Big data analytics on Amazon product reviews

Team: Classyfiers
-Satyanarayan Iyengar
-Vaibhav Joshi
-Amritha Venkataramana

Agenda

1) Goal
2) Datasets
3) Data pre-processing
4) Data management in SQL
5) Data mining component
6) Exploratory Analysis and Visualization
7) Results
8) Tools used

Goal

1) To analyse reviews on books purchased on Amazon from datasets obtained via two different sources using data mining and visualization techniques.
2) Why product reviews?
    a) They reveal customer sentiments
    b) Help manufactures decide constraints that could make the product a success.
3) To study and implement industry standard practices for data mining

Datasets and specifications

1) The datasets chosen for the project are
   a) **Stanford Amazon Reviews Dataset-** a collection of customer reviews written in the Amazon.com marketplace. (http://jmcauley.ucsd.edu/data/amazon/links.html)

Specifications:

Data format - JSON

Attributes: "reviewerID" - the id of the reviewer
"asin" - Amazon product ID
"reviewerName" - name of the reviewer
"helpful" - the number of times the review was thought to be helpful
"reviewText" - the content of the review
"overall" - the product rating (from 1 to 5)
"summary" - title of the review
"unixReviewTime" - the time of the review in UNIX format
"reviewTime" - the time of the review

b) **AWS Amazon Customer Reviews Dataset-** This dataset is divided into product reviews dataset and product metadata dataset. The reviews dataset includes ratings, text and helpfulness. The product metadata dataset includes product category, descriptions, price etc.

(https://s3.amazonaws.com/amazon-reviews-pds/readme.html)

Specifications:

Data format - Tab Separated File (.tsv)

Attributes: "marketplace"- 2 letter country code of the marketplace where the review was written.

"Customer_id"- Random identifier that can be used to aggregate reviews written by a single author.

"review_id" - The unique ID of the review.

"product_id" - The unique Product ID

"product_parent" - Random identifier that can be used to aggregate reviews for the same product.

"product_title" - Title of the product.

"product_category" - category of the product

"star_rating" - The rating of the review (from 1-5)

Attributes continued..

"helpful_votes" - Total number of helpful votes of the review
"total_votes" - total votes the review received.
“vine" - Review was written as part of the Vine program.
"verified_purchase" - The review is on a verified purchase.
"review_headline" - The title of the review.
"review_body" - The review text.
"review_date" - The date of the review

Data pre-processing

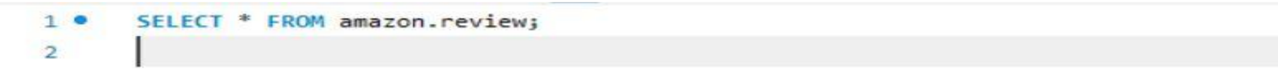
1) Data processing constitutes about 80% of a data mining task. It serves as a basis for a strong analysis.

Pre-processing tasks performed on the datasets:

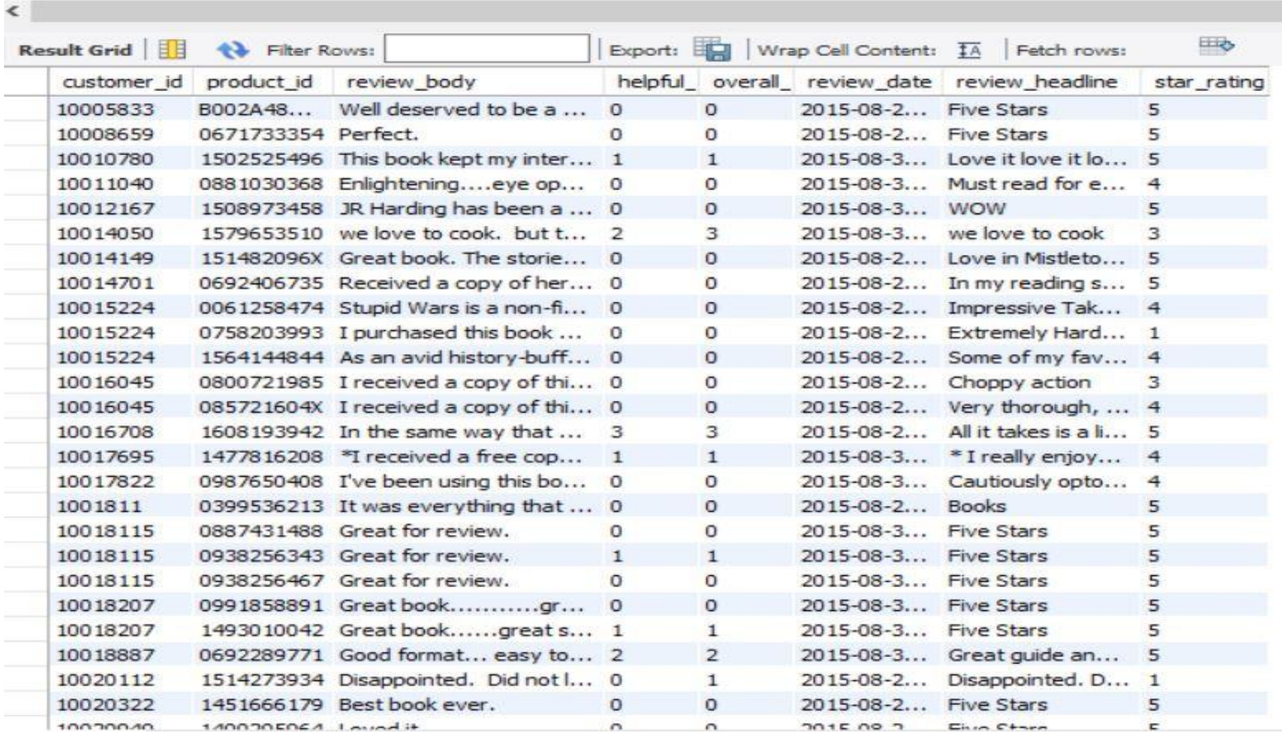
1) Data Loading and Formatting
2) Data Conversion
3) Dropping unnecessary columns/attributes
4) Handling missing values

Data management

1) Data management is useful in storing and querying data as well as keeping the data separate from the analysis. Typically done by database management systems (DBMS)
2) As part of the data management component, a base schema was designed using the attributes from the combined dataset.
3) Data management done in MySQL using SQL Workbench and Python.
4) A representation of the data management component is shown in the figure that follows



```
SELECT * FROM amazon.review;
```



customer_id	product_id	review_body	helpful_	overall_	review_date	review_headline	star_rating
10005833	B002A48...	Well deserved to be a ...	0	0	2015-08-2...	Five Stars	5
10008659	0671733354	Perfect.	0	0	2015-08-2...	Five Stars	5
10010780	1502525496	This book kept my inter...	1	1	2015-08-3...	Love it love it lo...	5
10011040	0881030368	Enlightening.....eye op...	0	0	2015-08-3...	Must read for e...	4
10012167	1508973458	JR Harding has been a ...	0	0	2015-08-3...	WOW	5
10014050	1579653510	we love to cook. but t...	2	3	2015-08-3...	we love to cook	3
10014149	151482096X	Great book. The storie...	0	0	2015-08-2...	Love in Mistleto...	5
10014701	0692406735	Received a copy of her...	0	0	2015-08-2...	In my reading s...	5
10015224	0061258474	Stupid Wars is a non-fi...	0	0	2015-08-2...	Impressive Tak...	4
10015224	0758203993	I purchased this book ...	0	0	2015-08-2...	Extremely Hard...	1
10015224	1564144844	As an avid history-buff...	0	0	2015-08-2...	Some of my fav...	4
10016045	0800721985	I received a copy of thi...	0	0	2015-08-2...	Choppy action	3
10016045	085721604X	I received a copy of thi...	0	0	2015-08-2...	Very thorough, ...	4
10016708	1608193942	In the same way that ...	3	3	2015-08-2...	All it takes is a li...	5
10017695	1477816208	*I received a free cop...	1	1	2015-08-3...	* I really enjoy...	4
10017822	0987650408	I've been using this bo...	0	0	2015-08-3...	Cautiously opto...	4
1001811	0399536213	It was everything that ...	0	0	2015-08-2...	Books	5
10018115	0887431488	Great for review.	0	0	2015-08-3...	Five Stars	5
10018115	0938256343	Great for review.	1	1	2015-08-3...	Five Stars	5
10018115	0938256467	Great for review.	0	0	2015-08-3...	Five Stars	5
10018207	0991858891	Great book...........gr...	0	0	2015-08-3...	Five Stars	5
10018207	1493010042	Great book......great s...	1	1	2015-08-3...	Five Stars	5
10018887	0692289771	Good format... easy to...	2	2	2015-08-3...	Great guide an...	5
10020112	1514273934	Disappointed. Did not l...	0	1	2015-08-2...	Disappointed. D...	1
10020322	1451666179	Best book ever.	0	0	2015-08-2...	Five Stars	5
[illegible]	[illegible]	Loved it	0	0	[illegible]	Five Stars	5

Data Mining

1) Why data mining?

