Robots 75.70 - Trabajo Practico 1

Aymeric Cousaert (105464)

8 de junio de 2020

Índice

1. Cambios realizados relacionados al enunciado

Tenemos 4 categorías para la clasificación de los tweets : alegría, tristeza-miedo, enojo, apoyo-esperanza.

En la categoría tristeza-miedo se borro el emoticón U0001F621 (pouting face).

En la categoría enojo se agrego los emoticones U00012620 (skull and crossbones), U0001F480 (skull) y U0001F92C (face with symbols on mouth).

2. Clasificador Bayes Naïve con sklearn

2.1. Cantidad de tweets

Los pasos fueron los siguientes :

1. Creamos un gran dataframe que contiene todos los tweets desde el día 16 de abril 2020 hasta el día 30 de abril 2020.

   Este dataframe contiene 3 306 413 tweets. Entre esos tweets, tenemos 107 686 tweets teniendo a lo mínimo un emoticón de la lista de los emoticones en cada categoría y 31 987 27 tweets non clasificados.

2. Eliminamos todos los tweets que tienen emoticones en 2 o más categorías. Logramos a tener de esta manera :

   - 32 849 tweets de la categoría alegría
   - 10 876 tweets de la categoría tristeza-miedo
   - 7 386 tweets de la categoría enojo
   - 51 230 tweets de la categoría apoyo-esperanza.

   5345 tweets fueron eliminados por estar en 2 o más categorías.

   Finalmente, tenemos 102341 tweets clasificados y 3204072 tweets non clasificados.

2.2. Precisión, Recall y Medida F1

Sobre el dataset de test, tenemos :

1. Accuracy = 0.55191637630662016
2. Precisión = 0.56462842770562371
3. Recall = 0.55204766553797802
4. Medida F1 = 0.55222224970632916

2.3. Resultados obtenidos

Teniendo en cuenta todos los tweets incluyendo los tweets non clasificados, tenemos :

- 700 406 tweets de la categoría alegría, i.e. 33,6 % del total;
- 666 967 tweets de la categoría tristeza-miedo, i.e. 25,0 % del total;
- 1 112 163 tweets de la categoría enojo, i.e. 21,2 % del total;

– 826 877 tweets de la categoría apoyo-esperanza, i.e. 20,2 % del total.

Tenemos :

– 1 527 283 tweets positivos, i.e. 46,1 % del total;

– 1 779 130 tweets negativos, i.e. 53,9 % del total.

2.4. Análisis sobre día 16 de Abril solamente

Teniendo en cuenta que la parte siguiente pide mucho tiempo de ejecución, hacemos el análisis con solo el día 16 para poder hacer comparación.

Tenemos para el model entrenado :

1. Accuracy = 0.52413793103448281
2. Precisión = 0.52924417493067422
3. Recall = 0.52399920808874678
4. Medida F1 = 0.52520052009011242

Los resultados son :

– 142054 tweets positivos, i.e. 47,6 % del total;

– 156159 tweets negativos, i.e. 52,4 % del total.

3. Análisis con SentiWordNet

Después de cambiar el código java para tratar un solo archivo .txt con todos los tweets, el tiempo de ejecución quedo muy elevado. En 24 horas pude tener el resultado para solo los 163051 primeros tweets, lo que es la mitad de los tweets del día 16 de abril.

Los resultados son :

– 101500 tweets positivos, i.e. 62,2 % del total;

– 61551 tweets negativos, i.e. 37,7 % del total.

Son muy diferentes de los resultados que conseguimos en la parte anterior : la mayoría de los tweets son acá positivos.

El código java cambiado ha sido agregado al repositorio github : `https://github.com/aymericcousaert/tweets-covid`


tp1 — java -jar fastSentimentClassifier.jar -f tweets.txt — 146×20
~ — jupyter-notebook ▸ python
~/Desktop/robots/tp1 — java -jar fastSentimentClassifier.jar -f tweets.txt
Number of negative tweets : 61385
Number of positive tweets : 101240
Number of negative tweets : 61411
Number of positive tweets : 101276
Number of negative tweets : 61425
Number of positive tweets : 101308
Number of negative tweets : 61443
Number of positive tweets : 101338
Number of negative tweets : 61463
Number of positive tweets : 101371
Number of negative tweets : 61480
Number of positive tweets : 101405
Number of negative tweets : 61496
Number of positive tweets : 101434
Number of negative tweets : 61517
Number of positive tweets : 101466
Number of negative tweets : 61535
Number of positive tweets : 101500
Number of negative tweets : 61551