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.

There are mainly three stages in our system :

Stage 1 Pro-processing the dataset.

Stage 2 Using MapReduce, Find the first hundred k-shingles , $1 \leq k \leq 5$, of each rating, $0 \leq$ rating ≤ 5.

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

Stage 1 Pre-processing

Pre-processing :
Data: Reviews.csv
Result: Result.csv
while *There exists next row inside Reviews.csv* **do**

  Extract **Rating** and **Text**;
  Lower **Text** and remove *some* stopwords and punctuation;
  Write resulting line to Result.csv 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...