Project Name：IMDb Spider

1. Functions

1.1 Crawling and storing movie review data;

1.2 Performing basic analysis on movie review data, such as identifying high-frequency words;

1.3 Presenting statistical results (displaying data directly in the first step, and displaying analyzed data in the second step).

2.Schedule

2.1 Selecting the target, completing web crawling (3 weeks).

2.2 Data analysis (3 weeks).

2.3 Data interpretation and document organization (2 weeks).

3.Steps

**It is recommended to use this video as the main framework for learning, and the other videos below as supplementary. (Link: https://www.bilibili.com/video/BV1Qr4y1b7M8?p=5&vd_source=80577c994ff26ccc24aeba51679119c7

3.1 Setting up Python development environment and learning how to use IDE, recommended IDE is PyCharm ; (Link: https://www.bilibili.com/video/BV1944y1x7SW/?spm_id_from=333.999.0.0&vd_source=80577c994ff26ccc24aeba51679119c7)

3.2 Learning and completing the development of web crawler, and normalizing the storage of result data ;(Link: https://www.bilibili.com/video/BV1Wr4y1B7Fq/?spm_id_from=333.337.search-card.all.click&vd_source=32666ac1a456f3c13150d8290a39684d)

4. API Related

4.1 legal cases

https://www.oschina.net/news/191991/linkedin-hiq-scraping-public-data

https://developer.aliyun.com/article/891291

http://tech.sina.com.cn/csj/2019-12-27/doc-iihnzhfz8598704.shtml

4.2 Normal process for applying APIs

https://blog.csdn.net/xu1988923/article/details/94622727

Linkedin: https://developer.linkedin.com/product-catalog

twitter： https://blog.csdn.net/rubinorth/article/details/52494833

4.3 Alternative approach for web scraping - crawling through third-party search engines

https://blog.csdn.net/Bone_ACE/article/details/71055153

https://github.com/LiuXingMing/LinkedinSpider/blob/master/linkedinSpider.py

5. Group Meeting Memo

Index	Date	Place	Topics	Feedback
Meeting 1	Feb 26	Zoom	Discuss and make agreement with project plan and detail ● Step in the beginning self-learning weeks ● Discuss process and problems in group chat	Resource sharing: 1 Gezi: JavaScript: https://m.youtube.com/watch?v=PkZNo7MFNFg 2. Jasmine: Web Scraping Job Postings Data from Boss Recruitment Website with Python: You can refer to its web scraping tools and the data it collects. https://www.bilibili.com/video/BV1pY4y147o6/?spm_id_from=333.1007.top_right_bar_window_default_collection.content.click&vd_source=32666ac1a456f3c13150d8290a39684d 3.Jasmine: Python's Latest and Most Comprehensive 100 Web Scraping Case Studies Tutorial, Data Analysis, Data Visualization. https://www.bilibili.com/video/BV1QZ4y1N7YA/?spm_id_from=333.788.video.desc.click&vd_source=80577c994ff26ccc24aeba51679119c7 4. Muyun: Undemy: https://www.udemy.com/course/scrapy-tutorial-web-scraping-with-python/ 5.Need to learn #css/xpath/re... and other syntaxes yourself. Python, Web Scraping, and Deep Learning (3) - Extra Episode (1) HTML Formatting and HTML Parsing with Python. https://lisper517.top/index.php/archives/41/
Meeting 2	March 6	Zoom	1.Process checking: ● python grammar ● data scraping ● django: html, css, javascript 2.Project review 3.API: ● Jasmine: boss ● Gary: zillow/	Jasmine: LinkedIn's anti-crawling measures are too strict, making it difficult to obtain data. Switch to scraping job postings from Boss Zhipin and need to use new techniques: 1. Selenium: A tool for web application testing. https://blog.csdn.net/qq_34337272/article/details/79594809 2. Chromedriver: It is an automation testing interface provided by Google for web developers, serving as a bridge for communication between Selenium 2 and Chrome bro

			● Muyun: twitter ● Gezi: 58 4.Learning resources 5.Distribution of work	wser. https://www.jianshu.com/p/31c8c9de8fcd 3. Installation of Selenium and Chromedriver in Mac environment. https://blog.csdn.net/weixin_28844235/article/details/113051669 4. Python + Selenium Web Scraping in-depth tutorial. https://developer.aliyun.com/article/951721 Muyun: Twitter feedback ask us wait and apply again, so we’d better use other possible API. （Thank you for applying for access to the Twitter API. We’re working on exciting updates including new access types and will have more to share soon. Please stay tuned to @TwitterDev and resubmit your application as soon as we launch our new API.) Gary: Stockx is a resell platform for trendy clothing, watches, game consoles, etc. It can scrape the prices and sales volume of a certain product on different dates. Here is the public API: https://developer.stockx.com/openapi/reference/overview/ Stockx API download. Reddit requires company information, but it wasn't successful. Not sure how to use Wikipedia's public API: https://en.wikipedia.org/api/rest_v1/#/Page%20content Gezi: I have completed the application for 58.com's used car API, but haven't tried calling it yet. Based on research, many domestic applications in China, such as Bilibili, WeChat, Tencent, Douban, etc., have already closed API application.
Meeting 3	March 21	NEU 225-306	IMDb Spider content: short and long reviews filter Reference: gary https://zhuanlan.zhihu.com/p/101515068 1. Spoiler Alert Filter: Muyun 2. word frequency analysis	1.user case: The Shawshank Redemption https://www.imdb.com/title/tt0111161/reviews?sort=curated&dir=desc&ratingFilter=0 2. Approach: Everyone should proceed in sequence and provide feedback in the group after completing each step.

