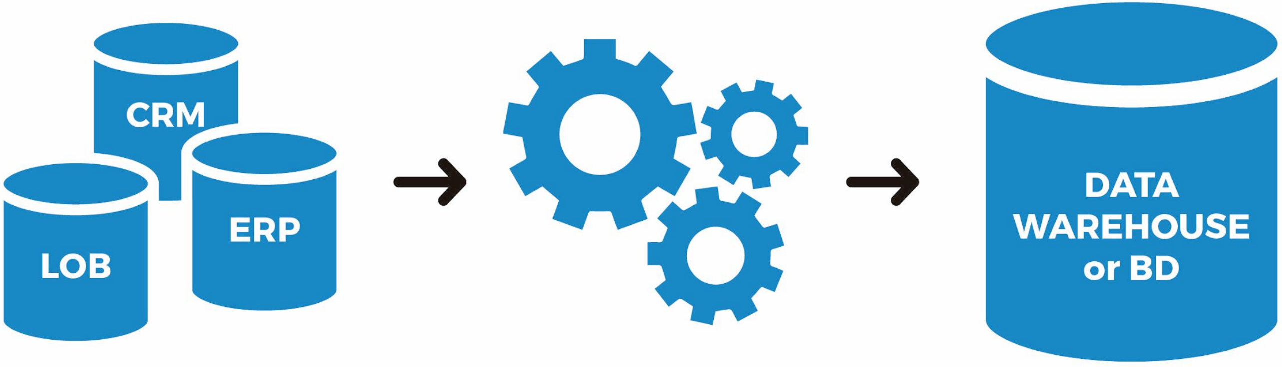
Universidad
Católica del Norte

Proyecto Analítico de Datos “Movie Data”

Johan Ordenes
Leonel Villagra

Movie Data



Export

“Lectura de los datos entregados”

