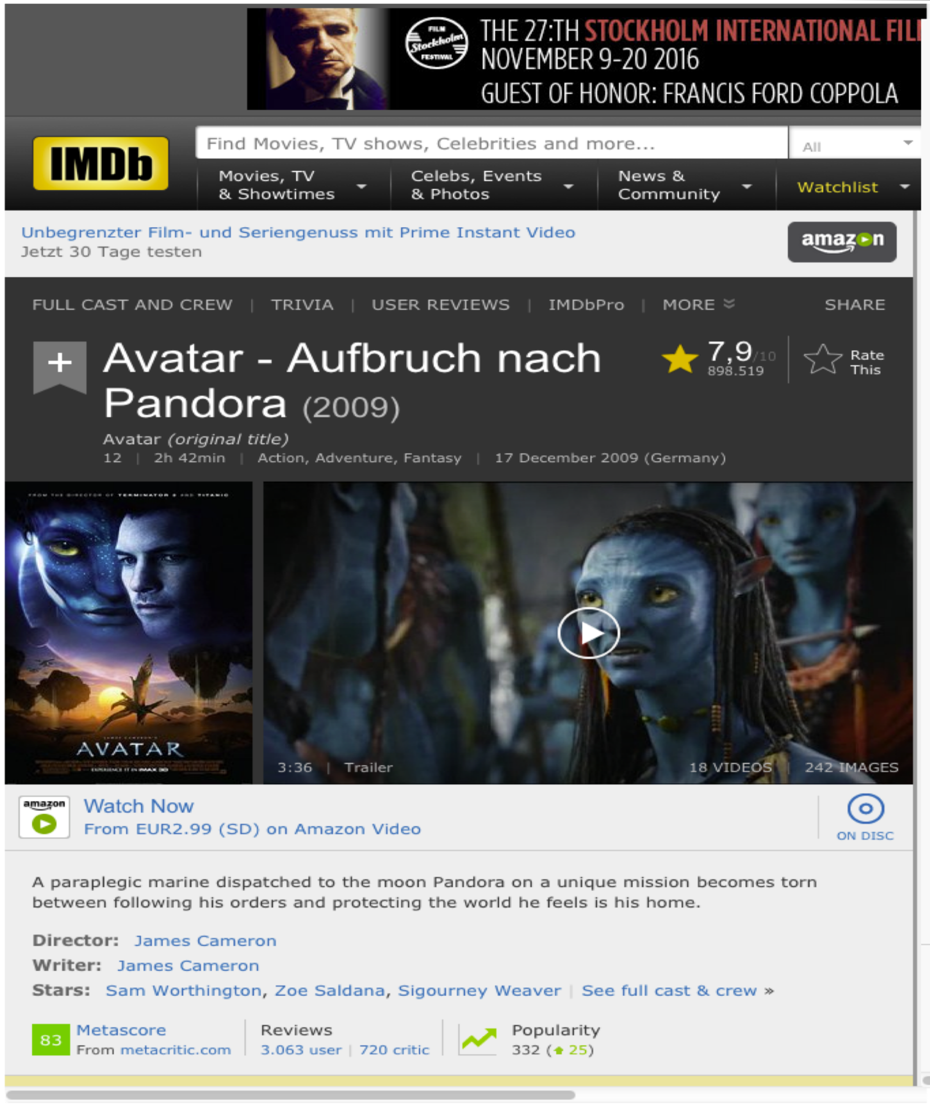
Information Retrieval Exercises

Assignment 1:

IMDB Spider

Mario Sänger (saengema@informatik.hu-berlin.de)

IMDB: Internet Movie Database



THE 27:TH STOCKHOLM INTERNATIONAL FIL
NOVEMBER 9-20 2016
GUEST OF HONOR: FRANCIS FORD COPPOLA
IMDb
Find Movies, TV shows, Celebrities and more...
All
Movies, TV & Showtimes
Celebs, Events & Photos
News & Community
Watchlist
amazon
FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE
SHARE
Rate This
AVATAR
3:36 | Trailer
18 VIDEOS | 242 IMAGES
ON DISC

Unbegrenzter Film- und Seriengenuss mit Prime Instant Video
Jetzt 30 Tage testen

Avatar - Aufbruch nach Pandora (2009)

