

E6893 Big Data Analytics:

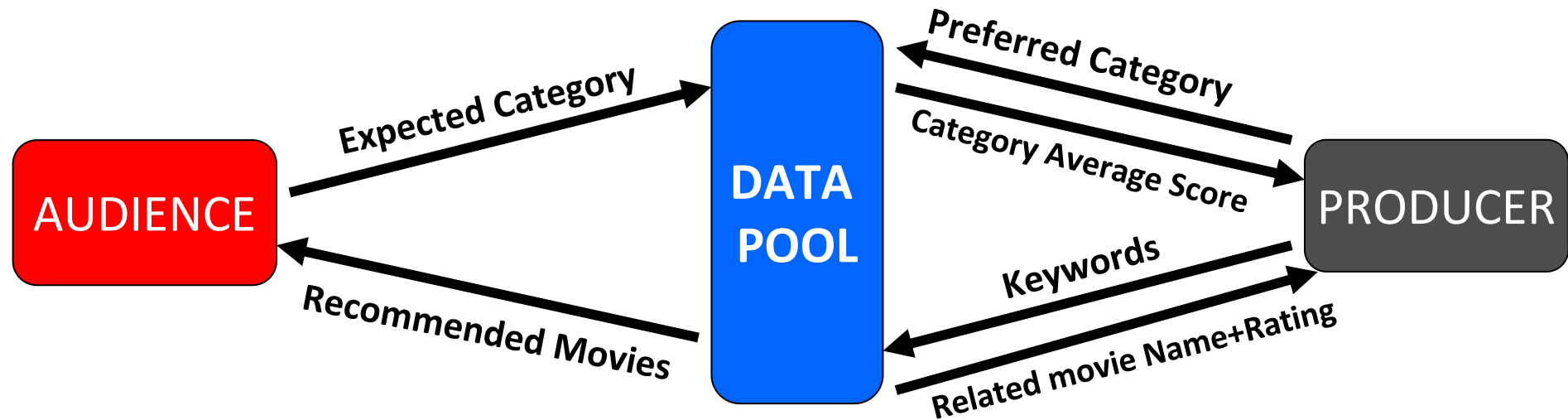
Movie Exploration

Team Members: *Yongjie Cao yc2978*
 Fengyi Song fs2523
 Yuzhe Shen ys2821
 Hui Zou hz2361



Dec 21, 2014

Introduction



Dataset:

Name	Source	Quantity	Content
movie_review	Amazon	7911684 reviews	movie ID, score, comments

Methodology Overview

➤ Process/Extract the original dataset

➤ Recommendation

➤ Cluster

➤ Evaluation and Prediction



➤ Dataset Field

Commenter ID, Movie ID, helpfulness,
movie rating, word comments...

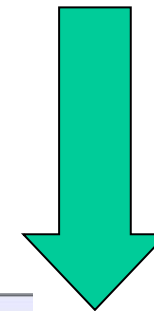
```
product/productId: B003AI2VGA
review/userId: A141HP4LYPWMSR
review/profileName: Brian E. Erland "Rainbow Sphinx"
review/helpfulness: 7/7
review/score: 3.0
review/time: 1182729600
review/summary: "There Is So Much Darkness Now ~ Come For The Miracle"
review/text: Synopsis: On the daily trek from Juarez, Mexico to El Paso,
ever increasing number of female workers are found raped and murdered in
surrounding desert. Investigative reporter Karina Danes (Minnie Driver)
Los Angeles to pursue the story and angers both the local police and the
```







➤ Movie

movie number, movie ID
movie number, user number, rating
movie ID, all comments about it

1	1	3.0
2	1	3.0
3	1	5.0
4	1	3.0
5	1	3.0

A141HP4LYPWMSR	1
A328S9RN3U5M68	2
A1I7QGUDP043DG	3
A1M5405JH9THP9	4
ATXL536YX71TR	5



 073600193X.txt
 076013393X.txt
 076780046X.txt
 076782184X.txt
 078401115X.txt
 080010255X.txt

