Introduction

In this assignment we work on the online review dataset released by Amazon and experience three analytics phases including summarizing statistics , removing unnecessary data and performing similarity analysis.

The complete dataset we use contains reviews and metadata from Amazon between 1995 and 2015, there are five categories data we can choose one from them which are respectively Music, PC, Video_DVD, Wireless, and Digital_Ebook_Purchase. In this project we analyse data from Music whose features involve customer id, product id,star rating, review id and review body.

Stage 1&2

Stage1:

Implements description：Given questions, I calculate the every reviews by customer_id, and count all of them. Using the dataframe methods to get the results. Unique users or unique can use ‘distinct’ method. TOP N can use ‘filter’ method. Median number method is achieved by custom method. The largest one : after sorting data, the result multiply -1 and the first one is the largest one.

Result：

The total number of reviews	4751191
The number of unique users	1940631
The number of unique products	782281
The largest number of reviews published by a single user	7168
The top 10 users ranked by the number of reviews they publish the median number of reviews published by a user	50736950, 38214553, 51184997, 18116317, 23267387, 50345651, 14539589, 15725862, 19380211, 20018062
The median number of reviews published by a user	1.0
The largest number of reviews written for a single product	2.0
The top 10 products ranked by the number of reviews they have	['B00008OWZG', 'B0000AGWEC', 'B00MIA0KGY', 'B00NEJ7MMI', 'B000089RVX', 'B004EBT5CU', 'B0026P3G12', 'B00009PRZF', 'B00004XONN', 'B00006J6VG']
The median number of reviews a product has	2.0

table 1. stage 1 running result

Satge 1 Performance Analysis：79s.

Stage2:

Implements description：Reviews can be divided into several sentences. After sort and filter, we can get results. Inner-join method can join two tables, we can get sub-results of each table.

Result：

reviews with less than two sentences in the review body.	-
reviews published by users with less than median number of reviews published.	-
reviews from products with less than median number of reviews received	-
top 10 users ranked by median number of sentences in the reviews they have pub- lished;	[(51865782, 440), (25628286, 251), (42072921, 243), (46097534, 228), (37118941, 227), (29580246, 201), (50476169, 200), (21809895, 194), (50595705, 191), (10794201, 185)]
top 10 products ranked by median number of sentences in the reviews they have received;	[('B00LTQ5EVY', 984), ('B00T7TYTCK', 503), ('B009SF2GZU', 321), ('B009SF2IRG', 308), ('B005ZHBBU6', 293), ('B0000020FQ', 270), ('B0002IJNGC', 268), ('B000003G29', 267), ('B000RY431G', 267), ('B00IOQSW7A' 256)] 1238.620638370514

table 2. stage 2 running result

Performance Analysis：1283s

Stage 3&4

stage 3: Details of Design

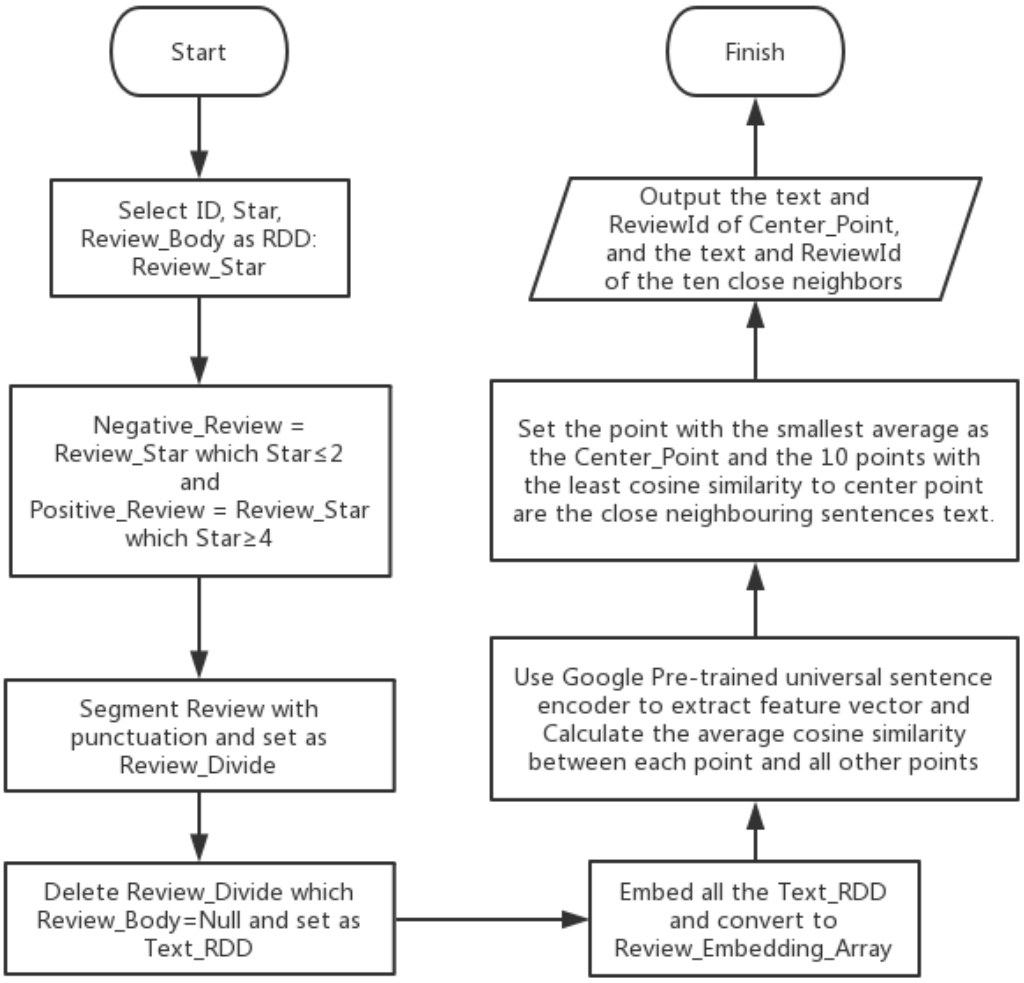
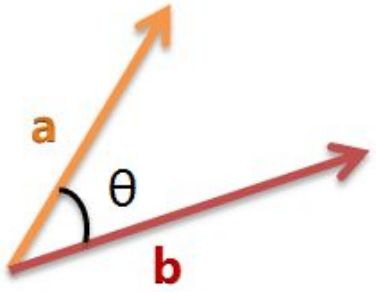


fig 1.Details of design

In this stage, since the text features are sparse features, cosine similarity is used to calculate the similarity between the texts. The degree of similarity of the vectors is judged by the size of the angle. The smaller the angle, the more similar it is.



2-D Law of Cosine

$$\cos\theta = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}}$$

n-D Law of Cosine

$$\cos\theta = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} = \frac{A \cdot B}{|A| \times |B|}$$

performance optimization features

Cache Applied In	Text RDD
Map Partition used In	Word Embedding
Execution Statictics	129.8s

table 3. performance optimization features

Result

Positive:

```
==================================Start==================================
positive center cluster :
---------------------------------------------------------
center review id:R1VR2BOC4IT29H
center sentence:Every song is great!
review id:R2L30H6HKZXUK7
sentence:every song is really good.
review id:R144NZND4C5S5A
sentence:Every single song is GREAT!
review id:R2N0NUDL7GEOCN
sentence:i LOVE every single song.
review id:R22YSMYT6ZBKWH
sentence:all the songs are great.
review id:RK0YKYW86PK2N
sentence:all the songs are great.
review id:R1ZAV0KB4C8FY5
sentence:The songs are all great!
review id:R32U2AEZWXU5X0
sentence:Almost all the songs on here are great.
review id:R71OTQFQVOIVL
sentence:The BEST SONG ON THE ALBLUM IS HOLD ON!
review id:R2H1YNHUCT31TS
sentence:It's got a great variation of songs, and every one of them is worth
listening to.
review id:R1GQBYL3HBOYSZ
sentence:I love every song in this C.D.
==================================End==================================
```

fig2. Stage3_Positive

Negtive:

```
==================================Start=====================================
positive center cluster : ------------------------------------------------------------
center review id:R2EKY5I5KJW2PB
center sentence:6.Girls & Boys-5/5-Awesome song!
review id:R26OJ99Q1J1B6Q
sentence:Riot Girl - 5/510.
review id:R26OJ99Q1J1B6Q
sentence:Hold On - 5/59.
review id:R2W54S6JP18GBZ
sentence:My Bloody Valentine 5/5    8.
review id:RY3PCDQ80U0O8
sentence:Hold On: 5/5.
review id:RY3PCDQ80U0O8
sentence:My Bloody Valentine: 5/5.
review id:RY3PCDQ80U0O8
sentence:Riot Girl: 3/5.
review id:R26OJ99Q1J1B6Q
sentence:My Bloody Valentine - 4/58.
review id:RPRNHKLZG1OEY
sentence:Riot Girl~ 4/5 Pretty fast paced.
review id:R1QDBW2YOP42R4
sentence:My Bloody Valentine  3/5        8.)
review id:R18APJEHVV6NDF
sentence:My Bloody Valentine 8/10                      8.
==================================End=======================================
```

fig3.Stage3_Negtive

Stage 4

In this stage we use word2Vec to convert review sentences into vectors before carrying out the same similarity analysis as stage three.**word2Vec** is a group of related models that are used to produce word embeddings,which takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space.

details of the design

All of the specific stage 4 procedures are as fig4 below :