7,9/10 898.519

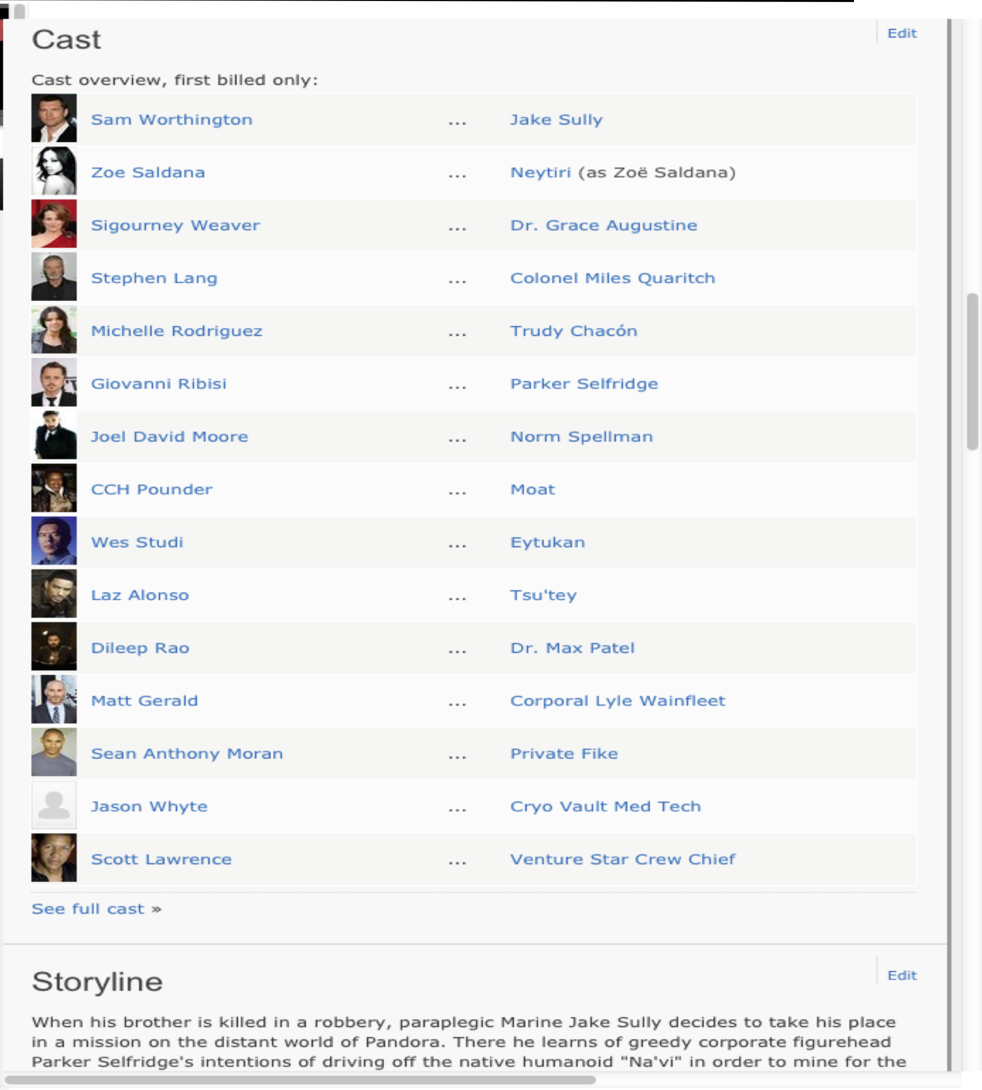
Avatar *(original title)*
12 | 2h 42min | Action, Adventure, Fantasy | 17 December 2009 (Germany)

Watch Now
From EUR2.99 (SD) on Amazon Video

A paraplegic marine dispatched to the moon Pandora on a unique mission becomes torn between following his orders and protecting the world he feels is his home.

Director: James Cameron
Writer: James Cameron
Stars: Sam Worthington, Zoe Saldana, Sigourney Weaver | See full cast & crew »

83 Metascore From metacritic.com | Reviews 3.063 user | 720 critic | Popularity 332 (▲ 25)



Cast

Edit

Cast overview, first billed only:

Sam Worthington	...	Jake Sully
Zoe Saldana	...	Neytiri (as Zoë Saldana)
Sigourney Weaver	...	Dr. Grace Augustine
Stephen Lang	...	Colonel Miles Quaritch
Michelle Rodriguez	...	Trudy Chacón
Giovanni Ribisi	...	Parker Selfridge
Joel David Moore	...	Norm Spellman
CCH Pounder	...	Moat
Wes Studi	...	Eytukan
Laz Alonso	...	Tsu'tey
Dileep Rao	...	Dr. Max Patel
Matt Gerald	...	Corporal Lyle Wainfleet
Sean Anthony Moran	...	Private Fike
Jason Whyte	...	Cryo Vault Med Tech
Scott Lawrence	...	Venture Star Crew Chief

