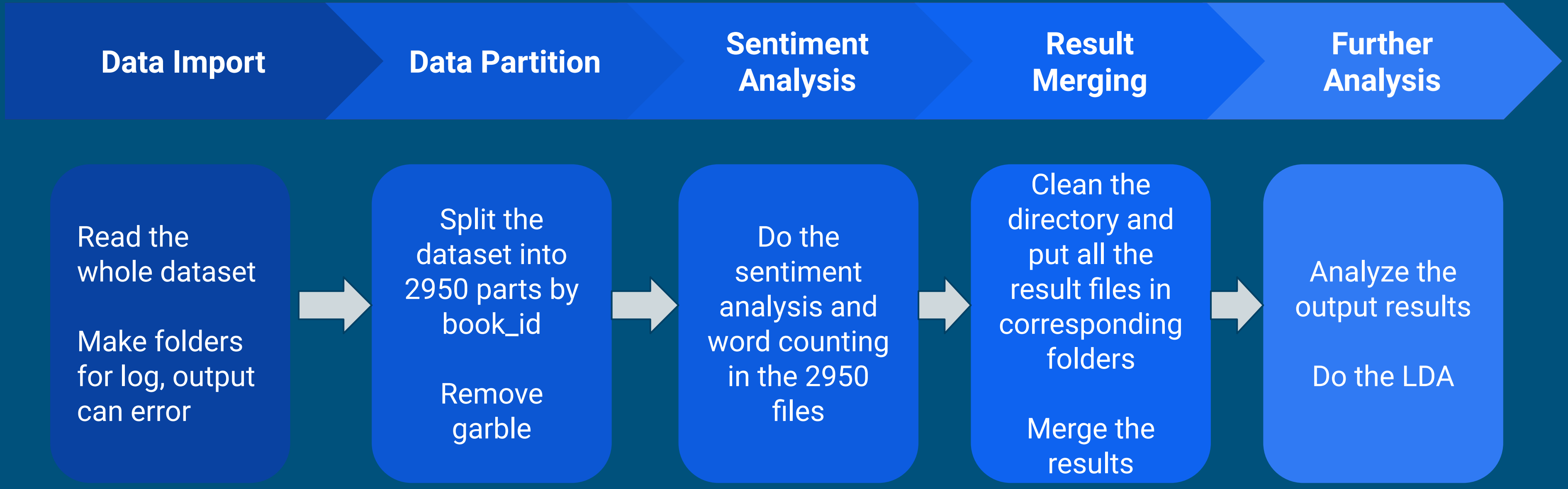
Good Reads Analysis

Project 7
Philip Lin, Yanrun Lu, Xiaoyang Wang, Ziang Zeng, Chixu Ni

Data Description

- goodreads_reviews_dedup.json
    - 15.5 GB
    - Contain general information about the review including the **user, book, and review id**.
    - Includes the text contained in the review and the **number of votes and comments**.
- goodreads_reviews_spoiler.json
    - 1.75 GB
    - Contains the user, book, and review id.
    - Includes sentences reviewed for **spoilers** and a binary column for whether the review has a spoiler or not.

