Movie Review Analysis

Using Machine Learning, Natural Language Processing, Locality Sensitive Hashing

Group 23
Md. Akber Hossain (267979)
Fatema Tuz Zohora (267981)

Introduction

- Every day millions and millions of reviews are posted on products of every kind.
- Important for both the service provider as well as for the consumer.
- Can be used for recommending, analyzing, getting a overview of products or services
- IMDb is an available and reliable source for information related to movies.

We mean to use the reviews collected for movies for different purposes after analysing.

Objectives

01. **Learning Haddop Architecture**

02. **Based on User Review**

- Generate Relevant Rating using Sentiment Analysis (NLP)
- Evaluate Generated Ratings using RMSE
- Classify Reviews of a movie whether Helpful or not using Naive Bayes Classifier
- Recommendation of movies on review similarity using LSH

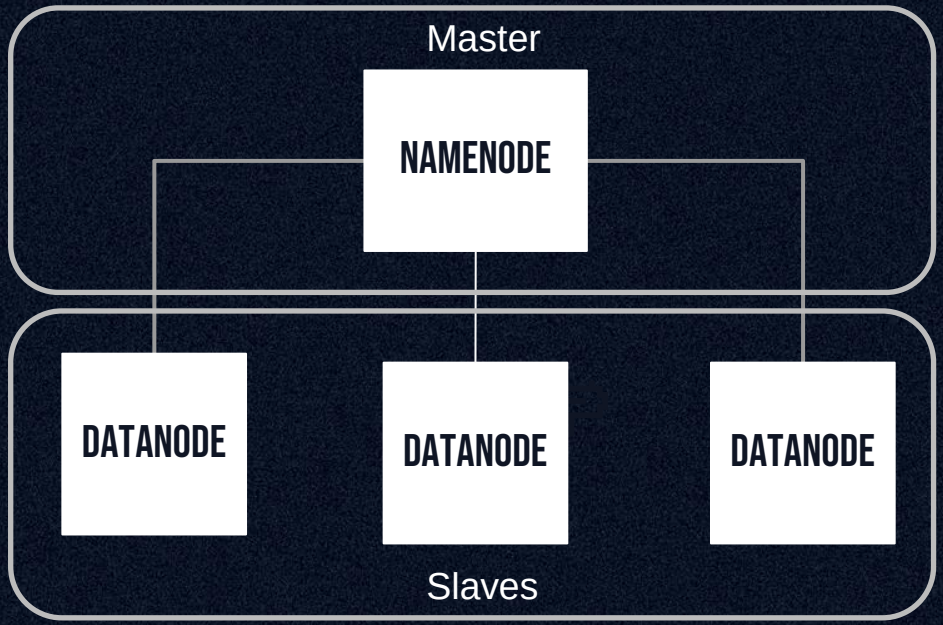
Data Set

Column Name	Column Description
Review_ID	ID of review
Reviewer	Reviewer User Name on Website of IMDb
Movie	Name of the Movie
Rating	Actual Rating Available in IMDb [on scale 1 – 10]
Review Summary	Title of each Review
Review_Date	Date of Posting the Review
Spoiler_Tag	Tags a review whether it contains spoiler
Review_Detail	Whole Review text
Helpful	Whether other finds the review as helpful

- Data Set: Kaggle
- Data volume – approximately 7 GB
- Data set generated from IMDb
- Dataset collected from Kaggle
- Total Number of Reviews – 10,10,293
- 9 attributes of the Movie reviews are included

Hadoop Architecture

- Apache Hadoop is an Open-source, Scalable Framework
- Why Hadoop?
  - can process large volume of data economically on cluster of commodity hardware
  - specifically suitable for parallel processing in distributed mode on cluster connected by a network



Specifications of our Cluster

- **Flavor** : m1.medium
- **RAM**: 4GB
- **VCPUs:** 2 VCPU
- **Disk:** 40GB
- **Source Image:** Ubuntu Focal 20.04
- **NameNode** Deamon on **Master**
- **DataNode** Deamon on **Workers**

Generating Relevant Rating from User Review

- Online reviews give power to the user to tell their observations and critical analysis.
- In case of movies, we can better understand the sentiment of the user/reviewer.