See full cast »

Storyline

Edit

When his brother is killed in a robbery, paraplegic Marine Jake Sully decides to take his place in a mission on the distant world of Pandora. There he learns of greedy corporate figurehead Parker Selfridge's intentions of driving off the native humanoid "Na'vi" in order to mine for the

Assignment

- Task:
  - Given a list of 500 movies, answer queries on movies
- Problem:
  - IMDB data is human-readable, but semi-structured
- Idea:
  - "Scrap"* data from each movie on IMDB
  - Then, we perform queries on the scrapped data.

* Data scrapping is a technique in which a computer program extracts data from human-readable output coming from another program.

Concrete tasks

1. Implement a JAVA program that reads a list of 500 movie titles from a JSON file

2. For each movie title, perform a web search on IMDB and retrieve movie's URL

3. For each movie, extract metadata (e.g. actors, budget, description) from the movie's URL and store them in a JSON file

4. Implement queries on movies' metadata

1. Read movie titles from JSON file

- Read movie titles from a JSON* file “movies.json”:

```
[
    {"movie_name":"Avatar"},
    {"movie_name":"Star Wars VII: The Force Awakens"},
    ...
]
```

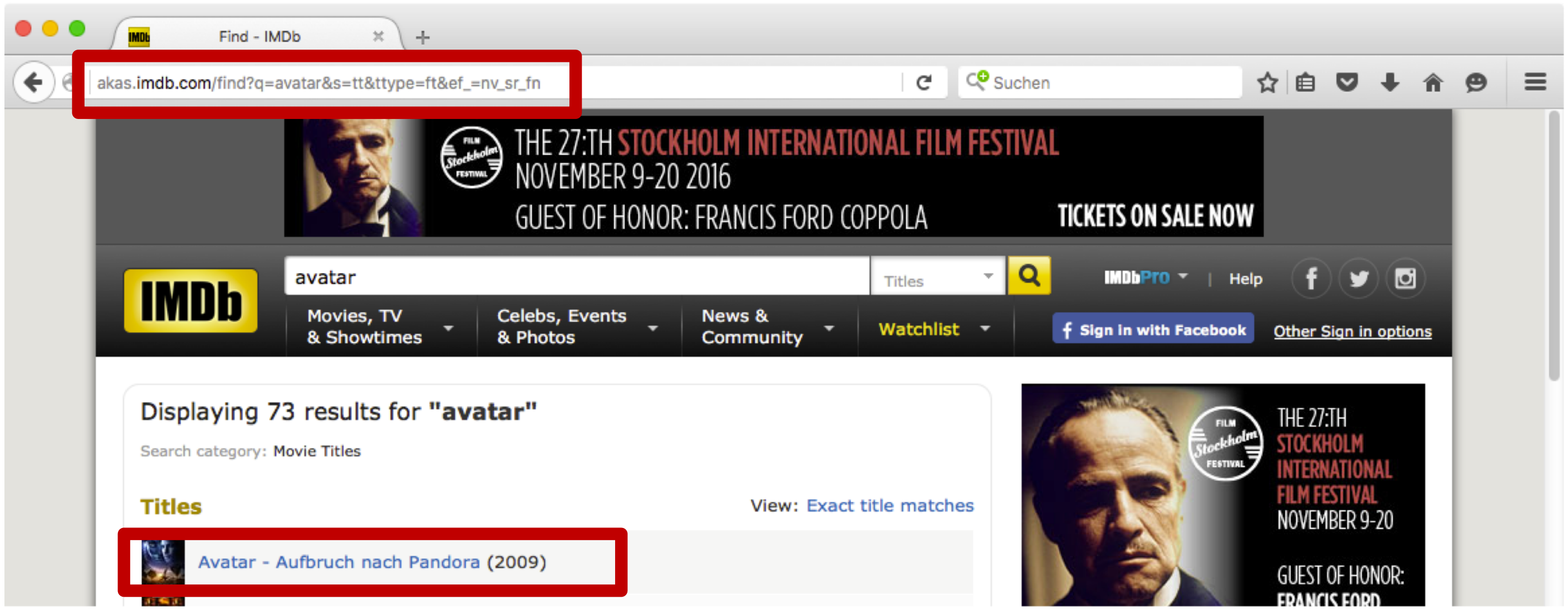
- You can use any Java library for parsing JSON files
  - Reference implementation: Oracle’s JSONP (https://jsonp.java.net/)
  - JSON.simple (https://github.com/fangyidong/json-simple)
  - GSON (https://github.com/google/gson)
  - Jackson Project (https://github.com/FasterXML/jackson)

