Data mayhem
Nearest Neighbor

Vaibhav karnam
Yatish kadam

Goal

- Give suggestions for books based on historical data.
- We find the k nearest neighbouring books for a book.and suggest that to a user.
- Suggestions based on the book that the user has rated before.

Project flow

Job-1

- Data Preprocessing
- Bootstrap sampling
- Ensemble training
- Classifying on the above training

Job -2

- Data preprocessing
- Knn classification

Data Preprocessing

User-ID; "ISBN"; "Book-Rating"
276725; "034545104X"; "0"

This is from the Bx-Book Rating.csv file

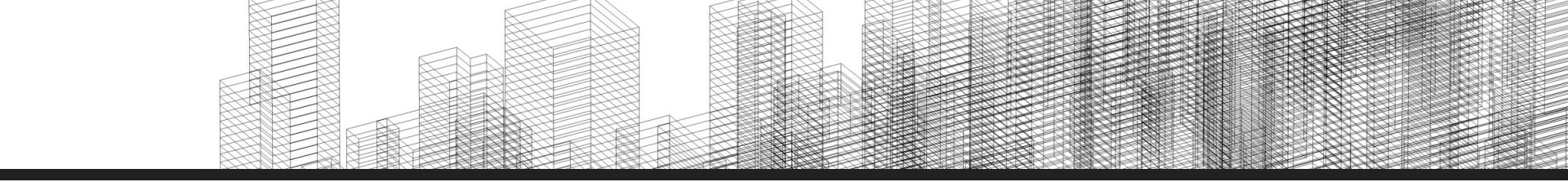
ISBN;"Book-Title";"Book-Author";"Year-Of-Publication";"Publisher";

0195153448;"Classical Mythology";"Mark P. O. Morford";"2002";"Oxford University Press";

This is from the BX-Books.csv file

User-ID; "Location"; "Age"

1; "nyc new york usa"; NULL
This is from the BX-Users.csv file



KNN

- Finding the nearest neighbours.
- The distance between yearOfPub, location, author, publisher, bookTitle

Analysis

Speed up

The following are the time we get for when
we ran on different number of nodes.

5 nodes		6854000
10 nodes		3864000

The following are the time we get for when we ran for a decreased data set.

10 nodes small data		4951000
10 nodes large data		3864000

Output

When k =3 , number of inputs = 2

For two books we get the following k nearest neighbours

2.0 Wish You Well,2000,Warner Books

3.0 The Joy Luck Club,1994,Harpercollins

5.0 Downtown,1995,Jove Books

2.0 Wish You Well,2000,Warner Books

3.0 Bleachers,2003,St. Martin's Minotaur

5.0 Bless The Beasts And Children : Bless The Beasts And Children,1995,St. Martin's Minotaur

Ensemble

Bootstrapping

Training

Classification

Bootstrapping

An example

The following numerical example will help to demonstrate how the process works. If we begin with the sample 2, 4, 5, 6, 6, then all of the following are possible bootstrap samples:

- 2 ,5, 5, 6, 6
- 4, 5, 6, 6, 6
- 2, 2, 4, 5, 5
- 2, 2, 2, 4, 6
- 2, 2, 2, 2, 2
- 4,6, 6, 6, 6

Training

We take the training data and save it as a model.

Classification and its results

Speed up

The following are the time we get for when we ran on different number of nodes.

Scale up

We reduced the input by 35% and ran the same on a cluster with 10 nodes the scale up is shown in the table below.

# Nodes		Classification milli seconds

5 nodes		9211000
10 nodes		6818000

10 nodes with large data		6818000
10 nodes with less data		4933000

Thank you