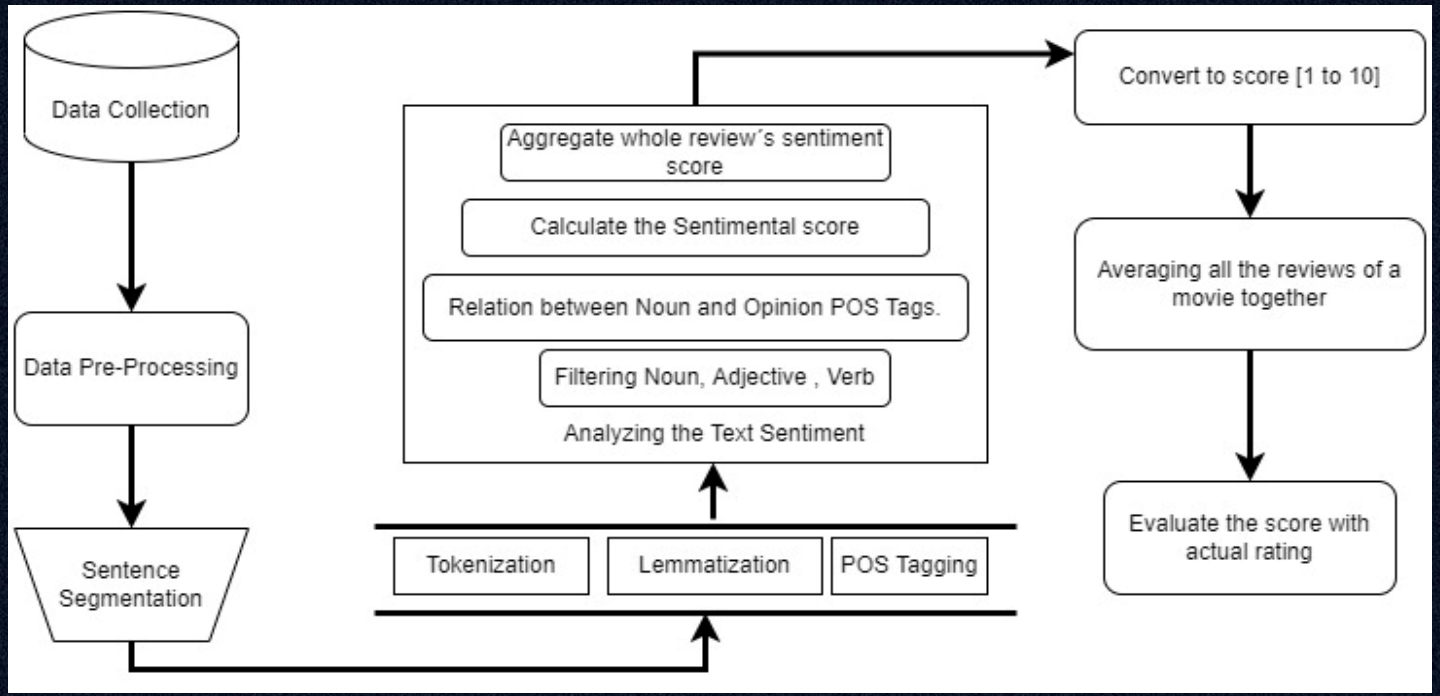
- Why Analyse User Reviews?

  - movie/product can have a huge amount review
  - not possible to go through all the review for anyone for taking decision
  - the audience won't miss the point of the review critical analysis

- What We did?