* JSON is a common syntax for storing and exchanging data. JSON is a widespread alternative to XML: http://www.w3schools.com/js/js_json_intro.asp

2. Perform a web search on IMDB

- Implement *IMDBSpider.java* that opens the URL: *https://www.imdb.com/find?q=<MOVIE>&s=tt&ttype=ft*

- From the results, extract the first element and its URL
- Use URL encoding of movie titles



Displaying 73 results for "**avatar**"

Search category: Movie Titles

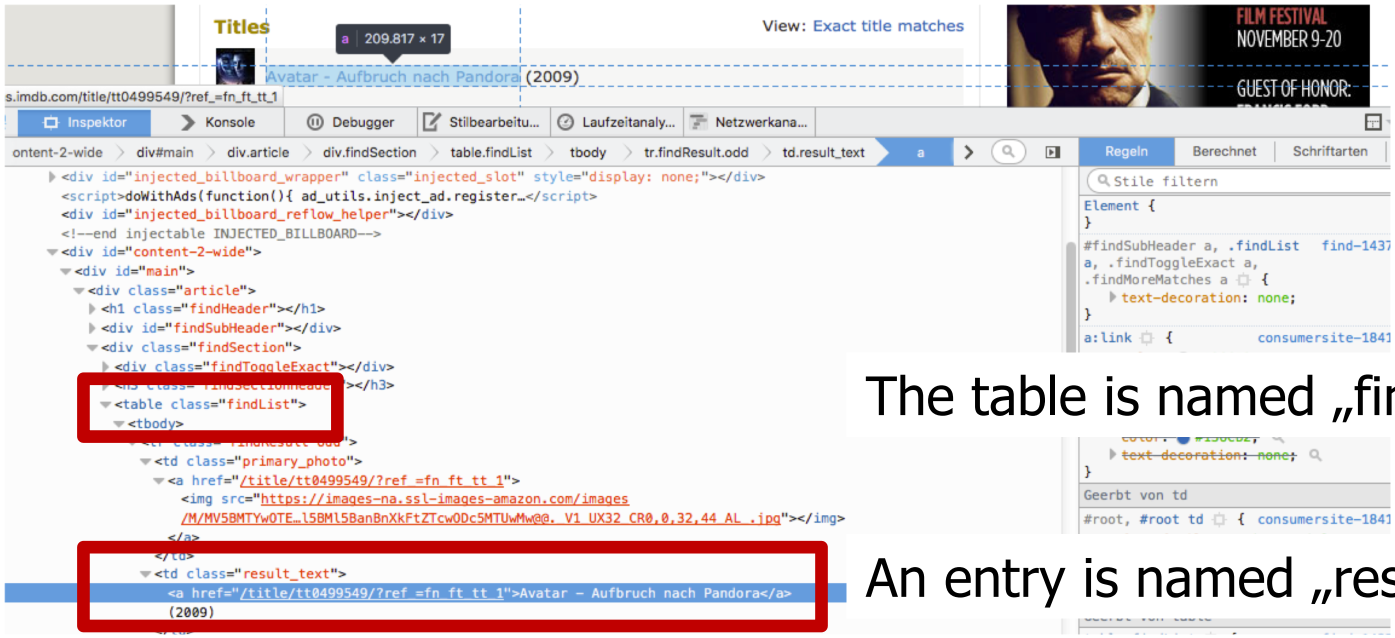
Titles

View: Exact title matches

Avatar - Aufbruch nach Pandora (2009)

2. Perform a web search on IMDB

- You have to parse the html file to extract the URL
- You can use any method. You could use XPATH and html cleaner:
  - http://htmlcleaner.sourceforge.net
  - http://htmlcleaner.sourceforge.net/doc/org/htmlcleaner/XPather.html