   -> Data management is useful in web-applications and query-based environments. It can execute complex queries however it cannot yield insights and it is difficult to perform visualizations. Thus, data mining is needed for predicting, modeling and visualizing data.

2) Customer reviews can be mined to generate trends as well analyse past history to improve future recommendations.

3) Cross Industry Standard Process for DataMining (or CRISPDM) is the most popular technique for Data mining tasks. It consists of the following steps:

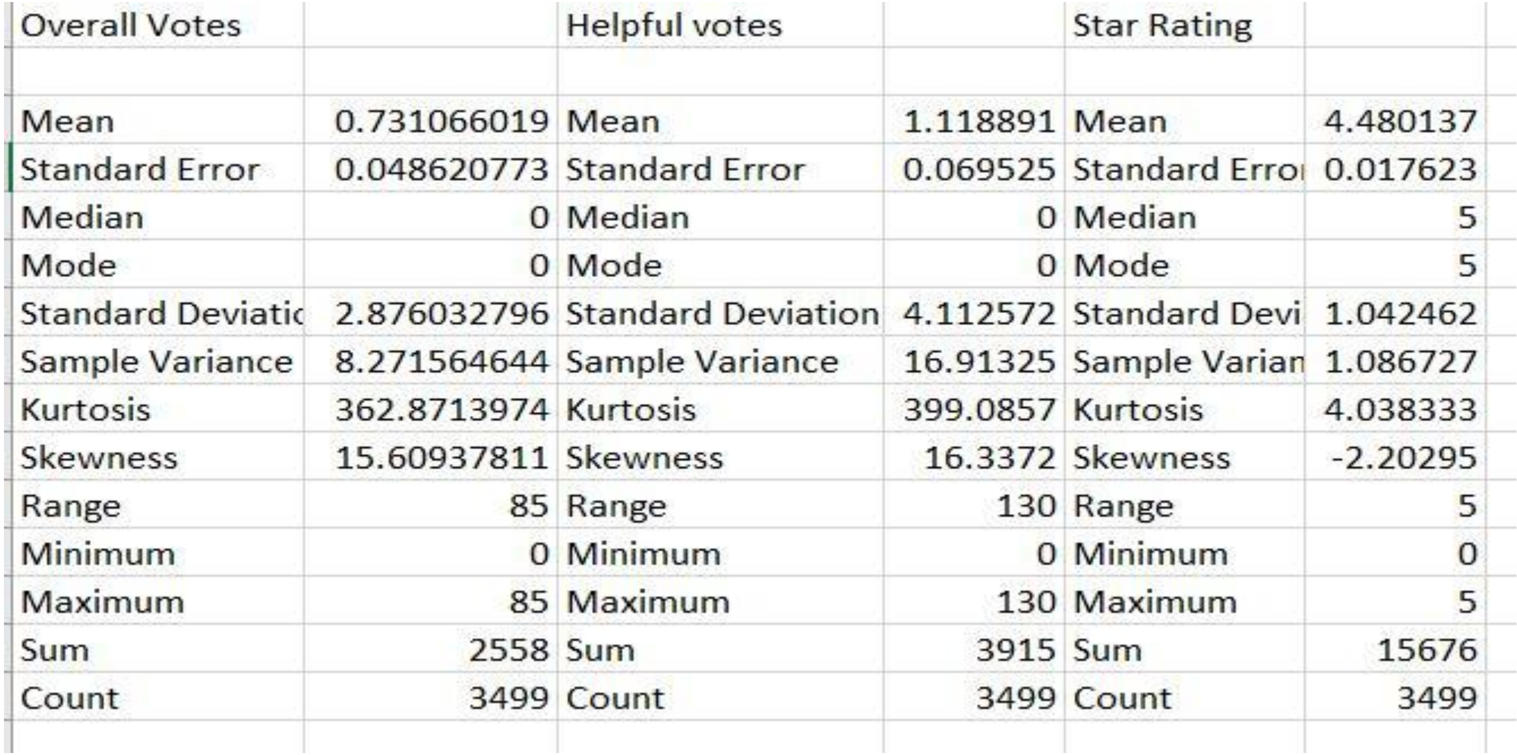
   a) Business Understanding
   b) Data Understanding
   c) Data Preparation
   d) Data Modeling
   e) Data Evaluation
   f) Deployment

