ABIYAANTRIX & SAPIENCE ACADEMY
INTERNSHIP + TRAINING

Internship Mini Project on

" MOVIE RECOMMENDATION SYSTEM "

Submitted in partial fulfillment towards Mini Project work of Internship

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE ENGINEERING

SUBMITTED BY

SANATH G	**4CA15CS023**
RIM NIKHATH	**4CA15CS022**
UM-E HANI	**4CA15CS029**
MANOJ KUMAR C S	**4CA15CS402**

UNDER THE GUIDANCE OF

Mrs. ANJANA SHASHI KIRAN	**Mr. SRIKRISHNAKASHYAP**
DIRECTOR	**INTERNSHIP TRAINER**

ACKNOWLEDGMENT

We sincerely owe our gratitude to all the persons who helped and guided us in Completing this mini-project.

We are thankful to **Mrs. Anjana Shashi Kiran**, *Honorary Director,* Abiyaantrix Tech Solutions, Mysuru, for having supported in our academic endeavours.

We are thankful to **Mr. Srikrishna S Kashyap**, *Trainer,* Abiyaantrix Tech Solutions, Mysuru, for all the support he has rendered.

We are extremely pleased to thank our parents, family members and friends for their continuous support, inspiration and encouragement, for their helping hand and also last but not the least, We thank all the members who supported directly or indirectly in this internship process.

SANATH G

RIM NIKHATH

UM-E HANI

MANOJ KUMAR C S

ABSTRACT

It is the process by which order, structure and meaning are given to the data (information).

It consists in transforming the collected data into useful and true conclusions and or lessons.

From the pre-established topics, the data are processed, looking for trends, differences and variations in the information obtained.

The processes, techniques and tools used are based on certain assumptions and as such have limitations.

The processes is used to describe and summarize the data, identify the relationships and differences between variables, compare variables and make predictions.

CONTENTS

INTRODUCTION

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to data mining.

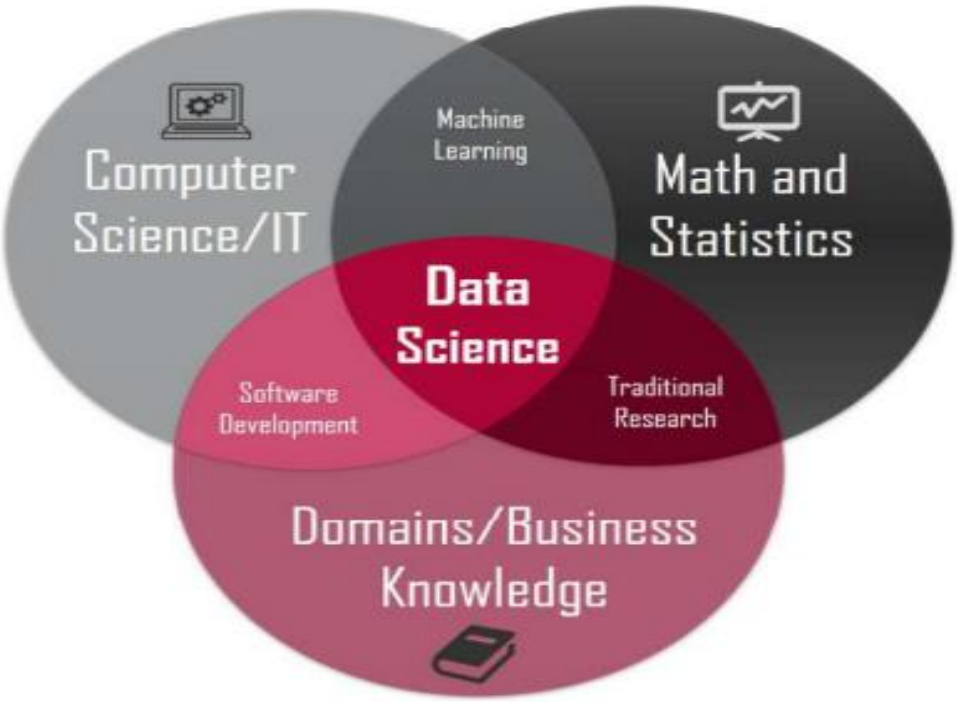
Data science is a "concept to unify statistics, data analysis, machine learn in gand their related methods" in order to "understand and analyse actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.