Research Process



Data Import

Read the whole dataset

Make folders for log, output can error

Data Partition

Split the dataset into 2950 parts by book_id

Remove garble

Sentiment Analysis

Do the sentiment analysis and word counting in the 2950 files

Result Merging

Clean the directory and put all the result files in corresponding folders

Merge the results

Further Analysis

Analyze the output results

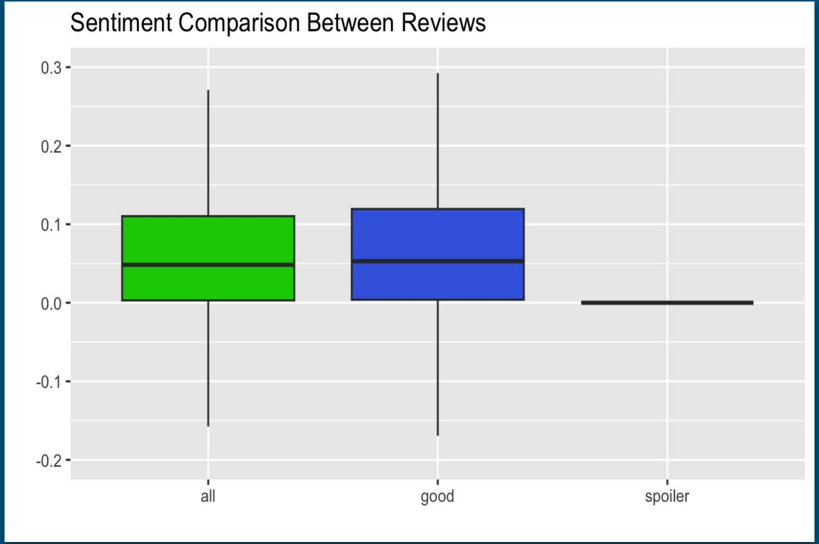
Do the LDA

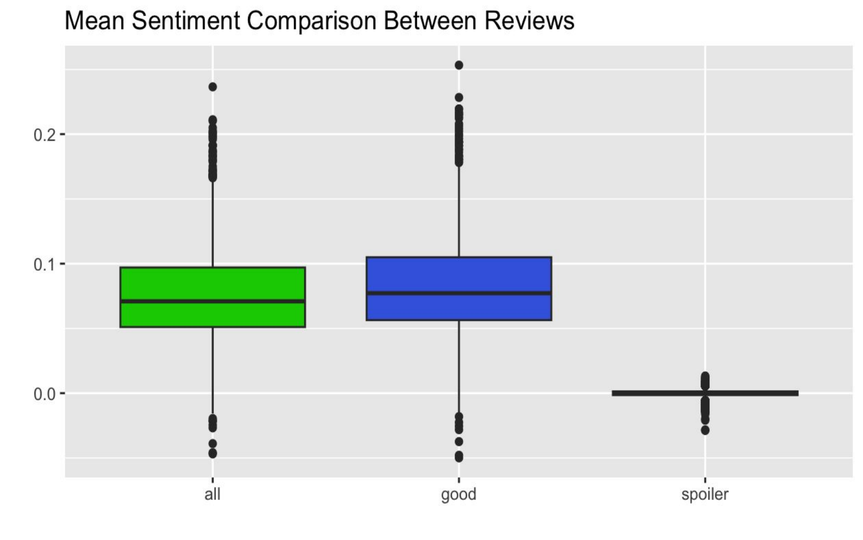
2950 parallel runs

2950 parallel runs

Sentiment Analysis – Spoiler

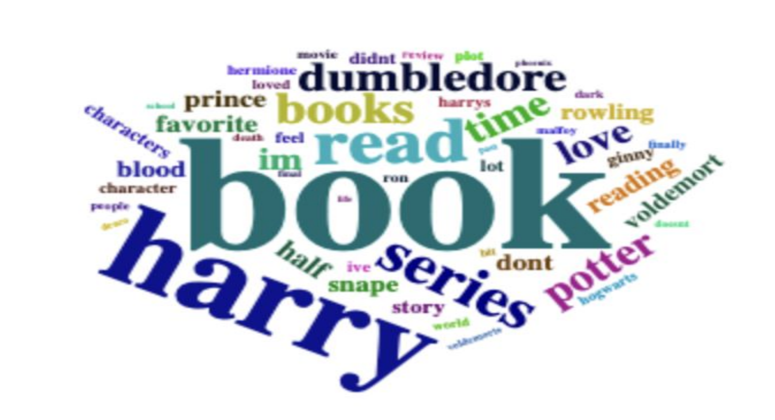
- **Spoiler reviews do have influence on the emotion of the comments:**