Exploratory analysis and Visualization

We have performed visualizations in Tableau to explore relationship between attributes as well as determine timelines and trends in the attributes.

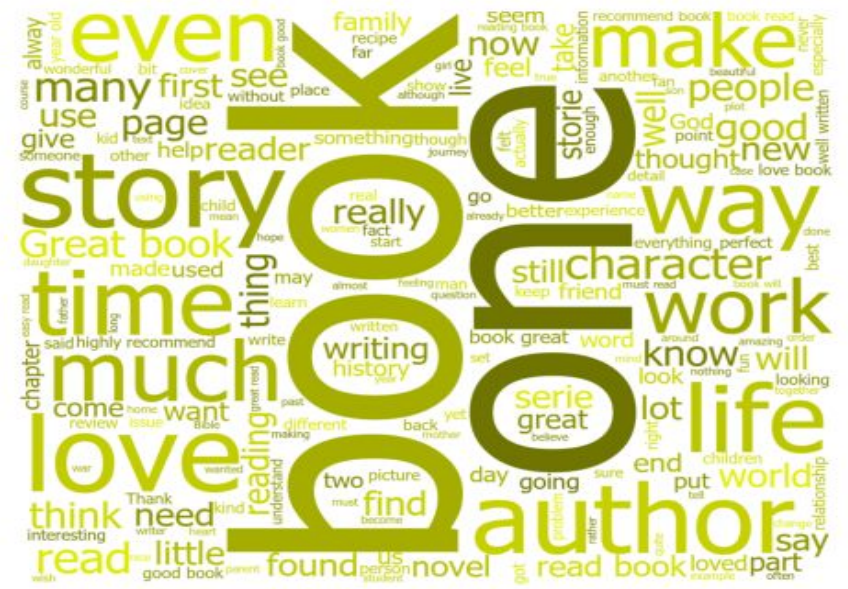
The visualizations follow in the next slides.

