UNIVERSITAT POLITÈCNICA DE VALÈNCIA

CAMPUS DE GANDIA

gti Grado en tecnologías interactivas

Big Data:

Proyecto despliegue

Raúl Santos López

Dataset: ibexunido.csv

Este dataset contiene los históricos de la bolsa del Ibex 35 de los últimos 5 años.

	DATE	OPEN	HIGH	LOW	CLOSE	ADJ CLOSE	VOLUME	NAME
1276	2022-04-22	3,1	3,17	3,062	3,09	3,09	10115198	CAIXABANK
1277	2022-04-25	3,03	3,075	3,01	3,023	3,023	16143684	CAIXABANK
1278	2022-04-26	3,057	3,067	2,932	2,935	2,935	24628811	CAIXABANK
1279	2017-04-27	14,146264	14,216775	14,049311	14,062532	13,849058	786724	CELLNEX TELECOM
1280	2017-04-28	14,084567	14,322541	14,036091	14,300507	14,08342	1706516	CELLNEX TELECOM
1281	2017-05-02	14,348983	14,578144	14,278472	14,520854	14,300424	1585118	CELLNEX TELECOM
1282	2017-05-03	14,503226	14,591365	14,423901	14,582551	14,361184	4044978	CELLNEX TELECOM
1283	2017-05-04	14,600178	14,776456	14,45475	14,776456	14,552143	932771	CELLNEX TELECOM
1284	2017-05-05	14,723573	14,855781	14,613399	14,846967	14,621585	933093	CELLNEX TELECOM
1285	2017-05-08	14,948326	15,027651	14,794084	14,807304	14,582525	1021499	CELLNEX TELECOM
1286	2017-05-09	14,807304	14,855781	14,666282	14,802897	14,609971	686109	CELLNEX TELECOM

Fuente: Yahoo Finances

Ejercicio 1: listado semanal IBEX 35

	DATE	OPEN	HIGH	LOW	CLOSE	ADJ CLOSE	VALUAE	NAME
1276	2022-04-22	3,1	3,17	3,062	3,09	3,09	10115198	CAIXABANK
1277	2022-04-25	3,03	3,075	3,01	3,023	3,023	16143684	CAIXABANK
1278	2022-04-26	3,057	3,067	2,932	2,935	2,935	24628811	CAIXABANK
1279	2017-04-27	14,146264	14,216775	14,049311	14,062532	13,849058	786724	CELLNEX TELECOM
1280	2017-04-28	14,084567	14,322541	14,036091	14,300507	14,08342	1706516	CELLNEX TELECOM
1281	2017-05-02	14,348983	14,578144	14,278472	14,520854	14,300424	1585118	CELLNEX TELECOM
1282	2017-05-03	14,503226	14,591365	14,423901	14,582551	14,361184	4044978	CELLNEX TELECOM
1283	2017-05-04	14,600178	14,776456	14,45475	14,776456	14,552143	932771	CELLNEX TELECOM
1284	2017-05-05	14,723573	14,855781	14,613399	14,846967	14,621585	933093	CELLNEX TELECOM
1285	2017-05-08	14,948326	15,027651	14,794084	14,807304	14,582525	1021499	CELLNEX TELECOM
1286	2017-05-09	14,807304	14,855781	14,666282	14,802897	14,609971	686109	CELLNEX TELECOM

ibexunido.csv

COLUMNAS
APERTURA
CIERRE
MÍNIMO
MÁXIMO

Script: /stocks/scripts/semanal.py **Origen:** HDFS(/ibex35/ibexunido.csv) **Destino:** pantalla

Ejercicio 2: listado mensual IBEX 35

	DATE	OPEN	HIGH	LOW	CLOSE	ADJ CLOSE	VOLUME	NAME
1276	2022-04-22	3,1	3,17	3,062	3,09	3,09	10115198	CAIXABANK
1277	2022-04-25	3,03	3,075	3,01	3,023	3,023	16143684	CAIXABANK
1278	2022-04-26	3,057	3,067	2,932	2,935	2,935	24628811	CAIXABANK
1279	2017-04-27	14,146264	14,216775	14,049311	14,062532	13,849058	786724	CELLNEX TELECOM
1280	2017-04-28	14,084567	14,322541	14,036091	14,300507	14,08342	1706516	CELLNEX TELECOM
1281	2017-05-02	14,348983	14,378144	14,278472	14,320854	14,300424	1585118	CELLNEX TELECOM
1282	2017-05-03	14,503226	14,591365	14,423901	14,582551	14,361184	4044978	CELLNEX TELECOM
1283	2017-05-04	14,600178	14,776456	14,45475	14,776456	14,552143	932771	CELLNEX TELECOM
1284	2017-05-05	14,723573	14,855781	14,613399	14,846967	14,621585	933093	CELLNEX TELECOM
1285	2017-05-08	14,948326	15,027651	14,794084	14,807304	14,582525	1021499	CELLNEX TELECOM
1286	2017-05-09	14,807304	14,855781	14,666282	14,802897	14,609971	686109	CELLNEX TELECOM

