National College of Ireland

# A recommender systems and social networking approach to alleviate the issue of coldstart - Configuration Manual

Research in Computing
MSc Data Analytics October 2020/21

Mahesh Arjun Matele
Student ID: 19179065

School of Computing
National College of Ireland

Supervisor: Paul Stynes

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**

National College of Ireland

| | |
|---|---|
| **Student Name:** | Mahesh Arjun Matele |
| **Student ID:** | 19179065 |
| **Programme:** | Msc. Data Analytics |
| **Year:** | September 2020/21 |
| **Module:** | Research in Computing |
| **Supervisor:** | Paul Stynes |
| **Submission Due Date:** | 16/08/2021 |
| **Project Title:** | A recommender systems and social networking approach to alleviate the issue of coldstart - Configuration Manual |
| **Word Count:** | XXX |
| **Page Count:** | 5 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | |
| **Date:** | 9th August 2021 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | ☐ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | ☐ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | ☐ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Contents

**1 Introduction** **1**

**2 System Specification** **1**
2.1 Hardware Specification . . . . . 1
2.2 Software Specification . . . . . 1

**3 Install Anaconda Navigator Environment (ANE)** **2**

**4 Packages and Libraries used** **2**

**5 Python version** **2**

**6 Download required files** **2**

**7 Steps to execute the code** **3**
7.1 Approach 1 - Jupyter Lab Steps for execution . . . . . 3
7.2 Approach 2 - Google Colab Steps for execution . . . . . 3

**8 Evaluation** **5**

# List of Figures

1 Jupyter NB - change path for movies file . . . . . 3
2 Jupyter NB - change path for ratings file . . . . . 3
3 Jupyter NB - change path for all_data file . . . . . 3
4 Google Colab - Upload required files to google colab environment . . . . 4
5 Google Colab - change path for all_data file . . . . . 4
6 Google Colab - change path for movies file . . . . . 4
7 Google Colab - change path for ratings file . . . . . 4
8 Output of the run . . . . . 5

# A recommender systems and social networking approach to alleviate the issue of coldstart - Configuration Manual

Mahesh Arjun Matele
19179065

## 1 Introduction

This configuration manual provides an insight into the prerequisites for conducting the experiment. It also acts as a step-by-step guide for making required changes in the code to run the experiment and receive expected results.

## 2 System Specification

Below details define the hardware and software configurations on which the experiment was conducted.

### 2.1 Hardware Specification

The experiment was conducted on the below h/w:

- Processor - Intel i5@2.3GHz
- RAM - 6.00 GB
- System - 64 bit Windows OS

### 2.2 Software Specification

- Python - Python is a programming language that lets you work more quickly and integrate your systems more effectively. [1]
- Anaconda Navigator - Anaconda Individual Edition is a free, easy-to-install package manager, environment manager, and Python distribution with a collection of 1,500+ open source packages with free community support. Anaconda is platform-agnostic. [2]
- Google Colab - Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with Zero configuration required, Free access to GPUs, Easy sharing.

[1] Python: `https://www.python.org/about/`
[2] Anaconda Navigator: `https://docs.anaconda.com/`

# 3 Install Anaconda Navigator Environment (ANE)

- The software for Anaconda Navigator can be installed by following the steps mentioned in the link here - https://docs.anaconda.com/anaconda/install/windows/
- Once the installation is completed, verify if the installation is done using the link here - https://docs.anaconda.com/anaconda/install/verify-install/
- Once the installation is verified, create an environment for python so that it can be used in Jupyter Lab using the link here - https://docs.anaconda.com/anaconda/navigator/tutorials/manage-environments/#creating-a-new-environment
- Click on the Jupyter Lab icon present on the ANE to start Jupyter Lab.

# 4 Packages and Libraries used

Once the Anaconda Navigator is installed make sure the below libraries are available in the ANE as they are used as part of the solution -

- matplotlib
- scipy
- sklearn
- itertools
- seaborn
- numpy
- pandas
- mpl_toolkits

# 5 Python version

The python version used for this experiment was 3.8.3.