Recommendation

- User based Recommendation
 - Similarity among filtered professional commenters
 - Recommender

```
import org.apache.mahout.cf.taste.impl.recommender.GenericUserBasedRecommender;
import org.apache.mahout.cf.taste.impl.similarity.PearsonCorrelationSimilarity;
import org.apache.mahout.cf.taste.impl.similarity.LogLikelihoodSimilarity;
import org.apache.mahout.cf.taste.model.DataModel;
import org.apache.mahout.cf.taste.neighborhood.UserNeighborhood;
import org.apache.mahout.cf.taste.recommender.RecommendedItem;
import org.apache.mahout.cf.taste.recommender.UserBasedRecommender;
import org.apache.mahout.cf.taste.similarity.UserSimilarity;

/**
 * Hell
 */
public class App {
    public static void main( String[] args ) throws IOException, TasteException
    {
        DataModel model = new FileDataModel(new File("/home/bigdata/Documents/eclips
        UserSimilarity similarity = new LogLikelihoodSimilarity(model);
```



```
RecommendedItem[item:1800183448, value:13.0]
RecommendedItem[item:1800106663, value:12.813395]
RecommendedItem[item:1800233344, value:12.642813]
RecommendedItem[item:1800340322, value:12.59749]
Hello World!
```

- Item based Recommendation
 - Similarity among filtered good movies
 - Recommender

```
6
[1800147492:12.0,1800353749:12.0,1802443561:12.0,1800121659:12.0,1800061638:12.0,1807858489:12.0,1800421139:12.0,18000225
81:12.0,1800042361:12.0,1802813191:12.0]
9
[1800367822:13.0,1800193490:13.0,1800018659:13.0,1800352372:13.0,1800169924:13.0,1804476866:13.0,1800034427:13.0,18001820
38:13.0,1800020333:12.670389,1800174422:12.668768]
15
[1808416489:10.506992,1807799951:10.506992,1808451090:10.502693,1800047114:10.502693,1800310203:10.502693,1808427977:10.5
02693,1800226298:10.502693,1807812168:10.467674,1800355643:10.455473,1808404537:10.451925]
19 [1807829081:8.645335,1808405757:8.360735,1800107744:8.345321,1808427564:8.345321,1808404470:8.264284]
22
[1800177920:12.0,1800111267:12.0,1800019246:12.0,1800042988:12.0,1800123803:12.0,1800235870:12.0,1800191133:12.0,18000497
01:12.0,1800294043:12.0,1800019098:12.0]
```


➤ Keywords processing

```
:VL-8545{n=2042 c=[0:0.028, 0.006913:0.005, 0.01:0.013, 0.05:0.004, 0.06:0.005, 0.1:0.003, 0.15:0.008
Top Terms:
said          => 2.5476569668202864
he            => 2.4511873209511497
would         => 2.2933491185876695
have          => 1.7944544532740385
trade         => 1.6707300046982425
which         => 1.5705148985757182
u.s           => 1.5347216478407557
new           => 1.5243509306146397
market        => 1.4609101353382854
has           => 1.4473888340472707
countries     => 1.3803546244203753
told          => 1.3565561558894383
meeting       => 1.3506609953653333
had           => 1.3393660824632785
minister      => 1.3153172033190377
from          => 1.3122743489576483
its           => 1.3029065475407824
banks         => 1.2825958407707476
we            => 1.2438235957755632
last          => 1.2385083851454657
```

```
while(*buf_end){
    if (
        buf_end - buf >= strlen(productId) && match_start(buf_end,
        productId, strlen(productId)) ||
        buf_end - buf >= strlen(userId) && match_start(buf_end, userId,
        strlen(userId)) ||
        buf_end - buf >= strlen(help) && match_start(buf_end, help,
        strlen(help)) ||
        buf_end - buf >= strlen(score) && match_start(buf_end, score,
        strlen(score))
    ) {
        unfinished = 1;
        buf_head = buf_end;
        byte_len_wr = 0;
        if (buf_end - buf >= strlen(score) && match_start(buf_end, score
        , strlen(score))) {
            is_score = 1;
        }
    }
}
```