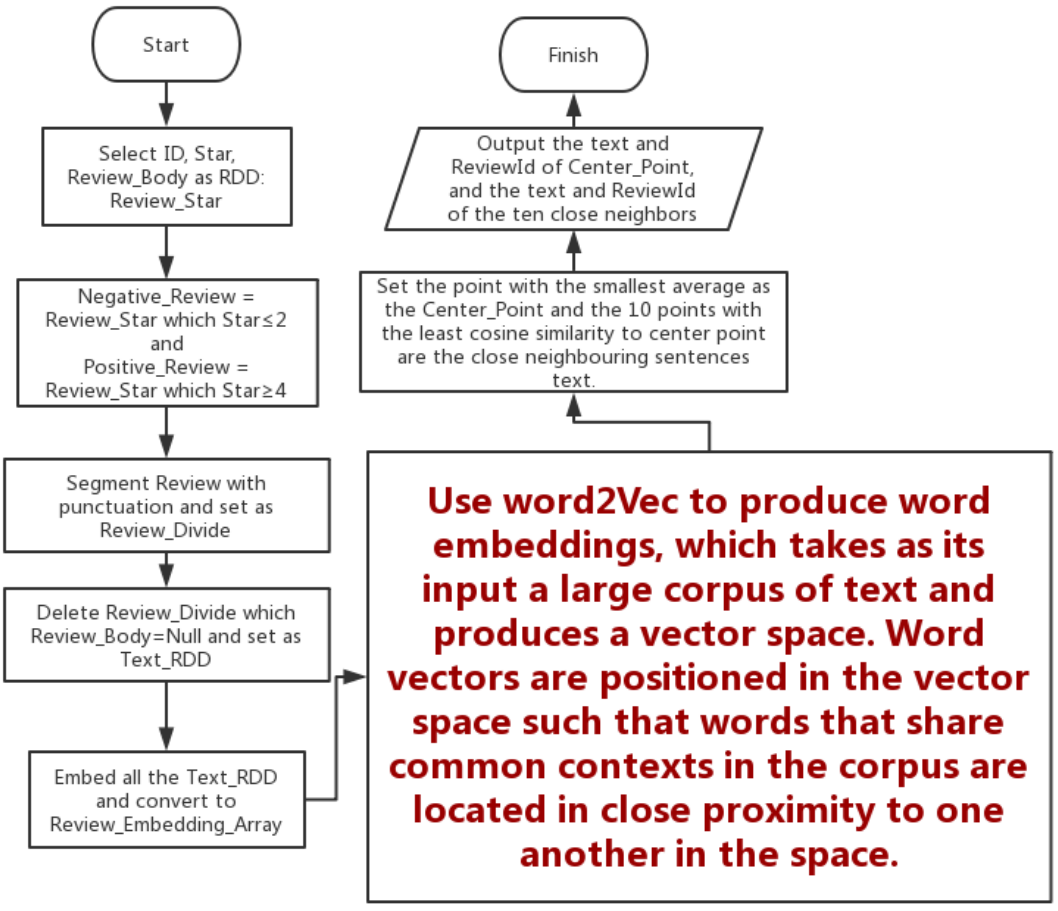


fig4.Word2Vec processing chart

performance analysis

After using word2Vec, this project's runtime decreased dramatically ,which means Spark's word embedding method is more advantageous in the speed of processing big data.

performance optimization features

Cache Applied In	Text RDD
Map Partition used In	word embedding
PCA	21
Normalization	0-1
Execution Statictics	13s

table 3.performance optimization features

The Result

Positive:

```
====================================Start=====================================
positive center cluster : ------------------------------------------------------------
center review id:R2EKY5I5KJW2PB
center sentence:6.Girls & Boys-5/5-Awesome song!
review id:R26OJ99Q1J1B6Q
sentence:Riot Girl - 5/510.
review id:RY3PCDQ80U0O8
sentence:Riot Girl: 3/5.
review id:R26OJ99Q1J1B6Q
sentence:Hold On - 5/59.
review id:R2W54S6JP18GBZ
sentence:My Bloody Valentine 5/5    8.
review id:RY3PCDQ80U0O8
sentence:My Bloody Valentine: 5/5.
review id:R26OJ99Q1J1B6Q
sentence:Emotionless - 3/514.
review id:R26OJ99Q1J1B6Q
sentence:My Bloody Valentine - 4/58.
review id:R26OJ99Q1J1B6Q
sentence:The Anthem - 5/53.
review id:RPRNHKLZG1OEY
sentence:Riot Girl~ 4/5 Pretty fast paced.
review id:RY3PCDQ80U0O8
sentence:Hold On: 5/5.
====================================End======================================
```

fig5. Stage4_Positive

Negtive:

```
====================================Start=====================================
negative center cluster : ------------------------------------------------------------
center review id:R1V5HHKRJNWUQM
center sentence:I still can't believe I saw Benji in good charlotte wear a &quot;Rancid radio
&quot; shirt.
review id:R3H1SO2YOK5GFB
sentence:I guess I could say &quot;punk is dead&quot;, but really it is all about a mindset t
hat many &quot;punk followers&quot; (yes even you hardcore kids that listen to &quot;undergro
und punk bands&quot; like the germs, mc5 and agnostic front) can't even begin to comprehend -
but I'm not saying i can either.
review id:R3JYNCM627LBF2
sentence:No one seems to see that the music that they all play isn't &quot;punk&quot;, it's j
ust &quot;rock n roll&quot;.
review id:R3RHJFESD45TA4
sentence:I'm not a &quot;trust fund kid&quot; and i have a good idea &quot;what punk is&quot;
and it sure as f**k isn't your poison for the ears.
review id:RU7CZWN099THN
sentence:I recomend looking at evanscenes new release it's really good you've probably heard
the song &quot;bring me to life&quot; from the daredevil movie.
review id:R2KN2QAOASMI3U
sentence:&quot;Boys like girls but girls like cars and money.&quot; That's deep.
review id:R2AQNU1AWT0HM5
sentence:I mean if you listen to the track &quot;Girls and Boys&quot; and even look at the vi
deo it's a total mess!
review id:R386QBTULFUGV
sentence:Come on, some idiot said &quot;if you have so much knowledge of punk then why don't
I see your face on my TV everyday?&quot; Well.
review id:R1P6AETM3HJIDO
sentence:I suggest Wakefield's new CD &quot;American Made&quot; Its awesome.
review id:R51IS66I169AG
sentence:I laugh when I see kids turn all &quot;punk&quot; (spike arm bands and ripped pants
don't make anyone punk) and only to find out their favourite band is GC.
review id:RJGXY7ZUOB9WB
sentence:Don't you know that &quot;pop&quot; is short for &quot;popular&quot;?Congratulations
!
====================================End======================================
13.088354587554932
```

