# The task

We are interested in looking for movies that pass the Bechdel test (<https://en.wikipedia.org/wiki/Bechdel_test>).

We've found a website that has a large collection of movie scripts - <http://www.imsdb.com/>.

We would like you to scrape a few hundred scripts from this website (doesn't have to be all of them), and look for instances where a female character is talking to another female character. Produce some statistics for the movies you have data for, that you think would be relevant to the Bechdel test.

The task is quite open-ended but is intended more as a programming exercise than an NLP/ML one. We would prefer you use Python for the task but you can use another programming language if you wish. If you do use Python, please submit in the form of a python script or Jupyter notebook. If you use multiple files, submit a working folder as a zip file, if you use any external libraries make sure that this is clearly documented.

Please do not spend more than a few hours on this and please get in touch if you have any questions. Have fun!

# Scraping movie scripts

Objective: scrape >= a few hundred movie scripts from <http://www.imsdb.com/>.

What is the structure of the site?  
The scripts can be found at:  
<https://imsdb.com/scripts/Joker.html>  
<https://imsdb.com/scripts/Danish-Girl,-The.html>  
<https://imsdb.com/scripts/A-Quiet-Place.html>

A possible way to proceed:   
Look up the Genres. From each, get the first 100 movies by alphabetical order.  
Note: some movies (e.g. “A quiet place”->sci-fi, horror) belong to more than one genre. For the purpose of the analysis, we should put them in a set() to remove the duplicates, and only later do analytics by genre.

## The genres links

In the Document Object Model, we do not have a specific class, but, instead, a <a> URL inside the tree structure:

<html>

<head>…</head>

<body>

<table>

<table> // 1st row, 1st cell  
 <tbody>

<tr>

<td>  
 <table>  
 <table>  
 <table>  
 <tbody>

<tr>  
 <td colspan="3" class="heading">Genre</td>  
 </tr>  
 <tr>   
 <td><a href="/genre/Action">Action</a></td>  
 <td><a href="/genre/Adventure">Adventure</a> </td>  
 <td><a href="/genre/Animation">Animation</a></td>  
 </tr>  
 <tr>  
 <td><a href="/genre/Comedy">Comedy</a> </td>  
 <td><a href="/genre/Crime">Crime</a> </td>  
 <td><a href="/genre/Drama">Drama</a> </td>  
 </tr>  
 …

</td></tr></tbody>

Retrieval method: get all the links (<a> elements) in the page and check whether their href attribute contains the substring “genre”.

## Movies from a genre page

In <html> -> <body>, 2nd table, 1st row, 3rd data cell:

<td valign="top">

<br>

<h1>Sci-Fi Movie Scripts</h1>

<p><a href="/Movie Scripts/12 Monkeys Script.html" title="12 Monkeys Script">12 Monkeys</a> (1994-06 Draft)<br><i>Written by David Peoples,Janet Peoples</i><br></p>

<p><a href="/Movie Scripts/2001: A Space Odyssey Script.html" title="2001: A Space Odyssey Script">2001: A Space Odyssey</a> (1989-02 Draft)<br><i>Written by Stanley Kubrick,Arhur C. Clarke</i><br></p>

<p><a href="/Movie Scripts/2012 Script.html" title="2012 Script">2012</a> (2008-02 Second draft)<br><i>Written by Roland Emmerich,Harald Kloser</i><br></p>

<p><a href="/Movie Scripts/28 Days Later Script.html" title="28 Days Later Script">28 Days Later</a> (Undated Draft)<br><i>Written by Alex Garland</i><br></p>

…  
<p><a href="/Movie Scripts/X-Men Origins: Wolverine Script.html" title="X-Men Origins: Wolverine Script">X-Men Origins: Wolverine</a> (Undated Draft)<br><i>Written by David Benioff,Skip Woods</i><br></p>

</td>

Again, for the URLs get all <a> elements with “Scripts” in the href attribute.

Problem:

|  |  |  |
| --- | --- | --- |
| **Latest Comments** | | |
| [Reservoir Dogs](https://imsdb.com/Movie%20Scripts/Reservoir%20Dogs%20Script.html) | 10/10 |  |
| [How to Train Your Dragon](https://imsdb.com/Movie%20Scripts/How%20to%20Train%20Your%20Dragon%20Script.html) | 10/10 |  |
| [Scream](https://imsdb.com/Movie%20Scripts/Scream%20Script.html) | 10/10 |  |
| [Groundhog Day](https://imsdb.com/Movie%20Scripts/Groundhog%20Day%20Script.html) | 10/10 |  |
| [Black Panther](https://imsdb.com/Movie%20Scripts/Black%20Panther%20Script.html) |  |  |

This section also contains links to movies, and thus gets added for every genre. How to exclude it?