# 6 Download required files

The files required for running the experiment are available in the link here - https://grouplens.org/datasets/movielens/100k/ For running the experiment make sure the below files are available in the local environment.
To know the metadata of the dataset use the below link: https://studentncirl-my.sharepoint.com/:t:/g/personal/x19179065_student_ncirl_ie/EY8DvCng0YxNopi5aUoUOJYBa6yb8P6wHqJov8KyJXKFw?e=UB37jA
Once the files are downloaded and unzipped and available for use in the local environment, edit the code to point to these datasets which is elaborated in section 7.

# 7 Steps to execute the code

There are two approaches in which the experiment can be replicated namely -

- Using Jupyter NB
- Using Google Colab

Below is the elaboration for the same.

## 7.1 Approach 1 - Jupyter Lab Steps for execution

Before running the code the only requirement is to change the path where the file is being read. Below is the screenshot for places to change the source file path to the one where the user has stored the source files:

```
# movie data
movies = pd.read_csv("C:\\Users\\Mahesh\\Desktop\\nci_references\\3) SEM 3\\RIC\\dataset\\another_dataset\\movies.tsv", sep='\t')
#print(f"Shape: {df.shape}")
movies.head(1)
```

| | movieID | name | year | genre1 | genre2 | genre3 |
|---|---|---|---|---|---|---|
| 0 | 1 | Toy Story | 1995 | Animation | Children's | Comedy |

Figure 1: Jupyter NB - change path for movies file

```
df_ratings = pd.read_csv('C:\\Users\\Mahesh\\Desktop\\nci_references\\3) SEM 3\\RIC\\dataset\\another_dataset\\ratings.csv')
df_ratings.head(1)
```

| | userID | movieID | rating |
|---|---|---|---|
| 0 | 747 | 1193 | 5 |

Figure 2: Jupyter NB - change path for ratings file

```
# all data
df = pd.read_csv("C:\\Users\\Mahesh\\Desktop\\nci_references\\3) SEM 3\\RIC\\dataset\\another_dataset\\allData.tsv", sep='\t')
print(f"Shape: {df.shape}")
df
```

```
Shape: (31622, 10)
```

| | userID | age | gender | movieID | name | year | genre1 | genre2 | genre3 | rating |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 747 | 1 | F | 1193 | One Flew Over the Cuckoo's Nest | 1975 | Drama | NaN | NaN | 5 |
| 1 | 747 | 1 | F | 661 | James and the Giant Peach | 1996 | Animation | Children's | Musical | 3 |
| 2 | 747 | 1 | F | 2804 | Christmas Story, A | 1983 | Comedy | Drama | NaN | 5 |
| 3 | 747 | 1 | F | 919 | Wizard of Oz, The | 1939 | Adventure | Children's | Drama | 4 |
| 4 | 747 | 1 | F | 2791 | Airplane! | 1980 | Comedy | NaN | NaN | 4 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

Figure 3: Jupyter NB - change path for all_data file

## 7.2 Approach 2 - Google Colab Steps for execution

Before running the code the only requirement is to change the path where the file is being read. Below is the screenshot for places to change the source file path to the one where the user has stored the source files:

```
from google.colab import files
uploaded_files = files.upload()
```

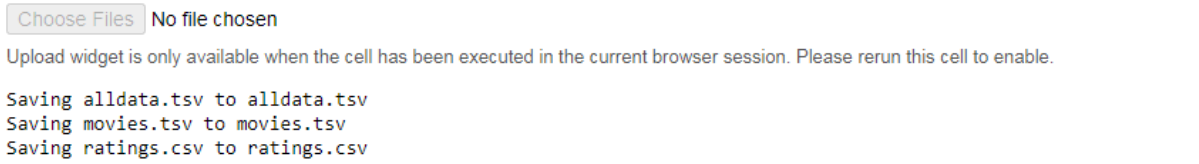



Figure 4: Google Colab - Upload required files to google colab environment

```
import io
df = pd.read_csv(io.StringIO(uploaded_files['alldata.tsv'].decode('utf-8')),sep='\t')
df.head()
```

| | userID | age | gender | movieID | name | year | genre1 | genre2 | genre3 | rating |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 747 | 1 | F | 1193 | One Flew Over the Cuckoo's Nest | 1975 | Drama | NaN | NaN | 5 |
| 1 | 747 | 1 | F | 661 | James and the Giant Peach | 1996 | Animation | Children's | Musical | 3 |
| 2 | 747 | 1 | F | 2804 | Christmas Story, A | 1983 | Comedy | Drama | NaN | 5 |