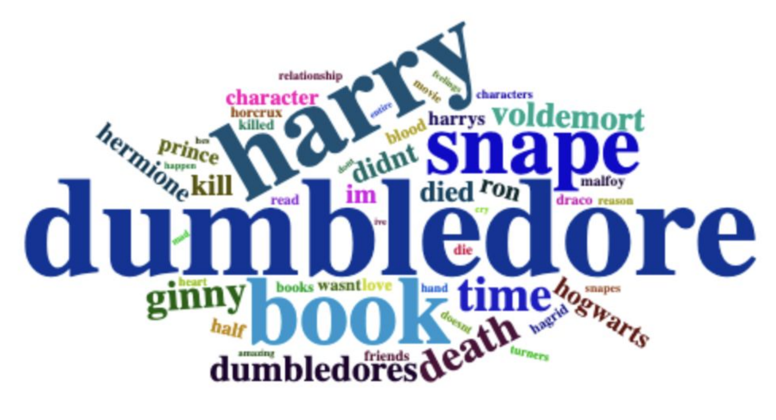


Word Cloud Example – Spoiler

- **Harry Potter and the Half-Blood Prince**



No Spoiler



Spoiler

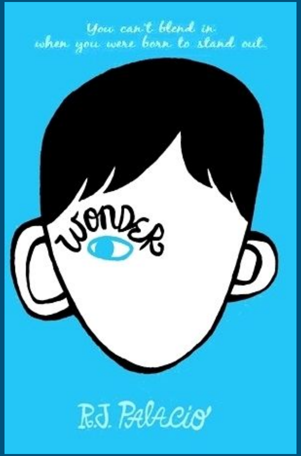
Sentiment Analysis

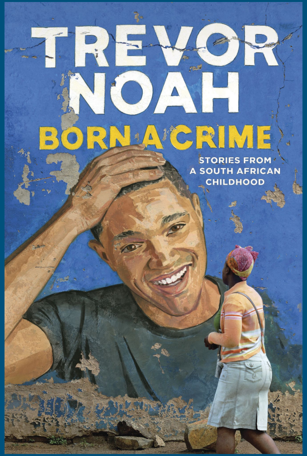
- **Books that are recommended from our perspectives:**

