Lab 7 – IMDB's Best Stars (with Pig)

CC5212-1

April 20, 2016

Today we will practice a little more with Pig. I will not give detailed hints this time.

- The data you need are on HDFS in the `/uhadoop/shared/imdb/` folder.
  - The `imdb-stars-g.tsv` file is about 1GB and contains 13 million roles; it contains a new column for gender (so do not confuse it with last week's `imdb-stars.tsv`). There is also a smaller file, `imdb-stars-g-100k.tsv` for testing.
  - There is a new file `imdb-reviews.tsv` that contains the IMDb rating and voting record for over 600,000 movies, TV episodes, etc. The first column gives the distribution of votes (we will not use this). The second column gives the number of votes. The third column gives the mean rating. The other columns give the name of the movie/TV series, the year, the number and the episode name (if any); this part is similar to `imdb-stars-g.tsv` (but without information on roles). You can use the full file for testing (and to make sure the join works).
- The goal of the lab is to identify the best actors/actresses in IMDb, which we will define as those who acted in the most GOOD MOVIES.
  - We define a GOOD MOVIE as one with at least ($\geq$) 1000 votes and a score $\geq$ 8.0.
  - The output will be a count of GOOD MOVIEs for actors and actresses in **two separate files**: one for males, one for females. The output should be in descending order of count.
  - Gender is given as `MALE`/`FEMALE` in the `gender` column of the `imdb-stars-g*` inputs.
  - As before, we will only consider entries of type `THEATRICAL_MOVIE`. Again note that in both inputs, `CONCAT(title,'##',year,'##',num)` is required for a unique movie key.
  - An actor/actress may play multiple roles in a movie. We want to make sure to only **count each movie once** per each actor/actress!
  - If an actor/actress does not star in a good movie, **a count of zero should be returned** (rather than omitting the actor/actress from the count).
- Download the code project from `http://aidanhogan.com/teaching/cc5212-1-2016/lab/07/mdp-lab07.zip` and open the script in the text editor of your choice. The inputs are already given for you.
- Check out the user guide at `http://pig.apache.org/docs/r0.14.0/basic.html` and/or last week's slides.
- Test the script over the small files (`100k`) first. You should see Dorothy and Ernest Adams on top with 5 movies each. If not, try debugging by putting a `STORE` after each line and checking the output.
- If it's working, start the script over the full files using `screen` (see last week's instructions: `http://aidanhogan.com/teaching/cc5212-1-2016/lab/06/mdp-lab06.pdf`). It's unlikely to finish before the end of the lab but make sure to come back to check out who the best actors/actresses are!
- Once the script works for the smaller file, feel free to submit to u-cursos as usual.
- OPTIONAL: Or how about worst actors/actresses (most movies < 3.0 for example)? Or more challenging: how about printing the name of the good movies for each actor/actress? ☺