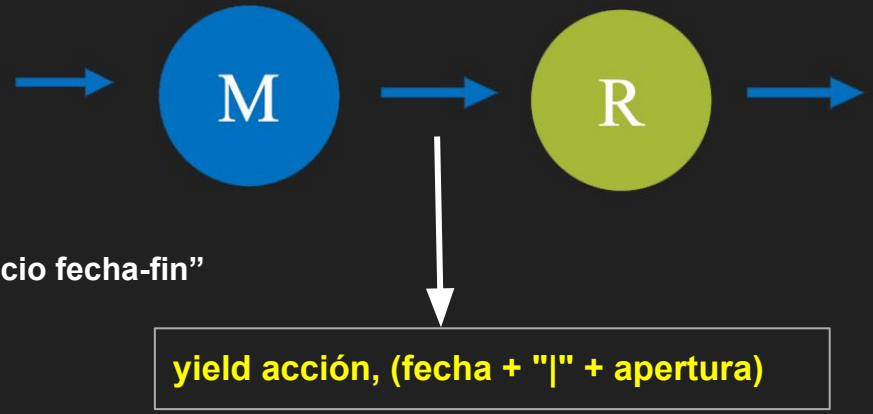
ibexunido.csv

COLUMNAS
APERTURA
CIERRE
MÍNIMO
MÁXIMO

Script: /stocks/scripts/mensual.py **Origen:** HDFS(/ibex35/ibexunido.csv) **Destino:** pantalla

Ejercicio 3: porcentaje de incremento y decremento entre fechas

	DATE	OPEN	HIGH	LOW	CLOSE	ADJ CLOSE	VOLUME	NAME
1275	2022-04-22	3,1	3,17	3,062	3,09	3,09	10115198	CAIXABANK
1277	2022-04-25	3,03	3,075	3,01	3,023	3,023	16143684	CAIXABANK
1278	2022-04-26	3,057	3,067	2,932	2,935	2,935	24628811	CAIXABANK
1279	2017-04-27	14,146264	14,216775	14,049311	14,062532	13,849058	786724	CELLNEX TELECOM
1280	2017-04-28	14,086457	14,322541	14,036091	14,309507	14,08042	1706516	CELLNEX TELECOM
1281	2017-05-02	14,348983	14,578144	14,278472	14,520854	14,300424	1585118	CELLNEX TELECOM
1282	2017-05-03	14,501626	14,591365	14,423901	14,582551	14,361184	4044078	CELLNEX TELECOM
1283	2017-05-04	14,600178	14,776456	14,45475	14,776456	14,552143	932771	CELLNEX TELECOM
1284	2017-05-05	14,723573	14,855781	14,613399	14,846967	14,621585	933093	CELLNEX TELECOM
1285	2017-05-08	14,948326	15,027651	14,794084	14,807304	14,582525	1021499	CELLNEX TELECOM
1286	2017-05-09	14,807304	14,855781	14,666282	14,802897	14,609971	686109	CELLNEX TELECOM



Args: “acción fecha-inicio fecha-fin”

ibexunido.csv

COLUMNAS
ACCIÓN
FECHA-APERTURA
VALOR-INICIAL
MÁXIMO
MÍNIMO
INCREMENTO
DECREMENTO

Script: /stocks/scripts/tres.py **Origen:** HDFS(/ibex35/ibexunido.csv) **Destino:** pantalla