Titles
View: Exact title matches
Avatar - Aufbruch nach Pandora (2009)
s.imdb.com/title/tt0499549/?ref_=fn_ft_tt_1
Inspektor
Konsole
Debugger
Stilbearbeitu...
Laufzeitanaly...
Netzwerkana...
ontent-2-wide › div#main › div.article › div.findSection › table.findList › tbody › tr.findResult.odd › td.result_text › a
Regeln
Berechnet
Schriftarten
FILM FESTIVAL
NOVEMBER 9-20
GUEST OF HONOR:

```
<div id="injected_billboard_wrapper" class="injected_slot" style="display: none;"></div>
<script>doWithAds(function(){ ad_utils.inject_ad.register…</script>
<div id="injected_billboard_reflow_helper"></div>
<!--end injectable INJECTED_BILLBOARD-->
<div id="content-2-wide">
  <div id="main">
    <div class="article">
      <h1 class="findHeader"></h1>
      <div id="findSubHeader"></div>
      <div class="findSection">
        <div class="findToggleExact"></div>
        <h3 class="findSectionHeader"></h3>
        <table class="findList">
          <tbody>
            <tr class="findResult odd">
              <td class="primary_photo">
                <a href="/title/tt0499549/?ref_=fn_ft_tt_1">
                  <img src="https://images-na.ssl-images-amazon.com/images
                  /M/MV5BMTYwOTE…l5BMl5BanBnXkFtZTcwODc5MTUwMw@@._V1_UX32_CR0,0,32,44_AL_.jpg"></img>
                </a>
              </td>
              <td class="result_text">
                <a href="/title/tt0499549/?ref_=fn_ft_tt_1">Avatar - Aufbruch nach Pandora</a>
                (2009)
```

```
Element {
}
#findSubHeader a, .findList    find-1437
a, .findToggleExact a,
.findMoreMatches a {
  text-decoration: none;
}
a:link {                       consumersite-1841
  text-decoration: none;
}
Geerbt von td
#root, #root td {              consumersite-1841
```

The table is named „findList“

An entry is named „result_text“

2. Perform a web search on IMDB

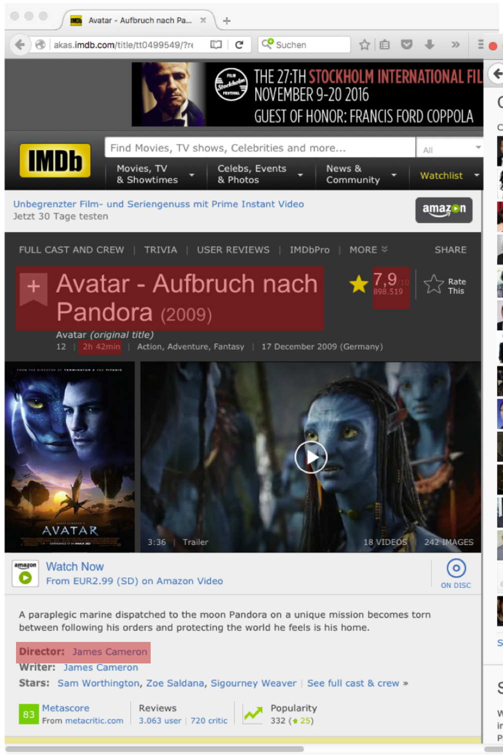
- XPATH is a syntax for navigating parts of an XML document
- Has a directory-path-like syntax

-
```
<table class=”list" >
    <tr>
          <td class=”result">Avatar</td>
    </tr>
</table>
```

- XPATH:
  `/table[@class='list']//td[@class='result']/text()`

=> Avatar

3. Extract metadata from movie’s URL


Avatar - Aufbruch nach Pa...
akas.imdb.com/title/tt0499549/?re
Suchen
THE 27:TH STOCKHOLM INTERNATIONAL FIL
NOVEMBER 9-20 2016
GUEST OF HONOR: FRANCIS FORD COPPOLA
IMDb
Find Movies, TV shows, Celebrities and more...
All
Movies, TV & Showtimes
Celebs, Events & Photos
News & Community
Watchlist
amazon
FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE
SHARE
Rate This
3:36 | Trailer
18 VIDEOS | 242 IMAGES
ON DISC

Unbegrenzter Film- und Seriengenuss mit Prime Instant Video
Jetzt 30 Tage testen

Avatar - Aufbruch nach Pandora (2009)

7,9/10
898.519

Avatar *(original title)*
12 | 2h 42min | Action, Adventure, Fantasy | 17 December 2009 (Germany)