Descriptive Statistics for numeric attributes

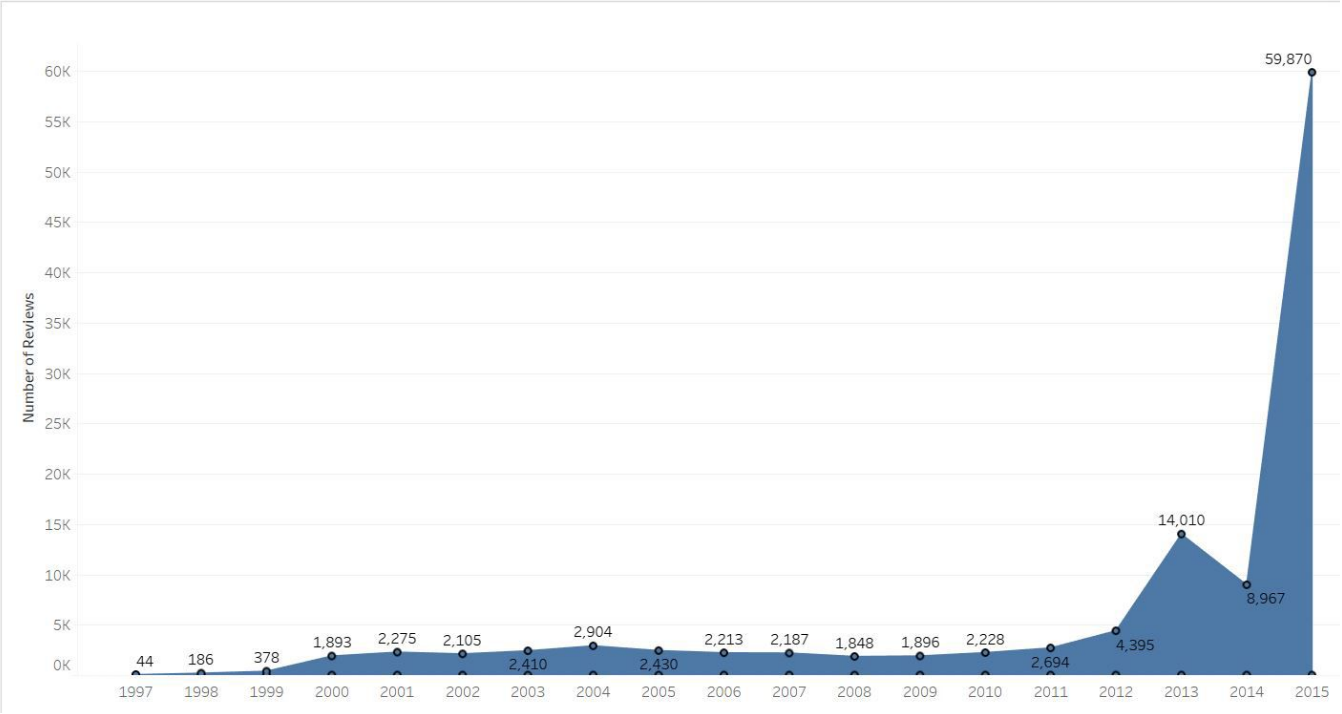


Overall Votes		Helpful votes		Star Rating	
Mean	0.731066019	Mean	1.118891	Mean	4.480137
Standard Error	0.048620773	Standard Error	0.069525	Standard Erroi	0.017623
Median	0	Median	0	Median	5
Mode	0	Mode	0	Mode	5
Standard Deviatic	2.876032796	Standard Deviation	4.112572	Standard Devi	1.042462
Sample Variance	8.271564644	Sample Variance	16.91325	Sample Varian	1.086727
Kurtosis	362.8713974	Kurtosis	399.0857	Kurtosis	4.038333
Skewness	15.60937811	Skewness	16.3372	Skewness	-2.20295
Range	85	Range	130	Range	5
Minimum	0	Minimum	0	Minimum	0
Maximum	85	Maximum	130	Maximum	5
Sum	2558	Sum	3915	Sum	15676
Count	3499	Count	3499	Count	3499