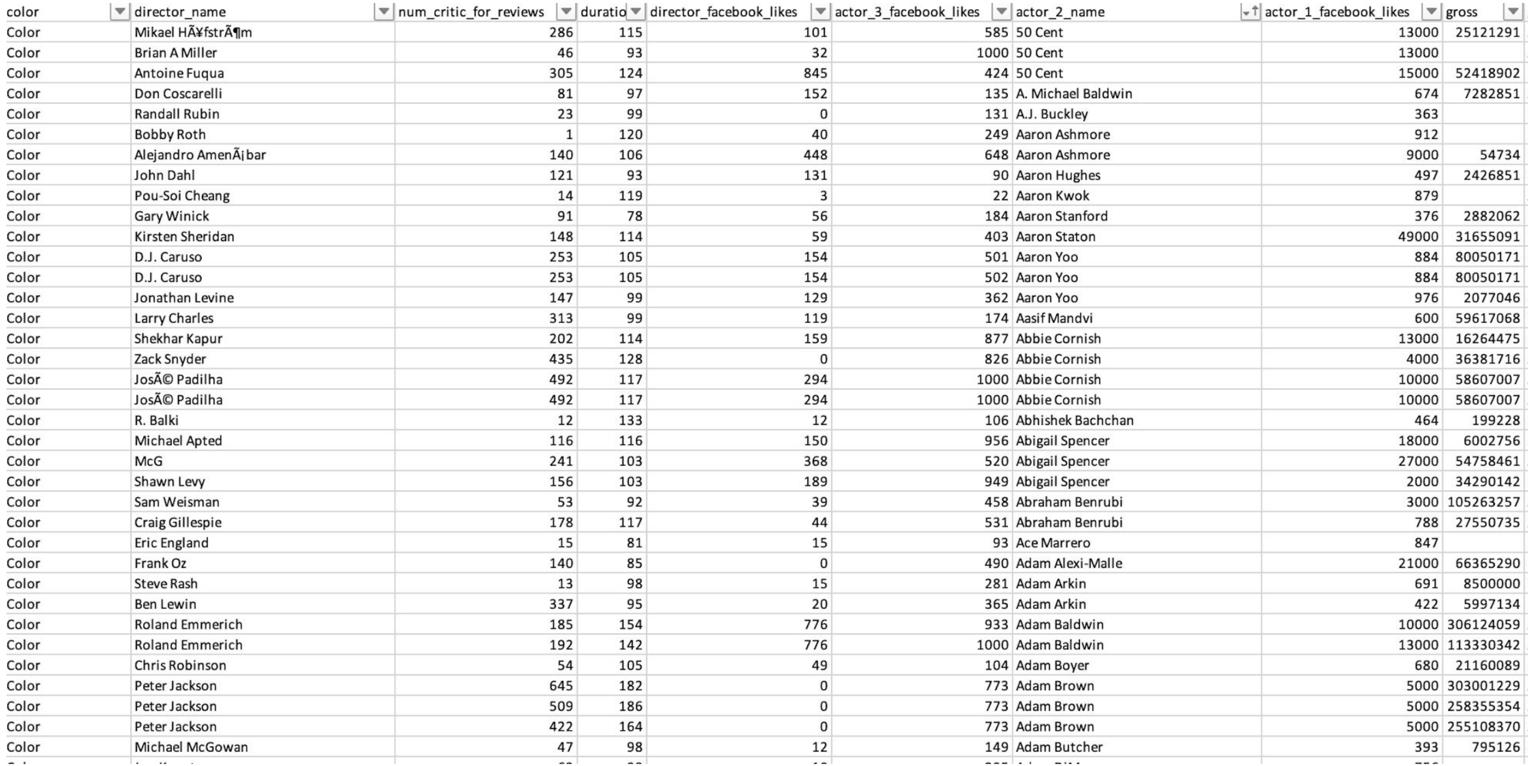
Transform

“Limpieza y eliminación de los datos inservibles”

Load

“Almacenamiento de los datos”

Lectura de archivo .xlsx



color	director_name	num_critic_for_reviews	duratio	director_facebook_likes	actor_3_facebook_likes	actor_2_name	actor_1_facebook_likes	gross
Color	Mikael HÃ¥fstrÃ¶m	286	115	101	585	50 Cent	13000	25121291
Color	Brian A Miller	46	93	32	1000	50 Cent	13000	
Color	Antoine Fuqua	305	124	845	424	50 Cent	15000	52418902
Color	Don Coscarelli	81	97	152	135	A. Michael Baldwin	674	7282851
Color	Randall Rubin	23	99	0	131	A.J. Buckley	363	
Color	Bobby Roth	1	120	40	249	Aaron Ashmore	912	
Color	Alejandro AmenÃ¡bar	140	106	448	648	Aaron Ashmore	9000	54734
Color	John Dahl	121	93	131	90	Aaron Hughes	497	2426851
Color	Pou-Soi Cheang	14	119	3	22	Aaron Kwok	879	
Color	Gary Winick	91	78	56	184	Aaron Stanford	376	2882062
Color	Kirsten Sheridan	148	114	59	403	Aaron Staton	49000	31655091
Color	D.J. Caruso	253	105	154	501	Aaron Yoo	884	80050171
Color	D.J. Caruso	253	105	154	502	Aaron Yoo	884	80050171
Color	Jonathan Levine	147	99	129	362	Aaron Yoo	976	2077046
Color	Larry Charles	313	99	119	174	Aasif Mandvi	600	59617068
Color	Shekhar Kapur	202	114	159	877	Abbie Cornish	13000	16264475
Color	Zack Snyder	435	128	0	826	Abbie Cornish	4000	36381716
Color	JosÃ© Padilha	492	117	294	1000	Abbie Cornish	10000	58607007
Color	JosÃ© Padilha	492	117	294	1000	Abbie Cornish	10000	58607007
Color	R. Balki	12	133	12	106	Abhishek Bachchan	464	199228
Color	Michael Apted	116	116	150	956	Abigail Spencer	18000	6002756
Color	McG	241	103	368	520	Abigail Spencer	27000	54758461
Color	Shawn Levy	156	103	189	949	Abigail Spencer	2000	34290142
Color	Sam Weisman	53	92	39	458	Abraham Benrubi	3000	105263257
Color	Craig Gillespie	178	117	44	531	Abraham Benrubi	788	27550735
Color	Eric England	15	81	15	93	Ace Marrero	847	
Color	Frank Oz	140	85	0	490	Adam Alexi-Malle	21000	66365290
Color	Steve Rash	13	98	15	281	Adam Arkin	691	8500000
Color	Ben Lewin	337	95	20	365	Adam Arkin	422	5997134
Color	Roland Emmerich	185	154	776	933	Adam Baldwin	10000	306124059
Color	Roland Emmerich	192	142	776	1000	Adam Baldwin	13000	113330342
Color	Chris Robinson	54	105	49	104	Adam Boyer	680	21160089
Color	Peter Jackson	645	182	0	773	Adam Brown	5000	303001229
Color	Peter Jackson	509	186	0	773	Adam Brown	5000	258355354
Color	Peter Jackson	422	164	0	773	Adam Brown	5000	255108370
Color	Michael McGowan	47	98	12	149	Adam Butcher	393	795126

