BOOK REVIEW PREDICTOR

WITH UNSUPERVISED CLUSTERING AND
SUPERVISED CLASSIFICATION

GROUP 1 MEMBERS

- Vũ Công Duy - 201476737
- Đào Hồng Quân - 20176850
- Trần Công Minh - 20176825

CONTENTS

INTRODUCTION

DATA SCRAPING AND CLEANING

UNSUPERVISED CLUSTERING

SUPERVISED CLASSIFICATION

BOOK REVIEW PREDICTOR

CONCLUSION

INTRODUCTION

- PROJECT OVERVIEW
- MACHINE LEARNING METHODS
- VARIABLES

VARIABLES

Unsupervised learning:

- List of authors
- Book's series (if exist)
- List of tags

Supervised learning:

- Number of raters
- Number of reviewers
- Number of pages
- Publication year

Response: 0 'negative', 1 'mixed', 2 'positive'

DATA SCRAPING AND DATA CLEANING

- 500,000 books for this project are crawled by using Scrapy from the website "goodreads.com" and saved in csv files

- The data are cleaned with the pandas and numpy libraries on Jupiter Notebook

- From the original 456,000 - row dataset, only over 97,000 books remain

UNSUPERVISED CLUSTERING

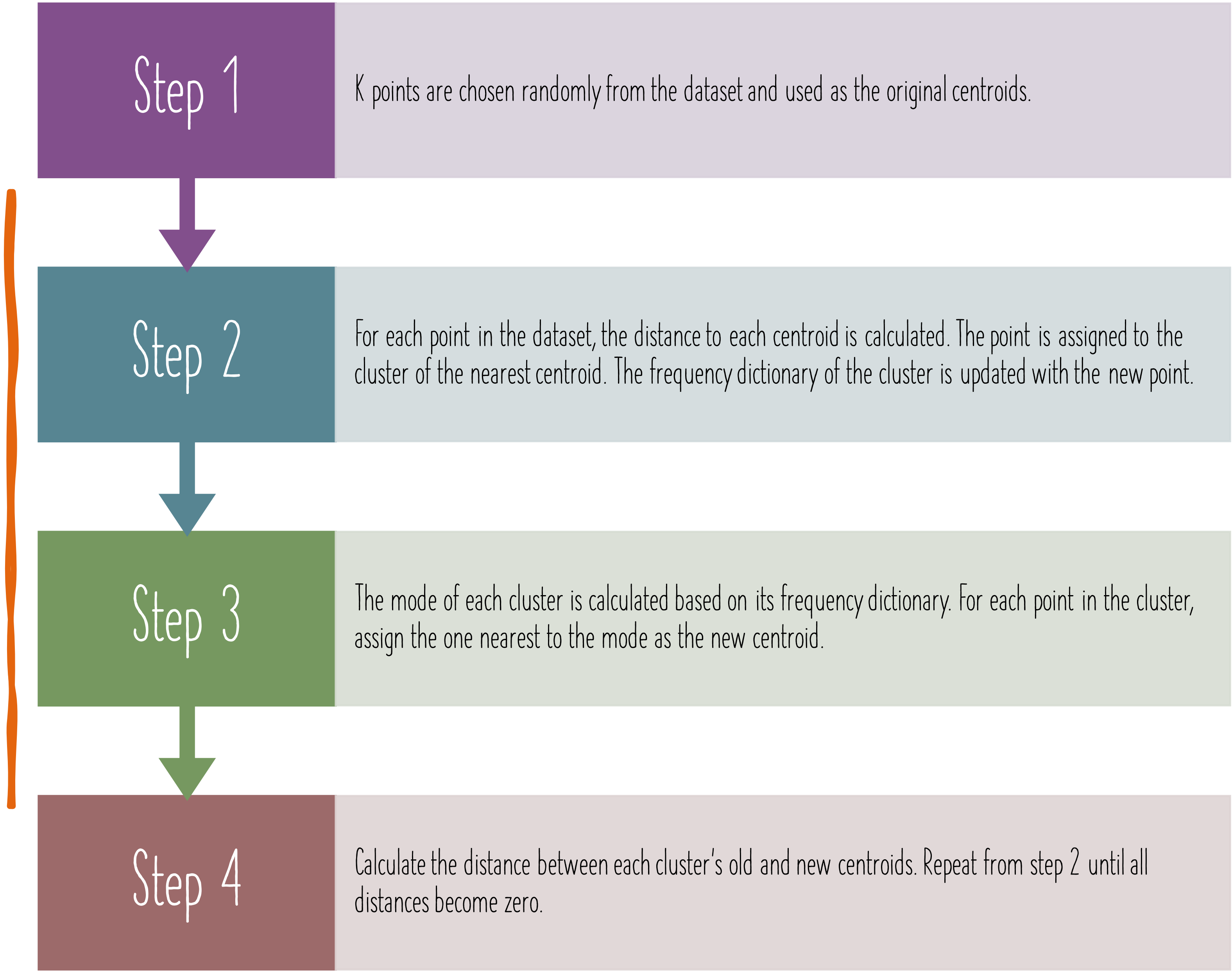
PROBLEM STATEMENTS

- A book has not only its official genre but is also frequently branded other tags by online readers
- There can be a hidden pattern in books with similar tags