  - Sentiment Analysis Score
  - Used Different NLP concepts for analyzing the reviews

Sentiment Analysis Score



Claculating and Converting Sentiment of Sentence

Rating = (s*10)-5(s-1)

1. The performance is good.

2. The movie has an amazing plot.

3. The script is well written.

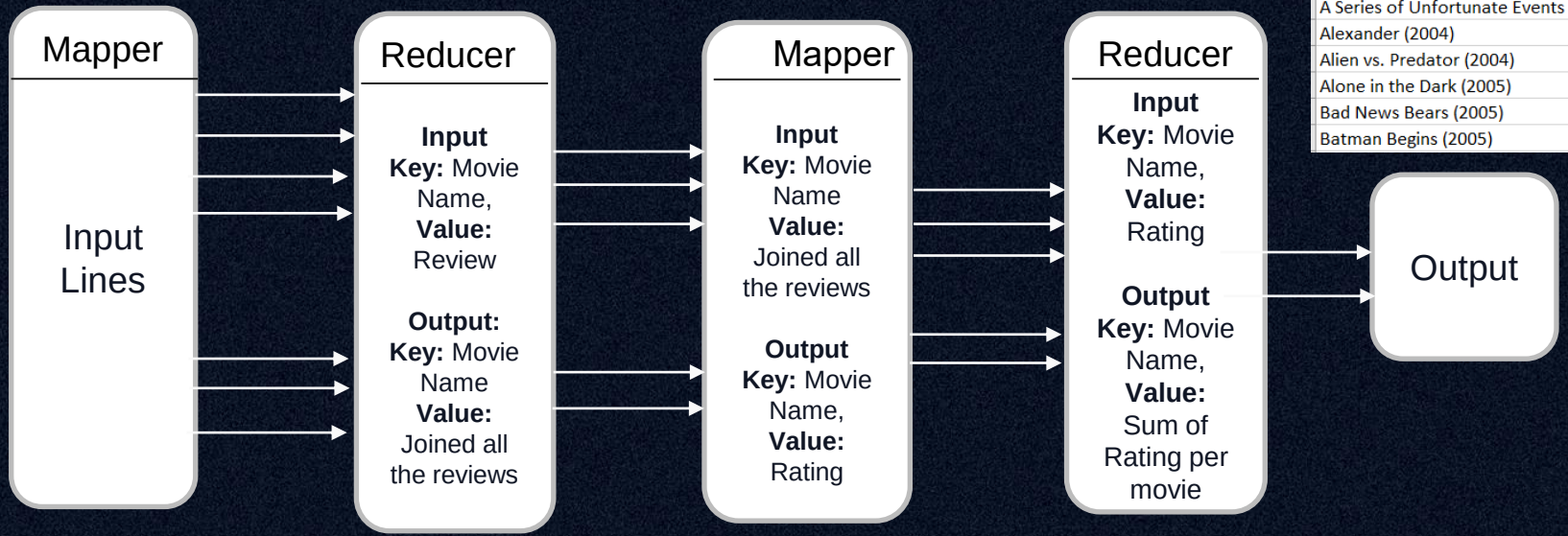
4. I love the story.

S = (+-)Polarity * subjectivity

Opinion word	Polarity (-1 to +1)	Subjectivity (0 to 1)
Good	+0.7	0.6

Score	Rating (out of 10)
-1	0
-0.5	2.5
0	5
0.5	7.5
1	10

Implementation in MapReduce



movie	rating
2001: A Space Odyssey (1968)	6
A Series of Unfortunate Events (2004)	7
Alexander (2004)	6
Alien vs. Predator (2004)	6
Alone in the Dark (2005)	6
Bad News Bears (2005)	6
Batman Begins (2005)	6

Evaluate Generated Rating

- **Evaluated** the generated rating with the Rating available.
- From the original dataset, We had movie rating provided by each user with their review and we average those rating for each movie by using **map reduce** approach. Here is the processed dataset snip.

movie	rating
2001: A Space Odyssey (1968)	8
A Series of Unfortunate Events (2004)	7
Alexander (2004)	6
Alien vs. Predator (2004)	6
Alone in the Dark (2005)	5
Bad News Bears (2005)	5
Batman Begins (2005)	6

