به نام خدا

گزارش تمرین پنجم بازیابی پیشرفته اطلاعات

ارزش‌گذاری مقالات علمی براساس ارجاعات

گروه ۳۰

پارسا حسینی

درسا مجدی

امیررضا باقری

چکیده روند پروژه

در این تمرین هدف آن است که مقالات را براساس ارجاعات، با استفاده از روش‌های تحلیل لینک ارزش‌گذاری کنیم. در مرحله اول، از دیتاست آماده شده در مرحله قبل استفاده کرده و پس از تمیزسازی، مقالاتی که زمینه علمی‌شان علوم کامپیوتر (Computer Science) و ریاضی (Mathematics) هستند، را جدا کردیم. سپس برای هر مقاله، k تا ارجاع اول آن را در نظر گرفته، و با استفاده از آنها گراف ارجاعات را می‌سازیم. پس از ساخت گراف ارجاعات، دو الگوریتم تحلیل لینک PageRank و HITS بر روی آنها اعمال و نتایج حاصل از آنها با یکدیگر مقایسه شدند.

بخش ۱. آماده‌سازی داده‌ها

از دیتاست آماده شده در تمرین قبلی استفاده می‌کنیم و ابتدا آن را براساس بخش‌های نویسنده (authors) و ارجاعات (references) به صورت زیر تمیز کردیم.

```
[ ]  def dropna_references(references):
         res = []
         for dic in references:
             if dic['paperId'] != None and dic['authors'] != None:
                 res.append(dic)
         return res
     def dropna_authors(authors):
         res = []
         for dic in authors:
             if dic['authorId'] != None and dic['name'] != None:
                 res.append(dic)
         return res
     df['references'] = df['references'].apply(dropna_references)
     df['authors'] = df['authors'].apply(dropna_authors)
     df.head()
```

Figure 1. تمیز کردن مقالات براساس بخش‌های نویسنده و ارجاعات

در مرحله بعدی، تصمیم‌گیری می‌شود که کدام یک از مقالات برای اجرای الگوریتم انتخاب شود. از بین کل مقالات، ابتدا آنهایی که زمینه علمی (Field of Study) آنها علوم کامپیوتر (Computer Science) یا ریاضی (Mathematics) است، جدا شد. سپس از بین مقالات جدا شده، تنها آنهایی نگه داشته می‌شوند که تعداد استنادها به آن (CitationCount) از ۵۰ بیشتر باشد. نهایتا با توجه به اینکه تنها به تعداد مشخصی از ارجاعات هر مقاله نیاز داریم، برای هر مقاله kتا ارجاع اول آن را نگه می‌داریم (مقدار پیش فرض برای k برابر با 10 است).

```
[ ]  df = df[df['fieldsOfStudy'].apply(str).str.contains('Computer Science') |
     df['fieldsOfStudy'].apply(str).str.contains('Mathematics')]
     df.shape

     (16226, 9)
```

Figure 2. جدا کردن مقالات با زمینه علمی Computer Science و Mathematics

```
[ ]  df = df[df['citationCount'] > 50]
     df.shape

     (7136, 9)
```

Figure 3. جدا کردن مقالات با تعداد Citation بیشتر از ۵۰

```
[ ]  def get_k_first_references(references, k=10):
         if len(references) <= k:
             return references
         return references[:k]
     df['k_references'] = df['references'].apply(get_k_first_references)
     df.head()

     /usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5: SettingWithCopyWarning:
     A value is trying to be set on a copy of a slice from a DataFrame.
     Try using .loc[row_indexer,col_indexer] = value instead

     See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
       """
```

Figure 4. نگه داشتن k ارجاع اول هر مقاله در ستون k_references

نهایتا داده‌های آماده شده در فایل clean_data.json ذخیره شده تا در مرحله بعدی برای اجرای الگوریتم‌های PageRank و HITS استفاده شود.

بخش ۲. تحلیل لینک

ابتدا با استفاده از کتابخانه‌ی Networkx و تابع Digraph گراف جهتدار ارجاعات مقالات مد نظر را ساختیم. توجه شود که در ساخت این گراف مقالاتی که در لیست مقالات هدف نبوده‌اند ولی به آنها ارجاع داده شده است هم به عنوان راس این گراف در نظر گرفته شده‌اند.

```
1 def get_edges():
2     edges = []
3     for index, paper in df.iterrows():
4         paperId = paper['paperId']
5         references = [dic['paperId'] for dic in paper['k_references']]
6         for reference in references:
7             edges.append((paperId, reference))
8     return edges
```

```
[ ]  1 G = nx.DiGraph()
     2 G.add_edges_from(get_edges())
     3 print(f"Number of nodes = {G.number_of_nodes()}")
     4 print(f"Number of edges = {G.size()}")

     Number of nodes = 39738
     Number of edges = 62510
```

برای تحلیل لینک هم از الگوریتم Page Rank و هم از الگوریتم HITS استفاده شده است. هردوی این الگوریتم-ها با توابع آماده‌ی Networkx محقق شده‌اند.