HISTORY

The term "data science" has appeared in various contexts over the past thirty years but did not become an established term until recently. In an early usage it was used as a substitute for computer science by Peter Naur in 1960. Naur later introduced the term "datalogy". In 1974, Naur published *Concise Survey of Computer Methods*, which freely used the term data science 1n its survey of the contemporary data processing methods that are used in a wide range of applications.

RELATIONSHIP TO STATISTICS



The popularity of the term "data science" has exploded in business environments and academia, as indicated by a jump in job openings. However, many critical academics and journalists see no distinction between data science and statistics. Writing in Forbes, Gil Press argues that data science is a buzzword without a clear definition and has simply replaced “business analytics” in contexts such as graduate degree programs. In the question-and answer section of his keynote address at the Joint Statistical Meetings of American Statistical Association, noted applied statistician Nate Silver said, “I think data-scientist is a sexed up term for a statistician....Statistics is a branch of science. Data scientist is slightly redundant in some way and people shouldn’t berate the term statistician.”Similarly, in business sector, multiple researchers and analysts state that data scientists alone are far from being sufficient in granting companies a real competitive advantage and consider data scientists as only one of the four greater job families companies require to leverage big data effectively, namely: data analysts, data scientists, big data developers and big data engineers. On the other hand, responses to criticism are as numerous. In a 2014 Wall Street Journal article, Irving Wladawsky Berger compares the data science enthusiasm with the dawn of computer science. He argues data science, like any other interdisciplinary field, employs methodologies and practices from across the academia and industry, but then it will morph them into a new discipline. He brings to attention the sharp criticisms computer science, now a well-respected academic discipline, had to once face. Likewise, NYUStern's Vasant Dhar, as do many other academic

proponents of data science, argues more specifically in December 2013 that data science is different from the existing practice of data analysis across all disciplines, which focuses only on explaining data sets. Data science seeks actionable and consistent pattern for predictive uses. This practical engineering goal takes data science beyond traditional analytics. Now the data in those disciplines and applied fields that lacked solid theories, like health science and social science, could be sought and utilized to generate powerful predictive models.

SYSTEM REQUIREMENT AND SPECIFICATION

2.1HardwareRequirements:

- Intel® Pentium 4 CPU and higher versions e 256 MBRAM,
- 80GBHDD Mouse
- QWERTYKeyboard
- Standard VGAMonitor

2.2SoftwareRequirements:

- OperatingSystem: WINDOWS 10,7
- Language :PYTHON e Tool : JUPYTER NOTEBOOK

TESTING AND RESULTS

The full creating and implementing Chicago Crime Dataset using python, in which we have used various python modules. Modules included pandas to handle dataframes, matplotlib to plot a graph ,numpy.

UNIT TESTING

Here the individual components are tested to ensure that they operate correctly. Each component is tested independently, without other system components.

MODULE TESTING

Module is a collection of dependent components such as procedures and functions. Since the module encapsulates related components can be tested with our other system modules. The testing process is concerned with finding errors which results from erroneous function calls from the main function to various individual functions.

SYSTEM TESING

The Modules are integrated to make up the entire system. The testing process is concerned with finding errors with the results from unanticipated interactions between module and system components. It is also concerned with validating that the system meets its functional and non-functional requirements.

