MARWADI UNIVERSITY
FACULTY OF TECHNOLOGY, GAURIDAD CAMPUS.
DEPARTMENT OF COMPUTER ENGINEERING

ASSIGNMENT - 1

AIM : Develop a Web Scraper [for minimum 3 pages] to mine structured data from any website using any method other than cURL and store that data to a WordPress Database.

TASKS :

1. Setup your WordPress on localhost.
2. Understand its database structure [Specifically Database Table for a POST].
3. Identify any Website having Pagination in it.
4. Mine Structured Data [Minimum 3 Elements (e.g. image, title, desc)].
5. Insert Extracted Data in the WordPress Database Table.
6. Your PHP Script must iterate mining for at-least 3 pages from Pagination.

1. Describe WordPress Database Structure in your Words.

→ A database is created whenever you build a WordPress website. Everything on your WordPress website, be it posts, custom post type, pages, comments, and even settings are stored in a database. It's like a warehouse of information. All your data is placed in an organized manner so that it's easy to find them. An image of a typical warehouse that comes to mind is that of rows and rows of cardboard boxes. The boxes are kept on storage shelves. In a WordPress database, the shelves are known as tables.

There are 11 tables by default on a new WordPress website. Every table can store only specific data. For instance, the wp comments table captures all information left by a person commenting on a post like IP Address, comment author slug, etc. Storing data in a specific table makes it faster and easier to find them.

In the next section, we are going to walk you through every WordPress table and show you which tables are responsible for the content on your WordPress website.

A New WordPress Website has 11 tables. Those are:

1. wp_posts
2. wp_postmeta
3. wp_options
4. wp_users
5. wp_usermeta
6. wp_term_taxonomy
7. wp_terms
8. wp_term_relationships
9. wp_links
10. wp_comments
11. wp_commentmeta

Many of the tables are related to one another. One piece of data can be related to other data, for instance, a post can be associated with categories and tags. So, the table that stored blog posts will share a relationship with the tables where category and tags are stored.

Before we start describing what the tables store and how they are connected to each other, here's a graphical structure of WordPress database:

1. **wp_posts**

   Since types of content from your posts and pages is stored in this table, it's arguably the most important table in the WordPress database. The content types include text, revisions, menu items, media attachments, and any custom items.

2. **wp_postmeta**

   It's an extension of the wp_posts table. It stores extra information from posts. Some plugins store data within this table. The social sharing plugin MashShare stores share counts of specific posts in this table.

   Note: Throughout the database, there are many such tables that enable the WordPress core or a theme or plugin to store extra information.

3. **wp_options**

   The options table is a different kind of table. Instead of storing the content of the website, it stores the settings of the site. This table stores your website configurations like the site title, tagline, and timezone. It typically stores the settings of plugins and themes as well. Unlike other tables, the wp_options table doesn't really share a relationship with any of the other tables.

4. **wp_users**

   The wp_users table stores the list of all registered users of your WordPress website. As a result, it saves basic information like their username, WordPress passwords, email ID, display name, time of registration, etc.

5. **wp_usermeta**

   WordPress stores extra information about users in the wp_usermeta table. For instance, the last name of a user is saved in the wp_usermeta table instead of the wp_users table.

6. **wp_terms**

   The wp_terms table stores categories for both posts and pages and tags for posts. Links related to categories are also present here. The wp_terms table shares a relationship with wp_term_taxonomy and wp_term_relationships table.

7. **wp_term_taxonomy**

   wp_term_taxonomy stores descriptions of categories, tags and certain links associated with categories.

8. **wp_term_relationships**

   The wp_term_relationships table helps maintain relationships. For instance, in this post, the one that you are reading, is associated with a few tags and a category. The wp_term_relationships table helps maintain this association.

9. **wp_links**

   The wp_links table stores information related to blogrolls. Since blogrolls are no longer in use, it's strange to still find the wp_links table. It's mainly for folks who are using older versions of WordPress.

