



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	VOLUME	NAME
1270	2022-04-22	3.1	3.17	3.002	3.09	3.09	1011539	CAIXABANK
1271	2022-04-25	3.03	3.075	3.01	3.023	3.023	1614984	CAIXABANK
1272	2022-04-26	3.057	3.087	2.992	2.995	2.995	1462801	CAIXABANK
1273	2017-04-27	14.14534	14.21675	14.04911	14.162632	13.849058	780724	CELLEX TELECOM
1274	2017-04-28	14.094567	14.20542	14.090901	14.300507	14.08342	1708516	CELLEX TELECOM
1275	2017-05-02	14.348963	14.578344	14.278472	14.520854	14.300404	1595118	CELLEX TELECOM
1276	2017-05-03	14.510126	14.591365	14.423901	14.582551	14.361184	4044978	CELLEX TELECOM
1277	2017-05-04	14.600178	14.776456	14.45475	14.77456	14.552143	932171	CELLEX TELECOM
1278	2017-05-05	14.723573	14.855781	14.613399	14.844967	14.621585	933093	CELLEX TELECOM
1279	2017-05-08	14.948326	15.027951	14.794084	14.807304	14.582525	1022499	CELLEX TELECOM
1280	2017-05-09	14.807304	14.855781	14.666282	14.802897	14.609971	686329	CELLEX 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
1270	2022-04-22	3.1	3.17	3.002	3.09	3.09	10115198	CAIXABANK
1271	2022-04-25	3.03	3.075	3.01	3.023	3.023	16149884	CAIXABANK
1272	2022-04-26	3.057	3.087	2.992	2.995	2.995	10623801	CAIXABANK
1273	2017-04-27	14.145354	14.216775	14.049111	14.162632	13.949058	786724	CELLNEX TELECOM
1280	2017-04-28	14.094567	14.237542	14.096991	14.309207	14.09342	1708516	CELLNEX TELECOM
1281	2017-05-02	14.348963	14.578344	14.278472	14.520854	14.300404	1595118	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	932171	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	15.027951	14.794084	14.807304	14.582525	1022499	CELLNEX TELECOM
1286	2017-05-09	14.807304	14.855781	14.666282	14.802897	14.609971	686329	CELLNEX TELECOM

ibexunido.csv



COLUMNAS

APERTURA

CIERRE

MÍNIMO

MÁXIMO

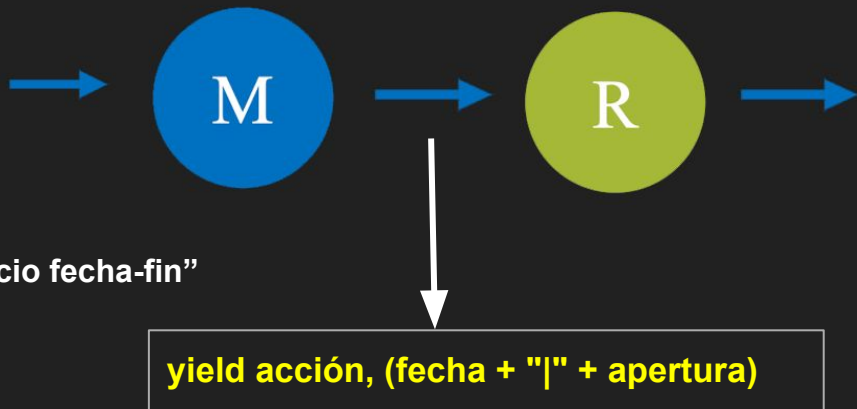
Script: /stocks/scripts/mensual.py **Origen:** HDFS(/ibex35/ibexunido.csv) **Destino:** pantalla