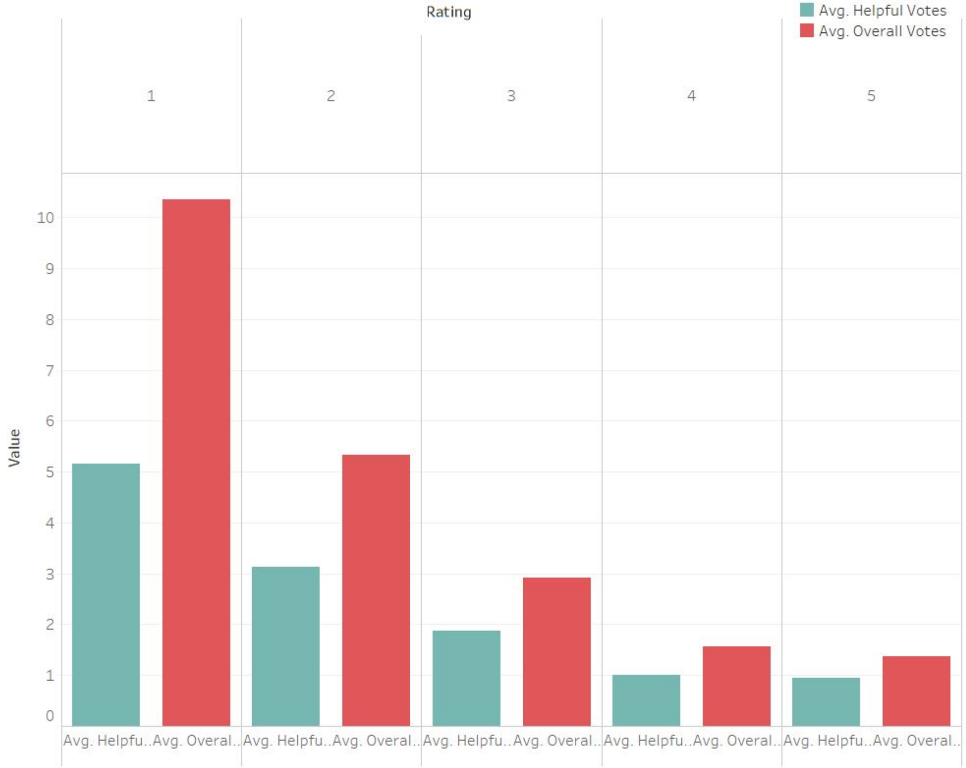
WORD CLOUD FOR ALL REVIEWS



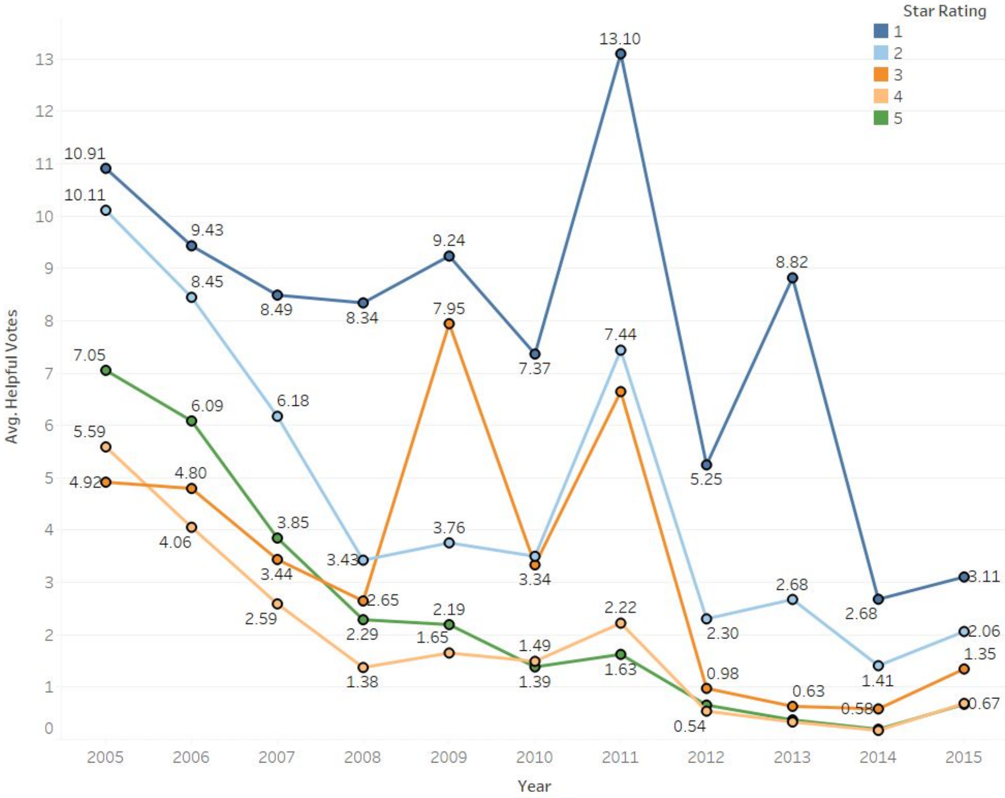
Total reviews for each year from 1997 - 2015.



Pairwise comparison of helpful votes and overall votes



Average helpful votes per rating



Modeling and Evaluation

Models Used:

Classification :

