COMP4434: Big Data Analytics

Final Project

ZHANG Caiqi 18085481d

The Hong Kong Polytechnic University

Content

- Introduction
- Data analysis
  - Many effective visualization methods are used to analyze the data.
- Data preprocessing
  - Feature engineering
  - Word embedding
- Data preprocessing by MapReduce

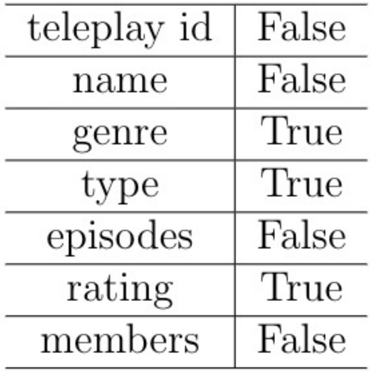
Content

- Models for Task 1
  - Ten traditional regression models
  - Neural network
- Models for Task 2
  - Contend-based recommendation system
  - Collaborative filtering-based recommendation system
  - Hybrid recommendation system using neural network

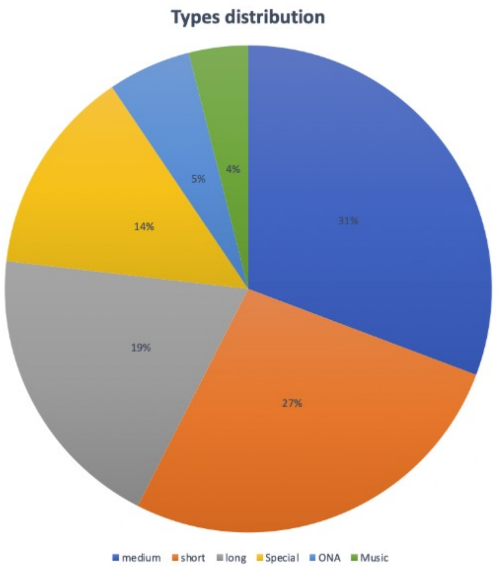
Introduction

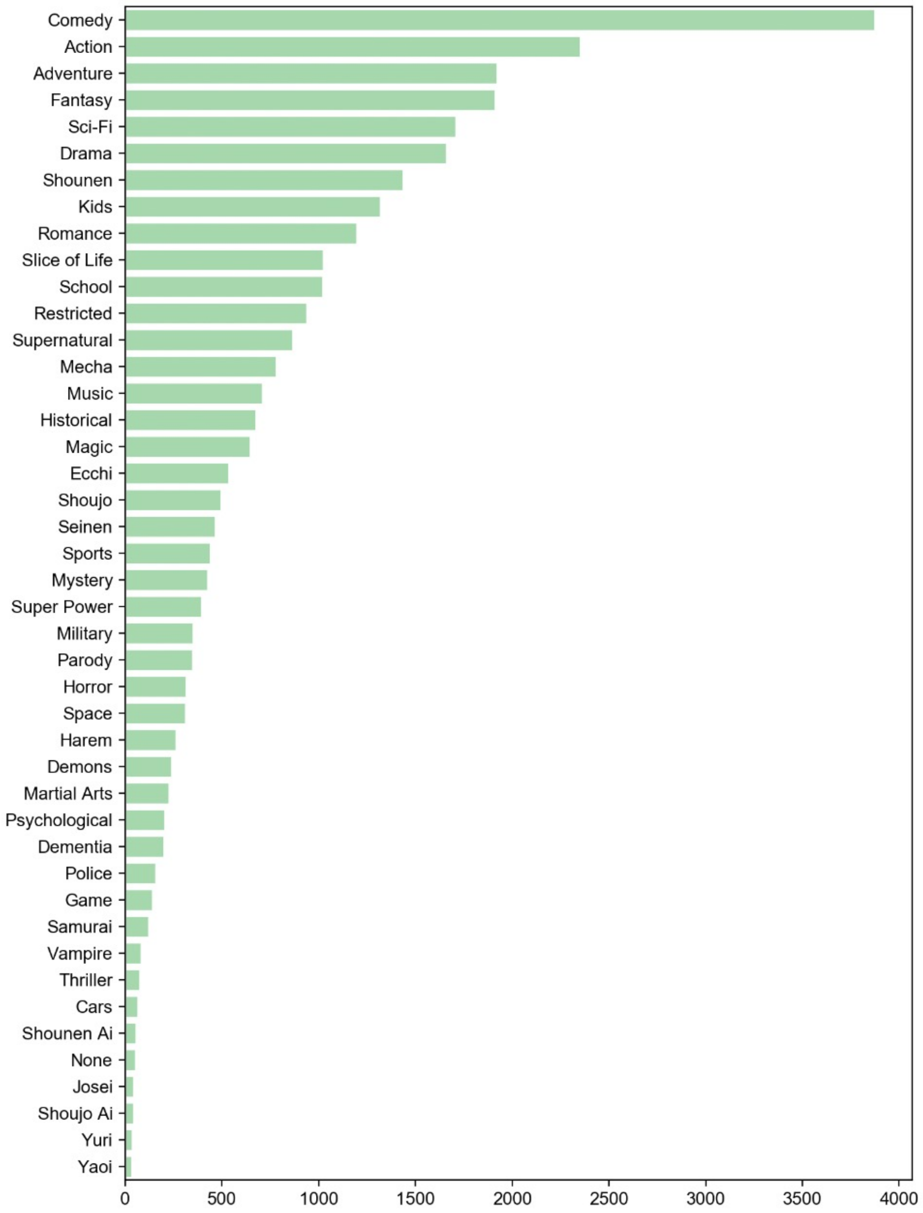
- In this project, we will complete two tasks.
    - Design a prediction model to predict the rating of recently published teleplays.
    - Design a recommendation system to provide personalized recommendation services.
- Details can be found in the report.