				Related resources: 3.1 Data analysis: IMDB movie analysis report. https://zhuanlan.zhihu.com/p/34757858 3.2 Python implementation of IMDB movie top data visualization. https://blog.csdn.net/weixin_43649691/article/details/121428709
Meeting 4	March28	zoom	Complete part: Muyun: crawled all reviews with title, review, and spoiler, and exported to CSV. Gary: Crawled multiple reviews, classified them as long reviews or short reviews, and printed them. Existing review data: ● Title ● Content (long and short) ● Spoiler alert ● Rate (movie rating) Additional data to be supplemented: ● Vote (number of votes received for reviews, Jas+) ● Date (year) - 10777 records (Jas+)	Overall data analysis approach: The Shawshank Redemption is a highly rated movie. What factors lead viewers to give it a high rating? Different focus areas of reviews in terms of dimensions (frequency of script/actor/director/production company-related terms appearing in content) - Gezi ● Overall situation summarized in 14 words across 4 categories ○ No comment about film company ○ Comparison of high-rated (>5 stars) and low-rated (<=5 stars) reviews in terms of percentage ○ Comparison of long reviews (>50 words) and short reviews (<=50 words) in terms of word count (median), presented in percentage ○ Time dimension and rate (high-rated/low-rated, average), numbers of reviews Popular vote for reviews with spoilers and without spoilers (this analysis is currently not possible as vote data has not been obtained yet) Spoiler alert analysis in a pie chart – Muyun Word cloud for titles and relationship between title and rate - Gary
Meeting 5	April 4	zoom	Code review for web crawler. Code review for data analysis. Create a shared cloud drive and upload team files.	Work Feedback: Jasmine: add vote, data, rate label and crawl all the data , save to cvs; Integrated crawler data and formulated data analysis approach.

				Gezi: Completed framework for data analysis and machine learning modeling. Muyun: Spoiler & vote visualization Gary: Words Cloud visualization PPT link: https://drive.google.com/drive/folders/1hBHt41-d2PisRGhHY4Gvqen7jbMpCn1Q?usp=share_link
Meeting 6	April 8	zoom	Completed parts: 1. Final crawling + CSV file Jasmine 2. Data visualization: GEZI: Core part Muyun: Spoiler & vote Gary: Words Cloud	Task for this stage: ● Modify own code based on suggestions from the meeting, update the final version in Jupyter notebook, export as .ipynb format, and add English comments (everyone should do this for their own part). ● Add annotations (fact+analysis) to the images: Muyun and Gary write their own images, Gezi and Jas share the task of writing Gezi's images (Muyun writes the first image, Gary writes the last two images, Jas writes 2 pie charts, and Gezi writes the rest). ● Integrate the three documents of data visualization into one (Jas). ● Complete before Tuesday, and we will discuss in the meeting after the class.
Meeting 7	April 14-15	zoom	Tasks for this stage: Write documents Create a PowerPoint presentation Schedule a one-on-one meeting with the instructor	● Project Description (Purpose, Content, Technology) ● Work Distribution Chart ● Progress Report ● Summary (Technical Achievements, Collaborative Achievements) ● Appendices (Formatting + Translation) Gary: PPT 1.intro1 2.problem1 3.solution2-3 4.conclusion1 https://docs.google.com/presentation/d/1Rpvh39ruQbXkIwkJyA-YhVsrvfce1ClUiRn0MJRbtzs/edit#slide=id.g118f3725c9d_0_315

Meeting 8	April 18	group chat	Final week tasks: 1. Finalize Summary 2. Organize and Translate Group Memo 3. Modify Project Idea in Excel to Final Version	meeting time: 12:30pm Wednesday(4.19) link: meet.google.com/ihp-zyqp-csc ● Run Final Version of Code (Spider + Data Analysis) on Individual Computers to Ensure No Bugs - All Team Members (Important!) ● Fill in Questions in Remarks Column for One on One Meeting Tomorrow - All Team Members, Reach Consensus ● Compile and Upload All Files, Conduct Final Check - Jasmine Responsible for Packaging and Uploading, Deadline: Friday 5pm

Attachment: Dimension of Analysis (Extraction of Movie Review Keywords)

1.writing credits	● Stephen King(based on the short novel "Rita Hayworth and the Shawshank Redemption" by) ● Frank Darabont (screenplay by)
2.Cast	Tim Robbins ... Andy Dufresne Morgan Freeman ... Ellis Boyd 'Red' Redding Bob Gunton ... Warden Norton William Sadler ... Heywood Clancy Brown ... Captain Hadley
3.director	Frank Darabont
4.film company	Castle Rock Entertainment