- K-Means clustering only works for numerical attributes

K - MODES

- K-Modes is a modified version of K-Means where the centroid is determined by the mode of each attribute of the data rather than the means
- Since the data fields are in string type, they will be measured by a variation of the hamming distance
- => the difference between books will not be too big and the complexity of the algorithm is better than only comparing the difference in each string

K - MODES ALGORITHM



Step 1

K points are chosen randomly from the dataset and used as the original centroids.

Step 2

For each point in the dataset, the distance to each centroid is calculated. The point is assigned to the cluster of the nearest centroid. The frequency dictionary of the cluster is updated with the new point.

Step 3

The mode of each cluster is calculated based on its frequency dictionary. For each point in the cluster, assign the one nearest to the mode as the new centroid.

Step 4

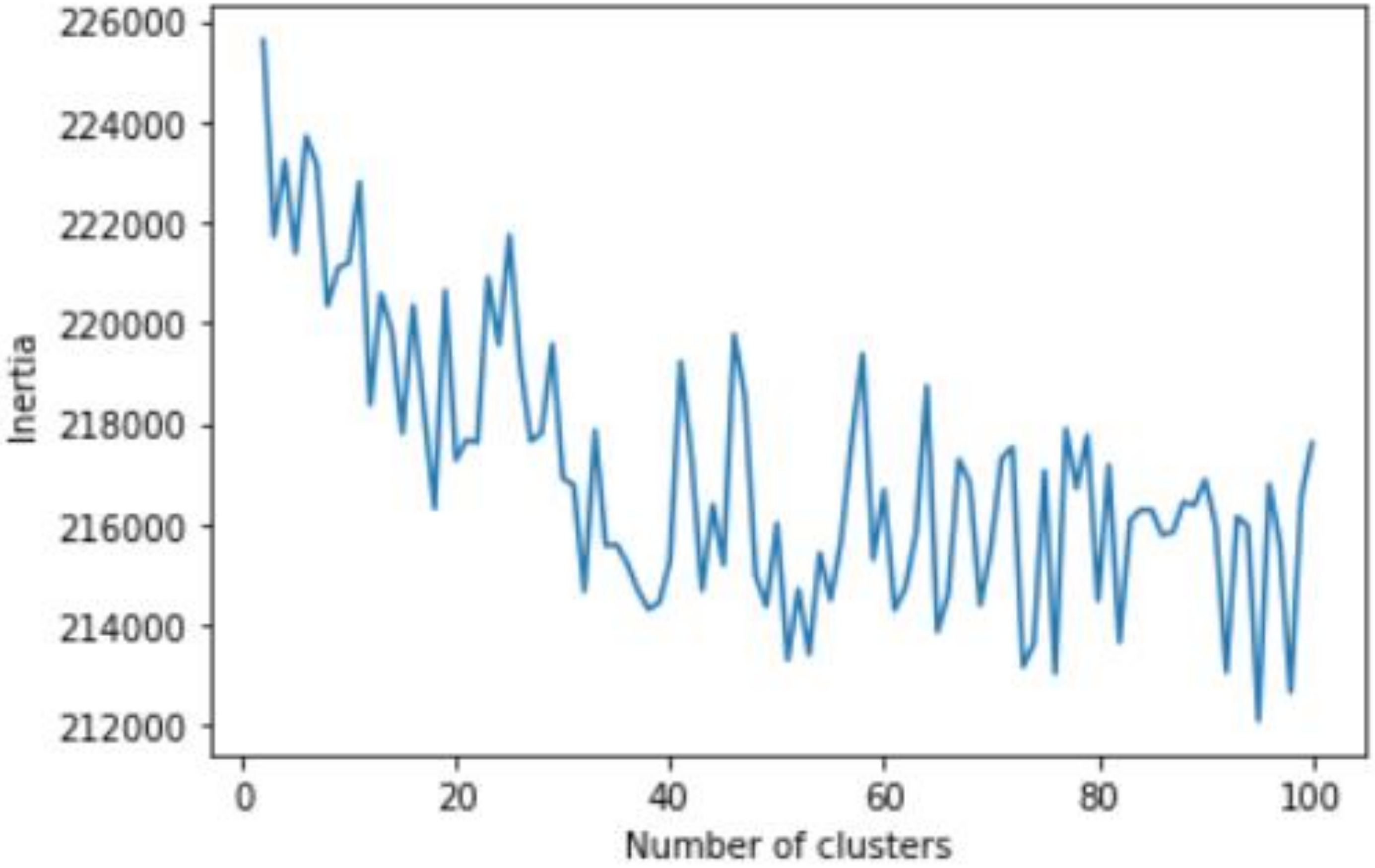
Calculate the distance between each cluster's old and new centroids. Repeat from step 2 until all distances become zero.