در الگوریتم Page Rank مقالات به ترتیب امتیازشان مرتب‌سازی شده و kتای بیشینه‌ی آن‌ها خروجی‌های ارزشگذاری از طریق Page Rank هستند.

	paperId	title	year	referenceCount	citationCount	fieldsOfStudy
0	204e3073870fae3d05bcbc2f6a8e263d9b72e776	Attention is All you Need	2017	44	34731	[Computer Science]
1	0b44fcbeea9415d400c5f5789d6b892b6f98daff	Building a Large Annotated Corpus of English: ...	1993	75	8163	[Computer Science]
2	05dd7254b632376973f3a1b4d39485da17814df5	SQuAD: 100,000+ Questions for Machine Comprehe...	2016	31	4240	[Computer Science]
3	077f8329a7b6fa3b7c877a57b81eb6c18b5f87de	RoBERTa: A Robustly Optimized BERT Pretraining...	2019	58	7161	[Computer Science]
4	0b544dfe355a5070b60986319a3f51fb45d1348e	Learning Phrase Representations using RNN Enco...	2014	39	14984	[Computer Science, Mathematics]
5	44d2abe2175df8153f465f6c39b68b76a0d40ab9	Long Short-Term Memory	1997	68	51403	[Computer Science, Medicine]
6	1af68821518f03568f913ab03fc02080247a27ff	Neural Machine Translation of Rare Words with ...	2015	53	4775	[Computer Science]
7	330da625c15427c6e42ccfa3b747fb29e5835bf0	Efficient Estimation of Word Representations i...	2013	43	21777	[Computer Science]
8	2c03df8b48bf3fa39054345bafabfeff15bfd11d	Deep Residual Learning for Image Recognition	2015	61	94458	[Computer Science]
9	084c55d6432265785e3ff86a2e900a49d501c00a	Foundations of statistical natural language pr...	1999	293	7781	[Computer Science]
10	1f6ba0782862ec12a5ec6d7fb608523d55b0c6ba	Convolutional Neural Networks for Sentence Cla...	2014	34	10029	[Computer Science]
11	0e6824e137847be0599bb0032e37042ed2ef5045	Aligning Books and Movies: Towards Story-Like ...	2015	52	1411	[Computer Science]
12	3febb2bed8865945e7fddc99efd791887bb7e14f	Deep Contextualized Word Representations	2018	65	7946	[Computer Science]
13	d2c733e34d48784a37d717fe43d9e93277a8c53e	ImageNet: A large-scale hierarchical image dat...	2009	24	27007	[Computer Science]
14	10f97f1fb4f5c2c8e6c44d4a33da46d331dd4aeb	Introduction to the CoNLL-2003 Shared Task: La...	2003	39	2952	[Computer Science]

در الگوریتم HITS مقالات به ترتیب امتیاز authority شان مرتب شده و k مقاله‌ی با authority بیشینه به عنوان خروجی این الگوریتم برگردانده می‌شوند.