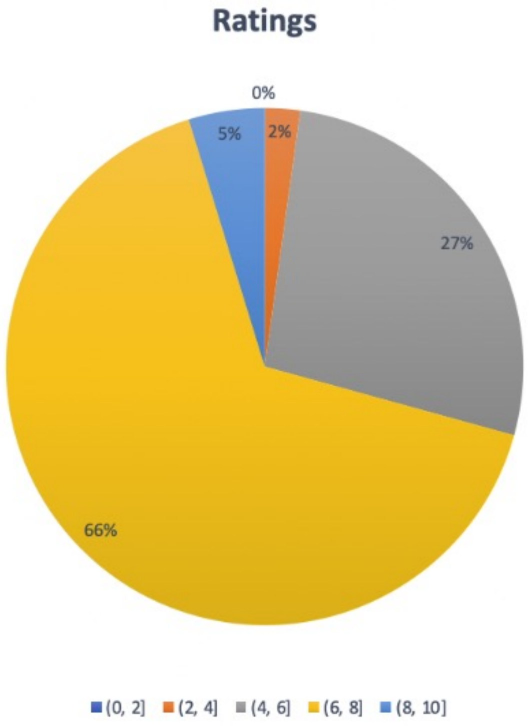
Data analysis

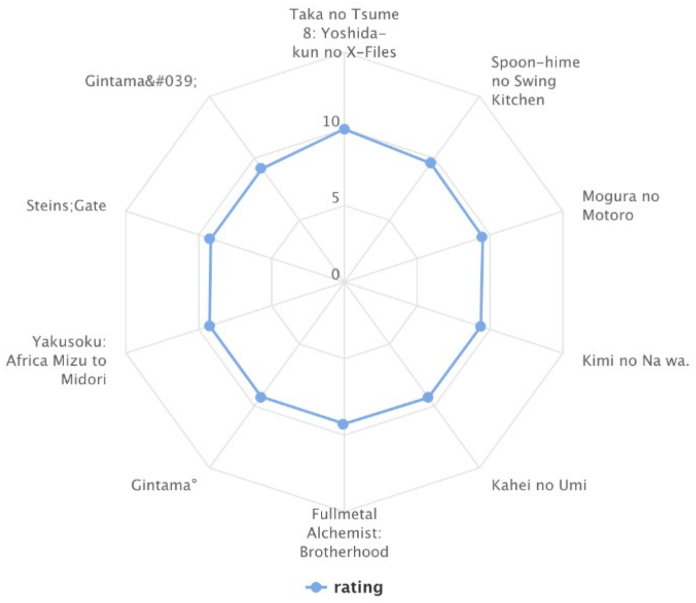


teleplay id	False
name	False
genre	True
type	True
episodes	False
rating	True
members	False









