Lexicon de sentimientos

El Lexicón es un diccionario donde en vez de tener el significado de la palabra, tenemos la carga de valor de la palabra.

Si no queremos entrenar una algoritmo de clasificacion podemos utilizar un lexicón de sentimientos

Ventajas

- Son de facil acceso ya que existen varios recursos disponibles de manera publica con alguna licencia.
- Son menos costosos porque no requieren la implementación de algoritmos avanzados de análisis de sentimientos.
- No es necesario disponer de datos de entrenamiento, si se utiliza un enfoque basado en diccionarios las etiquetas se determinan manualmente y se puede acceder rápidamente al significado de las palabras.

Desventajas

- No suelen identificar el sarcasmo, la negación, los errores gramaticales, las faltas de ortografía o la ironía. Por tanto, pueden no ser adecuados para analizar datos recogidos en plataformas de redes sociales.
- Como toda la clasificación se basa en etiquetas y reglas, se debe disponer de datos suficientes para crear un diccionario fiable.

- Una palabra se etiqueta igual independientemente del contexto. El problema es que un término puede ser positivo o negativo según el contexto.
- Como el etiquetado se realiza manualmente, la preparación de los datos puede llevar mucho tiempo.

Algunos diccionarios:

The General Inquirer

- 1915 palabras en la categoría: “positivas”
- 2291 palabras en la categoría “negativas”
- Clasificaciones complejas como por ejemplo: Fuerte vs. Débil o Activa vs. Pasiva
- Está en inglés
- Es gratis para su uso en investigación.

LIWC (Linguistic Inquiry and Word Count)

- 2300 palabras
- Más de 70 clases.
- Soporta idioma español.
- Tiene clasificaciones complejas.
- No es gratuito

MPQA Subjectivity Cues Lexicon

- Existe desde 2006
- 2718 palabras en la categoría “positivas”
- 4912 en la categoría “negativas”
- Está en idioma inglés
- Indica la intensidad de la palabra (fuerte/ debil)
- Se distribuye bajo licencia GNU GPL

Bing Liu Opinion Lexicon

- Existe desde 2004
- 2006 palabras en la categoría “positivas”
- 4783 en la categoría “negativas”
- Está en idioma ingles
- Solo tiene las categorías: positiva y negativa

SentiWordNet

- Clasificación: positiva, negativa u objetiva
  (pudiendo una palabra tener al mismo tiempo valores negativos y positivos)
- Está basada en WordNet (3.0 la última versión)
- Está en idioma inglés
- Se distribuye bajo una licencia: "ShareAlike" de Creative Commons

💡 **En general hay mas negativas que positivas**

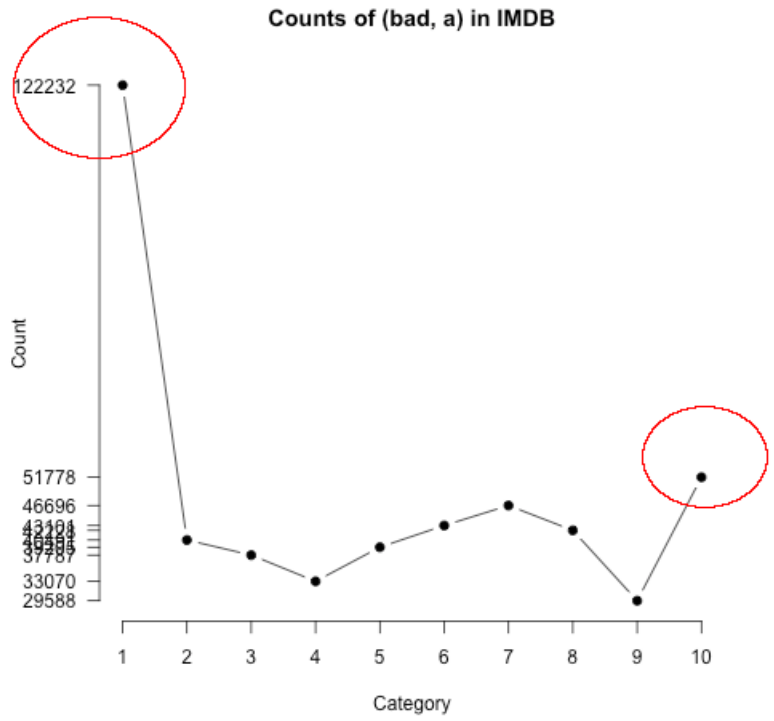
Desacuerdos entre distintos Lexicons

	MPQA	Opinion Lexicon	Inquirer	SentiWordNet	LIWC
MPQA	–	33/5402 (0.6%)	49/2867 (2%)	1127/4214 (27%)	12/363 (3%)
Opinion Lexicon		–	32/2411 (1%)	1004/3994 (25%)	9/403 (2%)
Inquirer			–	520/2306 (23%)	1/204 (0.5%)
SentiWordNet				–	174/694 (25%)
LIWC					–