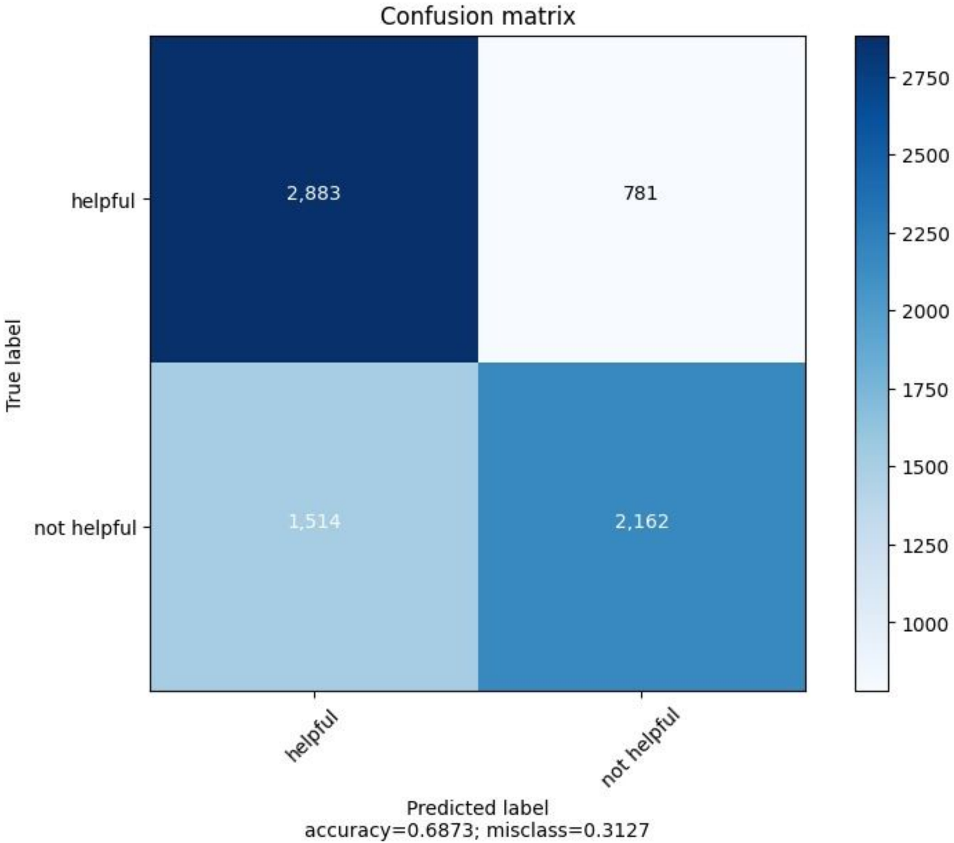
Support Vector Machine

Clustering :

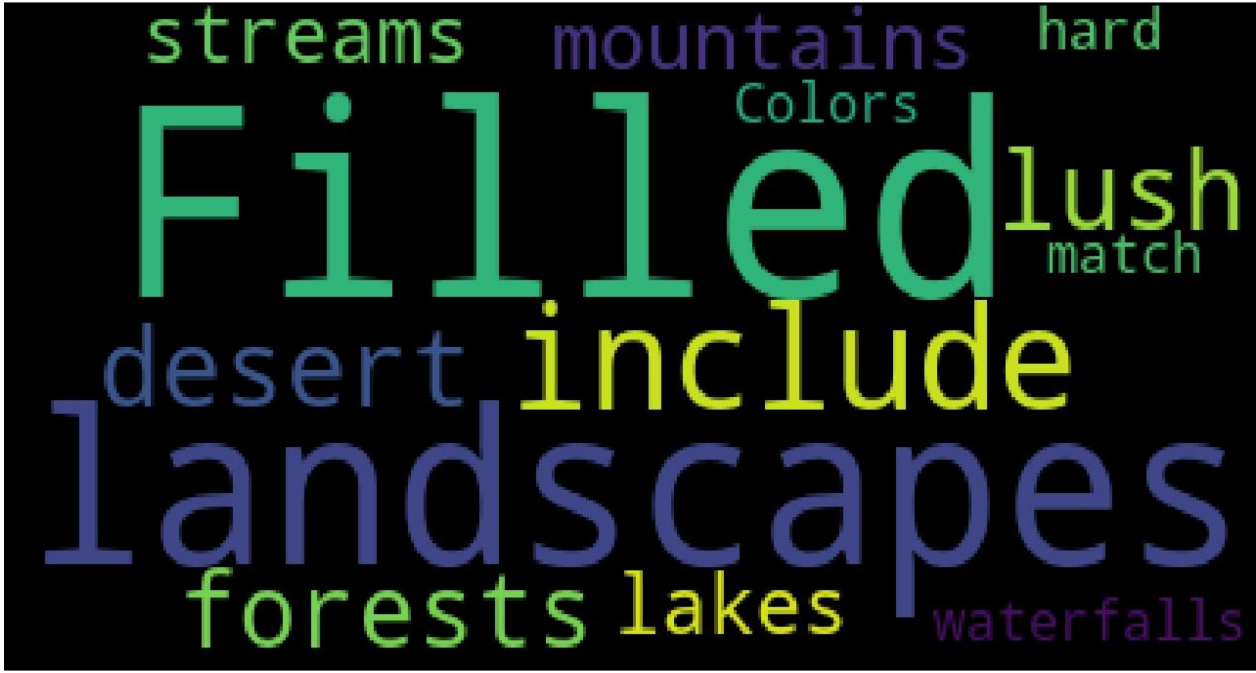
Agglomerative (Birch)