Figure 5: Google Colab - change path for all_data file

```
movies = pd.read_csv(io.StringIO(uploaded_files['movies.tsv'].decode('utf-8')),sep='\t')
movies.head()
```

| | movieID | name | year | genre1 | genre2 | genre3 |
|---|---|---|---|---|---|---|
| 0 | 1 | Toy Story | 1995 | Animation | Children's | Comedy |
| 1 | 2 | Jumanji | 1995 | Adventure | Children's | Fantasy |

Figure 6: Google Colab - change path for movies file

```
ratings = pd.read_csv(io.StringIO(uploaded_files['ratings.csv'].decode('utf-8')),sep=',')
ratings.head()
```

| | userID | movieID | rating |
|---|---|---|---|
| 0 | 747 | 1193 | 5 |
| 1 | 747 | 661 | 3 |
| 2 | 747 | 2804 | 5 |

Figure 7: Google Colab - change path for ratings file

Once the code is executed the same can be run with a single step by clicking on the :

- For Jupyter NB: Click on Kernel button present on the ribbon of Jupyter NB and then choosing Restart and Run All from the drop down to execute all the cells in which the code is present.
- For Google Colab: Click on Runtime button present on the ribbon of Google Colab and then choosing Run All from the drop down to execute all the steps.

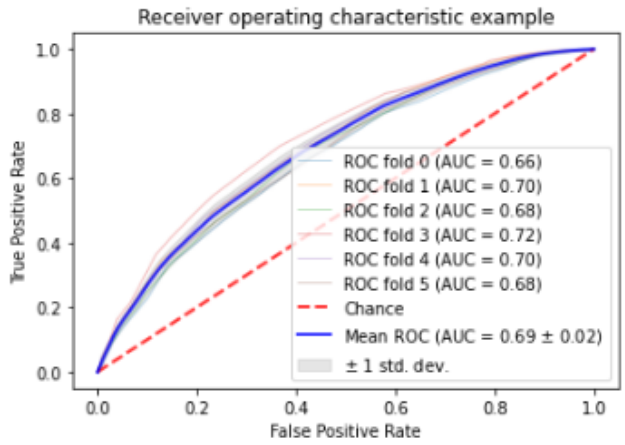
# 8 Evaluation

Once all the cells are executed the output of the code provides movie recommendations for a user whose demographics are Gender - Male and Age - 25.

If this has to be changed for some other demographic then make the change in the cell:

```
print(recommendation_without_user_info_knn(df=df, age=25, gender='M', number_recommendations=10))
## df = alldata.tsv
## age = some age of the new user
## gender = male or female only
## number_recommendations = the recommendations required based on the above
```

```
K Nearest Neighbours values :
Cross Validation Precision: 0.68 (+/- 0.03)
Cross Validation Recall: 0.76 (+/- 0.06)
Cross Validation roc_auc: 0.69 (+/- 0.02)
```




| movieID | name | year | genre1 |
|---|---|---|---|
| 2475 | 52 Pick-Up | 1986 | Action |
| 3654 | Guns of Navarone, The | 1961 | Action |
| 720 | Wallace & Gromit: The Best of Aardman Animation | 1996 | Animation |
| 1279 | Night on Earth | 1991 | Comedy |
| 1805 | Wild Things | 1998 | Drama |
| 21 | Get Shorty | 1995 | Action |
| 2948 | From Russia with Love | 1963 | Action |
| 1727 | Horse Whisperer, The | 1998 | Drama |
| 524 | Rudy | 1993 | Drama |

Figure 8: Output of the run

Here age is the age of the user and can be changed to any age,gender can be Male of Female and is represented as M or F respectively, number_recommendations represents the number of recommendations to be shown to the user. If a profile is present for this combination in the user-item matrix then a suitable recommendation is made else, default content based recommendation is chosen manually.