One of the elements out-of-place is:

<td><a href="/Movie Scripts/Reservoir Dogs Script.html">Reservoir Dogs</a></td>

Pick only the paragraphs (<p>) and then extract the <a>s. We achieve this by specifying the outer element that the link must be found in (it can be just <a>, or <p>).

Important note: since some movies are present in more than one genre, it is necessary to reunite all the lists into one and make it into a set() to eliminate duplicates.

## The script page for the movie

<a href="/scripts/Dances-with-Wolves.html">Read "Dances with Wolves" Script</a>, contained in a <td>

“Midnight in Paris” fails, and why? Because the request does not load the page. Which means, because the actual URL is  
<https://imsdb.com/Movie%20Scripts/Midnight%20in%20Paris%20Script.html>, i.e. with whitespaces as %20

Other problem: When “.next” is not opportune to pick the link inside the outer element, because it has a number of child elements (and not all of them are links)?  
Solution: consider the children nodes, check whether they are <a>.

Observation: some movies (“Casablanca”, “Harry Potter and the Prisoner of Azkaban”, etc.) actually have no script left on the site.  
Number of movies from the genre lists, including duplicates and no-script: 3205  
Number of script pages: 1177

Example scripts: 'https://imsdb.com/scripts/Matrix,-The.html', 'https://imsdb.com/scripts/Jane-Eyre.html', etc.etc.

# The Bechdel test

We are interested in looking for movies that pass the Bechdel test (<https://en.wikipedia.org/wiki/Bechdel_test>).

i.e. if it satisfies the following requirements:

* The movie has to have at least two women in it,
* who talk to each other,
* about something other than a man.[[10]](https://en.wikipedia.org/wiki/Bechdel_test#cite_note-10)[[11]](https://en.wikipedia.org/wiki/Bechdel_test#cite_note-11)[[12]](https://en.wikipedia.org/wiki/Bechdel_test#cite_note-AB-12)

We've found a website that has a large collection of movie scripts - <http://www.imsdb.com/>.

We would like you to scrape a few hundred scripts from this website (doesn't have to be all of them), and look for instances where a female character is talking to another female character. Produce some statistics for the movies you have data for, that you think would be relevant to the Bechdel test.

The script in the HTML, and the various fields:  
Inside <td class=scrtext>   
 <pre>  
we find the text:

<b> EXT. NIGHT. THE MOOR/MOOR HOUSE.</b>

Jane is toiling on through the lashing rain towards the light. It has become a window. A brief flash of lightning shows her a low stone cottage. Helen Burns is sitting on the gate.

<b> CUT TO: </b>

Jane knocking at the door. Hannah, an old servant answers.  
 She is suspicious; Jane looks like a wretch.

<b> HANNAH</b>

What do you want?  
 Jane manages to find her voice.

<b> JANE</b>

Shelter

Another excerpt, this time from “Dances with wolves”:

**INT. FIELD HOSPITAL TENT - DAY**

A black screen.

The sound of a knife cutting through boot leather.

Fade in on the waists of two men (THE SURGEONS) hovering around a crude operating table.

In the extreme background, TWO STRETCHER BEARERS are just leaving.

**<b> SURGEON 1 (O.S.)</b>**  
 Is this the last one?

One of the bearers stops and looks back. His face is numb.

And he nods hollowly.

We cannot see the patient stretched out on the table. But we do see that the first surgeon has succeeded in getting the

man's boot off.

**SURGEON 2 (O.S.)**

God, what a mess... at least there's

no gangrene.

**SURGEON 1 (O.S.)**

There will be if it doesn't come

off.

**SURGEON 2 (O.S.)**

Well I can't saw if I can't keep my

eyes open. Let's coffee up... he

can wait a few more minutes.

Last excerpt:

I'm inside. Anything to report?  
 We listen to the phone conversation as though we were on  
 a third line. The man's name is CYPHER. The woman,   
<b> TRINITY.</b>

<b> CYPHER (V.O.)</b>

Let's see. Target left work at <b> 5:01 PM.</b>

<b> SCREEN</b>

Trace program: running.  
The entire screen fills with racing columns of numbers.  
 Shimmering like green-electric rivets, they rush at a 10-  
 digit phone number in the top corner.