SVM Results

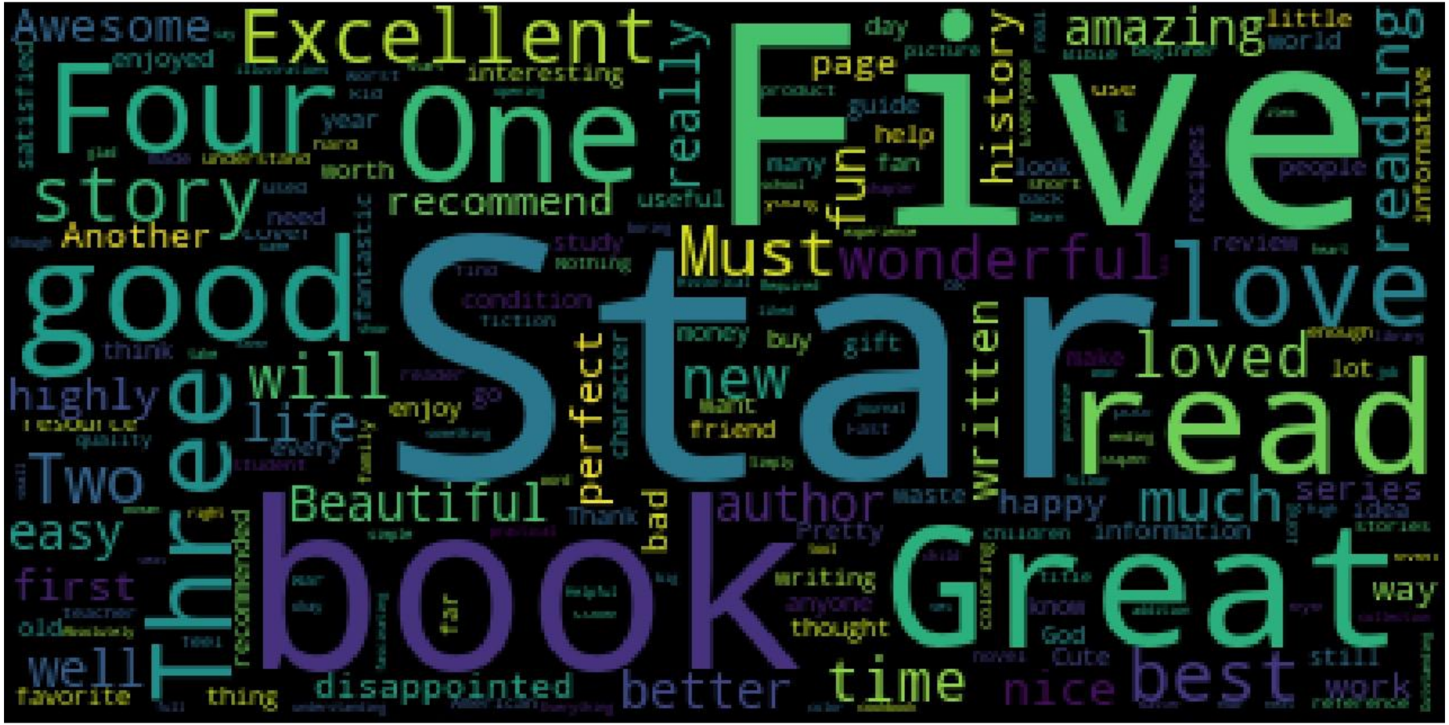
Accuracy : 68%



Birch results for a specific topic cluster



Birch results for a Generic topic cluster



Conclusion /Future Work

- Learnt industry standard data mining procedures
- Implemented SVM and Clustering. Got reasonable results
- Performed Visualization and explored relationship in attributes

Future Work

- More categories
- User identification

Tools used

1) Language: Python (Pycharm)
2) Softwares: Tableau
3) Frameworks: scikit-learn, matplotlib

THANK YOU