Limpieza de los datos

Para celdas vacías o erróneas:

- Datos tipo texto no conocidos como “UNKNOWN”
- Datos tipo número no conocidos como “0”

Para celdas ocupadas:

- Forman parte de datos que no se conocen “?,-1,0” (no se eliminaron)

Funciones:

- cleanColor(): Limpia los datos de la columna “color”
- cleanName(): Limpia los datos de las columnas tipo texto
- cleanMovie(): Limpia los datos de la columna “title_movie”
- cleanDigit(): Limpia los datos de las columnas que contienen dígitos enteros
- cleanDecimal(): Limpia los datos de las columnas que contienen dígitos flotantes

Limpieza con funciones

cleanColor():

- #ERROR
- nan
- lstrip()

Color
#ERROR
N/A
Black and White

ENTRADA

Color
UNKNOWN
UNKNOWN
Black and White

SALIDA

cleanName():

- #NODATA
- nan
- ?
- lstrip()

input.encode("latin-1", 'ignore').decode("utf-8", 'ignore').replace(u'\xa0', u").strip()

M. Night Shyamalan
Mikael HÃ¥fstrÃ¶m
Alejandro AmenÃ¡bar
JosÃ© Padilha

ENTRADA

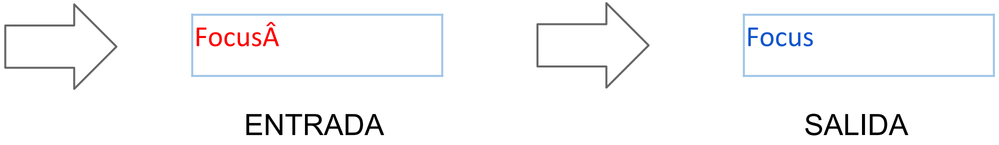
M. Night Shyamalan
Mikael HÃ¥fstrÃ¶m
Alejandro AmenÃ¡bar
JosÃ© Padilha

SALIDA

Limpieza con funciones

cleanMovie():

- input[:-1]



FocusÂ

ENTRADA

Focus

SALIDA

cleanDigit():

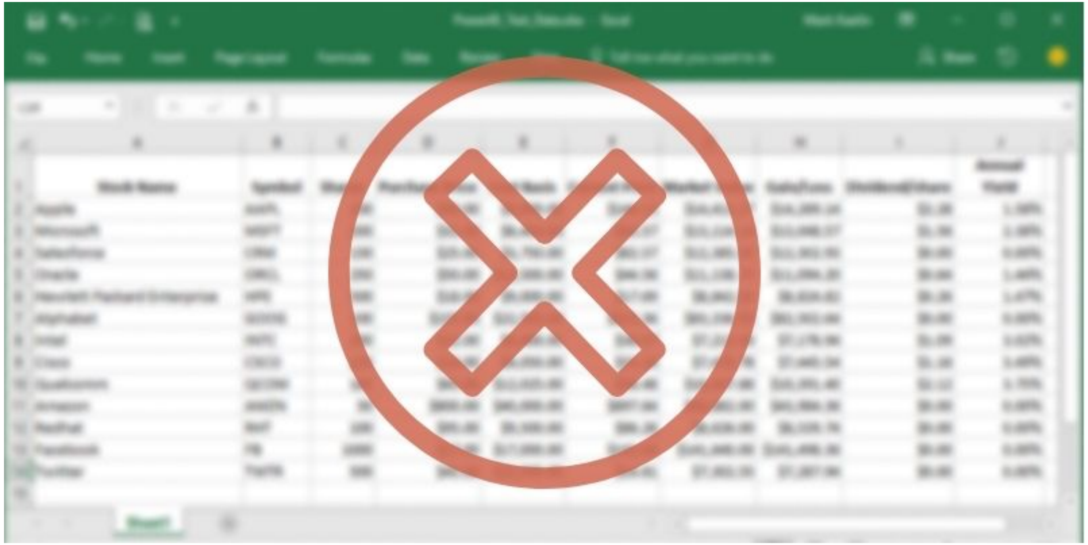
- nan
- #NODATA
- N/I

cleanDecimal():

- nan

Columnas sin limpiar

- genres
- num_voted_users
- cast_total_facebook_likes
- movie_imdb_link
- language
- country
- content_rating
- budget
- title_year
- imdb_score
- movie_facebook_likes