10. **wp_comments**

    Both approved and unapproved comments left on your posts and pages are stored in this table. Specific data about the author like the author name, email address, type of comment (whether it's a simple comment, pingback or trackback) are also saved in this table. Furthermore, it's important to note that if you are using a third-party comment service like Disqus, comments won't be stored in this table, they'll be saved on the commenting system's own server.

11. **wp_commentmeta**

    Extra data about the comments left on your website like which post is the comment associated with are stored here.

    That's the final table in the WordPress database of a new website.

    It's important to note that if you are checking the database of an old website, there's bound to be more than 11 tables.

    The more time you spent time running a website, the more information you add. As a result, your database becomes bigger. New tables are added to the database to support certain functions on a website. Gravity Forms, for instance, creates its own WordPress database table once you install it on your website. However, not all plugins add tables to the database. Many utilize those already present.

2. Describe Regular Expressions in PHP.

→ Regular expressions commonly known as a regex (regexes) are a sequence of characters describing a special search pattern in the form of text string. They are basically used in programming world algorithms for matching some loosely defined patterns to achieve some relevant tasks. Sometimes regexes are understood as a mini programming language with a pattern notation which allows the users to parse text strings. The exact sequence of characters are unpredictable beforehand, so the regex helps in fetching the required strings based on a pattern definition.

Regular Expression is a compact way of describing a string pattern that matches a particular amount of text. As you know, PHP is an open-source language commonly used for website creation, it provides regular expression functions as an important tool. Like PHP, many other programming languages have their own implementation of regular expressions. This is the same with other applications also, which have their own support of regexes having various syntaxes. Many available modern languages and tools apply regexes on very large files and strings. Let us look into some of the advantages and uses of regular expressions in our applications.

Advantages and uses of Regular expressions:

In many scenarios, developers face problems whenever data are collected in free text fields as most of the programming deals with data entries. Regular expressions are used almost everywhere in today's application programming.

- Regular expressions help in validation of text strings which are of programmer's interest.
- It offers a powerful tool for analysing, searching a pattern and modifying the text data.
- It helps in searching specific string pattern and extracting matching results in a flexible manner.
- It helps in parsing text files looking for a defined sequence of characters for further analysis or data manipulation.
- With the help of in-built regexes functions, easy and simple solutions are provided for identifying patterns.
- It effectively saves a lot of development time, which are in search of specific string pattern.
- It helps in important user information validations like email address, phone numbers and IP address.
- It helps in highlighting special keywords in a file based on search result or input.
- It helps in identifying specific template tags and replacing those data with the actual data as per the requirement.
- Regexes are very useful for creation of HTML template system recognizing tags.
- Regexes are mostly used for browser detection, spam filtration, checking password strength and form validations.

3. Describe a Method you have used to Fetch Data (Not cURL).

→ Simple HTML DOM Parser

Web Scraping is a technique used to extract large amounts of data from websites extracted and saved to a local file in your computer or to a database or can be used as API. Data displayed by most websites can be viewed by using a web browser only. They do not offer the functionality to save a copy of this data for use. Thus, the only option is to copy and paste the selected data that is required, which in reality, is a very tedious job and may take hours complete. In other terms Web Scraping is the technique of automating such a process, in place of manual work, the Web Scraping software performs the same task within seconds. The web scraping can be done by targeting the selected DOM components and then processing or storing the text between that DOM element of a web page. To do the same in PHP, there is an API which parses the whole page and looks for the required elements within the DOM. It is the Simple HTML DOM Parser.