K - MODES IMPLEMENTATION

- To determine the compactness of a clustering model, the inertia is calculated by adding up the distances of all data points to their respective centroid
- The model is coded as a class object with the aim to reuse in future project

RESULT

- The result fluctuates greatly, but slowly declines
- 30 is the most suitable number of clusters



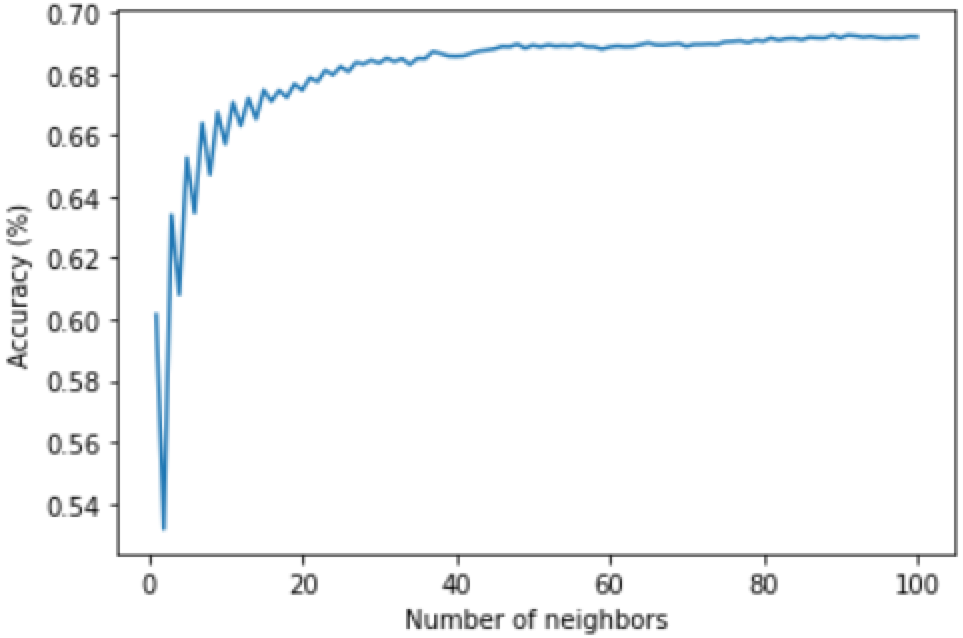
SUPERVISED CLASSIFICATION

PROBLEM STATEMENTS

- The main goal of project is to determine a book's general opinion based on its existing information
- The efficiencies of three basic classification methods are examined