- Evaluated Accuracy: 88%
- Algorithm Used : Root Mean Square Error (RMSE)
  - We programmed the method with the formula provided in Wikipedia, that is:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}.$$

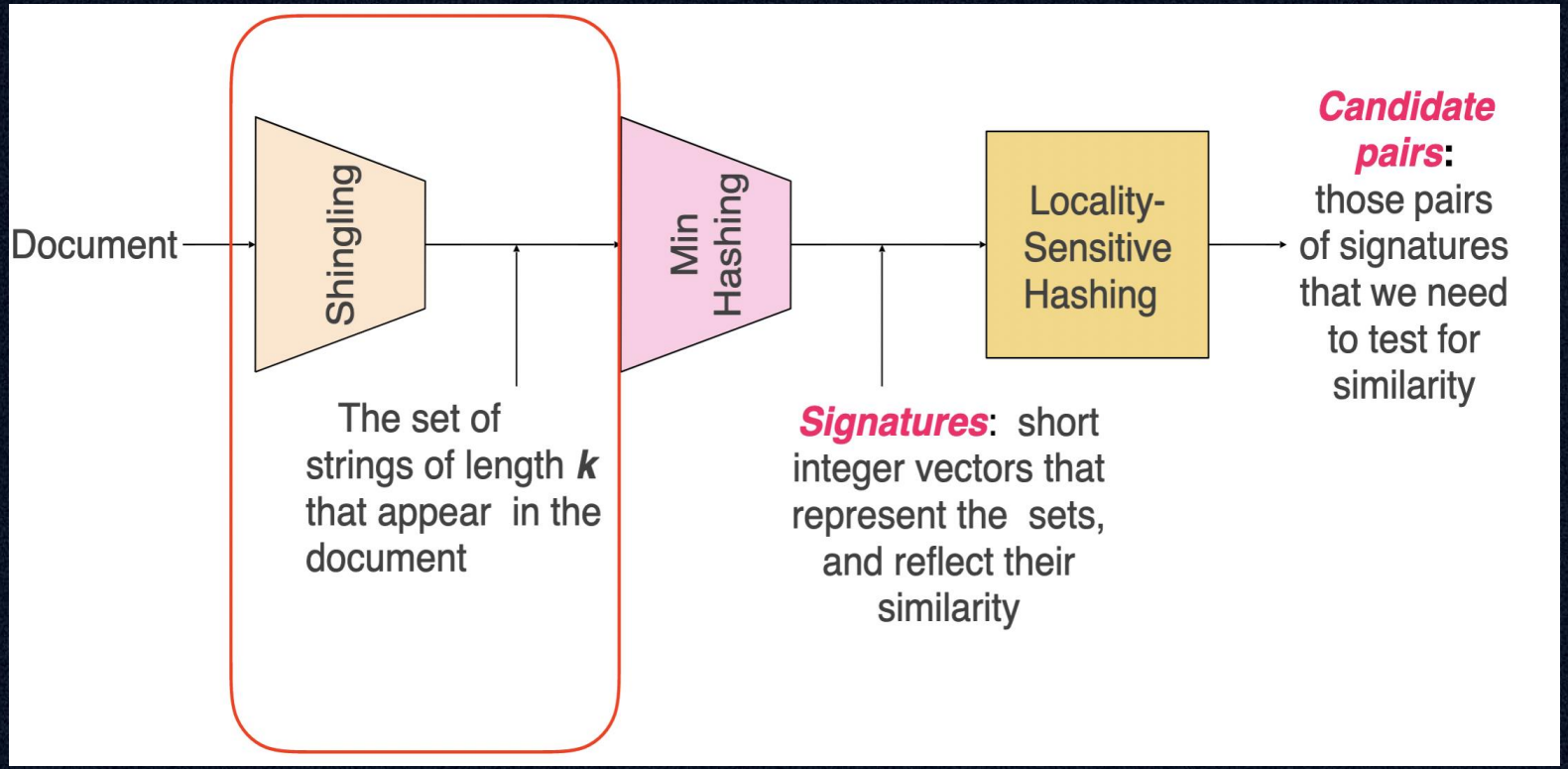
where, $\hat{y}$ is the predicted value
y is the actual value and
T is the size of the dataset.