Watch Now
From EUR2.99 (SD) on Amazon Video

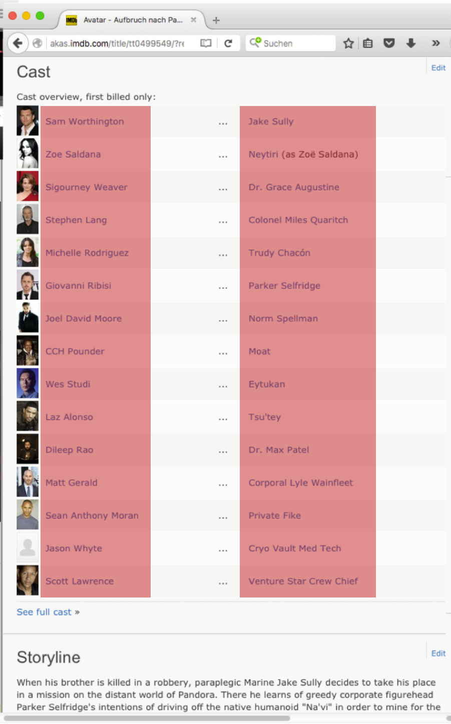
A paraplegic marine dispatched to the moon Pandora on a unique mission becomes torn between following his orders and protecting the world he feels is his home.

Director: James Cameron
Writer: James Cameron
Stars: Sam Worthington, Zoe Saldana, Sigourney Weaver | See full cast & crew »

83 Metascore
From metacritic.com

Reviews
3.063 user | 720 critic

Popularity
332 (↑ 25)


Avatar - Aufbruch nach Pa...
akas.imdb.com/title/tt0499549/?re
Suchen
Edit

Cast

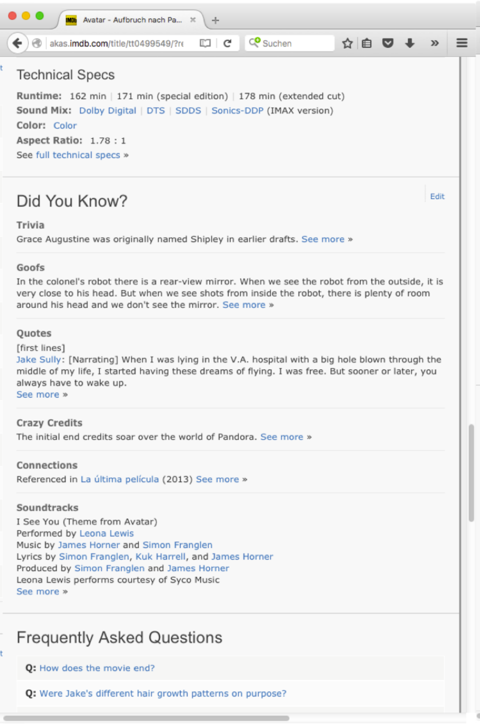
Cast overview, first billed only:

Sam Worthington	...	Jake Sully
Zoe Saldana	...	Neytiri (as Zoë Saldana)
Sigourney Weaver	...	Dr. Grace Augustine
Stephen Lang	...	Colonel Miles Quaritch
Michelle Rodriguez	...	Trudy Chacón
Giovanni Ribisi	...	Parker Selfridge
Joel David Moore	...	Norm Spellman
CCH Pounder	...	Moat
Wes Studi	...	Eytukan
Laz Alonso	...	Tsu'tey
Dileep Rao	...	Dr. Max Patel
Matt Gerald	...	Corporal Lyle Wainfleet
Sean Anthony Moran	...	Private Fike
Jason Whyte	...	Cryo Vault Med Tech
Scott Lawrence	...	Venture Star Crew Chief

See full cast »

Storyline

When his brother is killed in a robbery, paraplegic Marine Jake Sully decides to take his place in a mission on the distant world of Pandora. There he learns of greedy corporate figurehead Parker Selfridge's intentions of driving off the native humanoid "Na'vi" in order to mine for the


Avatar - Aufbruch nach Pa...
akas.imdb.com/title/tt0499549/?re
Suchen
Edit

Technical Specs

Runtime: 162 min | 171 min (special edition) | 178 min (extended cut)
Sound Mix: Dolby Digital | DTS | SDDS | Sonics-DDP (IMAX version)
Color: Color
Aspect Ratio: 1.78 : 1
See full technical specs »

Did You Know?

