Term Project Recommendation system item-item CF

Group 24 107062321 王劭元

Describe dataset and what I have done:

我們有 610 個 user，以及 9742 部 movie，我們有某些 user 對某些 movie 去做的評分。我們根據這些評分去算出 movie, movie 之間的 similarity。我的版本是，如果有(movie1, movie2) pair 沒有被任何一位 user 共同都有評過分的話，那我的 output 中就不會出現這些(movie1, movie2)的 pair。

Code Explanation:

我總共用到了五個 mapper

- mapper_input：

  在這個 mapper 裡，我們將 rating.csv 讀進來，由於 csv 檔讀進來每個 col 會以”,”隔開，故我也用”,” split 開，我們只會用到 userid, movieid, ratings，他們分別在[0], [1], [2]的位置，我們最後也就回傳(movie, (user, ratings))的 KV pair。

- mapper_cal_mean：

  在這個 mapper 裡，我們都可以拿到對此 movie 評分的所有 user，(movie, [(user1, rating1), (user2, rating2) ….))，故我們可以透過掃過後面的 list，得到這部 movie 的 rating 的 mean。並且可以用得到的 mean 去算出每個 user 對

此 movie 評分去減掉 mean，以及$\sqrt{\sum_{s\in S_{xy}}(r_{xs}-\overline{r_x})^2}$，最後，我就把每個(user, (movie, rating - mean, L2_norm))全部 append 到一個 list 後回傳。並在 main 裡用 flatMap 攤開。

- mapper_pair：

  在這裡面，我們可以拿到一個 user 評分過的所有 movie，也就是上面那個 mapper 做出的資訊，並 groupBykey 後的 list，我們就跑一個雙層迴圈，將每個(movie1, movie2)的可能都算出來，並且根據上面的 mapper 得到的資後，算出他們之間的$\sqrt{(r_{xs}-\overline{r_x})^2}\sqrt{(r_{ys}-\overline{r_y})^2}$以及$(r_{xs}-\overline{r_x})(r_{ys}-\overline{r_y})$，並且將每個 ((movie1, movie2), (長度相乘, 內積)) append 到一個 list，最後回傳此 list，並在 main 用 flatMap 攤開。

- mapper_div：

  在這裡，我們將 mapper_pair 得到的長度以及內積，相除之後相加，由於可能碰到 divide by zero，故我在這種情況我去把它 handle 掉，會直接去做下一個。

- mapper_adjust

  在這個 mapper 裡，我只是要要把 similarity 拿來當 key，並且做 sort，因此在這 mapper 裡，就只是為了調整位置，但這個是 debug 時用到的，最後的輸出中，並沒有用到這邊。

- 統整：

  在一開始將 data 讀進來後，經過 mapper_input 後，可以得到(movie, (user,

rating))的 pair，我們根據 movieid 去 reduceByKey 後，可以得到一個 movie 被評過分的紀錄，然後我會在這邊先 sortByKey 一次，以便我之後 debug。接著再把這個丟到 mapper_cal_mean 後，會得到很多(user, (movie, rating-mean, L2_norm))的 KV pair，接著對這些 pair 做 groupByKey，並 mapValues 成 list。這樣就可以得到一個 user 對所有他評過的 movie 的評分以及一些額外的資訊。

接著丟進 mapper_pair 後，我們就可以拿到(movie, movie)彼此之間的 $\sqrt{(\boldsymbol{r_{xs}}-\overline{\boldsymbol{r_x}})^2}\sqrt{\left(\boldsymbol{r_{ys}}-\overline{\boldsymbol{r_y}}\right)^2}$以及$(\boldsymbol{r_{xs}}-\overline{\boldsymbol{r_x}})\left(\boldsymbol{r_{ys}}-\overline{\boldsymbol{r_y}}\right)$，接著我們為了要 calculate sigma，故我們將(movie1, movie2)當成 key 後做 groupByKey，並 mapValues 成 list。

接著丟進 mapper_div 後，就可以得到(movie1, movie2)之間的 similarity。最後根據(movie1, movie2)去做 SortByKey 後，就可以 output 了。