.com/pysentimiento/pysentimiento (ml-based)

Bibliografía

- MEDHAT, Walaa; HASSAN, Ahmed; KORASHY, Hoda. **Sentiment analysis algorithms and applications: A survey**. *Ain Shams engineering journal*, 2014, vol. 5, no 4, p. 1093-1113.

- ZHANG, Lei; WANG, Shuai; LIU, Bing. **Deep learning for sentiment analysis: A survey.** *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018, vol. 8, no 4, p. e1253.

Clase 21, bis

Grafos de Palabras

¿Qué vimos?

- Estadística de palabras y su frecuencia, con modificaciones

- Análisis de sentimiento y emotividad

Nada sobre cómo se vinculan las palabras entre sí a través del contexto en el que aparecen en el texto

¿Qué veremos?

- Clasificación

- Tópicos

- Embedings

Todo con el objetivo de encontrar regularidad, similaridad a través de la coocurrencia de palabras en mismos textos.

Otra herramienta más, Grafos de Palabras

Una forma de indagar en cómo utilizamos el lenguaje, es prestando atención al orden en que enunciamos determinado discurso.

Otra herramienta más, Grafos de Palabras

Una forma de indagar en cómo utilizamos el lenguaje, es prestando atención al orden en que enunciamos determinado discurso.

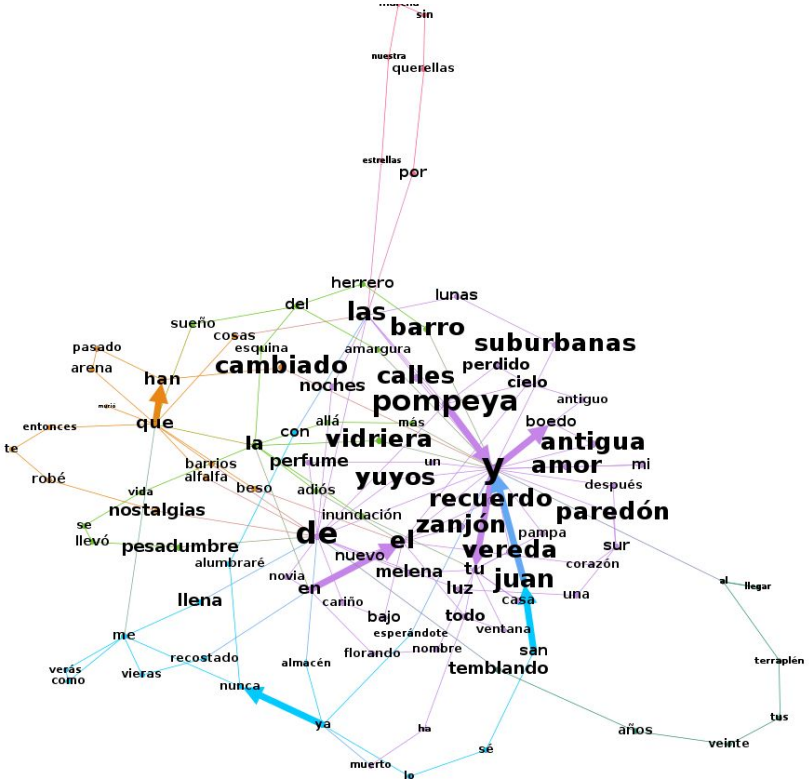
Una herramienta para estudiar este orden son los grafos/redes.

Otra herramienta más, Grafos de Palabras

Una forma de indagar en cómo utilizamos el lenguaje, es prestando atención al orden en que enunciamos determinado discurso.

Una herramienta para estudiar este orden son los grafos/redes.

Un grafo es un conjunto de nodos y enlaces que los vinculan. Por ejemplo, palabras vinculadas según su aparición sucesiva en un texto.



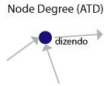
Topología y Discurso

De esta forma, estudiar la topología de la red nos permite hablar sobre cómo se enarbola un discurso.

Tiene grandes aplicaciones en estudio de efecto de drogas sobre las personas.

N (Nodos): Número de nodos

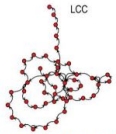
E (Ejes): Número de ejes



Grado: Cantidad promedio de ejes unidos a cada nodo

Densidad: Número de ejes dividido cantidad máxima posible de ejes

D = 2*E / N*(N-1)



Tamaño de la componente conexa más granda.

Cantidad de ejes paralelos

Diámetro: Longitud del camino corto más largo entre pares de nodos en el grafo

Camíno más corto promedio entre nodos

L1: Número de loops de orden 1

L2: Número de loops de orden 2

L3: Número de loops de orden 3