Data analysis

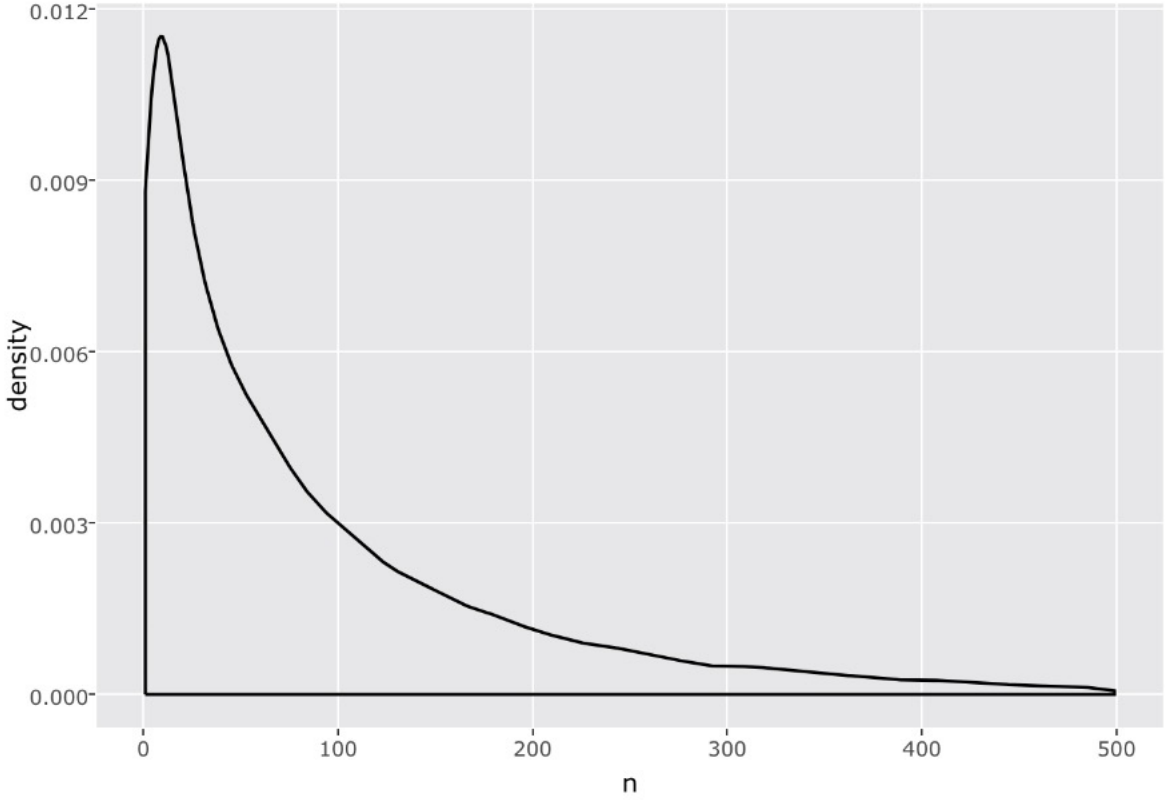


Figure 7: Users' ratings

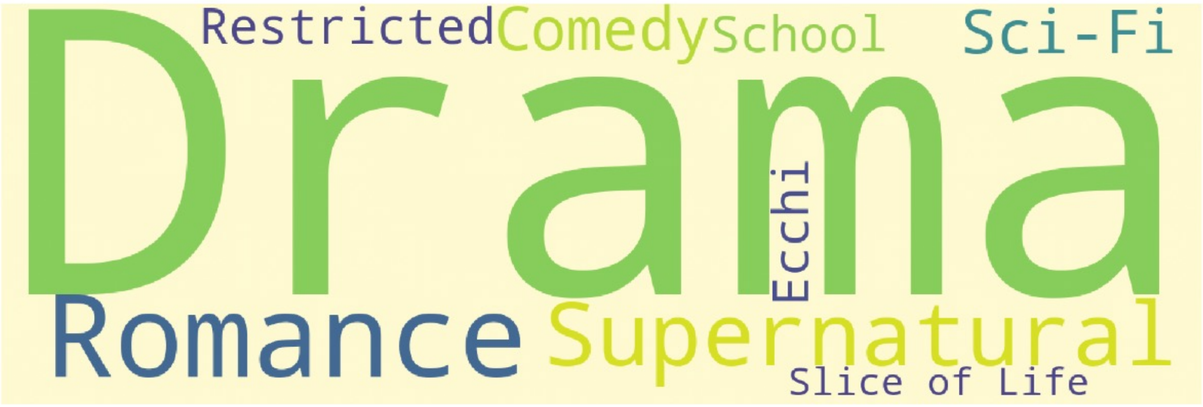
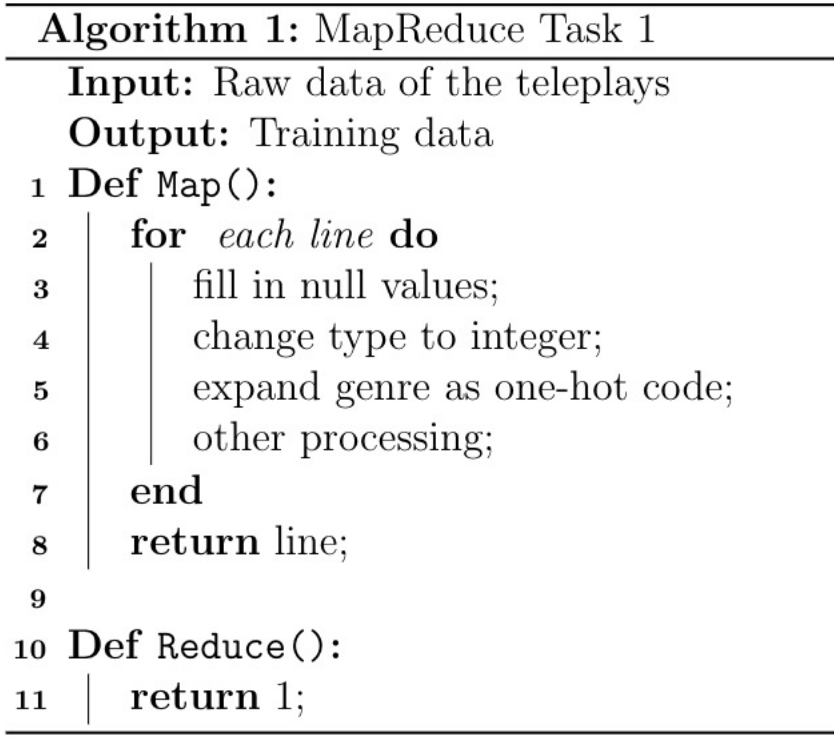


Figure 8: Word cloud of user 53698's preference

Data preprocessing

- Remove null values:
  - If there is no rating of a teleplay, we will remove the entire row.
  - If there is Unknown in the episodes, we will replace it as the average of the episodes.
  - If there is no value in the genre, as it will not influence the training result, we will ignore it.
- Feature engineering:
  - Type: we will use 0, 1, 2, 3, 4, 5
  - Genre: we will adopt the one hot coding
- Word embedding:
  - Every sentence will be transformed to a vector with length 768 using BERT.

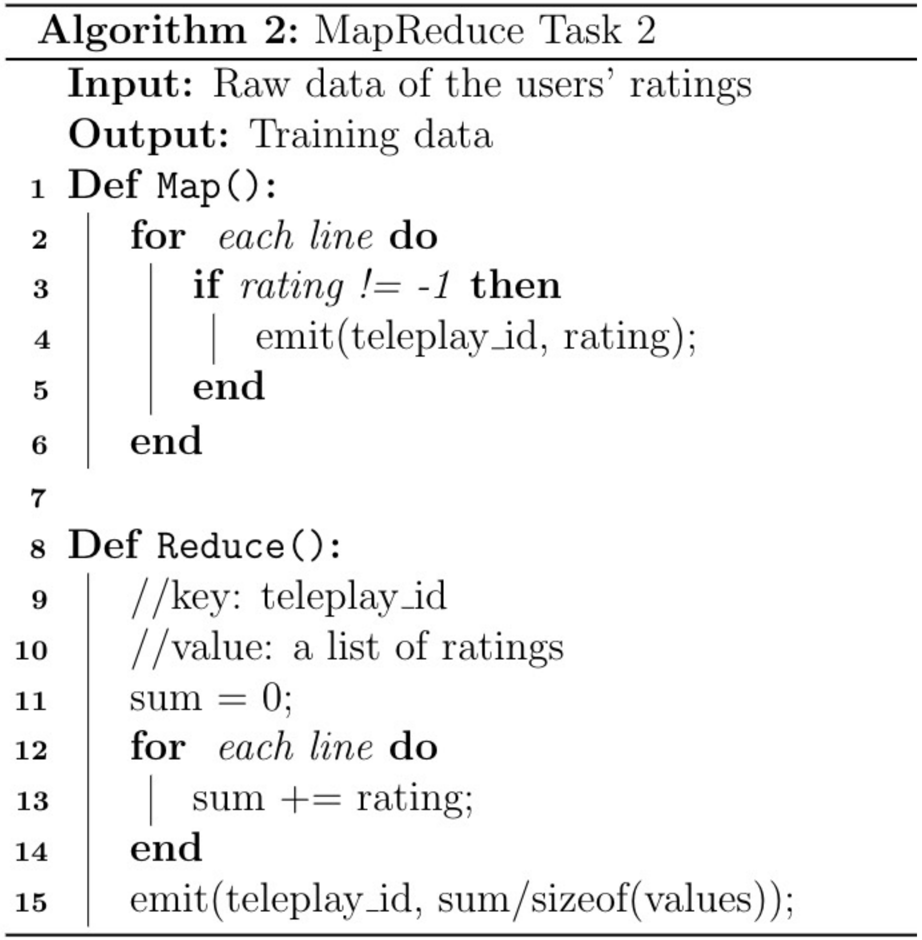
Data preprocessing by MapReduce



Algorithm 1: MapReduce Task 1

```
Input: Raw data of the teleplays
Output: Training data
Def Map():
    for each line do
        fill in null values;
        change type to integer;
        expand genre as one-hot code;
        other processing;
    end
    return line;

Def Reduce():
    return 1;
```



