Algoritmo de Pang y Lee

Detección de la polaridad

Criticas de pelicula

> Hace mucho tiempo no veía una peli tan buena como esta. Original y hermosa desde todo punto de vista, las canciones, los chistes y los efectos visuales. Increible como mejoraron en los últimos años con las expresiones de los personajes animados! Una obra maestra desde donde se la mire! No entendía cómo Monster University no había sido nominada al oscar, pero luego de ver Frozen, entiendo que Monster's no tenía nada que hacer. Es de las pelis que, para aquellos que crecimos con los clásicos de disney como el rey leon, tarzan o el jorobado de Notre dame, nos reavivan el niño que tenemos adentro. Si o si verla en 3D.

> La película es malísima. Está entre las peores que he visto en mi vida, sino es la peor. Es lineal, predecible, aburrida. Solo apta para menores de 13 años. Deberían dedicarse a otra cosa todos los que intervinieron para que exista. Realmente es muy recomendable que NO LA VEAN. Después no digan que no les advertí.

1. **Tokenización** del texto
2. **Extracción de características** (palabras o frases claves)
3. **Clasificación** utilizando distintos algoritmos de clasificación:
    a. Naïve Bayes
    b. MaxEnt (maxima entropia)
    c. SVM

Problemas comunes

● Lidiar con los tags XML o HTML

● Tener que reconocer las marcas de Twitter (si queremos sacar información de ahí) como los nombres de usuario y los hashtags

● El uso de mayúsculas. Generalmente nos va a interesar conservar las mayúsculas de las palabras en las distintas fases del algoritmo.

● Números de teléfono y fechas.

● Emoticones: es muy útil detectar los emoticones cuando se está haciendo análisis de sentimientos

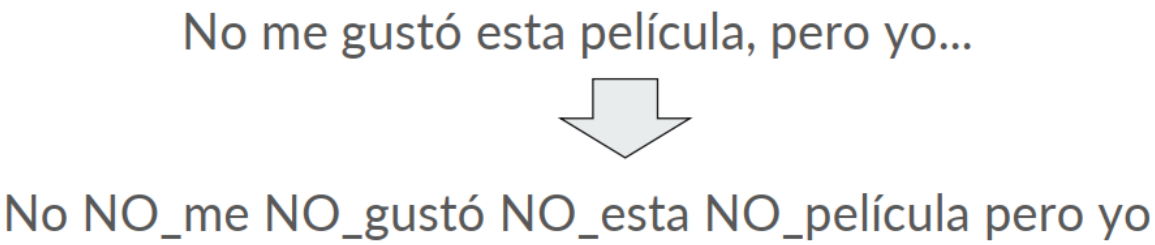
Extracción de características

En esta etapa tenemos dos problemas:

● **¿Cómo lidiar con la negación?**
    ○ No me gustó esta película.
    ○ Me gustó esta película.

El NO cambia todo el sentido de la oracion

Segun Bayes seria positiva porque no importa el orden



Entre un NO y signo de puntuacion (coma) cambiar las palabras

De esta manera aumento el vocabulario porque se crearon nuevas palabras, pero ahora en la critica negativa dejo de estar la palabra gustó que esta asociado a criticas positivas y aparecio la palabra NO_gustó que seguramente va a estar mas asociada a criticas negativas

- **¿Qué conviene usar?**
  - todas las palabras
  - solo los adjetivos

Problema de filtrado

Se demostró que al menos con la información de IMDB, es conveniente utilizar todas las palabras. Se obtienen así mejores resultados y en términos generales siempre conviene utilizar todas las palabras ya que a veces los sustantivos y los verbos nos dan información valiosa sobre el juicio de valor de una crítica.

Clasificación - Naïve Bayes

$$C_{map} = \underset{c_j \in C}{\operatorname{argmax}}\ P(c_j) \prod_{i \in \text{Posiciones}} P(x_i \mid c_j)$$

Los xi son las características del documento, es decir las palabras. Ya que un documento es representado como una "bolsa" de palabras.

Y cada una de las P(xi | cj) es calculada, usando Laplace smoothing:

$$P'(w \mid c) = \frac{\text{cantidad } (w \mid c) + 1}{\text{cantidad}(c) + |V+1|}$$

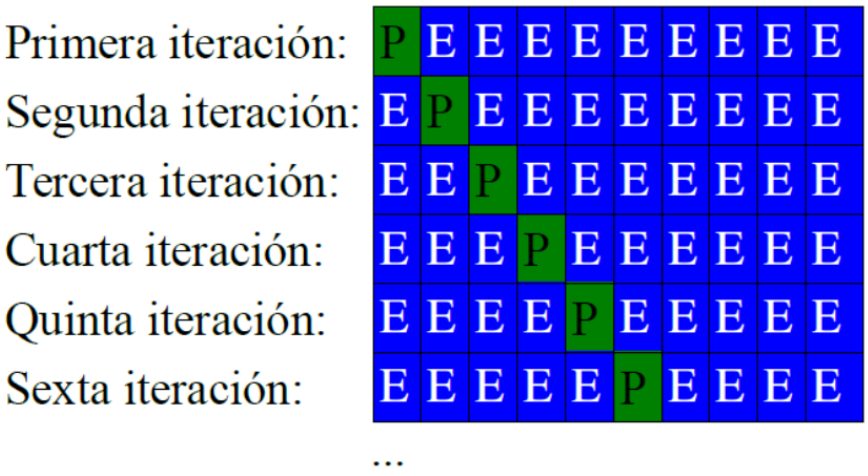
Clasificación - Naïve Bayes multinomial binarizada (o booleana)

Antes de comenzar a calcular las probabilidades de las clases y a contar las palabras vamos a recorrer uno por uno todos los documentos en el conjunto de entrenamiento y prueba, y vamos a eliminar las palabras duplicadas.

En términos generales, esta variante del algoritmo da mejores resultados que la versión tradicional que cuenta todas las ocurrencias de las palabras.

Entrenamiento

Hacemos Cross Validation para encontrar que conjunto de entrenamiento optimiza los valores y vamos partiendo el conjunto de entrenamiento en 10 partes y tomamos una de ellas como conjunto de validación para chequear cual es la que nos da mejores resultados



Buscar el que maximice la precisión, el recall y la medida F1

Problemas

Sutilezas:

Si usted está leyendo esto porque es su fragancia favorita, por favor úsela exclusivamente en su casa y cierre bien las ventanas.

Sarcasmo, ironia.

Expectativas frustradas:

La película debería ser excelente ya que cuenta con grandes actores y una banda sonora fantástica, sin embargo es terriblemente aburrida.

La critica es negativa, lo que importa es lo que esta despues de la coma. El problema es que en la primera parte esta usando palabras positivas para definir las expectativa.