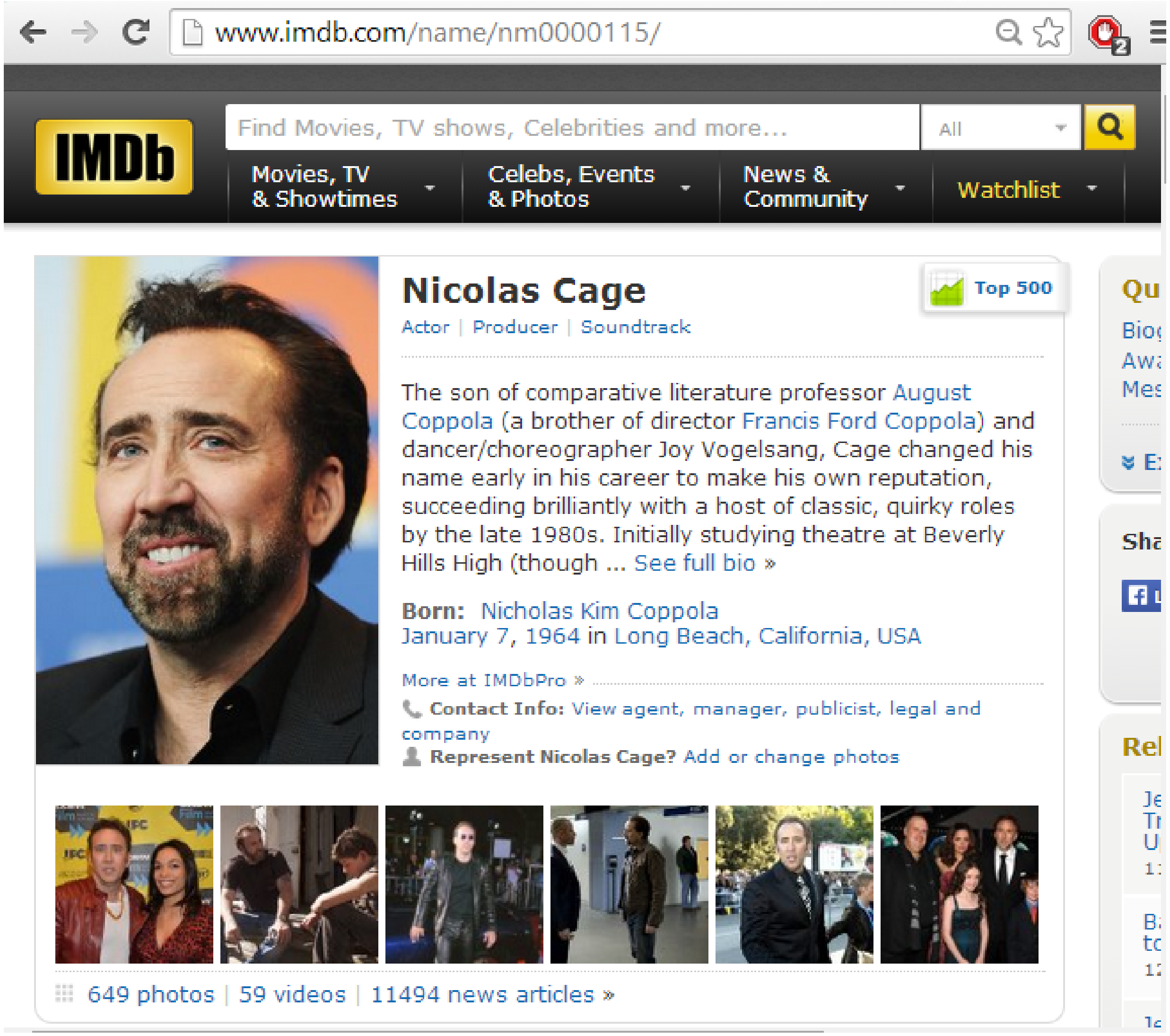
Counting (co-)Stars

Peligro

- Please be careful!

IMDb (I assume you all know?)



