Ranking and Prediction of Amazon Fine Food based on Costumer's rating and review

Ng Wing Hin
Lei Wen Fen
Cheng Tin Chu
Chan Hei Nam

Introduction

User's rating and review is one of the explicit ways to determine product's popularity. In this project, we are trying to process the Amazon Fine Food Reviews such that the result can be used to predict the rating based on review.

Basic statistics of the dataset

Basic statistics of the dataset

	Rating	Word(Processed)
Mean	4.18	255
Minimum	1	7
Maximum	5	14425

System Structure

There are mainly three stages in our system :

Stage 1 Pre-processing the dataset.

Stage 2 Using MapReduce, find the first hundred k-shingles, $1 \leq k \leq 5$, of each rating, $1 \leq \text{rating} \leq 5$.

Stage 3 Do prediction of rating based on the review input to the system.

Stage 1 Pre-processing

Algorithm 1: Pre-processing

while *There exists next row inside Reviews.csv* **do**
    Extract **Rating** and **Text**;
    Lower **Text** and remove *some* stopwords and punctuation;
    Returning line with format $Rating, word_1, word_2, ...$;
end

Example result from Stage 1

Original data :

Id, ProductId, UserId, ProfileName, HelpfullnessNumerator, HelpfullnessDenominator, Score, Time, Summary, Text
1,B001E4KFG0,A3SGXH7AUHU8GW,delmartian,1,1,5,1303862400,Good Quality Dog Food,I have bought...
2,B00813GRG4,A1D87F6ZCVE5NK,dll pa,0,0,1,1346976000,Not as Advertised,"Product arrived labeled...
3,B000LQOCH0,ABXLMWJIXXAIN,"Natalia Corres ""Natalia Corres""",1,1,4,1219017600,""Delight"" says it all","This is a confection...
4,B000UA0QIQ,A395BORC6FGVXV,Karl,3,3,2,1307923200,Cough Medicine,If you are looking...

Result data :

Score, Text
5,bought vitality canned dog food products good quality product looks...
1,product arrived labeled jumbo salted peanutsthe peanuts actually...
4,confection centuriesit light pillowy citrus gelatin nutsin case filberts...
2,looking secret ingredient robitussin believe iti got addition root...

Stage 2 MapReduce

Algorithm 2: MapReduce

Mapper :
while *There exists next row inside Preprocessed.csv* **do**
  for $k \leftarrow 1$ **to** 5 **do**
    Find all k-shingles;
    for *Each shingle found* **do**
      Return << Rating, k, shingle >, 1 >;
    end
  end
end
Reducer :
Sum up the value and return
< Rating, k, frequency, shingle >;

Exmaple result from Stage 2

Result from MapReduce :

Score	k	frequency	shingle
3	4	23	food freshly openedpi likes
1	2	238	br tried
2	1	1601	say
3	4	27	coffeetea love organic coffee

Stage 3 Prediction

Algorithm 3: prediction?

Stage 3 Prediction

Weighted formula for calculating prediction rating

Weighted formula

For each k :

t_i = k-shingles

$$\text{Weighted Score} = \frac{\sum_i \sum_j \text{score}_j(t_i) \times \text{freq}_j(t_i)}{\sum_i \sum_j \text{freq}_j(t_i)}$$

where $i = i^{th}$ k-shingle and $1 \leq j \leq 5$.

Regression Model

Regression model

System of linear equation :

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \cdots + \beta_5 x_5 \\ y_2 &= \beta_0 + \beta_1 x_1 + \cdots + \beta_5 x_5 \\ &\vdots \qquad\qquad\qquad \vdots \\ y_n &= \beta_0 + \beta_1 x_1 + \cdots + \beta_5 x_5 \end{aligned} \tag{1}$$

In matrix form :

$$\mathbf{Y} = \boldsymbol{\beta_0} + \boldsymbol{\beta} \cdot \mathbf{X}$$

where
$\mathbf{Y}$ is a $n \times 1$ matrix,
$\boldsymbol{\beta_0}$ is a $n \times 1$ matrix,
$\boldsymbol{\beta}$ is a 5×1 matrix,
$\mathbf{X}$ is a $n \times 5$ matrix.

Regression Model

Complexity

key computation step:

- $\mathbf{X}^T\mathbf{X}$ in $O(n \times k^2)$
- $(\mathbf{X}^T\mathbf{X})^{-1}$ in $O(k^3)$

As $n >> k$, the overall complexity is $O(n)$.

Demostration

Conculsion