Recommendation of Movies on Review Similarity

- Recommended similar movies which has similar reviews.
- Modified original data to have combinations of movie with their reviews.
- On the obtained reviews, implemented Locality Sensitive Hashing (LSH)

	review_id	movie_1	movie_2	review_1	review_2
0	0	'D' (2005)	'Gator Bait (1973)	I agree in most part with the first review especially when com	Seems a lot of viewers really don't get movies like GATOR BAIT.
1	1	'D' (2005)	'It's Alive!' (1969 TV Movie)	I agree in most part with the first review especially when com	my mom loved this movie, and would watch it all time when i wa
2	2	'D' (2005)	'R Xmas (2001)	I agree in most part with the first review especially when com	Our Christmas (2001) was a highly underrated film from street le
3	3	'D' (2005)	'Shitsurakuen': jôbafuku onna harakiri (1990 Video)	I agree in most part with the first review especially when com	This is NOT a horror-film, Nothing to be afraid of here, keep mov
4	4	'D' (2005)	'Way Out (1961)	I agree in most part with the first review especially when com	The "galaxybeing" did a good job of describing this series and c

Locality Sensitive Hashing



Locality Sensitive Hashing

- **Text Mining**
- Evaluated the candidate pairs with **Jaccard Similarity** Algorithm
- Several **MapReduce** steps were required to complete this Algorithm

Hostage (2005)	The Lord of the Rings: The Return of the King (2003)
Constantine (2005)	The Aristocrats (2005)
The Grudge (2004)	The Life Aquatic with Steve Zissou (2004)
Eternal Sunshine of the Spotless Mind (2004)	Submerged (2005 Video)
Fantastic Four (I) (2005)	The Shawshank Redemption (1994)
Alien vs. Predator (2004)	Team America: World Police (2004)
Bewitched (2005)	Team America: World Police (2004)

Classify Reviews of a movie whether Helpful

- From our original dataset we had helpful column [4, 6] like this which represent out of 6, 4 person marked this review as helpful. We convert this to 1 as helpful and 0 which are not helpful. By using MapReduce we have pre-processed our data for this format.

review_detail	help
About Clint s last movie I d it was something we h	1
I been waiting so long to get round to seeing this r	0
What the enormous interest in a sports hero ? Wh	0
I have to say this t as big a disappointment as peo	0
I d like to first comment on why i chose to give thi	0

- Then as a traditional Machine learning model apply approach, we have split into training and test dataset and then applied Naïve Bayes algorithm with help of count vector and TFID for binary classification problem on our processed review dataset (with help column).

- We are using spark for applying the ML algorithm in this task.

Related Works

- SentiWordNet, lexical resource for opinion mining, (http://sentiwordnet.isti.cnr.it/).
- V.K. Singh, R. Piryani, A. Uddin and P. Waila, “Sentiment Analysis of Movie Reviews A new Feature-based Heuristic for Aspect-level Sentiment Classification”, Proceedings of the 2013 International Muli-Conference on Automation, Communication, Computing, Control and Compressed Sensing, Kerala-India, March 2013.
- Exploring relationships between Attributes - https://minimaxir.com/2018/07/imdb-data-analysis/
- Predicting movie gross revenue - https://medium.com/@jae.huang111/imdb-data-machine-learning-predicting-movie-gross-2113513513bb
- Predicting review class of movies - http://ceur-ws.org/Vol-1365/paper12.pdf

Conclusion

- We have Analyzed and observed several use cases of Movie reviews:

  - Would Help to Understand probability of liking a movie in a more critical way
  - Have a great accuracy in predicting ratings
  - Recommends similar movies
  - Could be classified being helpful about the movie

Thank You