Nicolas Cage

Actor | Producer | Soundtrack

The son of comparative literature professor August Coppola (a brother of director Francis Ford Coppola) and dancer/choreographer Joy Vogelsang, Cage changed his name early in his career to make his own reputation, succeeding brilliantly with a host of classic, quirky roles by the late 1980s. Initially studying theatre at Beverly Hills High (though ... See full bio »

Born: Nicholas Kim Coppola
January 7, 1964 in Long Beach, California, USA

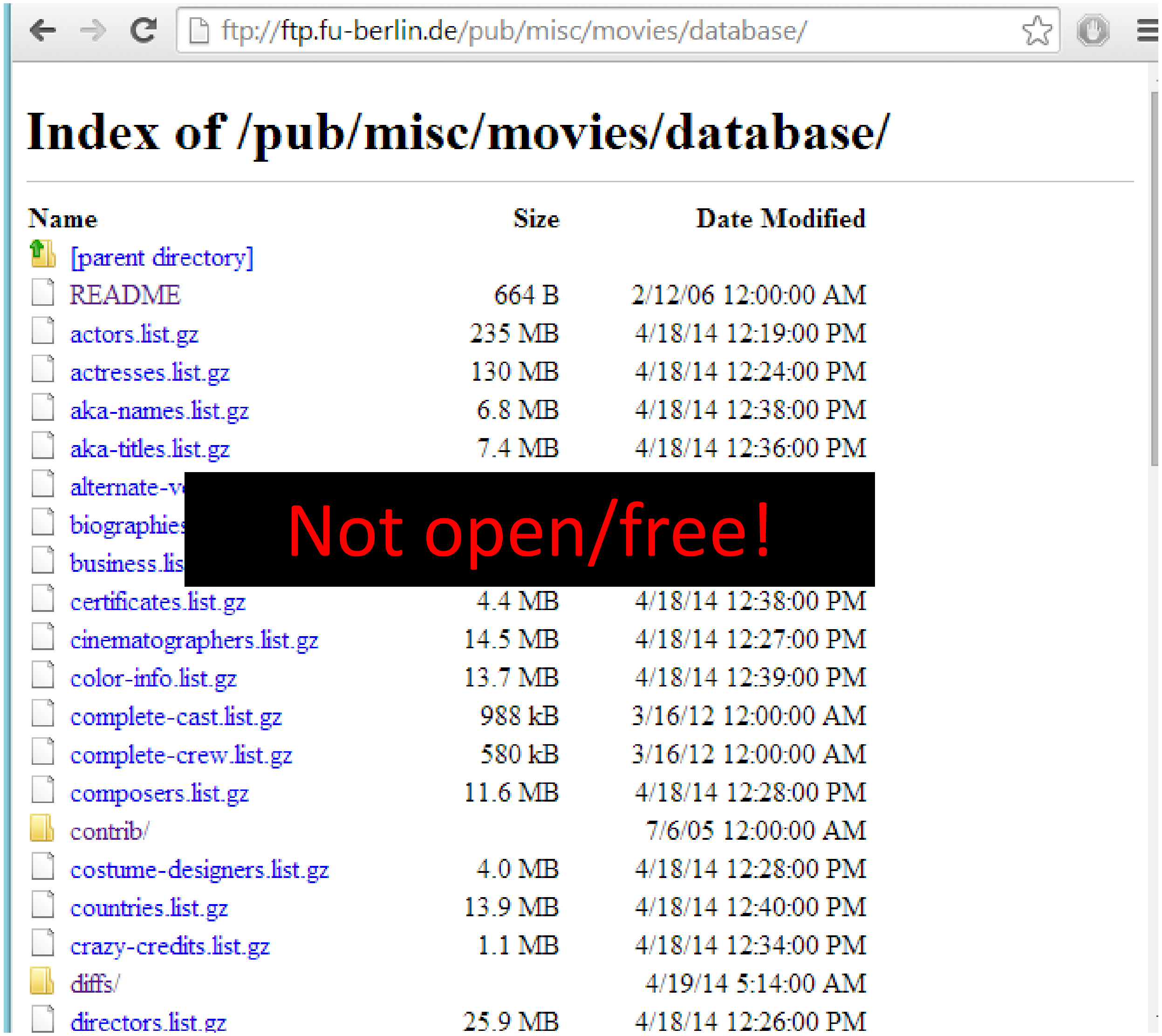
More at IMDbPro »

☎ **Contact Info:** View agent, manager, publicist, legal and company

Represent Nicolas Cage? Add or change photos

649 photos | 59 videos | 11494 news articles »