➤ Keywords selection

scientific: LASER , STAR , lightsaber, Skywalker, the Force, ROBOT
month, dragon, UFO, tank

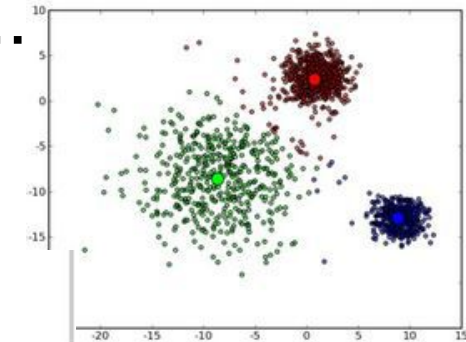
war: battlefield, troops, warfare, conflict, rival, military, wartime,
combat, weapon

Clustering

➤ Movie Cluster

Cluster movies of similar contents, category...

Cluster points: individual comments.



Main Film Genres (click here)		Film Sub-Genres (click here)	
	Action Films		Biographical Films (or "Biopics")
	Adventure Films		'Chick' Flicks
	Comedy Films		Detective & Mystery Films
	Crime & Gangster Films		Disaster Films
	Drama Films		Fantasy Films

Prediction

➤ Predict

movie score

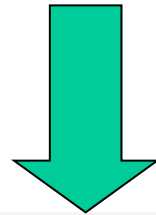
popularity trend

- ☐ adventure
- ☐ blackmail
- ☐ cartoon
- ☒ chaplin
- ☐ disney
- ☐ empires
- ☐ Epics
- ☐ gangsters
- ☐ ghosts
- ☐ horseback

submit

- ☐ kungfu
- ☐ legendary
- ☐ light-hearted
- ☒ musical
- ☐ poppins
- ☒ relish
- ☐ strength
- ☐ workout
- ☐ zombie
- ☐ flood

- ☐ historical
- ☐ mathematician
- ☐ monarchy
- ☐ monsters
- ☐ planets
- ☐ prince
- ☒ spy
- ☒ supernatural
- ☐ triumphs
- ☐ vies



Hello, hahaha!

You can use up to ten key words as you like. the score according to your keywords selection is 3.0625703

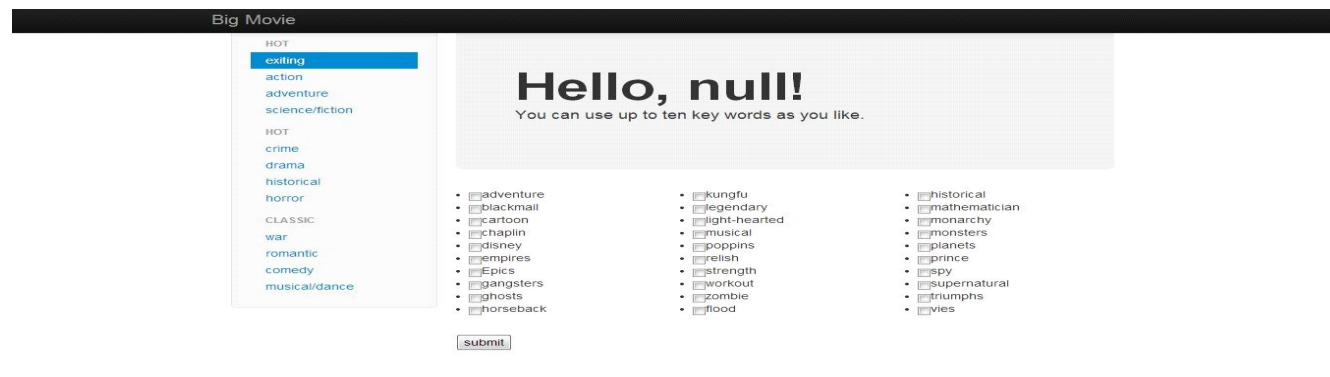


➤ Main Page



The screenshot shows the main page of the 'Super Movie Recommendation' website. At the top is a navigation bar with links: 'Big Movie', 'Key Word', 'Predict', 'Recommender', and 'Contact'. To the right of these links are input fields for 'username' and 'password', and buttons for 'Sign in' and 'Sign up'. Below the navigation bar, the title 'Super Movie Recommendation' is displayed in a large, bold font. Underneath the title is a subtitle: 'Awsome move recommendation, classification and key words search for both movie lovers and movie producers!'. The page is set against a white background with a black navigation bar.