fig6.Stage4_Positive

Comparison of Stage Three and Four

Difference	Stage Three	Stage Four
Word Embedding Method	Google Pre-trained universal sentence encoder	word2Vec
Sentence Vetor Dimensions	512 Dimensions	21 Dimensions(PCA)
Normalization	None	0-1
Execution Statictics	129.8s	13s

table 4.Comparison of Stage Three and Four

Appendix

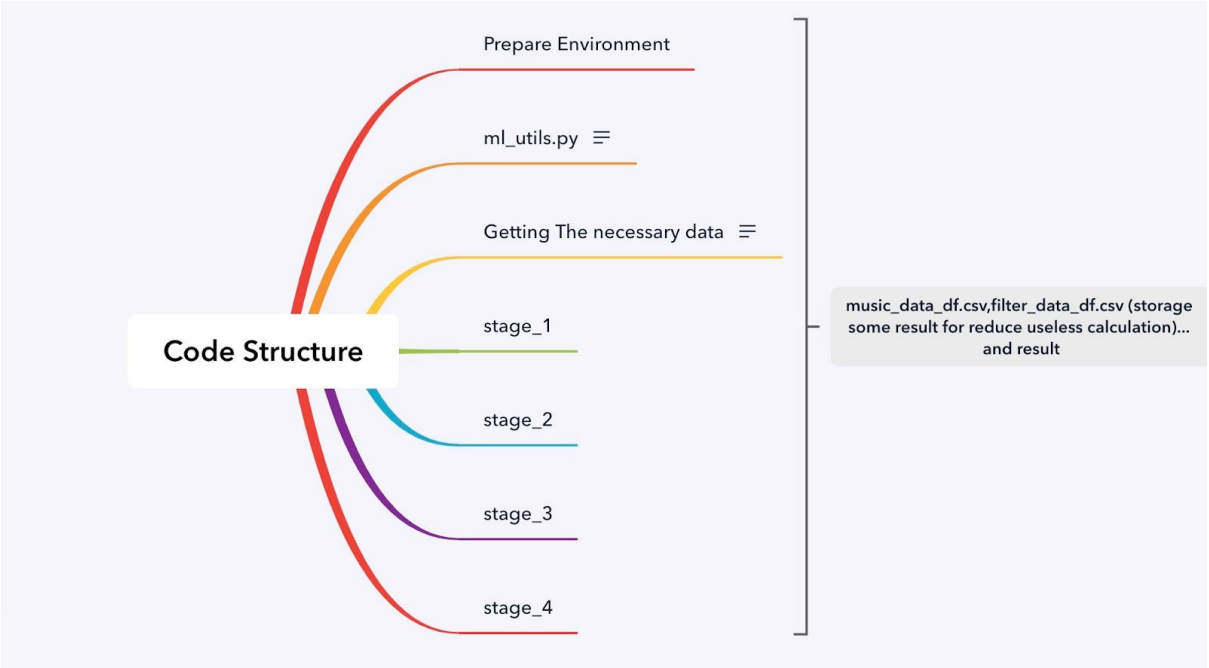


fig7.code frame chart