Trivia
Grace Augustine was originally named Shipley in earlier drafts. See more »

Goofs
In the colonel's robot there is a rear-view mirror. When we see the robot from the outside, it is very close to his head. But when we see shots from inside the robot, there is plenty of room around his head and we don't see the mirror. See more »

Quotes
[first lines]
Jake Sully: [Narrating] When I was lying in the V.A. hospital with a big hole blown through the middle of my life, I started having these dreams of flying. I was free. But sooner or later, you always have to wake up.
See more »

Crazy Credits
The initial end credits soar over the world of Pandora. See more »

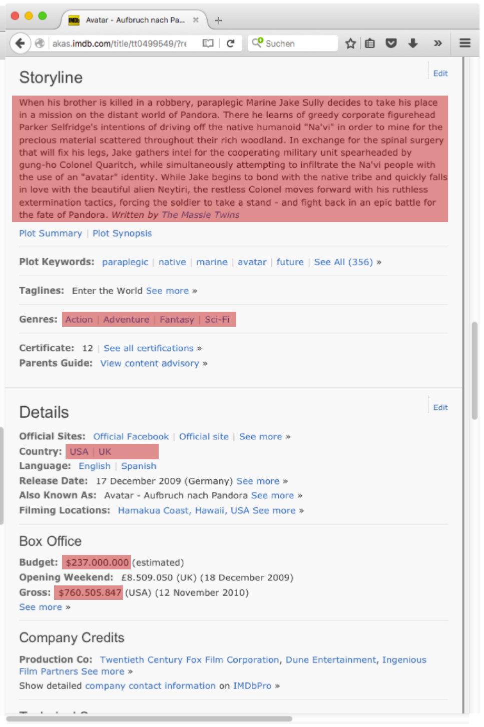
Connections
Referenced in La última película (2013) See more »

Soundtracks
I See You (Theme from Avatar)
Performed by Leona Lewis
Music by James Horner and Simon Franglen
Lyrics by Simon Franglen, Kuk Harrell, and James Horner
Produced by Simon Franglen and James Horner
Leona Lewis performs courtesy of Syco Music
See more »

Frequently Asked Questions

Q: How does the movie end?

Q: Were Jake's different hair growth patterns on purpose?


Avatar - Aufbruch nach Pa...
akas.imdb.com/title/tt0499549/?re
Suchen
Edit

Storyline

When his brother is killed in a robbery, paraplegic Marine Jake Sully decides to take his place in a mission on the distant world of Pandora. There he learns of greedy corporate figurehead Parker Selfridge's intentions of driving off the native humanoid "Na'vi" in order to mine for the precious material scattered throughout their rich woodland. In exchange for the spinal surgery that will fix his legs, Jake gathers intel for the cooperating military unit spearheaded by gung-ho Colonel Quaritch, while simultaneously attempting to infiltrate the Na'vi people with the use of an "avatar" identity. While Jake begins to bond with the native tribe and quickly falls in love with the beautiful alien Neytiri, the restless Colonel moves forward with his ruthless extermination tactics, forcing the soldier to take a stand - and fight back in an epic battle for the fate of Pandora. *Written by The Massie Twins*

Plot Summary | Plot Synopsis

Plot Keywords: paraplegic | native | marine | avatar | future | See All (356) »

Taglines: Enter the World See more »

Genres: Action | Adventure | Fantasy | Sci-Fi

Certificate: 12 | See all certifications »
Parents Guide: View content advisory »

Details

Official Sites: Official Facebook | Official site | See more »
Country: USA | UK
Language: English | Spanish
Release Date: 17 December 2009 (Germany) See more »
Also Known As: Avatar - Aufbruch nach Pandora See more »
Filming Locations: Hamakua Coast, Hawaii, USA See more »

Box Office

Budget: $237.000.000 (estimated)
Opening Weekend: £8.509.050 (UK) (18 December 2009)
Gross: $760.505.847 (USA) (12 November 2010)
See more »

Company Credits

Production Co: Twentieth Century Fox Film Corporation, Dune Entertainment, Ingenious Film Partners See more »
Show detailed company contact information on IMDbPro »

3. Extract Metadata from Movie's URL

- Extract the following information from each movie and store it to a separate JSON file:
  url, title, year, genreList, countryList, description, budget, gross, ratingValue, ratingCount, duration, castList, characterList, directorList

- Treat each attribute as a string and list names refer to JSON lists
  - Stick to exactly these names!

- Refer to example_*avatar.json* for an example.

```
[    {    "url":"https://www.imdb.com/title/tt0499549/?ref_=fn_ft_tt_1",
          "title":"Avatar - Aufbruch nach Pandora (2009)",
          "year":"2009",
          "genreList":["Action","Adventure", "Fantasy", "Sci-Fi"],    ...
          }
]
```

3. Extract Metadata from Movie’s URL

- Watch out: sometimes one or more information can be missing!
  - Use empty strings or empty arrays!