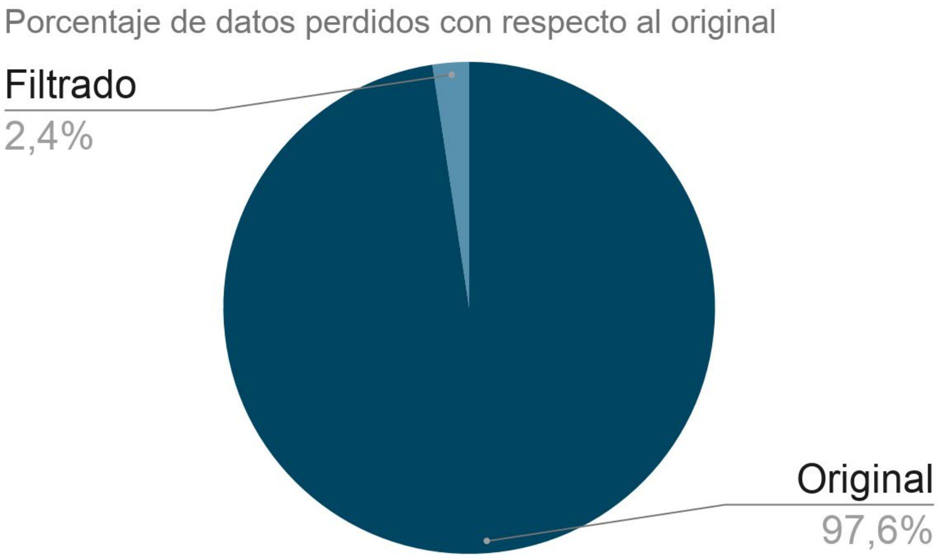


Análisis de datos

movie_title	num_voted_users	cast_total_facebook_likes
DisturbiaÂ	186984	2287
DisturbiaÂ	186982	2288
The WacknessÂ	27266	2748
The DictatorÂ	213863	1375
Elizabeth: The Golden AgeÂ	54787	16899
Sucker PunchÂ	197584	7067
RoboCopÂ	182899	14161
RoboCopÂ	182910	14160