No solo hay solapamiento entre ellos (cosa esperable) sino que hay desacuerdos respecto a la polaridad de una palabra.

¿Por qué observamos estas diferencias?

Cuantas veces aparecía la palabra: "bad" (malo en ingles) en criticas cinematográficas del sitio IMDB, discriminando según la cantidad de estrellas de la critica (van de 1 a 10)



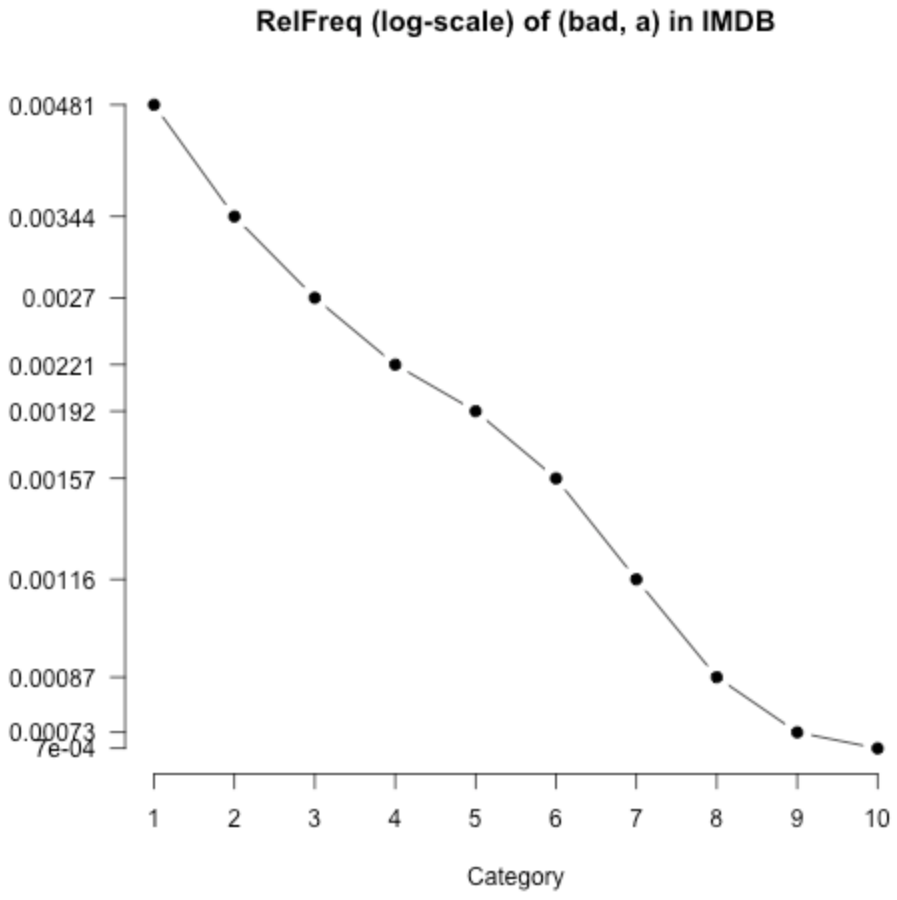
La palabra bad aparece mucho en criticas con 1 estrella y luego baja

Si la palabra bad aparece en una critica hay alta chances de que sea en una de 1 estrella, pero si no es de una estrella lo mas probable es que sea de una critica de 10 estrellas

→ Eso no tiene mucho sentido

Las personas en general hacen muchas criticas de 1 estrella o de 10, se van a los extremos

Si lo normalizamos y ponemos la frecuencia relativa segun la cantidad de criticas, entonces la frecuencia de bad empieza a disminuir como uno esperaria.



Creación de un lexicón propio

En ocasiones conviene armar un Lexicón propio para un dominio especifico antes que usar uno ya existente.