IMDb Dump



ftp://ftp.fu-berlin.de/pub/misc/movies/database/

Index of /pub/misc/movies/database/

Name	Size	Date Modified
[parent directory]		
README	664 B	2/12/06 12:00:00 AM
actors.list.gz	235 MB	4/18/14 12:19:00 PM
actresses.list.gz	130 MB	4/18/14 12:24:00 PM
aka-names.list.gz	6.8 MB	4/18/14 12:38:00 PM
aka-titles.list.gz	7.4 MB	4/18/14 12:36:00 PM
alternate-v		
biographie		
business.lis		
certificates.list.gz	4.4 MB	4/18/14 12:38:00 PM
cinematographers.list.gz	14.5 MB	4/18/14 12:27:00 PM
color-info.list.gz	13.7 MB	4/18/14 12:39:00 PM
complete-cast.list.gz	988 kB	3/16/12 12:00:00 AM
complete-crew.list.gz	580 kB	3/16/12 12:00:00 AM
composers.list.gz	11.6 MB	4/18/14 12:28:00 PM
contrib/		7/6/05 12:00:00 AM
costume-designers.list.gz	4.0 MB	4/18/14 12:28:00 PM
countries.list.gz	13.9 MB	4/18/14 12:40:00 PM
crazy-credits.list.gz	1.1 MB	4/18/14 12:34:00 PM
diffs/		4/19/14 5:14:00 AM
directors.list.gz	25.9 MB	4/18/14 12:26:00 PM

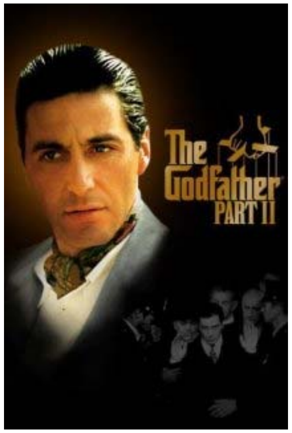
Not open/free!

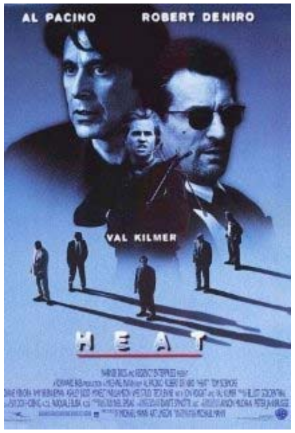
The Question You are Going to Answer …

Which pair of actors/actresses have acted together the most times?

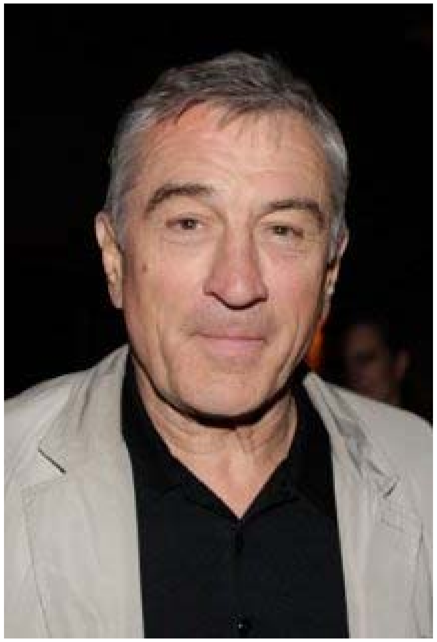
An Example

In how many movies have Al Pacino and Robert De Niro starred together in IMDb?