Cálculo de valor promedio y fusión de columnas en relación a la película

Reducción de datos

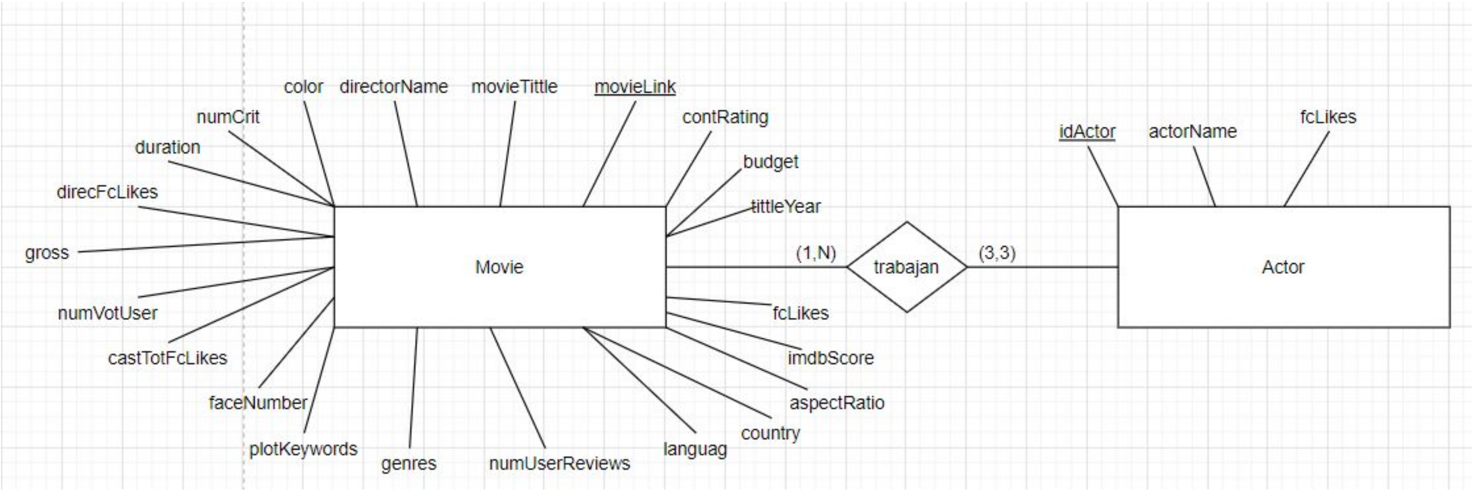


Cantidad de datos

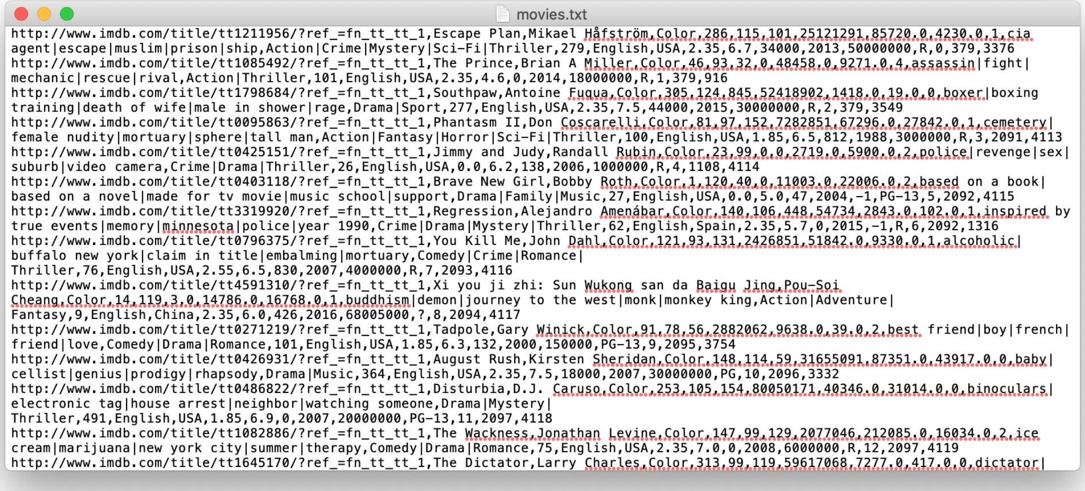
- Originales: 5043
- Filtrados: 4919

Elección de tablas

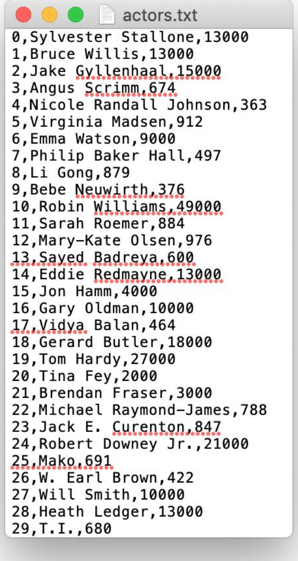
- Movie: Link IMDB de la película como clave primaria
- Actor: Claves foráneas de Movie