Para eso necesitamos:

● Un puñado de ejemplos previamente clasificados

● Algunas reglas escritas a mano que identifiquen ciertos patrones en una frase.

Algoritmo de Hatzivassiloglou y McKeown para la ampliación de un lexicón

Buscaron que otras palabras aparecían vinculadas a palabras ya conocidas. El razonamiento es el siguiente: si una palabra con polaridad conocida aparece unida por la conjunción "**y** " a una segunda palabra concluyo que la nueva palabra tendrá una polaridad similar. En cambio si vienen unidas por la conjunción "**pero** " la polaridad de la nueva palabra será opuesta.

Adjetivos unidos por "y" tienen la misma polaridad:

● Justo y legitimo

● corrupto y brutal

Adjetivos unidos por "pero" tienen distinta polaridad:

● justo pero brutal

● corrupto pero legitimo

● hermosa pero malvada

Teniendo esta idea en mente, idearon un algoritmo en 4 pasos:

1) Construyeron un Lexicón a mano con 1336 adjetivos:

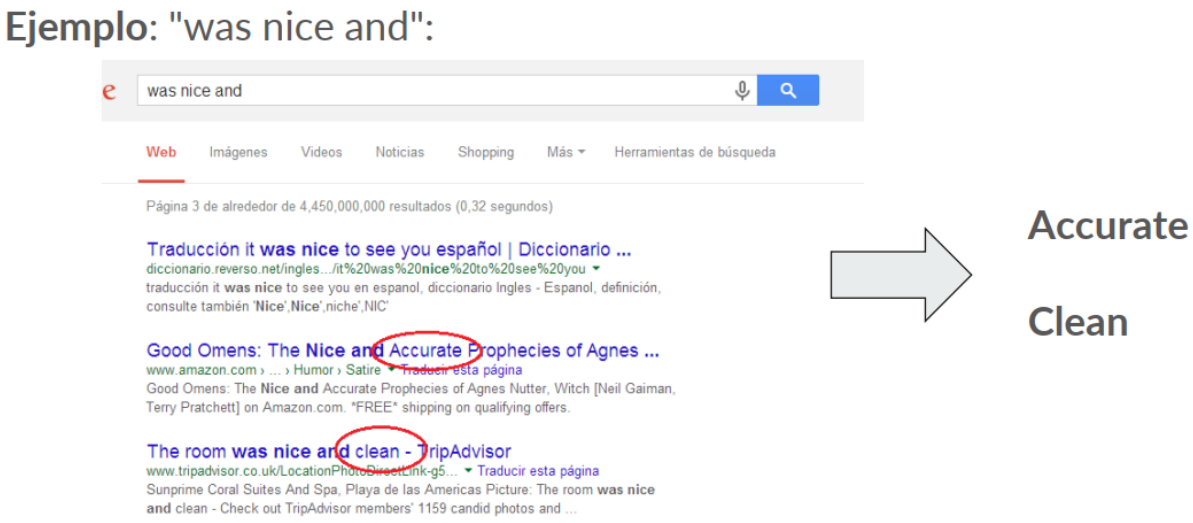
657 positivos y

679 negativos

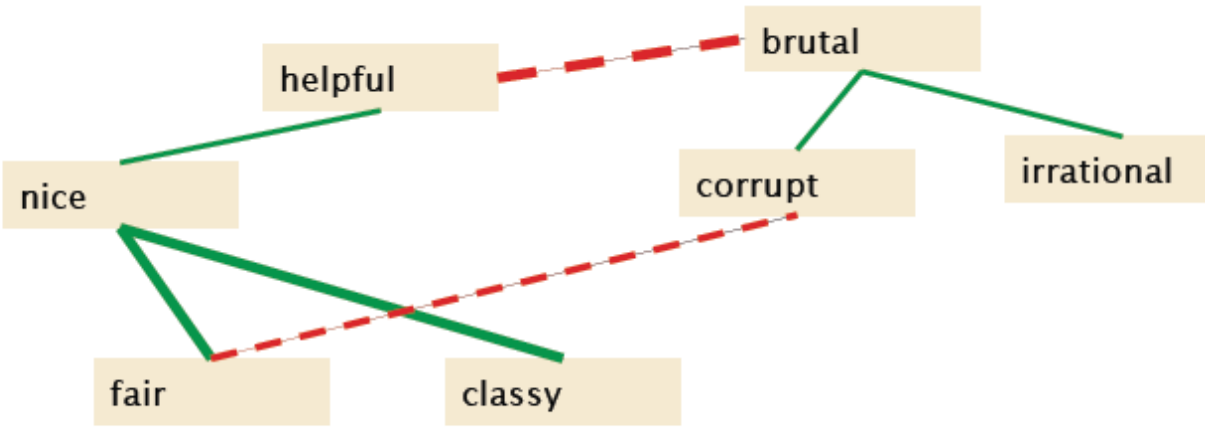
2) Buscaron en Google cada uno de los adjetivos con la formula:

“was <adjetivo> and” y recolectaron la palabra que seguía a continuación.

Luego lo repitieron con "but" en vez de "and".

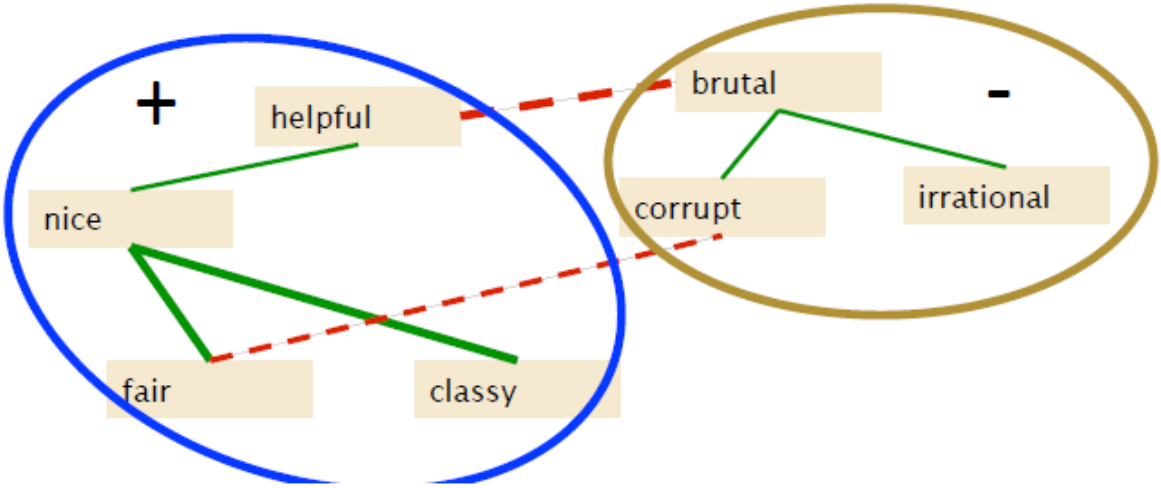


3) Construyeron un mapa que vinculaba las palabras similares entre sí, mostrando también las que tenían sentido opuesto:



En verde las de la misma polaridad y las de distinta con lineas rojas

4) Finalmente buscaron una forma de separar el mapa creado intentando que queden dos conjuntos bien diferenciados:



Si bien lograron ampliar considerablemente el Lexicón original, el nuevo Lexicón contenía algunos errores, es decir palabras mal catalogadas.

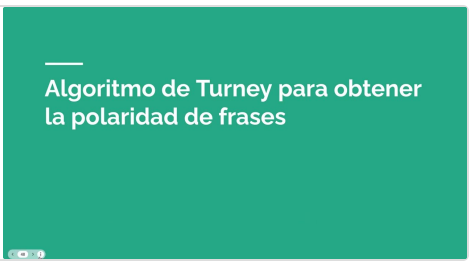
Es por ello que este algoritmo necesariamente necesita de un paso extra que consista en la revisión de los datos obtenidos.

Algoritmo de Turney para obtener la polaridad de frases

PLN Análisis de Sentimientos 02

Procesamiento de lenguaje natural: Análisis de sentimientos parte dos

https://www.youtube.com/watch?v=tHScfqHa7xI



Siempre estamos trabajando con palabras y vemos que las palabras tienen una carga de valor

Pero que pasa si tenemos tambien frases que tienen una carga de valor?