Ejercicio 4: valor mínimo y máximo de la última hora, mes y año

	DATE	OPEN	HIGH	LOW	CLOSE	ADJ CLOSE	VOLUME	NAME
1275	2022-04-22	3,1	3,17	3,062	3,09	3,09	10115198	CAIXABANK
1277	2022-04-25	3,03	3,075	3,01	3,023	3,023	16143684	CAIXABANK
1278	2022-04-26	3,057	3,067	2,932	2,935	2,935	24628811	CAIXABANK
1279	2017-04-27	14,146264	14,216775	14,049311	14,062532	13,849058	786724	CELLNEX TELECOM
1280	2017-04-28	14,084567	14,322541	14,036091	14,309507	14,08042	1706516	CELLNEX TELECOM
1281	2017-05-02	14,348983	14,578144	14,278472	14,520854	14,300424	1585118	CELLNEX TELECOM
1282	2017-05-03	14,501226	14,591365	14,423901	14,582551	14,361184	4044078	CELLNEX TELECOM
1283	2017-05-04	14,600178	14,779456	14,45475	14,779456	14,552143	932771	CELLNEX TELECOM
1284	2017-05-05	14,725573	14,855781	14,613399	14,846967	14,621585	933093	CELLNEX TELECOM
1285	2017-05-08	14,948326	15,027651	14,794084	14,807304	14,582525	1021499	CELLNEX TELECOM
1286	2017-05-09	14,807304	14,855781	14,666282	14,802897	14,609971	686109	CELLNEX TELECOM

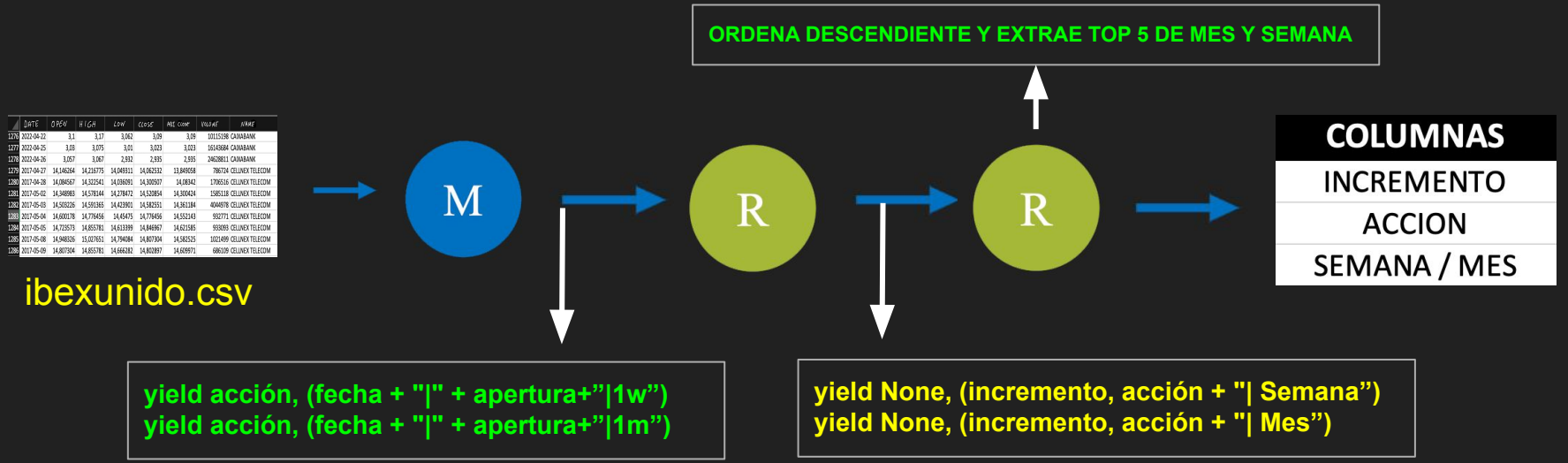
ibexunido.csv


```
yield acción, (fecha + "|" + apertura+"|1h")
yield acción, (fecha + "|" + apertura+"|1w")
yield acción, (fecha + "|" + apertura+"|1m")
```

