

هدف

کراتل کردن سایت Microsoft Academics برای واکنشی ۵۰۰۰ مقاله.
محاسبه معیار pagerank برای مقالات واکنشی شده.

شرح

پیاده سازی بارگذار صفحات

برای بارگذاری صفحه مربوط به یک مقاله و استخراج اطلاعات مورد نیاز از selenium در python استفاده کردیم.
برای بارگذاری یک صفحه وب نیاز به webdriver است. در این پروژه از chromedriver استفاده کردیم.
پس از ستاپ کردن درایور یک تابع پیاده سازی کردیم تا با ورودی گرفتن id یک مقاله (که در url قرار دارد) اطلاعات خواسته شده (title, abstract, date, authors, references) را استخراج کرده و به صورت یک آبجکت خروجی دهد.
به دلیل اینکه صفحات سایت از ریکوئست های ajax برای لود بخش های مختلف صفحه استفاده می کنند، پس از لود اولیه صفحه باید مقداری صبر کنیم تا المان های مورد نظر در صفحه قرار بگیرند. این سایت ابتدا اطلاعات اولیه مقاله را لود کرده و سپس لیست ارجاعات را بارگذاری کرده و به صفحه اضافه می کند. طبق این مشاهده معیار لود شدن یک صفحه را دیده شدن یک المان مربوط به ارجاعات قرار می دهیم. این المان را با CSS Selector زیر مشخص کرده و بعد از درخواست دادن برای لود صفحه تا زمانی که این المان به صفحه اضافه شود صبر می کنیم.

```
PAGE_SELECTOR = "#mainArea router-view router-view div.results  
div.results ma-card .primary_paper"
```

در صورتی که selenium نتواند در یک زمان مشخص شده المان ذکر شده را پیدا کند اکسپشن TimeoutException را ریز می کند. این اتفاق ممکن است به دلیل درخواست های زیاد و ریدایرک شدن به صفحه ی اصلی یا خالی بودن لیست ارجاعات رخ دهد. در این صورت پس گذشتن بازه زمانی مشخص شده (۲۰ ثانیه) درخواست را دوباره تکرار می کنیم و این کار را تا دو بار تکرار می کنیم و در صورت تکرار خطا سراغ مقاله ی بعدی می رویم. به این ترتیب بدون وقفه صفحات را لود کرده و فقط در صورت اعمال محدودیت از طرف سایت به مقدار لازم صبر می کنیم.

بعد از لود صفحه اطلاعات خواسته شده را از المان های مربوطه با استفاده از CSS Selector های زیر استخراج می کنیم.

```
TITLE_SELECTOR = "#mainArea h1.name"  
ABSTRACT_SELECTOR = "#mainArea > router-view > div > div > div >  
div > p"  
DATE_SELECTOR = "#mainArea > router-view > div > div > div > div  
> a > span.year"  
AUTHORS_SELECTOR = "#mainArea > router-view > div > div > div >  
div > ma-author-string-collection > div > div.authors  
.author-item.au-target a.au-target.author.link"
```

```
REFERENCES_SELECTOR = "#mainArea > router-view > router-view  
div.results > div > compose > div > div.results > ma-card  
div.primary_paper > a.title.au-target"
```

پیاده سازی خزشگر

خزشگر دارای یک صف است. این صف ابتدا با id سه مقاله داده شده مقدار دهی شده است. پس از لود هر صفحه ارجاعات آن مقاله در صورتی که قبلا به صف اضافه نشده باشند به صف اضافه می شود. برای بررسی این موضوع بعد از اضافه کردن یک مقاله به صف id آن به یک set اضافه می شود تا بررسی اینکه آیا یک مقاله تکراری است یا نه در $O(1)$ انجام شود. این فریاد تا رسیدن تعداد مقالات به تعداد مطلوب ادامه پیدا می کند.

محاسبه معیار pagerank

برای محاسبه این معیار ابتدا یک گراف جهت دار از روی لیست مقالات بدست آمده در بخش قبل ساختیم. در صورتی که مقاله A به مقاله B ارجاع داشته باشد. یک یال جهت دار از A به B در گراف وجود دارد. سپس با استفاده از تابع pagerank در کتابخانه networkx معیار pagerank را محاسبه کردیم.

لیست مقالات برتر به ازای $\alpha=0.85$ به این صورت بدست آمد:

rank	id	pagerank	title	# of refs to this paper
1	2310919327	0.001534219116286	Gradient-based learning applied to document recognition	133
2	2156909104	0.001505018049476	The Nature of Statistical Learning Theory	92
3	2618530766	0.001466219402838	ImageNet classification with deep convolutional neural networks	160
4	2049633694	0.001223988972697	Maximum likelihood from incomplete data via the EM algorithm	119
5	2154642048	0.000978804579047	Learning internal representations by error propagation	107
6	2136922672	0.00090482705138	A fast learning algorithm for deep belief nets	83
7	2097117768	0.000874230713392	Going deeper with convolutions	78

8	1652505363	0.000862894613811	Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations	62
9	2108598243	0.000856537684361	ImageNet: A large-scale hierarchical image database	77
10	2194775991	0.000846058969397	Deep Residual Learning for Image Recognition	103
11	2132260239	0.000818621927651	Identification of a novel coronavirus in patients with severe acute respiratory syndrome.	76
12	2166867592	0.000813673599856	Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia	82
13	2962835968	0.000802041241823	Very Deep Convolutional Networks for Large-Scale Image Recognition	94
14	2025170735	0.000766657046249	Coronavirus as a possible cause of severe acute respiratory syndrome	65
15	2104548316	0.000764705672361	A novel coronavirus associated with severe acute respiratory syndrome.	72
16	2102605133	0.000759526943601	Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation	60
17	2100495367	0.000758211565537	Reducing the Dimensionality of Data with Neural Networks	71
18	2116064496	0.000757817200963	Training products of experts by minimizing contrastive divergence	67
19	2148603752	0.000705306094271	Statistical learning theory	64
20	3017143921	0.000697825054301	Pattern classification and scene analysis	83