Chapter 5 – Advanced Pig Latin

Exercise

1. Find all movies that received a rating of 5. Use the files u.data and u.item. Only the movie name should be returned (and no duplicates). Store the results in HDFS.

```
ratings = LOAD 'u.data' AS (userid:int, itemid:int, rating:int,
timestamp:int);
only5s = FILTER ratings BY rating == 5;
movies = LOAD 'u.item' USING PigStorage('|') AS
(itemid:int,name:chararray);
joined = JOIN only5s BY itemid, movies BY itemid;
movie_names = FOREACH joined GENERATE name;
results = DISTINCT movie_names;
store results INTO 'good_movies';
```

2. Find the movies with an **average** score greater than 4.5. Hint: use `COGROUP`

```
ratings = LOAD 'u.data' AS (userid:int, itemid:int, rating:int,
timestamp:int);
movies = LOAD 'u.item' USING PigStorage('|') AS
(itemid:int,name:chararray);
grpd = COGROUP ratings BY itemid, movies BY itemid;
averages = FOREACH grpd GENERATE group, movies.name,
AVG(ratings.rating) AS score;
all5s = FILTER averages BY score > 4.5;
store results INTO 'very_good_movies';
```

3. Use `SPLIT` to read the u.item data and output two results. One should contain all the comedy movies (field $10) and the other should be documentaries (field $12).

```
movies = LOAD 'u.item' USING PigStorage('|');
SPLIT movies INTO comedies IF $10 == 1, documentaries IF $12==1;
```

4. Find the movies that were released in 1960. Hint: use `matches`

```
movies = LOAD 'u.item' USING PigStorage('|') AS
(itemid:int,name:chararray,releasedate:chararray);
released_in_1960 = FILTER movies BY releasedate matches '.*1960';
dump released_in_1960;
```