COLUMNAS
ACCIÓN
MÍMINO 1 HORA
MÁXIMO 1 HORA
MÍMINO 1 SEMANA
MÁXIMO 1 SEMANA
MÍMINO 1 MES
MÁXIMO 1 MES

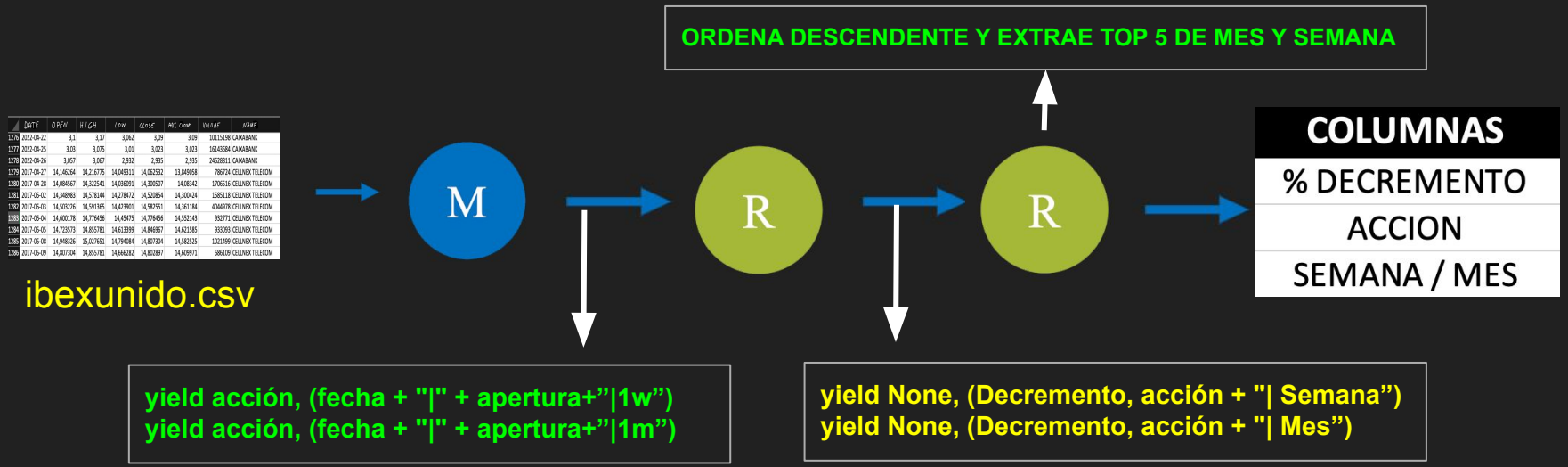
Script: /stocks/scripts/cuatro.py **Origen:** HDFS(/ibex35/ibexunido.csv) **Destino:** pantalla

Ejercicio 5: mostrar 5 acciones que más han subido en semana y mes



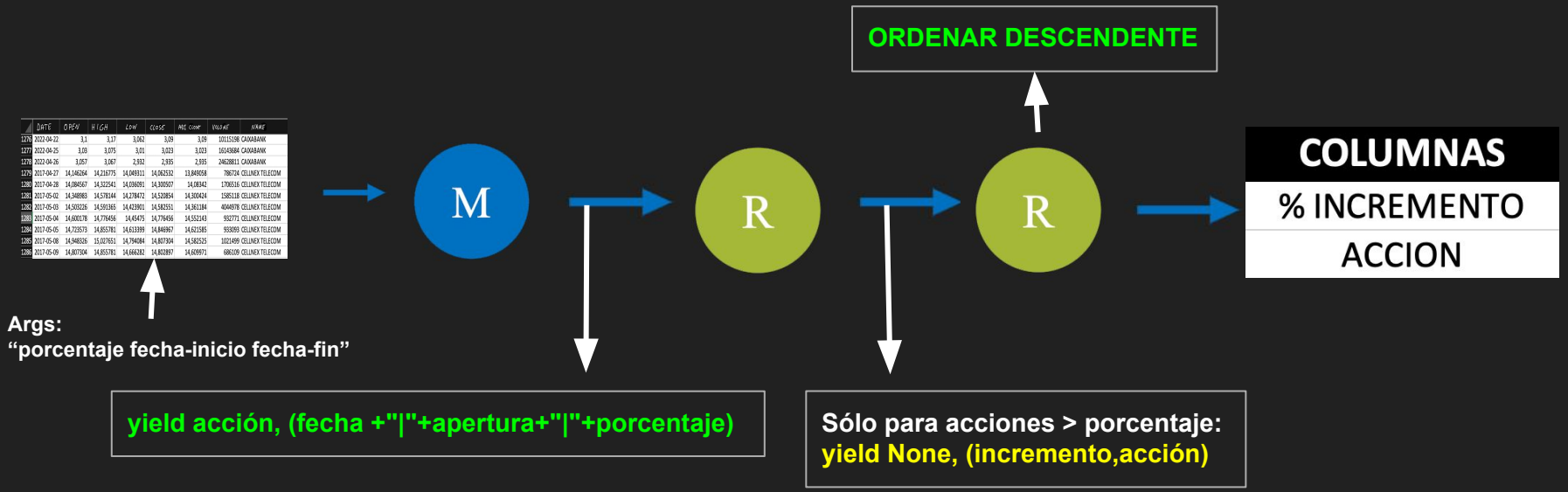
Script: /stocks/scripts/cinco.py **Origen:** HDFS(/ibex35/ibexunido.csv) **Destino:** pantalla