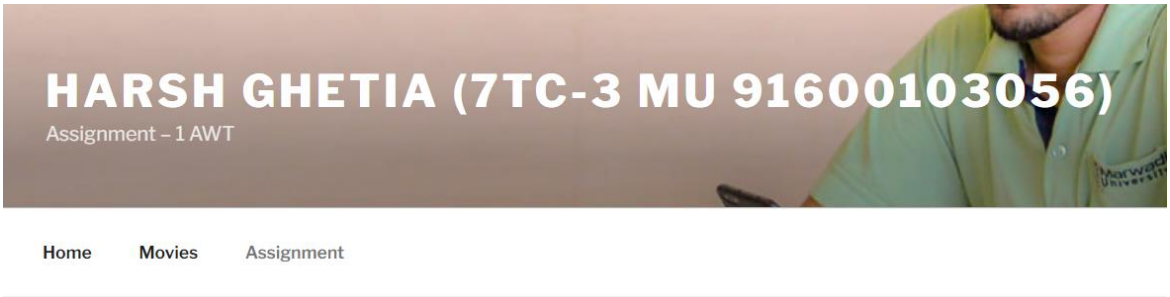
4. Your PHP Script Code for Web Scraper.

→ Target Website: https://moviesverse.com/

Code Snippet - 1:

```
<?php
include('simple_html_dom.php');
$servername = "localhost";
$username = "root";
$password = "";
$dbname = "harshawt";
$z = 1;
for ($x = 1; $x <= 3; $x++)
{
$target_url = 'https://moviesverse.com/page/'.$z.'/';
$html = new simple_html_dom();
$html->load_file($target_url);
foreach($html->find("div[id=content_box]") as $link)
{
for($i = 0; $i < 20; $i++)
{
$item['movieThumbnail'] = $link->find("img", $i)->src;
$movieThumbnail = html_entity_decode(trim($item['movieThumbnail']));
$item['movieName'] = $link->find("h2", $i)->plaintext;
$movieName = html_entity_decode(trim($item['movieName']));
$lastMovieName = substr($movieName, 0, strpos($movieName, ")"));
$finalMovieName = str_replace('Download', '', $lastMovieName);
echo '<img src="'.$movieThumbnail.'"/><br>';
$finalMovieName = $finalMovieName.")";
echo "<h3>Fetched Movie Name: </h3>".$movieName."<br>";
echo "<h3>Scraped Movie Name: </h3>".$finalMovieName."<br>";
$lastMovieDesc = strstr($movieName, '720');
$finalMovieDesc = $lastMovieDesc;
echo "<h3>Scraped Movie Description: </h3>".$finalMovieDesc."<br>";
echo "<br>";
```

```
$conn = new mysqli($servername, $username, $password, $dbname);
if ($conn->connect_error)
{
die("Connection Failed: " . $conn->connect_error);
}
$sql = "INSERT INTO wp_fetchdata_harsh (moviename, moviethumbnail, moviedesc)
VALUES ('$finalMovieName', '$movieThumbnail', '$finalMovieDesc')";
if ($conn->query($sql) === TRUE)
{
echo "Record Inserted Successfully!<br>";
}
else
{
echo "Error: " . $sql . "<br>" . $conn->error;
}
$conn->close();
}
$z++;
}
}
?>
```



Home Movies Assignment

CATEGORY: ASSIGNMENT

NOVEMBER 3, 2020

AWT Assignment – 1

Develop a Web scraper [for minimum 3 pages] to mine structured data from any website using any method other than cURL and store that data to a WordPress database.

Continue reading

Home Movies Assignment

NOVEMBER 3, 2020 BY SMIL07370

AWT Assignment – 1

Develop a Web scraper [for minimum 3 pages] to mine structured data from any website using any method other than cURL and store that data to a WordPress database.

TASKS:

1. Setup your word press on localhost.
2. Understand its database structure [Specifically database table for a POST].
3. Identify any website page having pagination in it.
4. Mine structured data [minimum 3 elements (e.g. image,title,desc)].
5. Insert extracted data in the word press database table.
6. Your php script must iterate mining for at-least 3 pages from pagination.