Este algoritmo se puede desglosar en **3 pasos principales**:

1. Extraer frases de opiniones/críticas (reviews) y armar un Lexicón de frases

2. Aprender la polaridad de cada frase

3. Puntuar las críticas según el promedio de las polaridades de sus frases

Extracción de frases

Turney no extrajo cualquier frase, sino que usó solo frases de 2 palabras, siempre que estas coincidiesen con alguno de los siguientes patrones:

Primer Palabra	Segunda Palabra	Tercer Palabra (no se extrajo)
Adjetivo	Sustantivo (plural o singular)	Cualquier palabra
Adverbio	Adjetivo	No Sustantivo
Adjetivo	Adjetivo	No Sustantivo
Sustantivo (plural o singular)	Adjetivo	No Sustantivo
Adverbio	Verbos	Cualquier palabra

Polaridad de las frases

Para verificar la polaridad de las frases, se verificó cuan cerca aparecían estas de palabras con polaridad ya conocida como por ejemplo: "excelente" y "pobre".

La idea detrás de esto es que la co-ocurrencia de una frase junto con la palabra "excelente" o bien con la palabra "pobre" no es casualidad, sino que la frase misma tiene una carga de valor, una polaridad.

Pointwise mutual information:

El PMI de un par de valores x e y pertenecientes a dos variables aleatorias discretas: X e Y respectivamente, cuantifica la discrepancia entre la probabilidad de su coincidencia dada su distribución conjunta y su distribución individual y asumiendo su independencia.

Matemáticamente:

$$PMI(X,Y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

En otras palabras: cuanto más posible es que el evento X aparezca vinculado al evento Y a que aparezcan ambos de forma independiente entre sí.

Pointwise mutual information entre dos palabras:

$$PMI(palabra1,palabra2) = \log_2 \frac{P(palabra1,palabra2)}{P(palabra1)P(palabra2)}$$

Turney utilizó el buscador Altavista.com para obtener estos valores, pero la forma que utilicemos para contar estos resultados dependerá de nuestro conjunto de datos de entrenamiento.

$$P(palabra) = \frac{\text{cantidad de resultados para "palabra"}}{\text{cantidad total}}$$

$$P(palabra) = \frac{\#palabra}{N}$$

Co-ocurrencia de palabra 1 y palabra 2

$$P(palabra1, palabra2) =$$

$$= \frac{\text{cantidad de resultados para "palabra1 } \textit{NEAR} \text{ palabra2"}}{(\text{cantidad total})^2}$$

$$P(palabra1, palabra2) = \frac{\#(palabra1\ \textit{NEAR}\ palabra2)}{N^2}$$

NEAR significa "cerca de", indica que palabra1 apareció a no más de N palabras de distancia de palabra2 en un texto dado.

$$PMI(palabra1, palabra2) = \log_2 \frac{\frac{\#(palabra1\ \textit{NEAR}\ palabra2)}{N^2}}{\frac{\#(palabra1)\ \#(palabra2)}{N*N}}$$

Los denominadores se cancelan

$$\textbf{PMI}(\text{palabra1, palabra2}) = \log_2 \frac{\#(\text{palabra1 NEAR palabra2})}{\#(\text{palabra1})\ \#(\text{palabra2})}$$

Calculando la polaridad de una frase:

$$\text{Polaridad(frase)} = \text{PMI(frase,"excelente")} - \text{PMI(frase,"pobre")}$$

En resumen:

$$\text{Polaridad (frase)} = \log_2 \frac{\#(\text{frase NEAR excelente})\ \ \#(\text{pobre})}{\#(\text{frase NEAR pobre})\ \ \#(\text{excelente})}$$

Ejemplos:

Frase	Polaridad
online service	2.8
online experience	2.3
low fees	0.33
inconveniently located	-1.5
Average	0.32

Puntuar las críticas

Armar una lista de frases y asociarles el valor obtenido con los cálculos anteriores para luego descomponer las críticas en sus frases y realizar el promedio.

Es decir, basicamente hacer el analisis de sentimientos sobre distintos textos utilizando las frases

Ejemplo: Sobre 410 opiniones de Epinions

- Exactitud promedio: 74%
  - criticas cinematográficas: 66%
  - críticas sobre bancos y automóviles: 80% y el 84%
  - críticas sobre viajes: intermedio.

Los resultados no fueron extremadamente positivos, pero si demostro que hay frases que tambien tienen carga de valor y nos estan dando pistas sobre si una critica es positiva o negativa

---

7 - 4 - Learning Sentiment Lexicons (14-45).mp4

https://www.youtube.com/watch?v=TrUo8qvd02I