RA 工作说明书

同学您好！非常感谢您愿意参与本次科研项目。您的加入为我们的研究团队注入了新的活力，相信通过我们的共同努力，此次合作将取得丰硕的成果。

为确保项目顺利推进，请您按照以下工作流程有序完成任务：

1、筛选 LinkedIn post 数据爬取名单

注册 LinkedIn 个人账号。

通过"LinkedIn_url"（"faculty_linkedin join date.xslx"文件 D 列）进入教授个人主页，点击"Show all posts"查看所有 post 具体情况。在"faculty_linkedin join date.xslx"文件中的 E 列登记教授最早发布的 post 的时间。示例如下。

	A	B	C	D	E
1	University	Name	Title	LinkedIn_url	join date
2	001_ASU	Adegoke Oke	Chair & Professor	https://www.linkedin.com/in/goke-oke-573525/	10yr
3	001_ASU	Ajay Vinze	Professor Emeritus	https://www.linkedin.com/in/ajay-vinze-5853526	6yr
4	001_ASU	Alejandro Manelli	Chair and Professor	https://www.linkedin.com/in/alejandro-manelli-b774562b	1yr
5	001_ASU	Allen Morrison	Professor of Global Strate	https://www.linkedin.com/in/allenmorrison	5yr
6	001_ASU	Amy Hillman	Professor and Rusty Lyon	https://www.linkedin.com/in/amy-hillman-8487b11b/	8yr
7	001_ASU	Andreas Kostol	Assistant Professor	https://www.linkedin.com/in/andreas-kostol-19b388b9/	3yr
8	001_ASU	Aneel Iqbal	Assistant Professor of Glo	https://www.linkedin.com/authwall?trk=gf&trkInfo=AQEIv	11yr
9	001_ASU	Anne Tsui	Emeritus Professor	https://www.linkedin.com/in/anne-tsui-605b1b14a?origin:	1yr
10	001_ASU	Benjamin Shao	Associate Dean for Asia-I	https://www.linkedin.com/in/benshao/	2yr

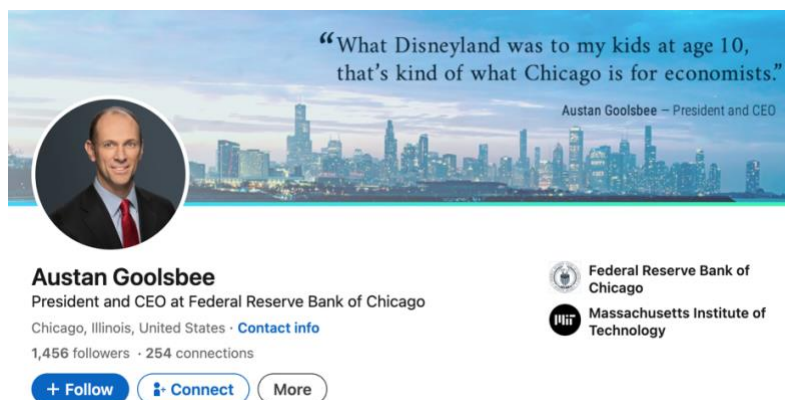接着，查看该教授的 post 是否满足以下条件：①3 年前或更早开始发布 post；②有≥3 个属于不同年份的发帖记录。

若同时满足，则进入后续的数据爬取环节；若没有同时满足，则直接跳过，不进行数据爬取。

2、代码爬取个人 post 基本信息

下载谷歌浏览器。

利用 selenium 驱动浏览器爬取 LinkedIn post 的相关信息。其中包括教授名称、发表 post 总数、post 发布时间、post 文本内容、携带媒体类型及其链接，以及每条 post 的 likes、comments、shares 的具体数量。

下面以教授 Austan Goolsbee 为例进行演示。

具体操作过程：

打开 python 代码文件"linkedin-post-scraper.py"，在代码中补充自己的 LinkedIn 账号的用户名和密码。

```python
# LinkedIn Credentials
username = " " # 填写LinkedIn账户用户名
password = " " # 填写LinkedIn账户密码
```

接着，在下面位置粘贴教授对应的 Linkedin_url，示例如下。

```python
# Set LinkedIn page URL for scraping
page = 'https://www.linkedin.com/in/austan-goolsbee-62b278277/'
```

运行代码，在.py 所属文件夹中生成.csv 文件，数据爬取结果如下。

Page	Number	Date	Post Text	Media Type	Likes	Comments	Shares	Media Link
Austan Goolsbee	20	1w • 1 week ago	The most important indicat	Image	27	2 comments	1 repost	None
Austan Goolsbee	20	2w • 2 weeks ago	Many thanks to Jack Lavin,	Image	63	1 comment	0	None
Austan Goolsbee	20	1mo • 1 month ago	Thrilled to host the 31st ann	Image	58	2 comments	3 reposts	None
Austan Goolsbee	20	5mo • 6 months ago	Looking forward to Monday	Article	44	0	1 repost	https://www.c
Austan Goolsbee	20	8mo • 9 months ago	Had a great time with Iowa	Image	43	0	1 repost	None
Austan Goolsbee	20	9mo • 10 months ago	Many thanks to the Econom	Article	64	2 comments	1 repost	https://www.c
Austan Goolsbee	20	10mo • 10 months ago	Excited to be in Minnesota	Article	48	1 comment	0	https://www.c
Austan Goolsbee	20	10mo • 11 months ago	Enjoyed today's conversatio	Image	52	0	2 reposts	None
Austan Goolsbee	20	10mo • 11 months ago	I'm looking forward to joinir	Article	27	1 comment	0	https://www.c
Austan Goolsbee	20	1yr • Edited • 1 year ago	This week, I had the opport	Image	56	0	0	None
Austan Goolsbee	20	1yr • 1 year ago	"After several quarters of st	Video	41	1 comment	2 reposts	None
Austan Goolsbee	20	1yr • 1 year ago	Tomorrow, November 14, I'	Image	23	2 comments	1 repost	None
Austan Goolsbee	20	1yr • 1 year ago	We can't wait to welcome vi	Article	49	0	2 reposts	http://bit.ly/4
Austan Goolsbee	20	1yr • 1 year ago	In partnership with The Sac	Article	37	0	0	https://bit.ly/
Austan Goolsbee	20	1yr • 1 year ago	Fed Governor Michelle Bow	Image	52	1 comment	1 repost	None
Austan Goolsbee	20	1yr • 1 year ago	Last week, I met with busin	Image	83	1 comment	0	None
Austan Goolsbee	20	1yr • 1 year ago	Our Supervision & Regulati	Image	35	0	2 reposts	None
Austan Goolsbee	20	1yr • 1 year ago	I enjoyed discussing fighting	Article	12	0	0	https://on.wsj
Austan Goolsbee	20	1yr • 1 year ago	Today is World Elder Abuse	Unknown	22	1 comment	0	None
Austan Goolsbee	20	1yr • 1 year ago	One of the great parts of my	Image	7	0	0	None

备注：代码有关问题可参考网页：

https://christophegaron.com/articles/mind/automation/scraping-

linkedin-posts-with-selenium-and-beautiful-soup/

3、数据整理

将所有爬取的 LinkedIn post 数据整理到"faculty_linkedin

post.xslx"文件中，名字顺序请依照"faculty_linkedin join date.xslx"文

件中的排序，不满足数据爬取条件的教授可直接跳过。