Algorithm 2: MapReduce Task 2

```
Input: Raw data of the users' ratings
Output: Training data
Def Map():
    for each line do
        if rating != -1 then
            emit(teleplay_id, rating);
        end
    end

Def Reduce():
    //key: teleplay_id
    //value: a list of ratings
    sum = 0;
    for each line do
        sum += rating;
    end
    emit(teleplay_id, sum/sizeof(values));
```

Models for Task 1

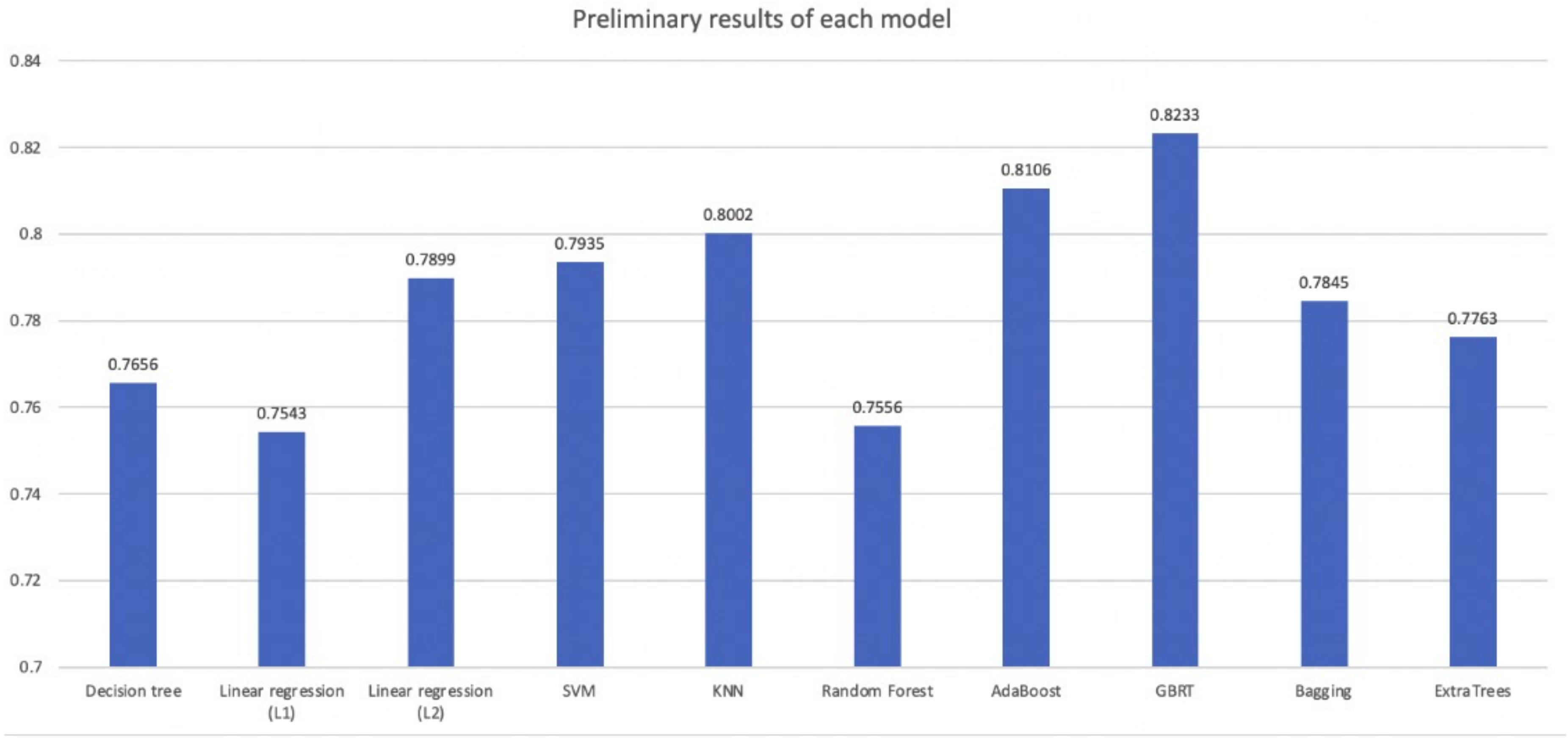
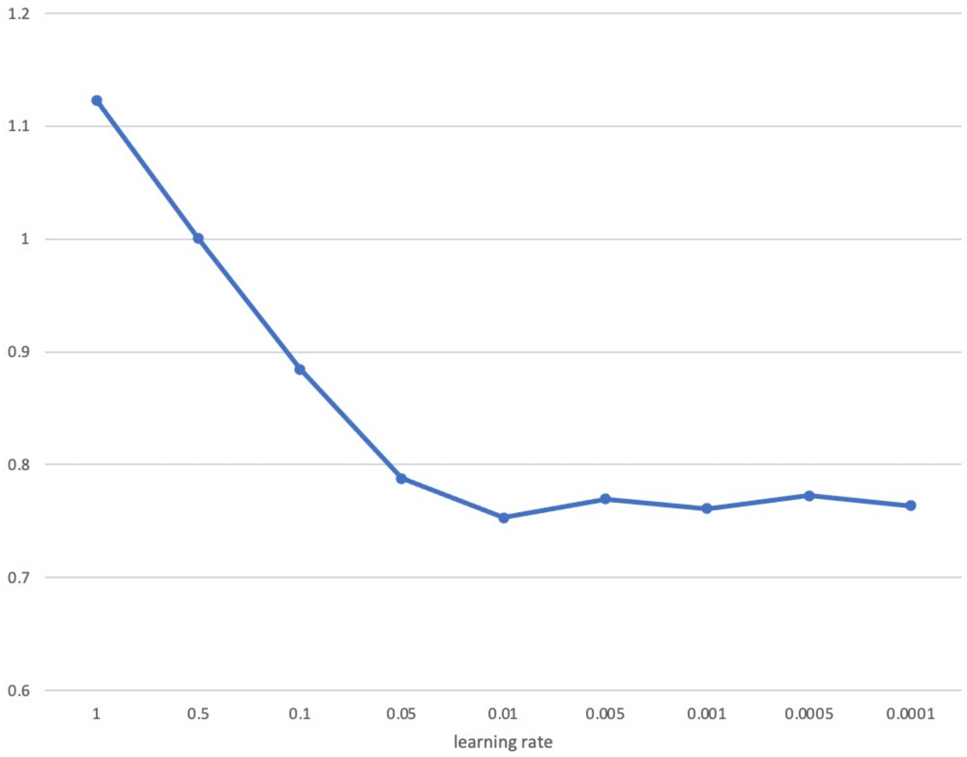
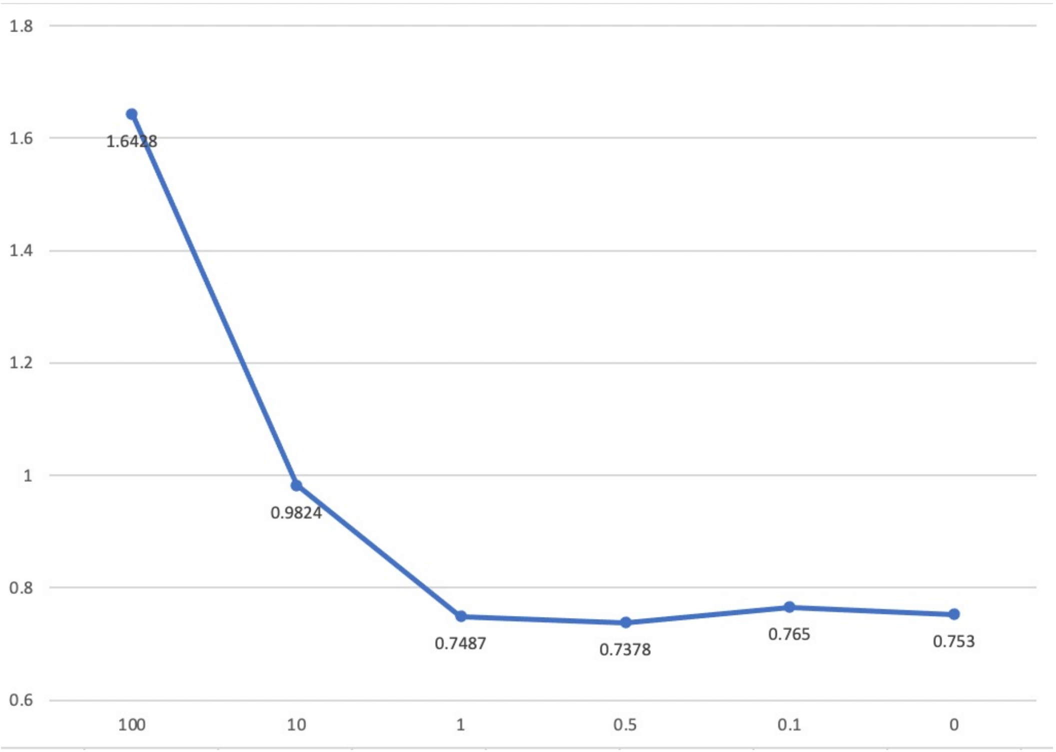