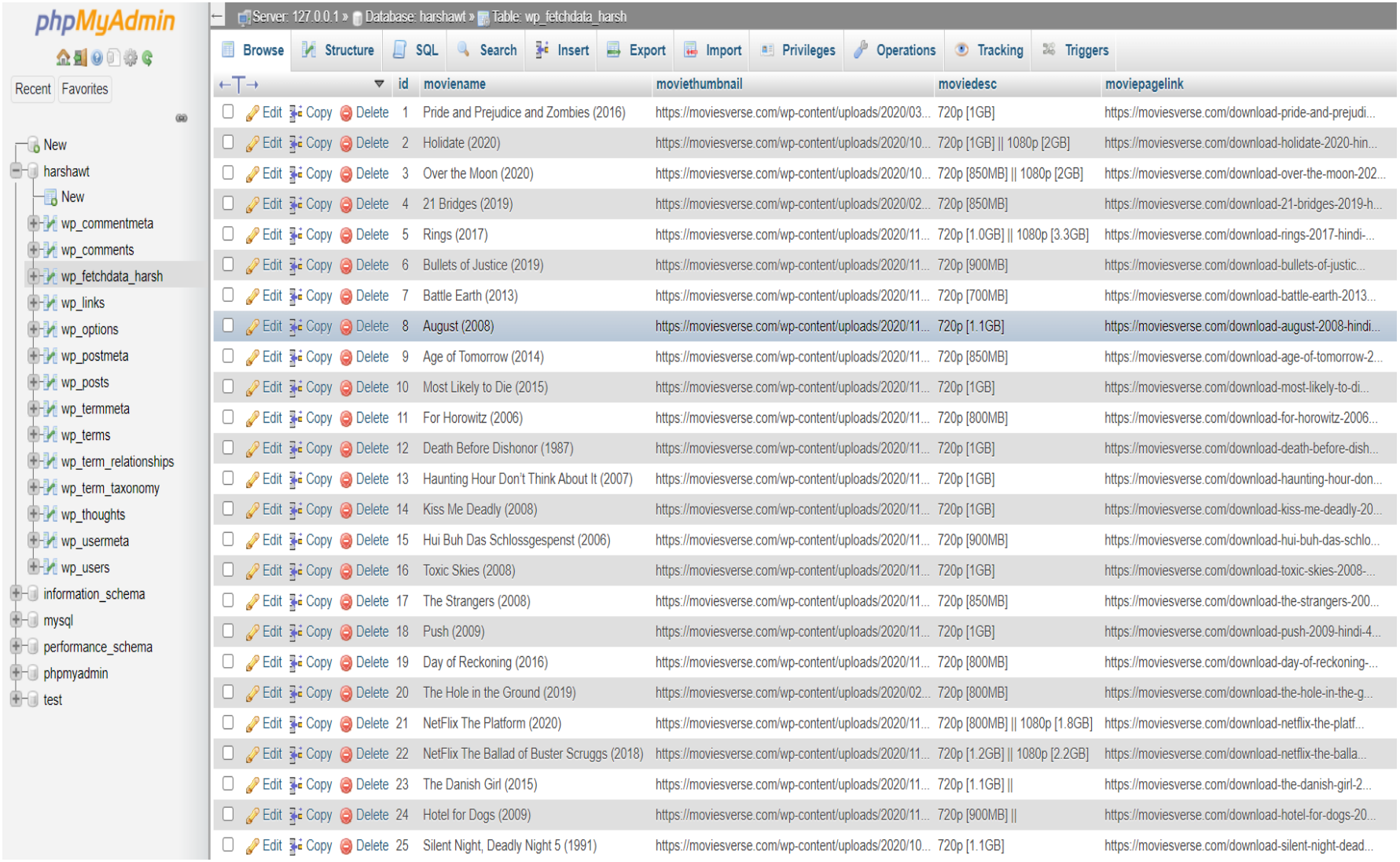
SUBMISSION:

Upload a word file which contains:

1. Describe WordPress database structure in your words.
2. Describe regular expressions in PHP.
3. Describe a method you have used to fetch data. [Not cURL]
4. Your PHP script code for web scraper.