- Special case ”gross”
  - First search for “Cumulative Worldwide Gross”
  - If missing, search for “Gross USA”
  - Only use dollar values (no Yen, Euros, etc)!

- Optional: extract further meta information

  languageList, keywordList, aspectRatio, contentRating, reviews, critics

4. Easy Queries I

- You have to correctly implement (at least) **three** basic / easy queries out of:

1. All-rounder: Determine all movies in which the director stars as an actor (cast). Return the top ten matches sorted by decreasing IMDB rating.

2. Under the radar: Determine the top ten US-American movies until (including) 2015 that have made the biggest loss despite an IMDB score above (excluding) 8.0, based on at least 1,000 votes. Here, loss is defined as gross minus budget.

3. The pillars of storytelling: Determine all movies that contain both (sub-)strings "kill" and "love" in their lowercase description (String.toLowerCase()). Sort the results by the number of appearances of these strings and return the top ten matches.

4. Easy Queries II

4. The red planet: Determine all movies of the Sci-Fi genre that mention "Mars" in their description (case-aware!). List all found movies in ascending order of publication (year).

5. Colossal failure: Determine all US-American movies with a duration beyond 2 hours, a budget beyond 1 million and an IMDB rating below 5.0. Sort results by ascending IMDB rating.

4. Harder Queries (Aggregation & Join) I

- You have to correctly implement (at least) **two** hard queries out of:

6. Uncreative writers: Determine the ten most frequent character names of all times ordered by frequency of occurrence. Filter any name containing "himself", "doctor", and "herself" from the result.

7. Workhorse: Provide a ranked list of the top ten most active actors (cast), i.e., those actors which have starred in most movies.

8. Must see: List the best rated movie of each year starting from 1990 until (including) 2010 with more than 10,000 ratings. Order the movies by increasing year.

4. Harder Queries (Aggregation & Join) II

9. Rotten Tomatoes: List the worst rated movie of each year starting from 1990 till (including) 2010 with an IMCB score larger than 0. Order the movies by increasing year.

10. Magic Couples: Determine those couples that feature together in the most movies (e.g. Adam Sandler and Allen Covert feature together in multiple movies). Report the top 10 pairs of actors and sort the result by the number of movies.

4. Optional: Custom Queries

- Come up with a fancy custom query
- Give a text description of the query, the implementation and the result

- Very creative queries can earn an extra point for the competition

Caveats

- Crawler:
  - You must implement the Java class *IMDBSpider.java*, which reads the movie titles from a JSON file and stores each movie in a separate JSON file
- Queries:
  - You must implement five queries in *IMDBQueries.java*
  - Optional Custom Query: You can implement one fancy custom query. Give a description of the query, source code and the result.
  - A query counts as implemented if it is correct. So, implement more than five to be sure ☺
  - No query result caching!

Competition

- Queries should not only be correct but as fast as possible
- While you have 500 movies - I will execute your queries with 5000+ movies

- Evaluation:
  - A correctly implemented query
  - Bonus for faster implementation

- The 5 best teams get points!

Submission

- Java source codes and **two** executable JARs
  - IMDBSpider must be callable with
    java -jar IMDBSpider.jar movies.json <moviesDir>
  - IMDBQueries must be callable with
    java -jar IMDBQueries.jar <moviesDir>
- Stick to the specified class interfaces
  - Don’t change class or method signatures
  - Don’t move the classes into other packages
  - But: You can create additional classes for your implementation

Submission

- **Group 1: Monday, 07.05., 23:59 (midnight)**
- **Group 2: Wednesday, 09.05., 23:59 (midnight)**

- Upload a ZIP archive named *ass1_<group-name>.zip* to https://hu.berlin/ire18_assignment1

- Test your JAR archives on gruenau!

Presentation of solutions

- You are be able to pick when and what you‘d like to present (first-come-first-served):

  Monday:

  https://dudle.inf.tu-dresden.de/zluz35py/

  Wednesday:

  https://dudle.inf.tu-dresden.de/8ks22llb/

- Presentation has to be given on 14.05./16.05.

Next week (attendance optional)

- Q/A session for assignment 1
- If desired: Live coding session
  - JSON parsing, Opening URLs, XPATH
  - Maven, Executable Jars
  - …

- If you have questions about topics from the lecture, write me an email in advance!

Questions?