Exportación de archivos



```
http://www.imdb.com/title/tt1211956/?ref_=fn_tt_tt_1,Escape Plan,Mikael Håfström,Color,286,115,101,25121291,85720.0,4230.0,1,cia
agent|escape|muslim|prison|ship,Action|Crime|Mystery|Sci-Fi|Thriller,279,English,USA,2.35,6.7,34000,2013,50000000,R,0,379,3376
http://www.imdb.com/title/tt1085492/?ref_=fn_tt_tt_1,The Prince,Brian A Miller,Color,46,93,32,0,48458.0,9271.0,4,assassin|fight|
mechanic|rescue|rival,Action|Thriller,101,English,USA,2.35,4.6,0,2014,18000000,R,1,379,916
http://www.imdb.com/title/tt1798684/?ref_=fn_tt_tt_1,Southpaw,Antoine Fuqua,Color,305,124,845,52418902,1418.0,19.0,0,boxer|boxing
training|death of wife|male in shower|rage,Drama|Sport,277,English,USA,2.35,7.5,44000,2015,30000000,R,2,379,3549
http://www.imdb.com/title/tt0095863/?ref_=fn_tt_tt_1,Phantasm II,Don Coscarelli,Color,81,97,152,7282851,67296.0,27842.0,1,cemetery|
female nudity|mortuary|sphere|tall man,Action|Fantasy|Horror|Sci-Fi|Thriller,100,English,USA,1.85,6.5,812,1988,3000000,R,3,2091,4113
http://www.imdb.com/title/tt0425151/?ref_=fn_tt_tt_1,Jimmy and Judy,Randall Rubin,Color,23,99,0,0,2719.0,5900.0,2,police|revenge|sex|
suburb|video camera,Crime|Drama|Thriller,26,English,USA,0.0,6.2,138,2006,1000000,R,4,1108,4114
http://www.imdb.com/title/tt0403118/?ref_=fn_tt_tt_1,Brave New Girl,Bobby Roth,Color,1,120,40,0,11003.0,22006.0,2,based on a book|
based on a novel|made for tv movie|music school|support,Drama|Family|Music,27,English,USA,0.0,5.0,47,2004,-1,PG-13,5,2092,4115
http://www.imdb.com/title/tt3319920/?ref_=fn_tt_tt_1,Regression,Alejandro Amenábar,Color,140,106,448,54734,2843.0,102.0,1,inspired by
true events|memory|minnesota|police|year 1990,Crime|Drama|Mystery|Thriller,62,English,Spain,2.35,5.7,0,2015,-1,R,6,2092,1316
http://www.imdb.com/title/tt0796375/?ref_=fn_tt_tt_1,You Kill Me,John Dahl,Color,121,93,131,2426851,51842.0,9330.0,1,alcoholic|
buffalo new york|claim in title|embalming|mortuary,Comedy|Crime|Romance|
Thriller,76,English,USA,2.55,6.5,830,2007,4000000,R,7,2093,4116
http://www.imdb.com/title/tt4591310/?ref_=fn_tt_tt_1,Xi you ji zhi: Sun Wukong san da Baigu Jing,Pou-Soi
Cheang,Color,14,119,3,0,14786.0,16768.0,1,buddhism|demon|journey to the west|monk|monkey king,Action|Adventure|
Fantasy,9,English,China,2.35,6.0,426,2016,68005000,?,8,2094,4117
http://www.imdb.com/title/tt0271219/?ref_=fn_tt_tt_1,Tadpole,Gary Winick,Color,91,78,56,2882062,9638.0,39.0,2,best friend|boy|french|
friend|love,Comedy|Drama|Romance,101,English,USA,1.85,6.3,132,2000,150000,PG-13,9,2095,3754
http://www.imdb.com/title/tt0426931/?ref_=fn_tt_tt_1,August Rush,Kirsten Sheridan,Color,148,114,59,31655091,87351.0,43917.0,0,baby|
cellist|genius|prodigy|rhapsody,Drama|Music,364,English,USA,2.35,7.5,18000,2007,30000000,PG,10,2096,3332
http://www.imdb.com/title/tt0486822/?ref_=fn_tt_tt_1,Disturbia,D.J. Caruso,Color,253,105,154,80050171,40346.0,31014.0,0,binoculars|
electronic tag|house arrest|neighbor|watching someone,Drama|Mystery|
Thriller,491,English,USA,1.85,6.9,0,2007,20000000,PG-13,11,2097,4118
http://www.imdb.com/title/tt1082886/?ref_=fn_tt_tt_1,The Wackness,Jonathan Levine,Color,147,99,129,2077046,212085.0,16034.0,2,ice
cream|marijuana|new york city|summer|therapy,Comedy|Drama|Romance,75,English,USA,2.35,7.0,0,2008,6000000,R,12,2097,4119
http://www.imdb.com/title/tt1645170/?ref_=fn_tt_tt_1,The Dictator,Larry Charles,Color,313,99,119,59617068,7277.0,417.0,0,dictator|
```



```
0,Sylvester Stallone,13000
1,Bruce Willis,13000
2,Jake Gyllenhaal,15000
3,Angus Scrimm,674
4,Nicole Randall Johnson,363
5,Virginia Madsen,912
6,Emma Watson,9000
7,Philip Baker Hall,497
8,Li Gong,879
9,Bebe Neuwirth,376
10,Robin Williams,49000
11,Sarah Roemer,884
12,Mary-Kate Olsen,976
13,Sayed Badreya,600
14,Eddie Redmayne,13000
15,Jon Hamm,4000
16,Gary Oldman,10000
17,Vidya Balan,464
18,Gerard Butler,18000
19,Tom Hardy,27000
20,Tina Fey,2000
21,Brendan Fraser,3000
22,Michael Raymond-James,788
23,Jack E. Curenton,847
24,Robert Downey Jr.,21000
25,Mako,691
26,W. Earl Brown,422
27,Will Smith,10000
28,Heath Ledger,13000
29,T.I.,680
```

Importación de archivos

- Importar y exportar para evitar la esperar a que se procesen los datos mientras hacemos pruebas al programar
- El tiempo mejora con el uso de hilos

psycopg

PostgreSQL