  Choose the Top 20 rating

  Analyze the sentiment

  Choose the book has sentiment > mean sentiment var < mean var

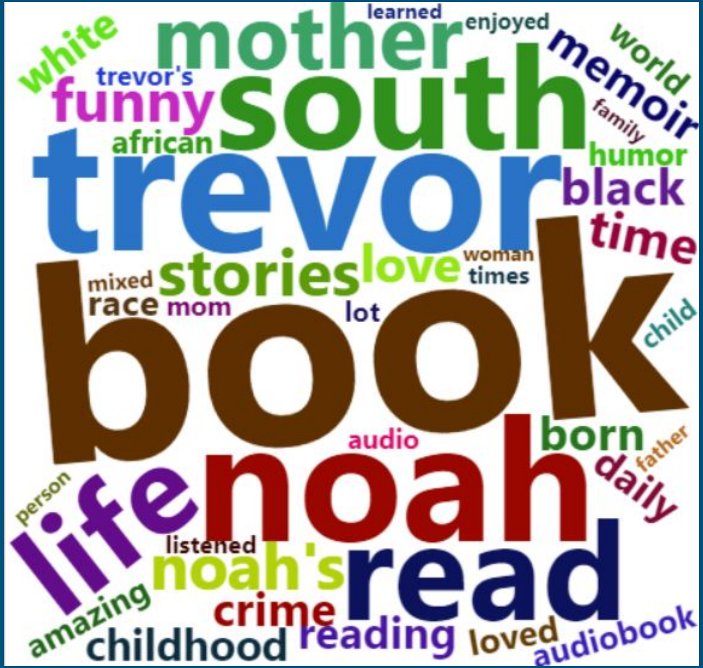




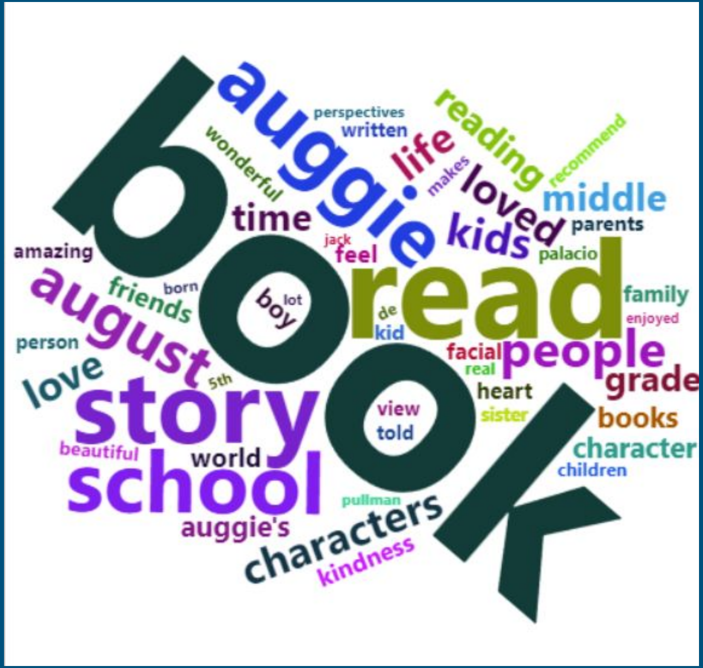
Name:	**Wonder**	**Born a Crime**
Author:	**R.J.Palacio**	**Trevor Noah**
Rating:	**4.38**	**4.49**

Word Cloud Example

Born a Crime



Wonder



Latent Dirichlet Allocation (LDA)

- LDA is an unsupervised algorithm which uncovers hidden topics or themes in large collections of textual data.
- The model implemented attempts to discover 5 main topics.
- The naming of the topics will need to be done manually since the learning for this model was not guided towards any specific topics.

LDA Output (Most to Least Influential)

- **Topic 0: General Book Appreciation**
  - Words: 0.036*"book" + 0.017*"like" + 0.016*"read" + 0.014*"love" + 0.012*"one" + 0.010*"know" + 0.009*"really" + 0.008*"even" + 0.007*"much" + 0.007*"would"
- **Topic 1: Character and Story Analysis**
  - Words: 0.040*"book" + 0.021*"really" + 0.017*"read" + 0.017*"like" + 0.015*"story" + 0.014*"characters" + 0.011*"one" + 0.009*"much" + 0.008*"good" + 0.008*"first"
- **Topic 2: Broader Literary Discussion**
  - Words: 0.008*"story" + 0.007*"book" + 0.006*"one" + 0.006*"world" + 0.006*"novel" + 0.005*"people" + 0.004*"time" + 0.004*"life" + 0.003*"read" + 0.003*"like"
- **Topic 3: Personal and Romantic Themes**
  - Words: 0.008*"story" + 0.007*"love" + 0.007*"one" + 0.007*"life" + 0.005*"family" + 0.005*"girl" + 0.005*"relationship" + 0.004*"get" + 0.004*"like" + 0.004*"romance"
- **Topic 4: Series and Sequels**
  - Words: 0.023*"series" + 0.013*"love" + 0.010*"book" + 0.009*"one" + 0.007*"world" + 0.006*"see" + 0.005*"review" + 0.005*"new" + 0.005*"next" + 0.005*"get"

Understanding LDA

- The weights in front of each words represent the probability of the word belonging to that topic with higher weights correlated to the word being more strongly associated with the topic.

- LDA can help with understanding which aspects of books resonate most with readers: emotional connection, narrative style, thematic depth, series sequel...

- Authors can use this information to understand what themes to highlight for different reader segments.

Summary

- With the help of CHTC, we can do the parallel jobs to improve the efficiency of operation, which helps us to deal with the big data in a short time.

- From the sentiment analysis of the spoilers, we found that these reviews actually impact other readers' experience.

- Our insight from sentiment analysis and LDA aids in tailoring book themes for different audiences and enhancing online bookstores' and libraries' recommendation systems.

Future Work

- Build a model for each book to determine whether there exists spoilers in the review and catch them.

- Improve sentiment analysis methods, such as using phrases rather than words to calculate ratings.

- Combine the results of sentiment analysis and LDA to filter the critical reviews, so that we can provide readers or other customers with more accurate information about the books.