Ejercicio 3: porcentaje de incremento y decremento entre fechas

	DATE	OPEN	HIGH	LOW	CLOSE	MEAN	WMA	NAME
1770	2022-04-22	3.1	3.17	3.062	3.08	3.08	3.081158	CANABANK
1771	2022-04-25	3.03	3.075	3.01	3.023	3.023	3.023458	CANABANK
1772	2022-04-26	3.067	3.067	2.992	2.995	2.995	2.992811	CANABANK
1773	2022-04-27	3.146264	3.421075	3.140811	3.162512	3.160958	3.162724	CELINEX TELECOM
1774	2022-04-28	3.104057	3.425241	3.101051	3.101057	3.101042	3.101051	CELINEX TELECOM
1775	2022-05-02	3.140063	3.157044	3.127072	3.132054	3.130742	3.130118	CELINEX TELECOM
1776	2022-05-03	3.1503216	3.1591385	3.1421801	3.1542531	3.1531384	3.1546078	CELINEX TELECOM
1777	2022-05-04	3.1600176	3.1774656	3.145475	3.1734956	3.1532143	3.172771	CELINEX TELECOM
1778	2022-05-05	3.1723573	3.1855781	3.1613189	3.1848967	3.1821585	3.183993	CELINEX TELECOM
1779	2022-05-06	3.1848326	3.1877651	3.1794384	3.1807304	3.1830252	3.1821499	CELINEX TELECOM
1780	2022-05-09	3.1807504	3.1855781	3.1668101	3.1801087	3.1805971	3.180105	CELINEX 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	ME close	VOLUME	NAME
1773	2022-04-22	3.11	3.17	3.062	3.09	3.09	10115108	CANABANK
1772	2022-04-25	3.03	3.075	3.01	3.023	3.023	16145884	CANABANK
1771	2022-04-26	3.067	3.067	2.992	2.995	2.995	24028811	CANABANK
1770	2022-04-27	34.146264	34.216775	34.048331	34.062332	33.849658	788724	CELINEX TELECOM
1769	2022-04-28	34.094507	34.202541	34.034951	34.103807	34.083842	1706535	CELINEX TELECOM
1768	2022-05-02	34.348983	34.578144	34.278472	34.520854	34.387424	1585118	CELINEX TELECOM
1767	2022-05-03	34.503216	34.591385	34.423801	34.542531	34.361384	4046078	CELINEX TELECOM
1766	2022-05-04	34.600176	34.776456	34.454375	34.754456	34.552343	932771	CELINEX TELECOM
1765	2022-05-05	34.723573	34.855781	34.613389	34.848967	34.621585	933693	CELINEX TELECOM
1764	2022-05-06	34.948326	35.027651	34.794384	34.807304	34.582525	1021499	CELINEX TELECOM
1763	2022-05-09	34.807504	34.853783	34.646332	34.808387	34.659571	686105	CELINEX 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