	paperId	title	year	referenceCount	citationCount	fieldsOfStudy
0	204e3073870fae3d05bcbc2f6a8e263d9b72e776	Attention is All you Need	2017	44	34731	[Computer Science]
1	077f8329a7b6fa3b7c877a57b81eb6c18b5f87de	RoBERTa: A Robustly Optimized BERT Pretraining...	2019	58	7161	[Computer Science]
2	05dd7254b632376973f3a1b4d39485da17814df5	SQuAD: 100,000+ Questions for Machine Comprehe...	2016	31	4240	[Computer Science]
3	1af68821518f03568f913ab03fc02080247a27ff	Neural Machine Translation of Rare Words with ...	2015	53	4775	[Computer Science]
4	2c03df8b48bf3fa39054345bafabfeff15bfd11d	Deep Residual Learning for Image Recognition	2015	61	94458	[Computer Science]
5	0b544dfe355a5070b60986319a3f51fb45d1348e	Learning Phrase Representations using RNN Enco...	2014	39	14984	[Computer Science, Mathematics]
6	3febb2bed8865945e7fddc99efd791887bb7e14f	Deep Contextualized Word Representations	2018	65	7946	[Computer Science]
7	1e077413b25c4d34945cc2707e17e46ed4fe784a	Universal Language Model Fine-tuning for Text ...	2018	56	2237	[Computer Science]
8	0e6824e137847be0599bb0032e37042ed2ef5045	Aligning Books and Movies: Towards Story-Like ...	2015	52	1411	[Computer Science]
9	44d2abe2175df8153f465f6c39b68b76a0d40ab9	Long Short-Term Memory	1997	68	51403	[Computer Science, Medicine]
10	0c908739fbff75f03469d13d4a1a07de3414ee19	Distilling the Knowledge in a Neural Network	2015	14	8617	[Mathematics, Computer Science]
11	d2c733e34d48784a37d717fe43d9e93277a8c53e	ImageNet: A large-scale hierarchical image dat...	2009	24	27007	[Computer Science]
12	1fa9ed2bea208511ae698a967875e943049f16b6	HuggingFace's Transformers: State-of-the-art N...	2019	80	2709	[Computer Science]
13	128cb6b891aee1b5df099acb48e2efecfcff689f	The Winograd Schema Challenge	2011	46	690	[Computer Science]
14	330da625c15427c6e42ccfa3b747fb29e5835bf0	Efficient Estimation of Word Representations i...	2013	43	21777	[Computer Science]
15	0b44fcbeea9415d400c5f5789d6b892b6f98daff	Building a Large Annotated Corpus of English: ...	1993	75	8163	[Computer Science]

در هر دو الگوریتم مقالاتی که خروجی ارزشگذاری هستند برای نمایش مشخصات مبسوط‌تر دوباره طریق IDشان از طریق تابع request_papers_by_id از API سایت semanticscholar گرفته می‌شوند.

بخش 3. ارزیابی و نتیجه‌گیری

همان‌طور که در خروجی‌های الگوریتم Page Rank و HITS مشاهده می‌شود، مقاله‌ی مهمی مانند Attention is all you need در رتبه‌ی اول قرار گرفته است و این نشان از کارایی این الگوریتم‌ها دارد. البته باید توجه داشت که چون این مقاله در مجموعه‌ی مقالات اولیه‌ی ما نبوده و در نتیجه یال‌های خروجی آن در نظر گرفته نشده، ارزش نسبی احتمالا بالاتری از حالتی که داخل مقالات می‌بود به خود اختصاص داده است.

همچنین برای مقایسه‌ی خروجی‌های دو الگوریتم با هم، رتبه‌ی هر مقاله‌ی مشترک در خروجی‌های Page Rank و HITS در کنار هم قرار داده شده‌اند. مشاهده می‌شود 16 مورد از 20 خروجی این دو الگوریتم با هم مشترک بوده، همچنین مقاله‌ی اول نیز در هردو یکسان بوده است. این موضوع هم راستایی کلی این دو الگوریتم در ارزش دادن به مقاله‌های پرارجاع را تصدیق می‌کند.

	paperId	title	year	referenceCount	citationCount	fieldsOfStudy	HITS rank	Page Rank rank
0	204e3073870fae3d05bcbc2f6a8e263d9b72e776	Attention is All you Need	2017	44	34731	[Computer Science]	1	1
1	077f8329a7b6fa3b7c877a57b81eb6c18b5f87de	RoBERTa: A Robustly Optimized BERT Pretraining...	2019	58	7161	[Computer Science]	2	4
2	05dd7254b632376973f3a1b4d39485da17814df5	SQuAD: 100,000+ Questions for Machine Comprehe...	2016	31	4240	[Computer Science]	3	3
3	1af68821518f03568f913ab03fc02080247a27ff	Neural Machine Translation of Rare Words with ...	2015	53	4775	[Computer Science]	4	7
4	2c03df8b48bf3fa39054345bafabfeff15bfd11d	Deep Residual Learning for Image Recognition	2015	61	94458	[Computer Science]	5	9
5	0b544dfe355a5070b60986319a3f51fb45d1348e	Learning Phrase Representations using RNN Enco...	2014	39	14984	[Computer Science, Mathematics]	6	5
6	3febb2bed8865945e7fddc99efd791887bb7e14f	Deep Contextualized Word Representations	2018	65	7946	[Computer Science]	7	13
7	1e077413b25c4d34945cc2707e17e46ed4fe784a	Universal Language Model Fine-tuning for Text ...	2018	56	2237	[Computer Science]	8	19
8	0e6824e137847be0599bb0032e37042ed2ef5045	Aligning Books and Movies: Towards Story-Like ...	2015	52	1411	[Computer Science]	9	12
9	44d2abe2175df8153f465f6c39b68b76a0d40ab9	Long Short-Term Memory	1997	68	51403	[Computer Science, Medicine]	10	6