Figure 11: Preliminary experiments

Models for Task 1

Linear regression with L1 regularization

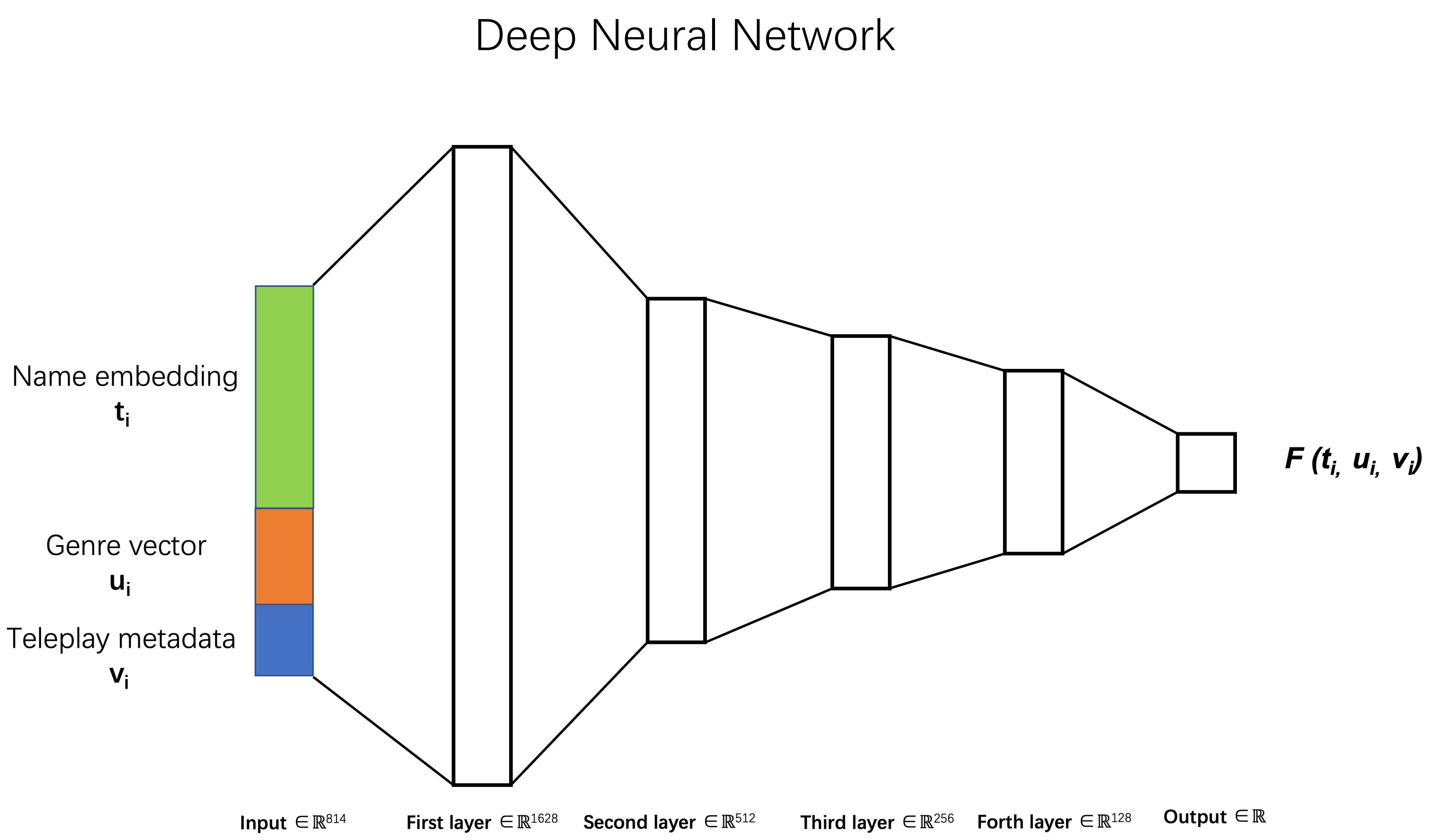




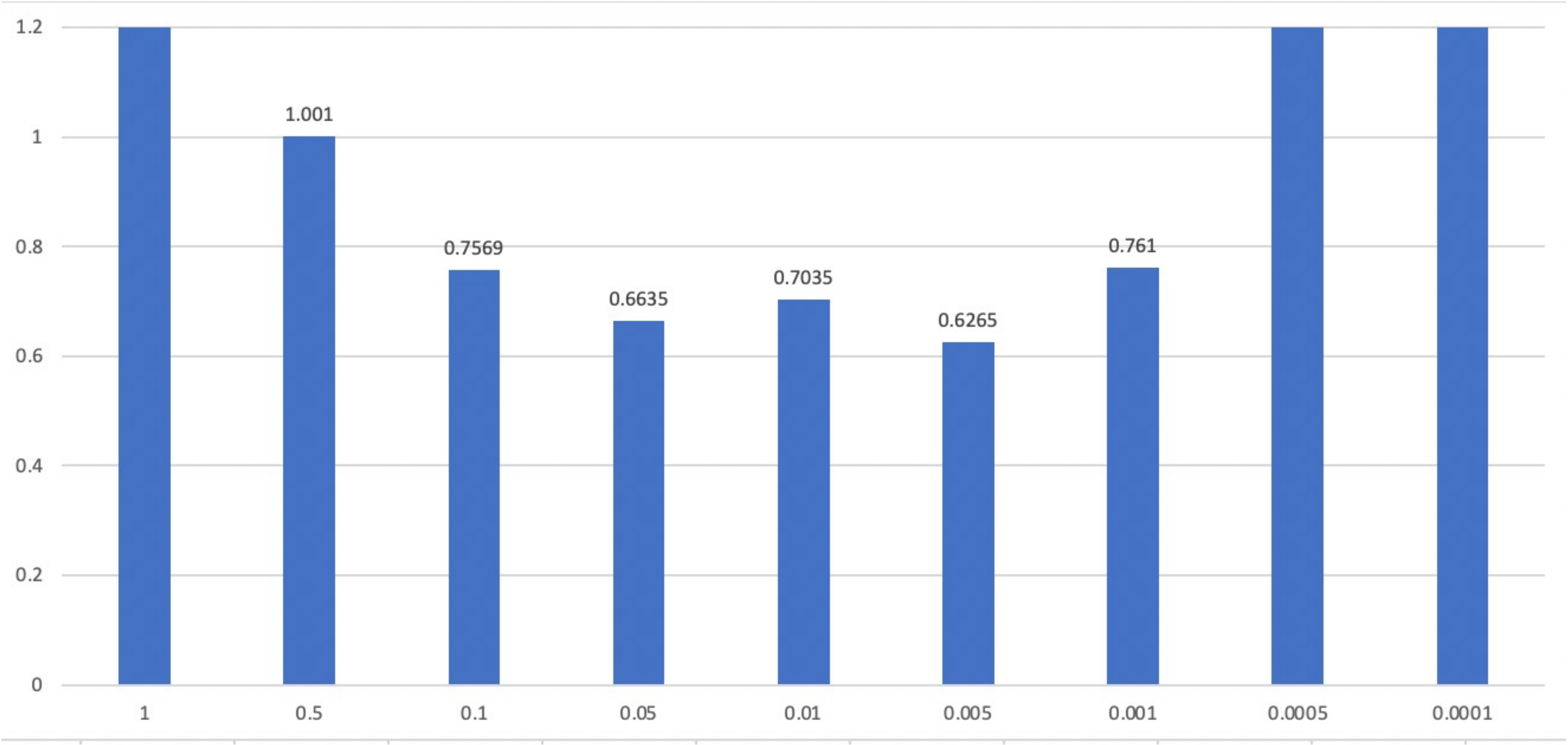
α = 0.01

λ = 0.5

Models for Task 1


Deep Neural Network
Name embedding
$\mathbf{t}_i$
Genre vector
$\mathbf{u}_i$
Teleplay metadata
$\mathbf{v}_i$
$F\ (t_{i,}\ u_{i,}\ v_i)$
Input $\in \mathbb{R}^{814}$
First layer $\in \mathbb{R}^{1628}$
Second layer $\in \mathbb{R}^{512}$
Third layer $\in \mathbb{R}^{256}$
Forth layer $\in \mathbb{R}^{128}$
Output $\in \mathbb{R}$

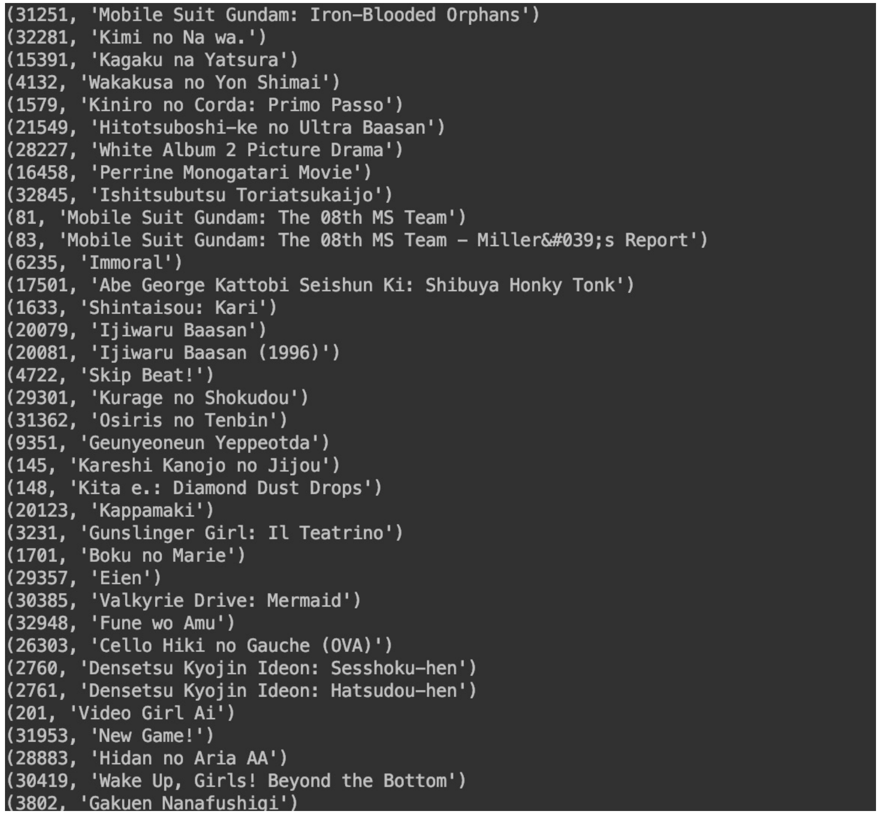
Models for Task 1



Learning rate is set to 0.005.

Models for Task 2: Contend-based

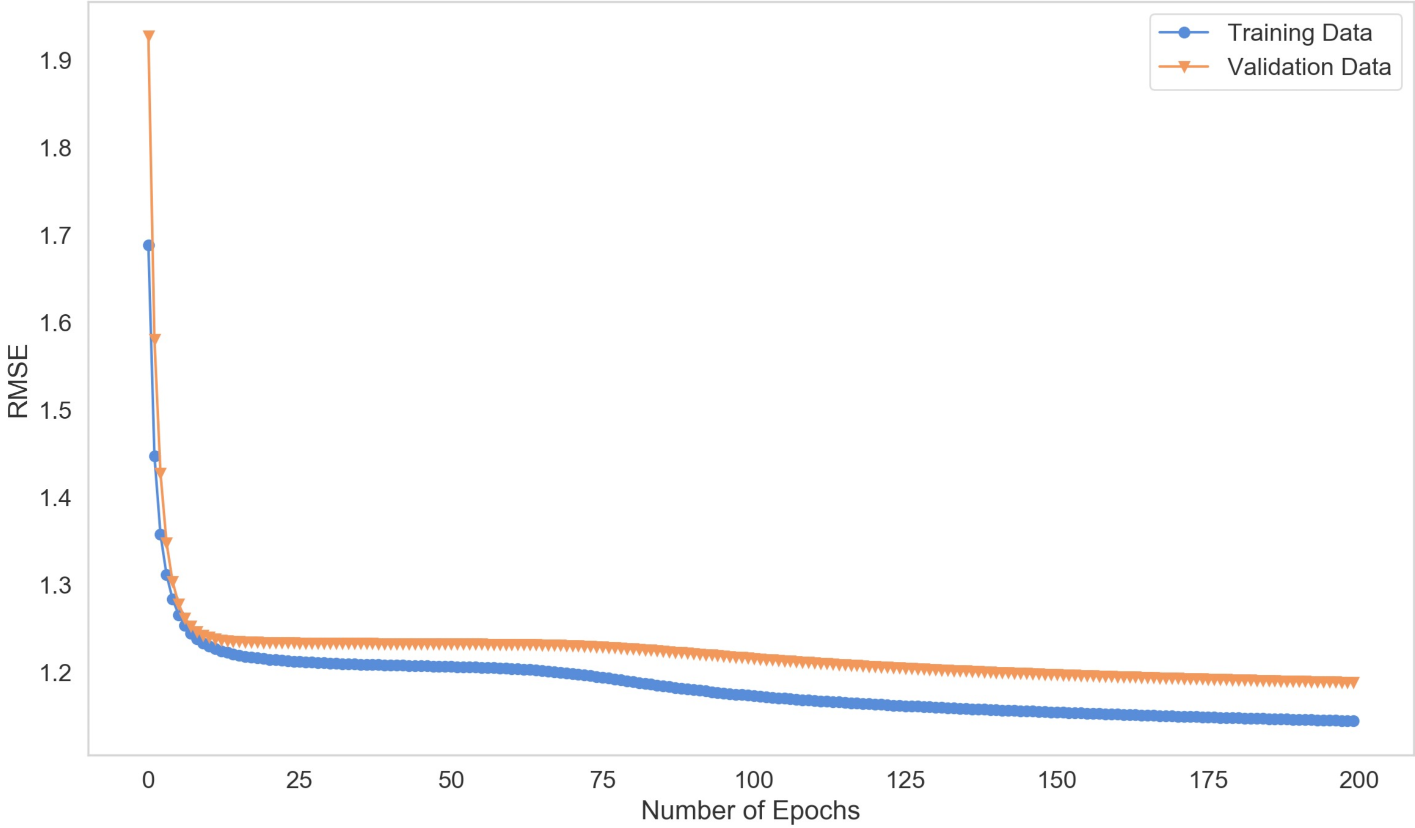
- Calculate the similarity between teleplays. In this case, we use the cosine similarity to calculate the similarities.
  - m2m = cosine_similarity(df_movies_tf_idf_described)
- From existing dataset, find user 53698's favorite teleplay, and based on that, recommend similar teleplays for user 53698.