Screen-shot of your localhost wordpress webpage listing all those POSTs populated using script.
[Do not edit/crop your screen-shot]

ALL RECORDS SUCCESSFULLY INSERTED INTO WORDPRESS DATABASE! PLEASE CHECK wp_fetchdata_harsh TABLE FOR OUTPUT!



	id	moviename	moviethumbnail	moviedesc	moviepagelink
Edit Copy Delete	1	Pride and Prejudice and Zombies (2016)	https://moviesverse.com/wp-content/uploads/2020/03...	720p [1GB]	https://moviesverse.com/download-pride-and-prejudi...
Edit Copy Delete	2	Holidate (2020)	https://moviesverse.com/wp-content/uploads/2020/10...	720p [1GB] || 1080p [2GB]	https://moviesverse.com/download-holidate-2020-hin...
Edit Copy Delete	3	Over the Moon (2020)	https://moviesverse.com/wp-content/uploads/2020/10...	720p [850MB] || 1080p [2GB]	https://moviesverse.com/download-over-the-moon-202...
Edit Copy Delete	4	21 Bridges (2019)	https://moviesverse.com/wp-content/uploads/2020/02...	720p [850MB]	https://moviesverse.com/download-21-bridges-2019-h...
Edit Copy Delete	5	Rings (2017)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [1.0GB] || 1080p [3.3GB]	https://moviesverse.com/download-rings-2017-hindi-...
Edit Copy Delete	6	Bullets of Justice (2019)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [900MB]	https://moviesverse.com/download-bullets-of-justic...
Edit Copy Delete	7	Battle Earth (2013)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [700MB]	https://moviesverse.com/download-battle-earth-2013...
Edit Copy Delete	8	August (2008)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [1.1GB]	https://moviesverse.com/download-august-2008-hindi...
Edit Copy Delete	9	Age of Tomorrow (2014)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [850MB]	https://moviesverse.com/download-age-of-tomorrow-2...
Edit Copy Delete	10	Most Likely to Die (2015)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [1GB]	https://moviesverse.com/download-most-likely-to-di...
Edit Copy Delete	11	For Horowitz (2006)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [800MB]	https://moviesverse.com/download-for-horowitz-2006...
Edit Copy Delete	12	Death Before Dishonor (1987)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [1GB]	https://moviesverse.com/download-death-before-dish...
Edit Copy Delete	13	Haunting Hour Don't Think About It (2007)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [1GB]	https://moviesverse.com/download-haunting-hour-don...
Edit Copy Delete	14	Kiss Me Deadly (2008)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [1GB]	https://moviesverse.com/download-kiss-me-deadly-20...
Edit Copy Delete	15	Hui Buh Das Schlossgespenst (2006)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [900MB]	https://moviesverse.com/download-hui-buh-das-schlo...
Edit Copy Delete	16	Toxic Skies (2008)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [1GB]	https://moviesverse.com/download-toxic-skies-2008-...
Edit Copy Delete	17	The Strangers (2008)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [850MB]	https://moviesverse.com/download-the-strangers-200...
Edit Copy Delete	18	Push (2009)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [1GB]	https://moviesverse.com/download-push-2009-hindi-4...
Edit Copy Delete	19	Day of Reckoning (2016)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [800MB]	https://moviesverse.com/download-day-of-reckoning-...
Edit Copy Delete	20	The Hole in the Ground (2019)	https://moviesverse.com/wp-content/uploads/2020/02...	720p [800MB]	https://moviesverse.com/download-the-hole-in-the-g...
Edit Copy Delete	21	NetFlix The Platform (2020)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [800MB] || 1080p [1.8GB]	https://moviesverse.com/download-netflix-the-platf...
Edit Copy Delete	22	NetFlix The Ballad of Buster Scruggs (2018)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [1.2GB] || 1080p [2.2GB]	https://moviesverse.com/download-netflix-the-balla...
Edit Copy Delete	23	The Danish Girl (2015)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [1.1GB] ||	https://moviesverse.com/download-the-danish-girl-2...
Edit Copy Delete	24	Hotel for Dogs (2009)	https://moviesverse.com/wp-content/uploads/2020/11...	720p [900MB] ||	https://moviesverse.com/download-hotel-for-dogs-20...
Edit Copy Delete	25	Silent Night, Deadly Night 5 (1991)	https://moviesverse.com/wp-content/uploads/2020/10...	720p [1.1GB]	https://moviesverse.com/download-silent-night-dead...