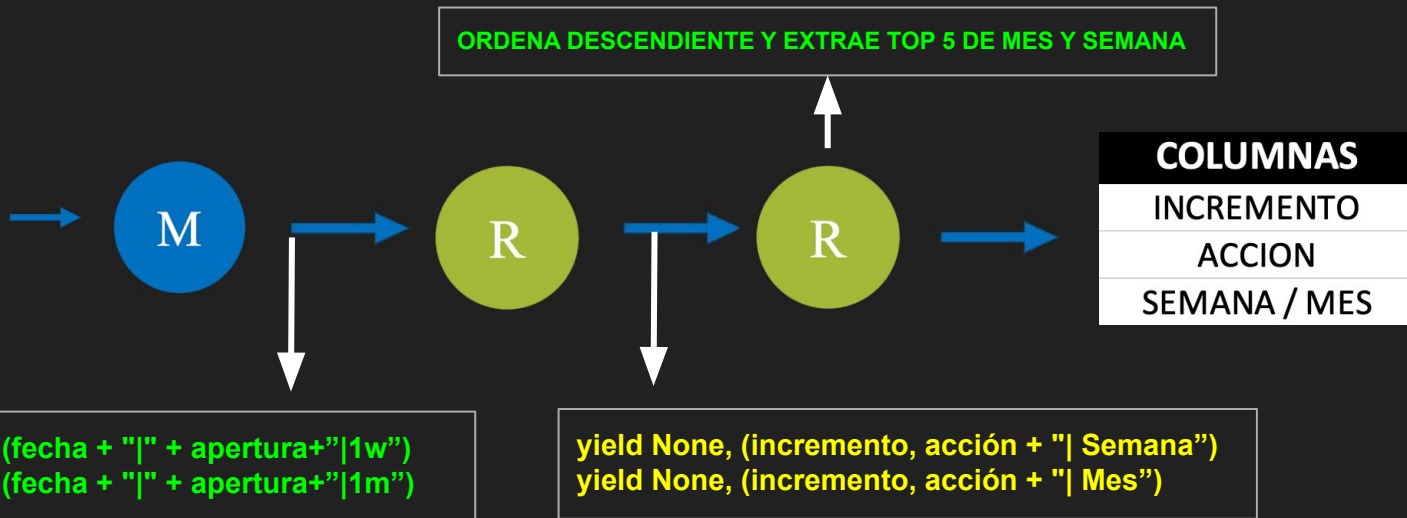
MÁXIMO 1 MES

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

#	DATE	OPEN	HIGH	LOW	CLOSE	ME close	VOLUME	NAME
1773	2022-04-22	3.1	3.17	3.042	3.08	3.08	10115108	CANABANK
1772	2022-04-25	3.03	3.075	3.01	3.023	3.023	16145684	CANABANK
1771	2022-04-26	3.067	3.067	2.992	2.995	2.995	24028811	CANABANK
1770	2022-04-27	34.146264	34.216775	34.048331	34.062532	33.849658	786724	CELINEX TELECOM
1769	2022-04-28	34.094507	34.252541	34.034051	34.101007	34.08942	17065135	CELINEX TELECOM
1768	2022-05-02	34.348983	34.578144	34.278472	34.530854	34.307424	1585138	CELINEX TELECOM
1767	2022-05-03	34.503216	34.591385	34.423801	34.542531	34.361384	4046078	CELINEX TELECOM
1766	2022-05-04	34.600176	34.776456	34.454375	34.759456	34.552343	932771	CELINEX TELECOM
1765	2022-05-05	34.723573	34.855781	34.613389	34.848967	34.621585	938993	CELINEX TELECOM
1764	2022-05-06	34.948326	35.027651	34.794384	34.807304	34.582525	1021499	CELINEX TELECOM
1763	2022-05-09	34.807504	34.855781	34.646332	34.680897	34.659591	980105	CELINEX TELECOM

ibexunido.csv



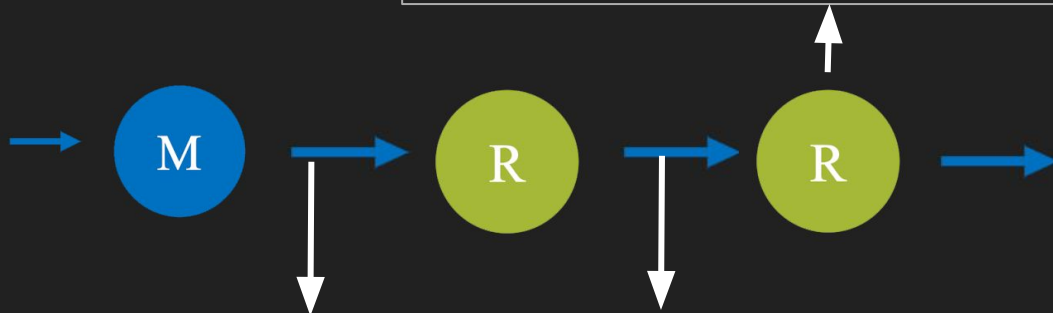
Script: /stocks/scripts/cinco.py **Origen:** HDFS(/ibex35/ibexunido.csv) **Destino:** pantalla