IMPLEMENTATION

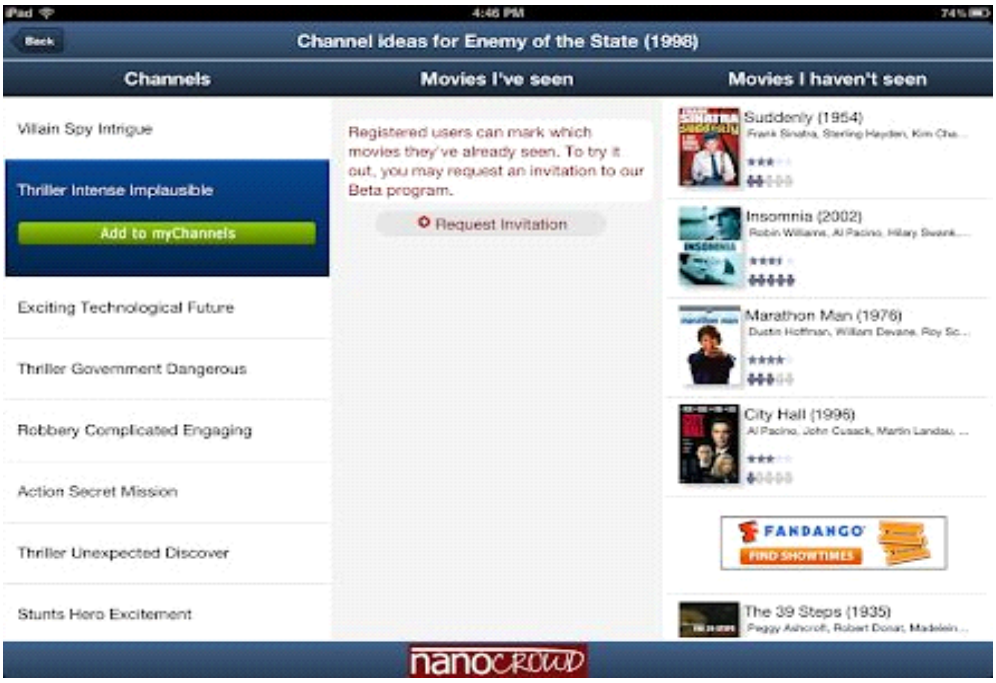
```
import numpy as np
import pandas as pd


names = ['user_id', 'item_id', 'rating', 'timestamp']
df =pd.read_csv('ml-100k/u.data', sep='\t', names=names)
df.head()
n_users = df.user_id.unique().shape[0]
n_items = df.item_id.unique().shape[0]
print str(n_users) + ' users'
print str(n_items) + ' items'
ratings = np.zeros((n_users, n_items))
for row in df.itertuples():
    ratings[row[1]-1, row[2]-1] = row[3]
ratings
sparsity = float(len(ratings.nonzero()[0]))
sparsity /= (ratings.shape[0] * ratings.shape[1])
sparsity *= 100
print 'Sparsity: {:4.2f}%'.format(sparsity)
def train_test_split(ratings):
    test = np.zeros(ratings.shape)
    train = ratings.copy()
    for user in xrange(ratings.shape[0]):
        test_ratings = np.random.choice(ratings[0, :].nonzero()[0],
                                        size=10,
                                        replace=False)
        train[user, test_ratings] = 0.
```

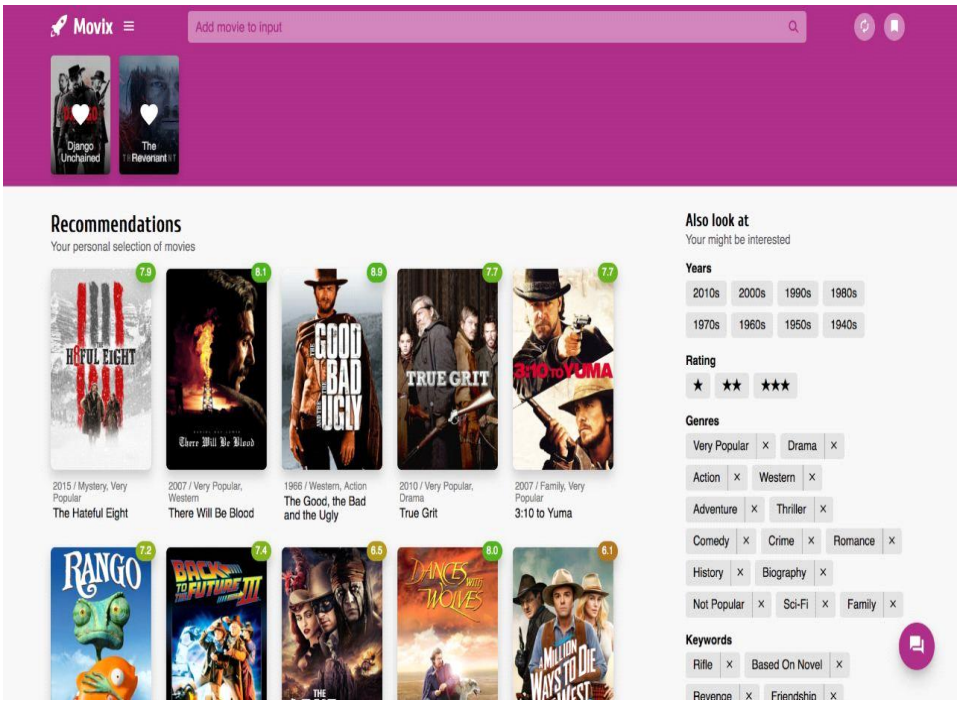
```
    test[user, test_ratings] = ratings[user, test_ratings]
```

