

Trabajo Práctico 3: Map-Reduce

Martín Arjovsky 683/12
Ezequiel Darío Gambaccini 715/13
Silvio Vileriño 106/12

Julio 2014

1 Encontrar las mejores películas

Se requería obtener la lista de las 12 películas mejor rankeadas con más de 20 reviews. Para esto, lo que decidimos hacer fue que en el map se emitiera el id del producto con un objeto que contenía el puntaje de la review y una variable count inicializada a 1.

En el siguiente paso, reduce, lo que se hace es sumar todos los objetos puntaje, cantidad y emitir el resultado.

Finalmente, en el finalize, se filtra aquellas películas que tienen count ≥ 19 . Para las películas que tienen más de 19 reseñas se devuelve la sumatoria de puntajes dividida la cantidad de reseñas.

Para las que tienen menos de 20 reseñas se devuelve -1.

Luego, usando un script de python para parsear la salida de runner.py, se filtra aquellas películas que su puntaje promedio es distinto de -1, y luego se las ordena en orden decreciente por puntaje, quedándose con las primeras 12.

Las películas que obtuvimos con sus respectivos puntajes y nombres (conseguidos a través del buscador de Amazon) fueron:

- B00000JSJ5, 5.0, All Creatures Great and Small, Series 2: Volumes 1-6
- B00004YKS6, 5.0, Genghis Blues
- B00004YKS7, 5.0, Genghis Blues
- B0002NY7UY, 5.0, Live in Concert (Dion)
- B0007GAEXK, 5.0, The Mole - The Complete First Season
- B0007Z4HAC, 5.0, Salsa Crazy Presents: Learn to Salsa Dance, Intermediate Series, Volume 1
- B000AOEPU2, 5.0, WWE: Bret "Hitman" Hart - The Best There Is, The Best There Was, The Best There Ever Will Be
- B000MCIADA, 5.0, A Reiki 1st, Aura and Chakra Attunement Performed

- B003YBGJ4S, 5.0, WELL worked out with Tannis
- B004LK24BI, 5.0, 50/50 Cardio and Weights with Angie Gorr
- B006JN87UC, 5.0, Transformers: Prime - Season One
- B008COIZHQ, 5.0, Genghis Blues
- B005FY0FPG, 4.987012987012987, Dream With Me in Concert
- B000M7XRC4, 4.977777777777778, The Venture Bros. - Season Two

A continuación se muestra el código de map, reduce y finalize respectivamente.

```

1 function () {
2   data = {};
3   data.score = parseFloat(this.score,10);
4   data.count = 1;
5   emit(this.productId, data);
6 }
7
8 function (key, values) {
9   var res = {};
10  res.score = 0;
11  res.count = 0;
12  for (var i = values.length - 1; i >= 0; i--) {
13    var data = values[i];
14    res.score += data.score;
15    res.count += data.count;
16  };
17  return res;
18 }
19
20 function (key, reducedVal) {
21   if (reducedVal.count > 19) {
22     return reducedVal.score/reducedVal.count;
23   } else{
24     return -1;
25   };
26 }

```

2 Palabras más frecuentes

Para obtener las 5 palabras más usadas de cada puntaje, en la función de map se realizan varias operaciones:

1. Se filtran los caracteres del texto de la review para que queden solo los alfanumericos, y se los pasa a minúscula.
2. Se divide el texto en un array de palabras y luego se aplica una función para filtrar aquellas palabras que están en el array de stop words.

3. Finalmente, se cuentan las apariciones de cada palabra en un objeto, y se emite el valor del puntaje y el objeto que cuenta ocurrencias.

En el reduce, simplemente se genera un nuevo objeto con las ocurrencias de cada palabra de ese puntaje particular.

En el finalize, se invierten los valores del objeto, es decir, se devuelve un objeto que contiene numeros como claves y listas de palabras como valor. Los numeros representan la cantidad de apariciones de las palabras de la lista.

Finalmente, las palabras más usadas para las reviews, divididas por puntaje, ocurrencias, y palabras, es: