

Kaggle Team: NTF Team

INTEGRANTES

* Florencia Montero xxxxx
* Nicolás Poncet Carrasco 94983
* Tomas Mussi xxxxx

75.06 Organización de Datos

Informe de diseño

Índice

1. Introducción 2  
   - Descripción de resolución 2
2. Objetivos 2

- Objetivo **………………………………………………………………………………………………………2**  
- Explicación del proceso de resolución…………………………………………………………………………3

1. Pruebas y análisis de resultados………………………………………..................3
2. Bibliografía………………………………………………………………………….5

# Introducción

**Descripción breve de la solución**

La resolución del programa se basará en el análisis de palabras mediante la extracción de información utilizando n-gramas, métodos de hashing, bag of words y combinar otras posibilidades para procesarse mediante el uso del clasificador Maximum Entropy las posibles soluciones.

# Objetivos

**Objetivo**

El objetivo principal del grupo es alcanzar de la manera más eficiente el mejor resultado posible. Para esto se realizan diversas pruebas con diferentes condiciones buscando las mejores soluciones con el mejor manejo de recursos para la resolución.

**Explicación del proceso de resolución**

El grupo realizó diversas pruebas de análisis de los datos para poder encontrar un modelo lo más óptimo posible (ver cuadro de pruebas Página 3). De la información recolectada se tomó la decisión de utilizar como método de clasificación de sentimiento al Maximum Entropy, que dio uno de los mejores resultados y que puede mejorarse para aumentar su precisión.

¿Por qué Maximun Entropy?

Maximum Entropy (MaxEnt) es un clasificador avanzado funcionalmente que utiliza modelos matemáticos y estadísticos complejos para llegar a un resultado. Para utilizar MaxEnt se variarán los “features” o características y atributos de los reviews probando distintas combinaciones mediante la refinación de los “features extractors” encargados de generar esas características.

En las pruebas realizadas, MaxEnt tuvo buenos resultados de performance y buenos resultados de acierto en la predicción de los sentimientos buscados. Se espera que estos modelos se puedan complementar hasta alcanzar un óptimo resultado modificando y ampliando las características a analizar. Se espera obtener un acierto de resultados entre 90 y 95 %

El objetivo principal del grupo es hacer foco en mejorar el análisis y procesamiento de las características de los reviews para crear “features” con información más precisa y exacta. Se realizarán procesos de limpieza, procesos de hashing que puedan servir para futura identificación de palabras, se utilizará, si es posible el proceso de Stemming, un proceso de reducción de palabras derivadas en otras, donde por ejemplo aquellas palabras que se desprendan de otras serán asociadas a la misma palabra (greatest -> great, disgusting -> disgust, etc). Esto permitirá que si en el proceso de clasificación no se conoce una palabra, se pueda asociar a otra que se conozca su valoración (sea positiva o negativa). Después de este pre-procesamiento, sobre el texto, se excluirán stopwords y se procederá al uso de n-gramas para la transformación de la información resultante a vectores

# Pruebas y análisis de resultados

Se realizaron pruebas con distintos “features” y clasificadores para recolectar información y decidir cuál o cuáles serían los métodos a utilizar.

|  |  |  |  |
| --- | --- | --- | --- |
| Features | Clasificador | Resultado  (Kaglle submition) | |
| Bag Of Words (5000 fts.) | Random Forest (200 Trees) | 0.84516 | 84.5% |
| Bag Of Words (5000 fts.) | Random Forest (300 Trees) | 0.85292 | 85.2% |
| Bag Of Words (5000 fts.) | Random Forest (400 Trees) | 0.85396 | 85.3% |
| Bigram (5000 fts.) | Random Forest (300 Trees) | 0.85592 | 85.6& |
| Bigram (7000 fts.) | Random Forest (300 Trees) | 0.92834 | 92.8% |
| Bigram (7000 fts.) | Random Forest (251 Trees) | 0.93021 | 93% |
| Bigram (9000 fts.) | Random Forest (251 Trees) | 0.93127 | 93.1% |
| Bag Of Words (5000 fts.) | Gaussian Bayes Naive | 0.68852 | 68.8% |
| Bag Of Words (5000 fts.) | Gaussian Bayes Naive | 0.62976 | 62.9% |
| Bag Of Words (5000 fts.) | K Nearest Neighbors (dist 5) | 0.62476 | 62.5% |
| Bigram (5000 fts.) | Maximum Entropy | 0.92608 | 92.6% |

Sólo se incluyeron pruebas significativas.

Luego de la ejecución de las pruebas, se vieron diferencias notorias entre los distintos clasificadores y la relación entre los tipos de features que se generen para cada uno. En el caso particular de KNN se descartó principalmente por el uso desmedido de recursos que usa.

Es por esto que se decidió elegir aquellos dos generadores de features y clasificadores que dieron los mejores resultados.

Nota: estos resultados fueron subidos a Kaggle en las cuentas individuales de los alumnos. El usuario del grupo no tiene ningún submit realizado.

# Bibliografía

Materiales consultados:

*Twitter Sentiment Classification using Distant supervisión –* <http://s3.eddieoz.com/docs/sentiment_analysis/Twitter_Sentiment_Classification_using_Distant_Supervision.pdf>

Sentiment Analysis (Wikipedia) - <http://en.wikipedia.org/wiki/Sentiment_analysis>

Sentiment symposium - <http://sentiment.christopherpotts.net/classifiers.html>

Logistic Regression Model - <http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>

Naïve Bayes Classifier - <http://www.ic.unicamp.br/~rocha/teaching/2011s1/mc906/aulas/naive-bayes.pdf>

Kaggle Tutorial - <https://www.kaggle.com/c/word2vec-nlp-tutorial>

Stemming Algorith - <http://xapian.org/docs/stemming.html>

Stemming (Wikipedia) - <http://en.wikipedia.org/wiki/Stemming>