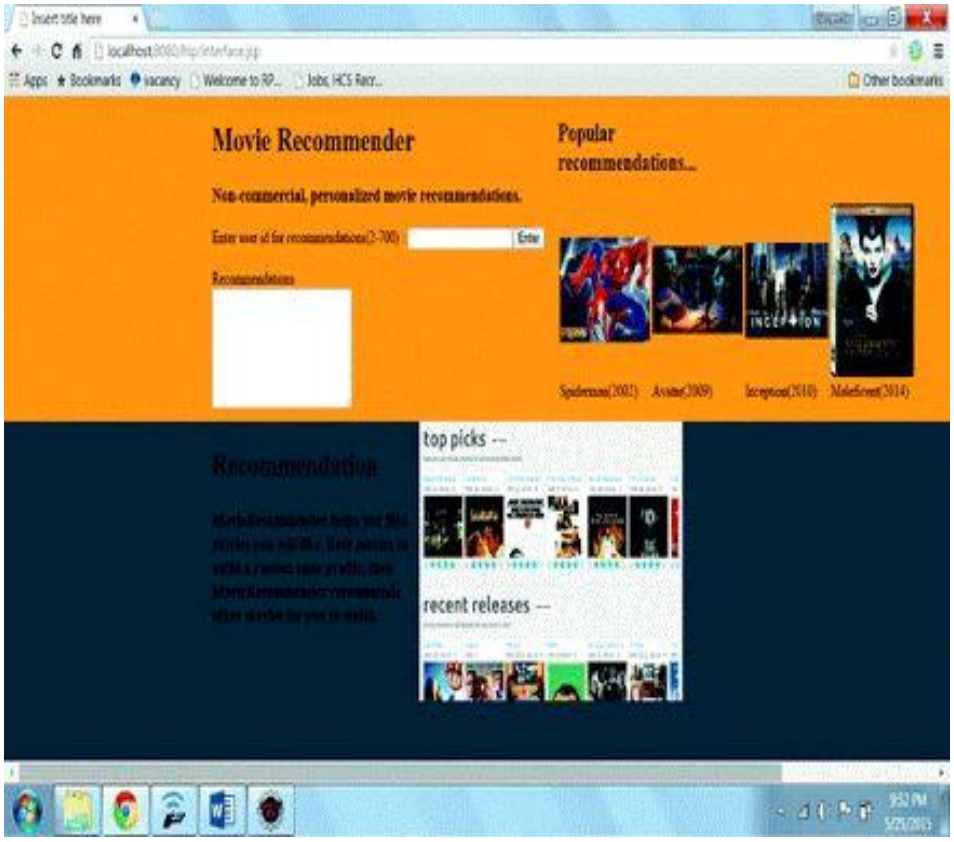
```
import requests
import json

response = requests.get('http://us.imdb.com/M/title-exact?Toy%20Story%20(1995)')
print response.url.split('/')[-2]

# Get base url filepath structure. w185 corresponds to size of movie poster.
headers = {'Accept': 'application/json'}
payload = {'api_key': 'Plz insert your key here '}
response = requests.get("http://api.themoviedb.org/3/configuration", params=payload, headers=headers)
response = json.loads(response.text)
base_url = response['images']['base_url'] + 'w185'
def get_poster(imdb_url, base_url):

 # Get IMDB movie ID
    response = requests.get(imdb_url)
    movie_id = response.url.split('/')[-2]

    # Query themoviedb.org API for movie poster path.
    movie_url = 'http://api.themoviedb.org/3/movie/{:}/images'.format(movie_id)
    headers = {'Accept': 'application/json'}
    payload = {'api_key': 'INSERT API_KEY HERE'}
    response = requests.get(movie_url, params=payload, headers=headers)
```

```
    try:
        file_path = json.loads(response.text)['posters'][0]['file_path']
    except:
        # IMDB movie ID is sometimes no good. Need to get correct one.
        movie_title = imdb_url.split('?')[-1].split('(')[0]
        payload['query'] = movie_title
        response = requests.get('http://api.themoviedb.org/3/search/movie', params=payload,
headers=headers)
        movie_id = json.loads(response.text)['results'][0]['id']
        payload.pop('query', None)
        movie_url = 'http://api.themoviedb.org/3/movie/{:}/images'.format(movie_id)
        response = requests.get(movie_url, params=payload, headers=headers)
        file_path = json.loads(response.text)['posters'][0]['file_path']
       return base_url + file_path
       toy_story = 'http://us.imdb.com/M/title-exact?Toy%20Story%20(1995)'


# Load in movie data
    idx_to_movie = {}
   with open('ml-100k/u.item', 'r') as f:
    for line in f.readlines():
        info = line.split('|')
        idx_to_movie[int(info[0])-1] = info[4]
   def top_k_movies(similarity, mapper, movie_idx, k=6):
   return [mapper[x] for x in np.argsort(similarity[movie_idx,:])[:-k-1:-1]]
   idx = 0 # Toy Story
  movies = top_k_movies(item_similarity, idx_to_movie, idx)
```