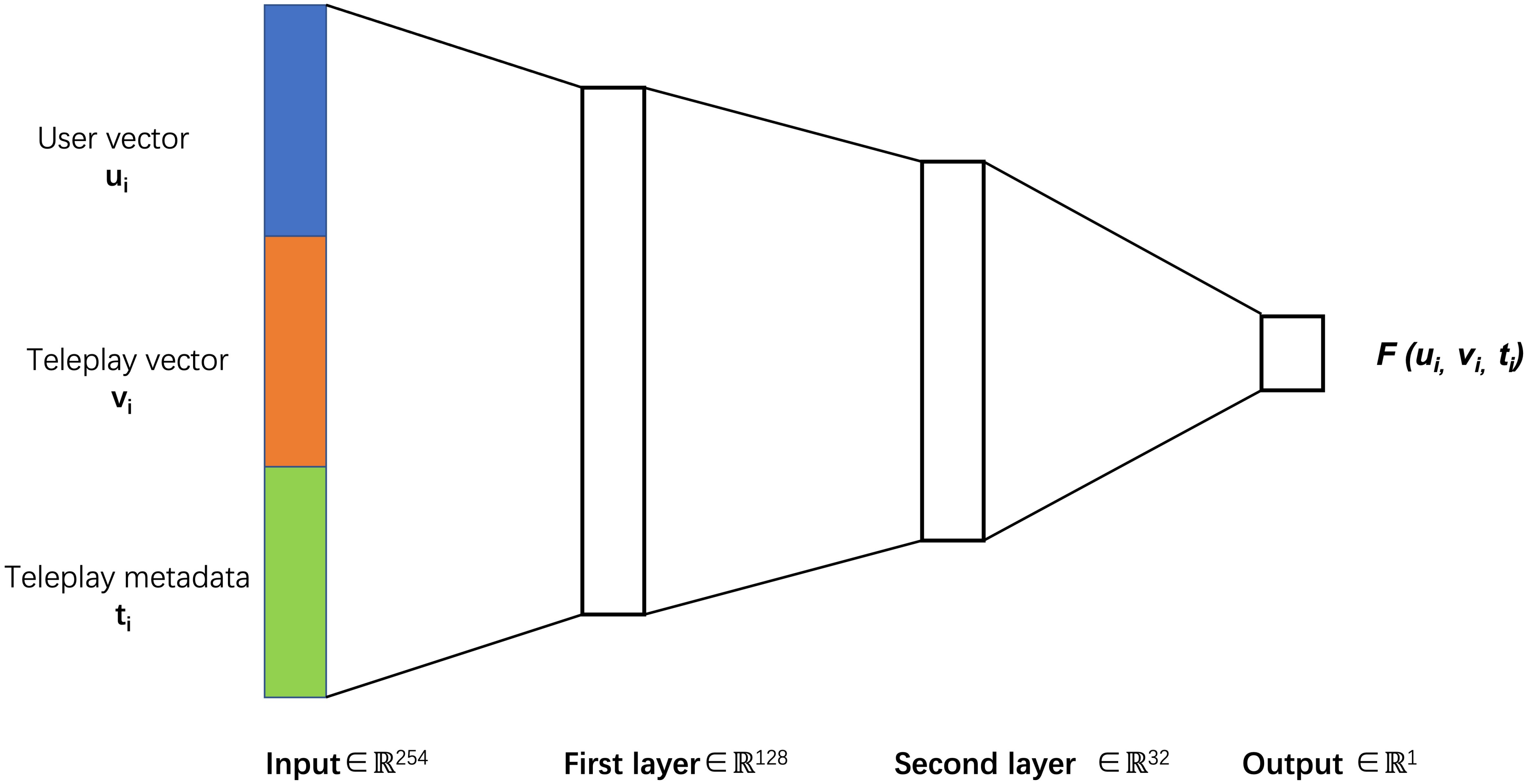
```
(31251, 'Mobile Suit Gundam: Iron-Blooded Orphans')
(32281, 'Kimi no Na wa.')
(15391, 'Kagaku na Yatsura')
(4132, 'Wakakusa no Yon Shimai')
(1579, 'Kiniro no Corda: Primo Passo')
(21549, 'Hitotsuboshi-ke no Ultra Baasan')
(28227, 'White Album 2 Picture Drama')
(16458, 'Perrine Monogatari Movie')
(32845, 'Ishitsubutsu Toriatsukaijo')
(81, 'Mobile Suit Gundam: The 08th MS Team')
(83, 'Mobile Suit Gundam: The 08th MS Team - Miller&#039;s Report')
(6235, 'Immoral')
(17501, 'Abe George Kattobi Seishun Ki: Shibuya Honky Tonk')
(1633, 'Shintaisou: Kari')
(20079, 'Ijiwaru Baasan')
(20081, 'Ijiwaru Baasan (1996)')
(4722, 'Skip Beat!')
(29301, 'Kurage no Shokudou')
(31362, 'Osiris no Tenbin')
(9351, 'Geunyeoneun Yeppeotda')
(145, 'Kareshi Kanojo no Jijou')
(148, 'Kita e.: Diamond Dust Drops')
(20123, 'Kappamaki')
(3231, 'Gunslinger Girl: Il Teatrino')
(1701, 'Boku no Marie')
(29357, 'Eien')
(30385, 'Valkyrie Drive: Mermaid')
(32948, 'Fune wo Amu')
(26303, 'Cello Hiki no Gauche (OVA)')
(2760, 'Densetsu Kyojin Ideon: Sesshoku-hen')
(2761, 'Densetsu Kyojin Ideon: Hatsudou-hen')
(201, 'Video Girl Ai')
(31953, 'New Game!')
(28883, 'Hidan no Aria AA')
(30419, 'Wake Up, Girls! Beyond the Bottom')
(3802, 'Gakuen Nanafushigi')
```

Models for Task 2: Collaborative filtering-based

- CF system will give recommendations to a user based on the preferences of “similar” users and recommendation is dependent on other users’ historical data.



Models for Task 2: Hybrid system using neural network


User vector
$\mathbf{u}_i$
Teleplay vector
$\mathbf{v}_i$
Teleplay metadata
$\mathbf{t}_i$
$F\ (u_i,\ v_i,\ t_i)$
Input $\in \mathbb{R}^{254}$
First layer $\in \mathbb{R}^{128}$
Second layer $\in \mathbb{R}^{32}$
Output $\in \mathbb{R}^{1}$

Thank you!