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

#	DATE	OPEN	HIGH	LOW	CLOSE	ME close	VOLUME	NAME
1773	2022-04-22	3.11	3.17	3.062	3.09	3.09	10115108	CANABANK
1772	2022-04-25	3.03	3.075	3.01	3.023	3.023	16145684	CANABANK
1771	2022-04-26	3.067	3.067	2.992	2.995	2.995	24028611	CANABANK
1770	2022-04-27	34.146264	34.216775	34.048311	34.062532	33.849658	786724	CELINEX TELECOM
1769	2022-04-28	34.094507	34.215241	34.034051	34.101037	34.08942	1786515	CELINEX TELECOM
1768	2022-05-02	34.340963	34.576144	34.278472	34.530854	34.307424	1585118	CELINEX TELECOM
1767	2022-05-03	34.503216	34.591385	34.423801	34.542531	34.361384	4046078	CELINEX TELECOM
1766	2022-05-04	34.600176	34.776456	34.65475	34.779456	34.552143	932771	CELINEX TELECOM
1765	2022-05-05	34.723573	34.855781	34.613389	34.848967	34.621585	939693	CELINEX TELECOM
1764	2022-05-06	34.948326	35.027651	34.794384	34.807304	34.582525	1021499	CELINEX TELECOM
1763	2022-05-09	34.807504	34.855781	34.646331	34.818887	34.659591	980105	CELINEX TELECOM

ibexunido.csv

ORDENA DESCENDENTE Y EXTRAE TOP 5 DE MES Y SEMANA



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

yield None, (Decremento, acción + "|" Semana")
yield None, (Decremento, acción + "|" Mes")

COLUMNS

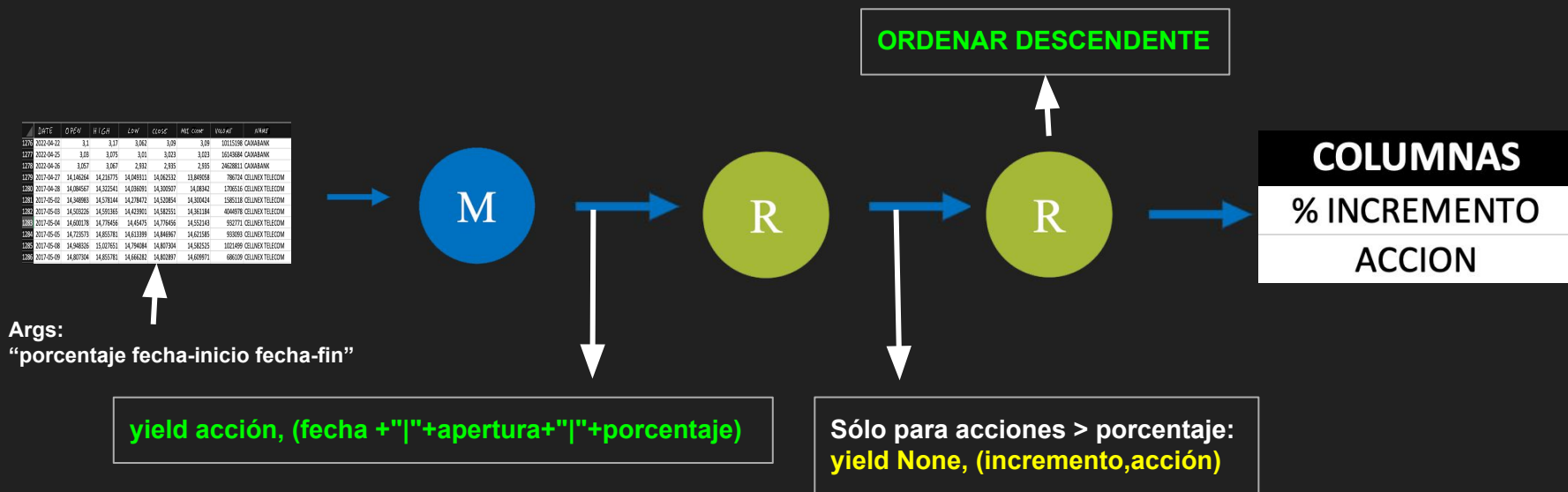
% DECREMENTO

ACCION

SEMANA / 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 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

```
2022-04-21,Reuters,"Crypto exchange Binance curbs services in Russia"
2022-04-21,Reuters,"China, South Korea protest over Japanese PM's of
2022-04-20,Reuters,"Russia to update its strategy in World Trade Orga
2022-04-20,Reuters,"Ukraine seeks Mariupol evacuation talks after sur
2022-04-20,Reuters,"EU preparing measures to prevent Russia from ev
```

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