K-NEAREST NEIGHBORS

-We use the KNeighborsClassifier model from the scikit-learn library



-The suitable number of neighbors should be between 20 and 40

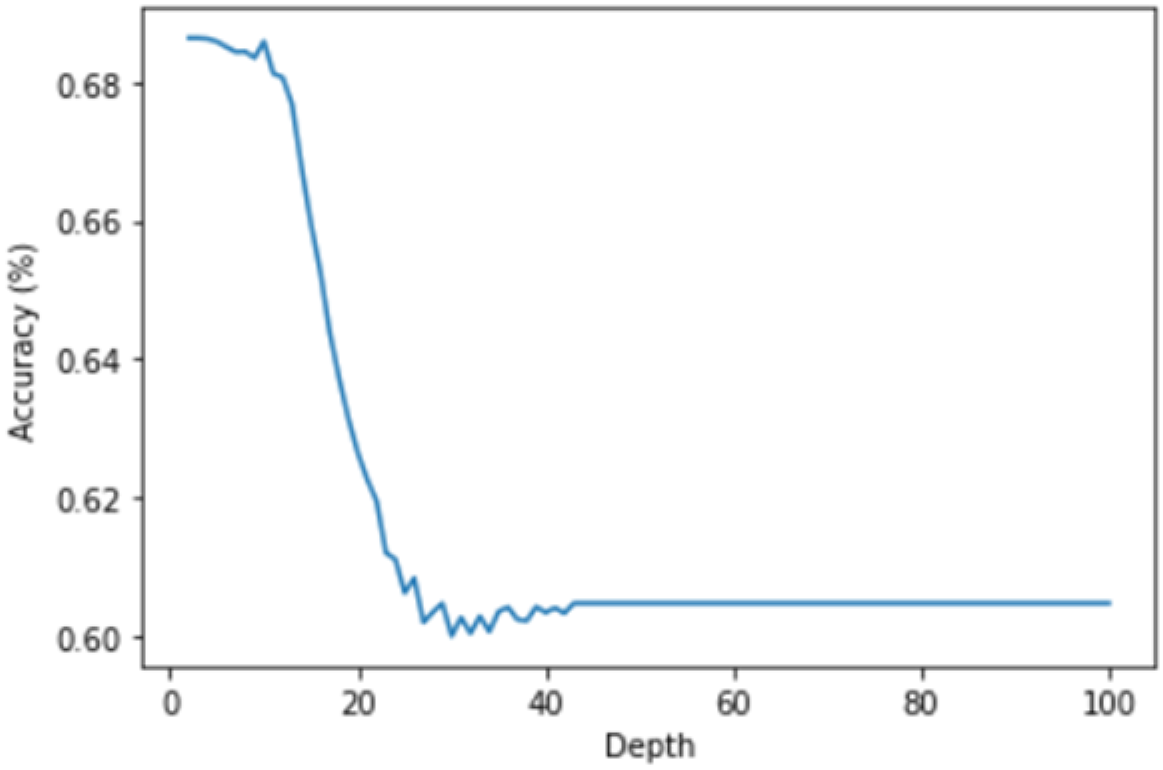
-However, since the number of negative books is too small, high number of neighbors will ignore the "negative" prediction

GAUSSIAN NAÏVE BAYES

- Since the data are numeric, the Gaussian model GaussianNB of the sklearn.naive_bayes library is used
- The model has extremely poor result of around 15% correct.
- However, the model has a fairly distributed prediction

DECISION TREE

We use the DecisionTreeClassifier model from sklearn library



Even at its best result, the model suffers great bias as it usually returns the result of most frequent label

BOOK REVIEW PREDICTOR

- The program is combination of the K-Modes model and a classification method of choice
- When a dataset is used to train the model, it is divided into clustering data for the K-Modes model and classification data for other classifiers to use
- When a classifier is passed into the model, each cluster will generate its own classifier from the original one and train it with the cluster data
- When the model's predict method is called, it receives the dataset and a classifier as arguments

RESULT

- At 3 clusters, overall, there is only slight improvement in comparison with the sole classifier. In Naïve Bayes case, the accuracy increases by over 10%
- With the ideal number of clusters as 30, the book model can reach up to 70% correct with the help of a good KNN model

CONCLUSION

- While a large number of data is crawled, most of them are noises and cannot be used
- The data itself is also majorly affected by users' bias

- => the accuracy can hardly get over 70%

THANKS FOR LISTENING