<b> TRINITY (V.O.)</b>

All right, you're relieved. Use the usual exit.

<b> CYPHER (V.O.)</b>

Do you know when we're going to make contact?

<b> TRINITY</b>

Soon.

Only two thin digits left.

<b> CYPHER (V.O.)</b>

Just between you and me, you don't believe it, do you? You don't believe this guy is the one?

<b> TRINITY (V.O.)</b>

I think Morpheus believes he is.

Observations and Idea 1:

Get the sequence of all the bold (<b>) elements in the script. We must determine the sequences where characters speak to each other.  
We imagine that all the bold fragments containing “INT” or “EXT” or “CUT” should be kept, but obviously not considered among the characters.

However, getting the bold segments excludes the text of the characters speaking. We should analyze it to determine whether it includes the name of a male character.

Thus, the steps become::

* Get character names: find all the bold segments, exclude INT, EXT, CUT, DRAFT, strings made of numbers & punctuation (e.g.17:01), ~~words tagged as adjectives in NLTK (e.g. desperate)~~ anything with parentheses. The remaining set should contain the character names
* Use the downloaded resource from BehindTheName to determine the gender of the speaker
* Read the script, and apply our chosen criteria for the Bechdel test:
  + there must be moments where 2 female characters talk with each other, without mentioning the name of a male character in any of the lines. The length of these exchanges can be: 4 lines, 8 lines.   
    This must be measured against the total number of 4-line and 8-line exchanges between characters.  
    All scene changes (INT, EXT, CUT) separate the different dialogues.

Character names, checking some examples:

|  |  |  |
| --- | --- | --- |
| Jane Eyre | The Matrix | The Fellowship of the Ring |
| ['ADELE', 'MISS TEMPLE', 'HELEN', 'UNDER COOK', 'JANE', 'MASON', 'GRACE', 'MARY', 'MADAME PIERROT', 'WOOD', 'LOUISA ESHTON', 'BRIGGS', 'HANNAH', 'MISS SCATCHERD', 'ROCHESTER', 'ST JOHN', 'JANE EYRE', 'LADY INGRAM', 'BROCKLEHURST', 'JOHN REED', 'BESSIE', 'MRS FAIRFAX', 'DR CARTER', 'JOHN', 'LEAH', 'LORD INGRAM', 'MISS ABBOT', 'DIANA', 'MRS REED', 'BLANCHE', 'SOPHIE'] | ['ORACLE', 'CYPHER', 'OLD MAN', 'MOJO', 'COMPUTER SCREEN', 'ON COMPUTER SCREEN', 'MOMMY', 'AGENT SMITH', 'AGENT JONES', 'TANK', 'THE END', 'SCREEN', 'BOY', 'DOZER', 'BIG COP', 'COPS', 'RHINEHEART', 'TALL EMPLOYEE', 'NEO', 'LIEUTENANT', 'MORPHEUS', 'DUJOUR', 'ANGLE ON NEO', 'FEDEX', 'CABLE', 'PRIESTESS', 'THE MATRIX', 'REX', 'TRINITY', 'SPOON BOY', 'ANTHONY', 'AGENT BROWN', 'APOC', 'SERGEANT', 'MOUSE', 'AGENT BROWTJ', 'COP', 'GIZMO', 'PILOT', 'SWITCH'] | ['', 'FRODO DISAPPEARS', 'MERRY', 'ODO PROUDFOOT', 'BLACK RIDER', 'PIPPIN', 'HOBBIT BOUNDER', 'GIMLI', 'ORC OVERSEER', 'FARMER MAGGOT', 'MERRY AND PIPPIN', 'VOICE OF SAURON', 'BILBO', 'GALADRIEL', 'ARAGORN', 'STRIDER', 'LEGOLAS', 'SARUMAN', 'LURTZ', 'THE END', 'GOLLUM', 'ELROND SURVEYS THE GROUP', 'HALDIR', 'ELROND', 'BOROMIR', 'SAM', 'FRO DO', 'GATEKEEPER', 'ARWEN', 'FRODO', 'CELEBORN', 'ISILDUR', 'FADE TO BLACK', 'BUTTERBUR', 'WITCH KING', 'GANDALF', 'FRODO SMILES'] |

If something like “screen” or “angle on Neo” remains, the sentence will simply not be found as a name on BehindTheName or similar sources.

Let us proceed with determining the names’ gender.