Ejercicio 6: mostrar 5 acciones que más han bajado en semana y mes



Script: /stocks/scripts/seis.py **Origen:** HDFS(/ibex35/ibexunido.csv) **Destino:** pantalla

Ejercicio 7: dado porcentaje y rango fechas mostrar las acciones



Script: /stocks/scripts/siete.py **Origen:** HDFS(/ibex35/ibexunido.csv) **Destino:** pantalla

Funcionalidad avanzada: extracción de noticias internacionales

1. Script para la extracción:
    a. **Un Request a Google search con las palabras “world news”, añadiendo en la query “&tbm=nws” para que los resultados tengan estatus de noticia y “&tbs=lr:lang_1en,cdr:1,cd_min:{dia-objetivo},cd_max:{dia-objetivo}” para ir extrayendo días completos ordenados por importancia.**
    b. **De los resultados anteriores filtro cuales provienen de la agencia de noticias Reuters (hipervínculo Wikipedia), para cada noticia hago un Request a Reuters con el hipervínculos que me da Google search y extraigo los 9 primeros párrafos de cada artículo más su fecha y titular.**

Script: /news_reuters/news2reuters.py

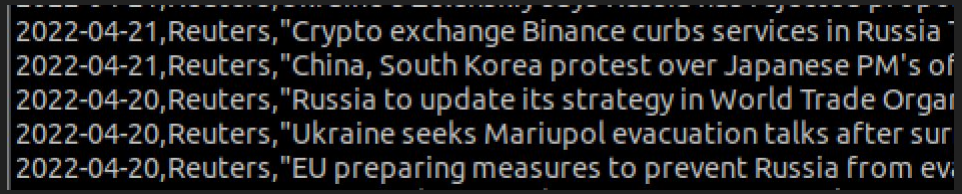
Funcionalidad avanzada: extracción de noticias internacionales

2. Ejemplo de registro que que voy almacenando con un máximo de 35 por día:

2022-04-20,Reuters,"Italy reports 99,848 coronavirus cases on Wednesday, 205 deaths Italy reported 99,848 COVID-19 related cases on Wednesday, against 27,214 the day before, the health ministry said, while the daily number of deaths rose to 205 from 127. Italy has registered 162,098 deaths linked to COVID-19 since its outbreak emerged in February 2020, the second-highest toll in Europe after Britain and the eighth highest in the world. The country has reported 15.86 million cases to date. Patients in hospital with COVID-19 - not including those in intensive care - stood at 10,207 on Wednesday, down from 10,214 a day earlier. There were 44 new admissions to intensive care units, up from 38 on Tuesday. The total number of intensive care patients stood at 413, decreasing from a previous 422. Some 610,600 tests for COVID-19 were carried out in the past day, compared with a previous 174,098, the health ministry said. ",1 week ago, https://www.reuters.com/world/europe/italy-reports-99848-coronavirus-cases-wednesday-205-deaths-2022-04-20/

Columnas por orden: "Fecha que yo decido", "nombre de la fuente artículo de Google", "párrafos del artículo de Reuters.com", "fecha artículo de Google", "hipervínculo al artículo de Google"

File: /news_reuters/reutersnews_{año}.csv



Funcionalidad avanzada: extracción de noticias internacionales

3. Filtrado de palabras:

a. Eliminar palabras que no aporten nada para obtener el objetivo, la palabra que se repite con mayor frecuencia en un día o un intervalo de fechas.

b. Generar un informe de los últimos 3 años semana a semana mostrando el porcentaje de incremento y decremento del IBEX 35 y las 12 palabras que hay en las noticias que se repiten con mayor frecuencia, evitando duplicados:

c. Haz click en el enlace para ver los resultados, muy curioso.

https://github.com/raul2222/big-data/blob/master/news_reuters/informe_final.csv

Script: /news_reuters/postproceso.py

Funcionalidad avanzada: extracción de noticias internacionales

4. Conclusiones:

a. He de reconocer que cuando he visto el informe final he sentido una gran satisfacción, el esfuerzo ha valido la pena. He disfrutado en todas las fases de su creación.

b. Ya he pensado un nuevo nivel de profundidad para este trabajo, la idea sería etiquetar cada palabra con un identificador único para poder llegar a la frase original de la que proviene y entender su contexto.

https://github.com/raul2222/big-data/blob/master/news_reuters/informe_final.csv