SNAPSHOTS


iPad
4:46 PM
74%
Back
Channel ideas for Enemy of the State (1998)
Channels
Movies I've seen
Movies I haven't seen
Villain Spy Intrigue
Thriller Intense Implausible
Add to myChannels
Exciting Technological Future
Thriller Government Dangerous
Robbery Complicated Engaging
Action Secret Mission
Thriller Unexpected Discover
Stunts Hero Excitement
Registered users can mark which movies they've already seen. To try it out, you may request an invitation to our Beta program.
Request Invitation
Suddenly (1954)
Insomnia (2002)
Marathon Man (1976)
City Hall (1996)
FANDANGO
FIND SHOWTIMES
The 39 Steps (1935)
nanoCROWD


Movix
Add movie to input
Django Unchained
The Revenant
Recommendations
Your personal selection of movies
2015 / Mystery, Very Popular
The Hateful Eight
2007 / Very Popular, Western
There Will Be Blood
1966 / Western, Action
The Good, the Bad and the Ugly
2010 / Very Popular, Drama
True Grit
2007 / Family, Very Popular
3:10 to Yuma
Also look at
Your might be interested
Years
2010s
2000s
1990s
1980s
1970s
1960s
1950s
1940s
Rating
Genres
Very Popular
Drama
Action
Western
Adventure
Thriller
Comedy
Crime
Romance
History
Biography
Not Popular
Sci-Fi
Family
Keywords
Rifle
Based On Novel
Revenge
Friendship


Insert title here
Apps
Bookmarks
vacancy
Welcome to RP...
Jobs, HCS Recr...
Other bookmarks
Movie Recommender
Non-commercial, personalized movie recommendations.
Enter user id for recommendations(2-700)
Enter
Recommendations
Popular recommendations...
Spiderman(2002)
Avatar(2009)
Inception(2010)
MaleficentI(2014)
Recommendation
top picks --
recent releases --

FUTURE ENHANCEMENTS

In order to achieve mastery over working with abundant data, this data set can serve as the ideal stepping stone in the pursuit of tackling mountainous data. We can implement this Dataset with R language which is more powerful compared to python.

CONCLUSION

An overwhelming expansion of data archives posed a challenge to various industries, as these are now struggling to make use of such enormous amount of information. Almost 90% of all data ever recorded worldwide has been created in the last decade alone.

In this project we have explored the data and it provides the insights and forecasts about movie recommendation. It extracts the data from the movie data set and helps in recommending movies to the user in an effective way.

BIBLIOGRAPHY

Newspaper article

Wikipedia.org

Techopedia.com Stackoverflow.com

Quora.com Elitedatascience.com/data-cleaning Trifecta.com

Mean.io Searchbusinessanalytics.techtarget.com Optimizely.com Geeksforgeeks.com

Youtube.com

Studytonight.com

Techterms.com

Google images

Codingdojo.com

Github.com

Hackermoon.com