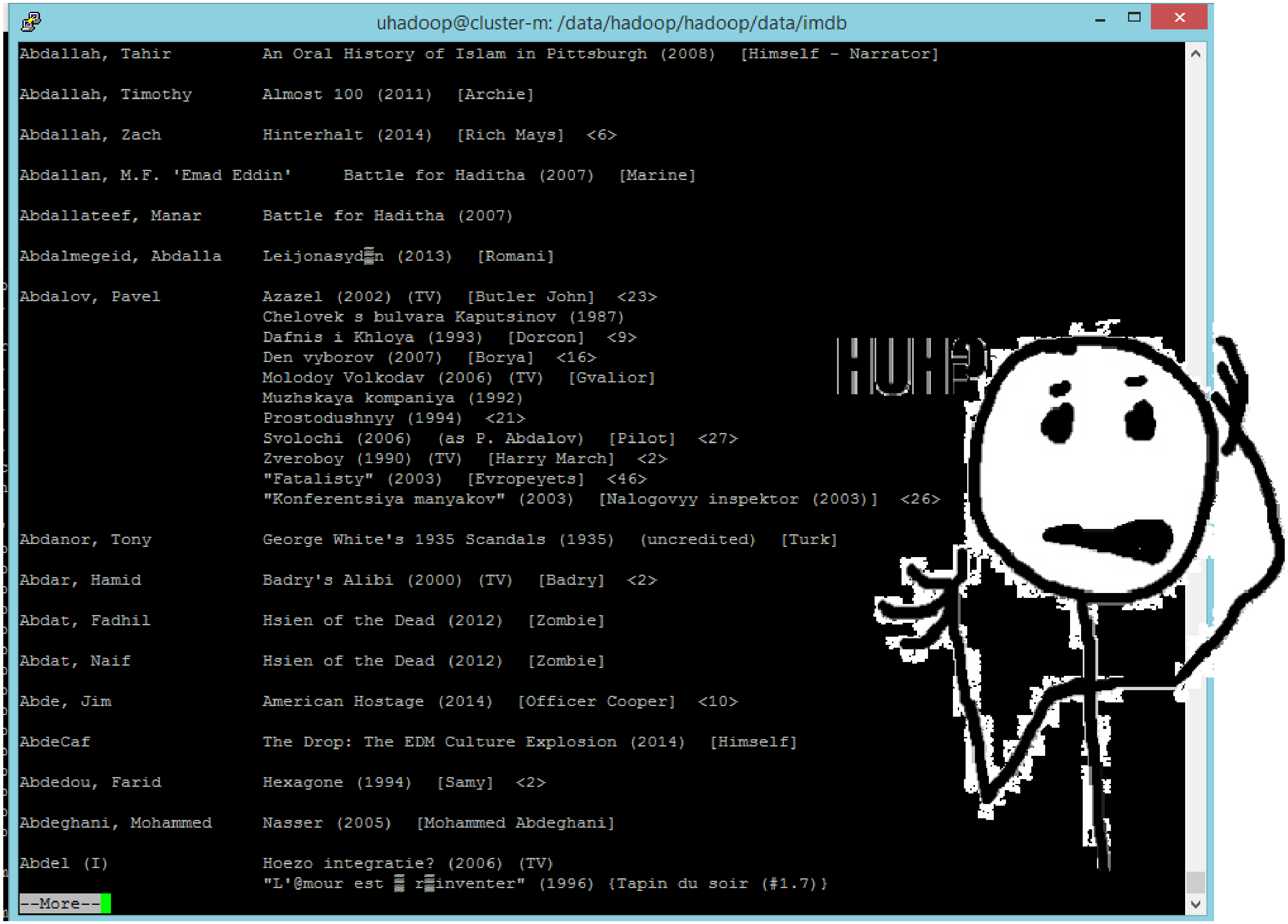




DENIRO
PACINO
Most people respect the badge.
Everybody respects the gun.
RIGHTEOUSKILL

