Por Pablo Llobregat Ruiz email: pllorui@etsinf.upv.es

Linkedin: https://www.linkedin.com/in/pablo-llobregat-ruiz-122a0419b/

CaixaBank

HACKATHON

RETO DATA SCIENCE

1. Preprocesamiento de datos

```
import pandas as pd
import numpy as np

#LECTURA DE DATOS y prelimpieza

tweets= pd.read_csv("tweets_from2015_#Ibex35.csv")
train= pd.read_csv("train.csv")
test= pd.read_csv("test_x.csv")

#Pretratamiento

train["Date"]= pd.to_datetime(train["Date"]).dt.date
train["Date"]=train["Date"].apply(str)
test["Date"]= pd.to_datetime(test["Date"]).dt.date
test["Date"]=test["Date"].apply(str)
train=train.set_index(train["Date"]).drop(["Date"],axis=1).dropna()
test=test.set_index(test["Date"]).drop(["Date"],axis=1).dropna()

# Limpieza de la fecha de los tweets
def to_date(date1):
    try:
        from datetime import datetime
        date=datetime.strptime(date1, '%a %b %d %H:%M:%S %z %Y')
        return date
    except: return None

tweets["tweetDate"]=tweets["tweetDate"].apply(to_date)
tweets["tweetDate"] = pd.to_datetime(tweets["tweetDate"]).dt.date
tweets= tweets.dropna()
```

Cargo los datasets y limpio la fecha de todos ellos. En especial la del dataset de twitter, que estaba en otro formato completamente distinto.

2. Análisis de los sentimientos de los tweets

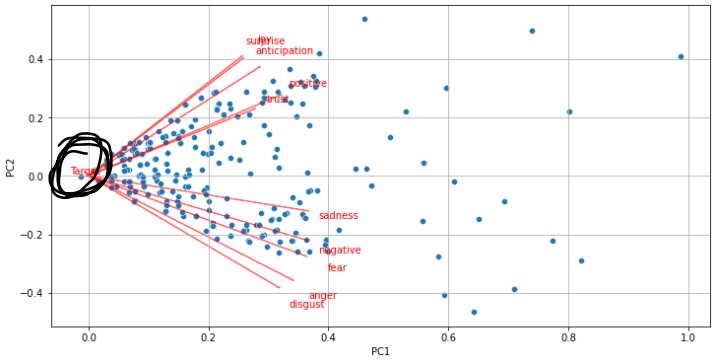
```
tweets = read.delim("Tweets.csv", sep=",", encoding="UTF-8")

library(syuzhet)

sent = data.frame()
for (i in 1:nrow(tweets)){
  sent = rbind(sent,get_nrc_sentiment(tweets$text[i]))
}
```

Agrupo los tweets por día. Seguidamente, en R con la librería “SYUZHET” extraigo todos los sentimientos de cada día de los tweets.

Por Pablo Llobregat Ruiz email: pllorui@etsinf.upv.es

Linkedin: https://www.linkedin.com/in/pablo-llobregat-ruiz-122a0419b/

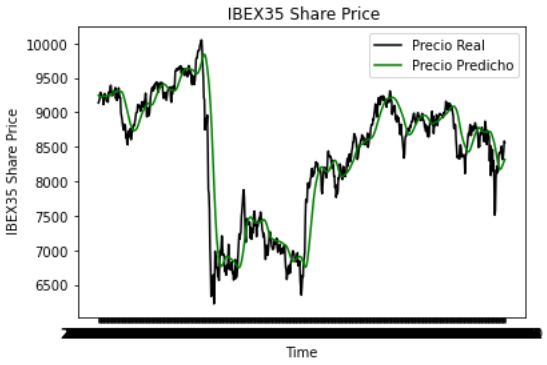
Una vez tengo los sentimientos de cada día junto el Dataset de Entrenamiento (TRAIN) con los sentimientos de cada día, para mediante un PCA intentar explicar la variable TARGET.



Como se puede apreciar en el PCA de 2 dimensiones que explica más del 64% de la variabilidad... La variable TARGET no está explicada por los sentimientos de los tweets, por tanto, no consideraremos los tweets para nuestro modelo predictivo. (Meter los sentimientos en el modelo predictivo solo añadiría ruido)

3. Análisis de los sentimientos de los tweets

Propongo crear un modelo basado en redes neuronales que dada la información de los 30 días anteriores sea capaz de predecir la cotización a los 3 días.

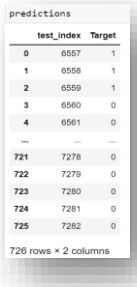



```
# CONSTRUCCIÓN DE LA RED NEURONAL

# consultando varias fuentes, las capas LSTM son las más adecuadas para estos casos

model = Sequential()

model.add(LSTM(units=50, return_sequences=True, input_shape=(x_train.shape[1], 1)))
model.add(Dropout(0.2))
model.add(LSTM(units=50, return_sequences=True))
model.add(Dropout(0.2))
model.add(LSTM(units=50))
model.add(Dropout(0.2))
model.add(Dense(units=1)) #prediction of the next closing value

model.compile(optimizer='adam', loss='mean_squared_error')  # Optimizador adam y error medio cuadrático
model.fit(x_train, y_train, epochs=5, batch_size=100)
```

(en el Notebook se explica cómo se ha creado)

Se generan las predicciones del dataset TEST

4. Exportación resultados variable TARGET



predictions

	test_index	Target
0	6557	1
1	6558	1
2	6559	1
3	6560	0
4	6561	0
...	...	...
721	7278	0
722	7279	0
723	7280	0
724	7281	0
725	7282	0

726 rows × 2 columns

A partir de las predicciones creamos los valores de TARGET muy fácilmente

Si la predicción es mayor al valor actual -> 1, si no 0.