➤ Keywords Selection



The screenshot shows the 'Keywords Selection' page. On the left side, there is a sidebar with a list of movie genres categorized under 'HOT', 'CLASSIC', and 'romantic'. The 'HOT' category is currently selected, showing a list of genres including 'action', 'adventure', 'science/fiction', 'crime', 'drama', 'historical', and 'horror'. Below this list is a 'submit' button. The main content area features a large, light gray box with the text 'Hello, null!' and 'You can use up to ten key words as you like.' Below this box, there are three columns of movie genres, each with a checkbox next to it. The genres listed are: adventure, blackmail, cartoon, chaplin, disney, empires, epics, gangsters, ghosts, horseback, kungfu, legendary, light-hearted, musical, poppins, relish, strength, workout, zombie, flood, historical, mathematician, monarchy, monsters, planets, prince, spy, supernatural, triumphs, and vines.

➤ Score input

[Big Movie](#) [Key Word](#) [Recommender ▼](#) [About](#) [Contact](#)

user_name:

password

password:

password

Chushingura

This film is one of the two best Samurai films of all time, the other being Kurosawa's SEVEN SAMURAI. CHUSHINGURA ("loyalty") is based on a real incident in 16th century Japan, wherein 47 loyal retainers of a disgraced lord take a vow of vengeance on the corrupt nobleman who caused his downfall and death.

not rated ▼

Westward Ho

The older surviving brother of a wagon train ambush, seeks revenge on the leader of the gang responsible for the murder of his parents and the kidnapping of his younger brother. Years later, he is the leader of a group of 'vigilantes' who oppose lawlessness and unknowingly confronts his 'long-lost-brother' who has since learned the tricks-of-the-trade from the outlaws.

not rated ▼

Cartoon Network Christmas 3

I Am Weasel: "The Legend of Big Butt" - Paranormalists Weasel and Baboon take a trip to the Himalayas in search of the elusive creature Big Butt.

not rated ▼

The Girl Next Door

There is a point in "The Girl Next Door - Unrated Edition" (which I received and promptly devoured, extra's and all, in one sitting) where the movie is in danger of changing in tone and focus from American Pie to Showgirls, but luckily for the movie - and viewers - it recovers itself and ends in a definite teen comedy spirit.

not rated ▼

Baby Einstein: World Music

Baby Einstein's newest DVD offers a whole new musical beat that exposes babies to music from around the world. The fun-damentals and beauty of music are explored through real world images and sounds collected around the globe and reorchestrated for baby's eyes and ears. Activities and simple children's songs encourage parents and babies to drum, shake, blow, strum and sing along with the musical fun.

not rated ▼

National Lampoon's Lost Reality 1 and 2

The first Lost Reality is funny, but the second is better. Lost Reality is a collection of mock reality show pilots that apparently didn't get chosen to become ongoing reality TV series. They are offensively hilarious, but Lost Reality is probably an acquired taste in it's vulgar humor. Both films are worth buying if you're into a little bit of funny shock value, especially Lost Reality 2.

not rated ▼

submit

➤ Movie link output

[Big Movie](#) [Key Word](#) [Recommender ▼](#) [About](#) [Contact](#) Logged in as hahaha

0. movie: B003AI2VGA rate: 3

<http://www.amazon.com/dp/B003AI2VGA>

1. movie: B00006HAXW rate: 5

<http://www.amazon.com/dp/B00006HAXW>

2. movie: B00004CQT3 rate: 5

<http://www.amazon.com/dp/B00004CQT3>