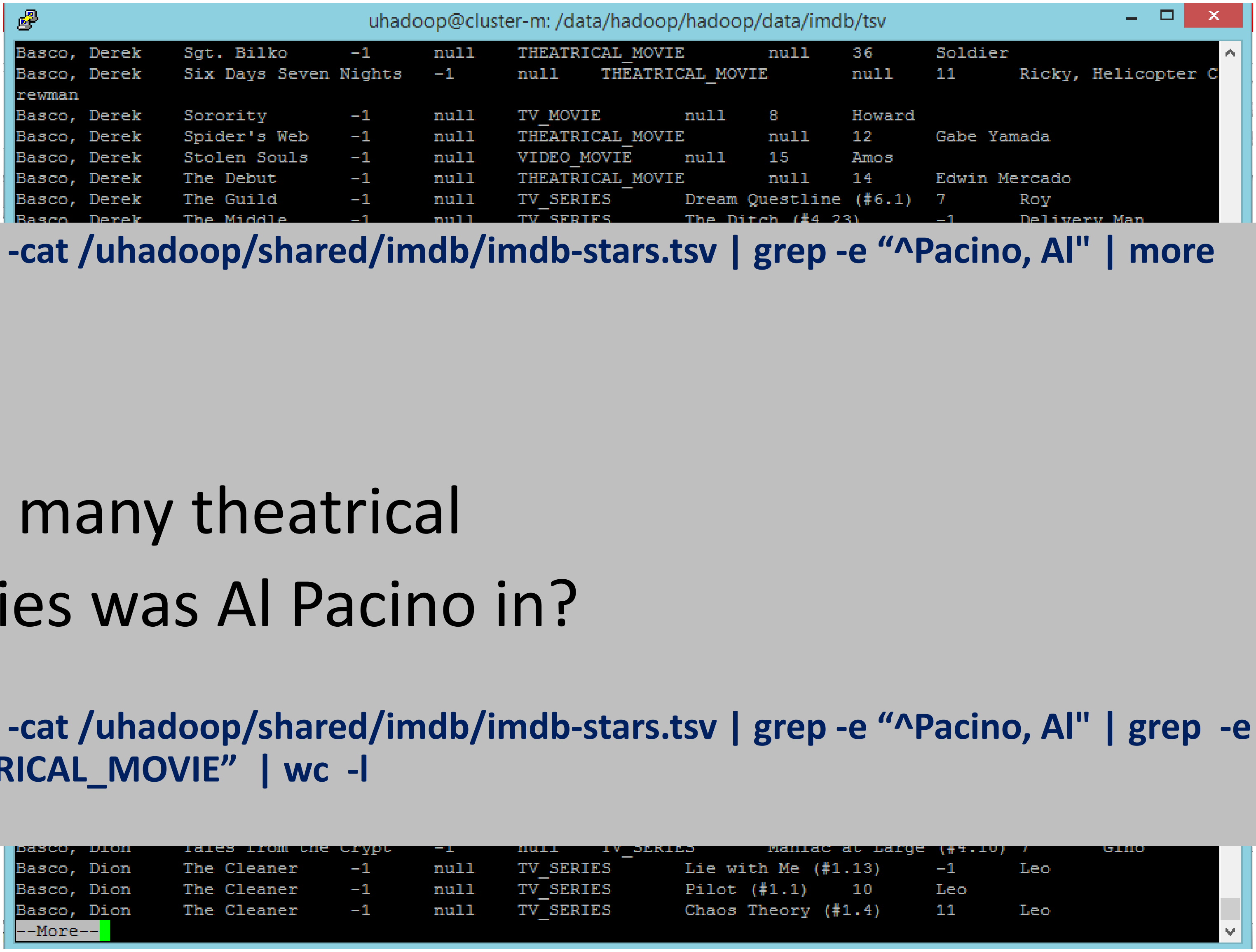


IMDB: Typical File



```
Abdallah, Tahir         An Oral History of Islam in Pittsburgh (2008)  [Himself - Narrator]

Abdallah, Timothy       Almost 100 (2011)  [Archie]

Abdallah, Zach          Hinterhalt (2014)  [Rich Mays]  <6>

Abdallan, M.F. 'Emad Eddin'     Battle for Haditha (2007)  [Marine]

Abdallateef, Manar      Battle for Haditha (2007)

Abdalmegeid, Abdalla    Leijonasyd▒n (2013)  [Romani]

Abdalov, Pavel          Azazel (2002) (TV)  [Butler John]  <23>
                        Chelovek s bulvara Kaputsinov (1987)
                        Dafnis i Khloya (1993)  [Dorcon]  <9>
                        Den vyborov (2007)  [Borya]  <16>
                        Molodoy Volkodav (2006) (TV)  [Gvalior]
                        Muzhskaya kompaniya (1992)
                        Prostodushnyy (1994)  <21>
                        Svolochi (2006)  (as P. Abdalov)  [Pilot]  <27>
                        Zveroboy (1990) (TV)  [Harry March]  <2>
                        "Fatalisty" (2003)  [Evropeyets]  <46>
                        "Konferentsiya manyakov" (2003)  [Nalogovyy inspektor (2003)]  <26>

Abdanor, Tony           George White's 1935 Scandals (1935)  (uncredited)  [Turk]

Abdar, Hamid            Badry's Alibi (2000) (TV)  [Badry]  <2>

Abdat, Fadhil           Hsien of the Dead (2012)  [Zombie]

Abdat, Naif             Hsien of the Dead (2012)  [Zombie]

Abde, Jim               American Hostage (2014)  [Officer Cooper]  <10>

AbdeCaf                 The Drop: The EDM Culture Explosion (2014)  [Himself]

Abdedou, Farid          Hexagone (1994)  [Samy]  <2>

Abdeghani, Mohammed     Nasser (2005)  [Mohammed Abdeghani]

Abdel (I)               Hoezo integratie? (2006) (TV)
                        "L'@mour est ▒ r▒inventer" (1996) {Tapin du soir (#1.7)}
--More--
```