Code Snippet - 2:

```
<?php
include('simple_html_dom.php');
$target_url = ' https://moviesverse.com/download-psycho-1960-hindi-480p-720p-1080p/';
$html = new simple_html_dom();
$html->load_file($target_url);
foreach($html->find("div[class=thecontent clearfix]") as $link)
{
$item['movieThumbnail'] = $link->find("img", 0)->src;
$movieThumbnail = html_entity_decode(trim($item['movieThumbnail']));
echo '<center><img src="'.$movieThumbnail.'"/></center><br>';
$item['movieStoryLine'] = $link->find("p", 2)->plaintext;
$movieStoryLine = html_entity_decode(trim($item['movieStoryLine']));
echo "<h3>Fetched Movie Story Line: </h3>".$movieStoryLine."<br>";
$item['movieInfo'] = $link->find("ul", 0);
$movieInfo = html_entity_decode(trim($item['movieInfo']));
echo "<h3>Fetched Movie Info: </h3>".$movieInfo."<br>";
echo "<center><h3>Fetched Movie Screenshots: </h3></center>";
for($i = 1 ; $i <= 3 ; $i++)
{
$item['movieScreenshots'] = $link->find("img", $i)->src;
$movieScreenshots = html_entity_decode(trim($item['movieScreenshots']));
echo '<center><img src="'.$movieScreenshots.'" width="400"
height="200"/></center><br></br>';
}
}
foreach($html->find("div[class=inline canwrap]") as $link)
{
for($i = 0 ; $i <= 2 ; $i++)
{
$item['movieDownload'] = $link->find("h4", $i);
$movieDownload = html_entity_decode(trim($item['movieDownload']));
echo $movieDownload."<br>";
$item['movieDownloadLink'] = $link->find("a", $i);
$movieDownloadLink = html_entity_decode(trim($item['movieDownloadLink']));
echo "<center>".$movieDownloadLink."</center><br>";
}
}
?>
```

Home Movies Assignment

CATEGORY: MOVIES

NOVEMBER 4, 2020

Psycho (1960)

Continue reading

NOVEMBER 3, 2020

Over the Moon (2020)

Continue reading

Home Movies Assignment

NOVEMBER 4, 2020 BY SMIL07370

Psycho (1960)

Fetched Movie Story Line:

Phoenix office worker Marion Crane is fed up with the way life has treated her. She has to meet her lover Sam in lunch breaks, and they cannot get married because Sam has to give most of his money away in alimony. One Friday, Marion is trusted to bank forty thousand dollars by her employer. Seeing the opportunity to take the money and start a new life, Marion leaves town and heads towards Sam's California store. Tired after the long drive and caught in a storm, she gets off the main highway and pulls into the Bates Motel. The motel is managed by a quiet young man called Norman who seems to be dominated by his mother.

Fetched Movie Info:

- **Full Name:** Psycho
- **Language:** Dual Audio (Hindi-English)
- **Subtitles:** Yes (English)
- **Released Year:** 1960
- **Size:** 350MB & 1.2GB & 2.1GB
- **Quality:** 480p & 720p & 1080p Bluray
- **Format:** Mkv

Fetched Movie Screenshots:

Download Psycho (1960) Dual Audio (Hindi-English) 480p [350MB]

Download Links

Download Psycho (1960) Dual Audio (Hindi-English) 720p [1.2GB]

Download Links

Download Psycho (1960) Dual Audio (Hindi-English) 1080p [2.1GB]

Download Links