IMDb: Already Parsed



uhadoop@cluster-m: /data/hadoop/hadoop/data/imdb/tsv

```
Basco, Derek    Sgt. Bilko      -1      null    THEATRICAL_MOVIE        null    36      Soldier
Basco, Derek    Six Days Seven Nights   -1      null    THEATRICAL_MOVIE        null    11      Ricky, Helicopter C
rewman
Basco, Derek    Sorority        -1      null    TV_MOVIE        null    8       Howard
Basco, Derek    Spider's Web    -1      null    THEATRICAL_MOVIE        null    12      Gabe Yamada
Basco, Derek    Stolen Souls    -1      null    VIDEO_MOVIE     null    15      Amos
Basco, Derek    The Debut       -1      null    THEATRICAL_MOVIE        null    14      Edwin Mercado
Basco, Derek    The Guild       -1      null    TV_SERIES       Dream Questline (#6.1)  7       Roy
Basco, Derek    The Middle      -1      null    TV_SERIES       The Ditch (#4.23)       -1      Delivery Man
Basco, Dion     Tales from the Crypt    -1      null    TV_SERIES       Maniac at Large (#4.10) 7       Gino
Basco, Dion     The Cleaner     -1      null    TV_SERIES       Lie with Me (#1.13)     -1      Leo
Basco, Dion     The Cleaner     -1      null    TV_SERIES       Pilot (#1.1)    10      Leo
Basco, Dion     The Cleaner     -1      null    TV_SERIES       Chaos Theory (#1.4)     11      Leo
--More--
```

hdfs dfs -cat /uhadoop/shared/imdb/imdb-stars.tsv | grep -e “^Pacino, Al" | more

How many theatrical
movies was Al Pacino in?

hdfs dfs -cat /uhadoop/shared/imdb/imdb-stars.tsv | grep -e “^Pacino, Al" | grep -e “THEATRICAL_MOVIE” | wc -l

The Question You are Going to Answer …

How many times each pair of co-stars has acted together in the IMDb database

… and perhaps if we have time, which pair has acted together the most number of times.

Instructions for DFS/Hadoop, etc.

- Same as last week:
- http://aidanhogan.com/teaching/cc5212-1-2016/lab/04/mdp-lab04.pdf

Instructions for this lab

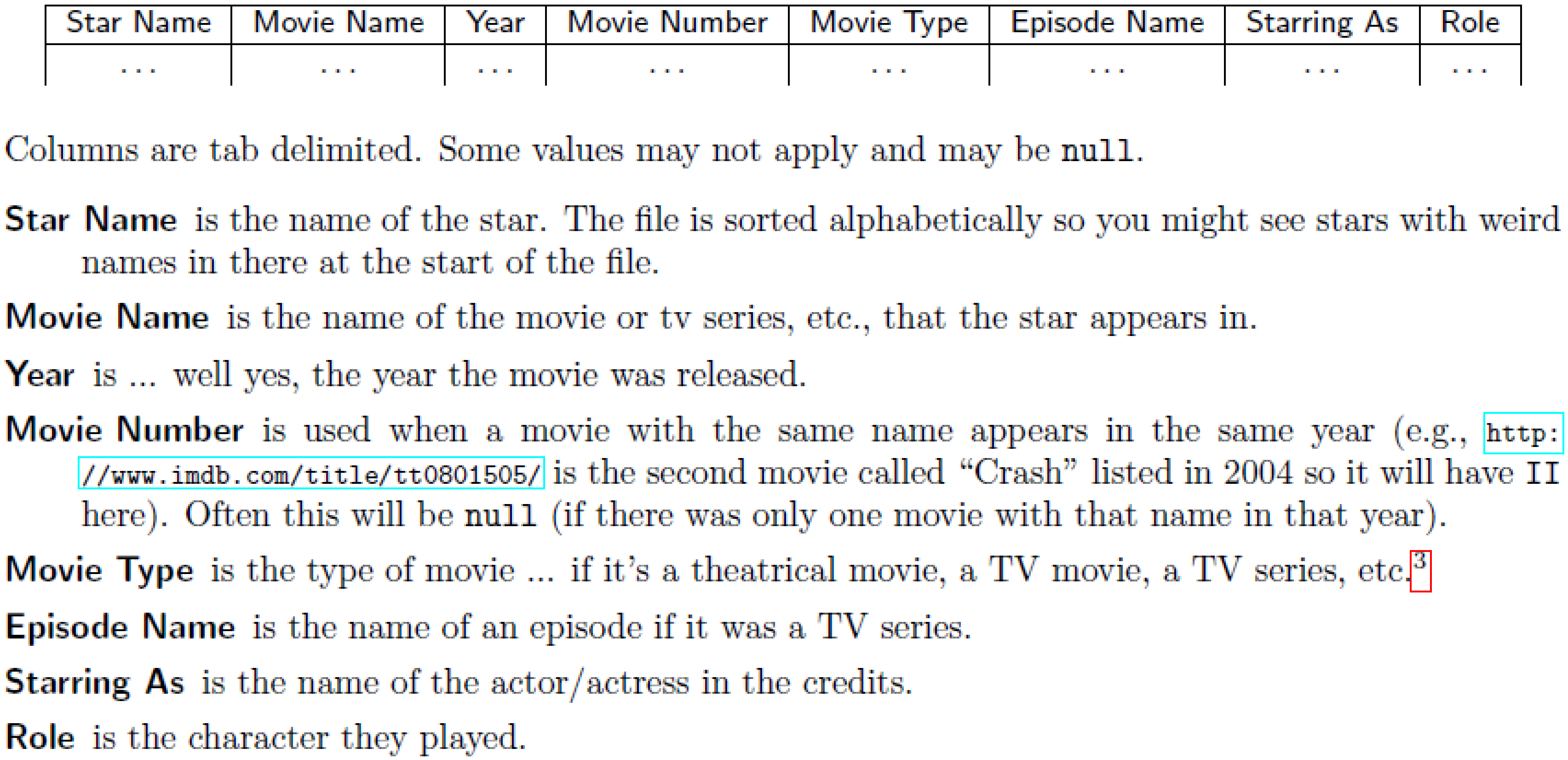
- http://aidanhogan.com/teaching/cc5212-1-2016/lab/05/mdp-lab05.pdf

Download code project:

- http://aidanhogan.com/teaching/cc5212-1-2016/lab/05/mdp-lab05.zip
- You do not need to change Main.java
- CitationCount.java left for reference
  - You can copy and adapt this to suit your needs

How to proceed

- The input format is described in the instructions of the lab



Star Name	Movie Name	Year	Movie Number	Movie Type	Episode Name	Starring As	Role
...	...	...	...	...	...	...	...

Columns are tab delimited. Some values may not apply and may be `null`.

Star Name is the name of the star. The file is sorted alphabetically so you might see stars with weird names in there at the start of the file.

Movie Name is the name of the movie or tv series, etc., that the star appears in.

Year is ... well yes, the year the movie was released.

Movie Number is used when a movie with the same name appears in the same year (e.g., `http://www.imdb.com/title/tt0801505/` is the second movie called "Crash" listed in 2004 so it will have `II` here). Often this will be `null` (if there was only one movie with that name in that year).

Movie Type is the type of movie ... if it's a theatrical movie, a TV movie, a TV series, etc.[3]

Episode Name is the name of an episode if it was a TV series.

Starring As is the name of the actor/actress in the credits.

Role is the character they played.

How to proceed

- Only consider lines where the movie type is the string “THEATRICAL_MOVIE”
- Movies are uniquely identified by Movie Name, Year AND Movie Number (Movie Names alone are not unique)
- You can use “##” to concatenate two things
  - E.g., “Actor1##Actor2”
- Output unique co-stars (n*(n-1)/2)
- You may need multiple Jobs:
  - Feel free to run them manually, passing output of one as input of another

Testing …

- Test each job seperately
- Test over a small file first:
  - /uhadoop/shared/imdb/imdb-stars**-100k**.tsv
- Use the shared input
- Output to a folder in /uhadoop/[username]/imdb/
- hadoop jar mdp-lab5.jar [ClassName] [Input] [Output]
- When your code works you can assume it will work okay for large file ☺
  - The large file might take some time.