آزمایشگاه فناوری وب
Web Technology Lab

From Local to Global

A GraphRAG Approach for Query-Focused Summarization

Microsoft Research

ارائه دهنده: علی معمری

اردیبهشت ماه ۱۴۰۴

From Local to Global: A GraphRAG Approach to Query-Focused Summarization

Darren Edge[1†] **Ha Trinh**[1†] **Newman Cheng**[2] **Joshua Bradley**[2] **Alex Chao**[3]

Apurva Mody[3] **Steven Truitt**[2] **Dasha Metropolitansky**[1] **Robert Osazuwa Ness**[1]

Jonathan Larson[1]

[1]Microsoft Research
[2]Microsoft Strategic Missions and Technologies
[3]Microsoft Office of the CTO

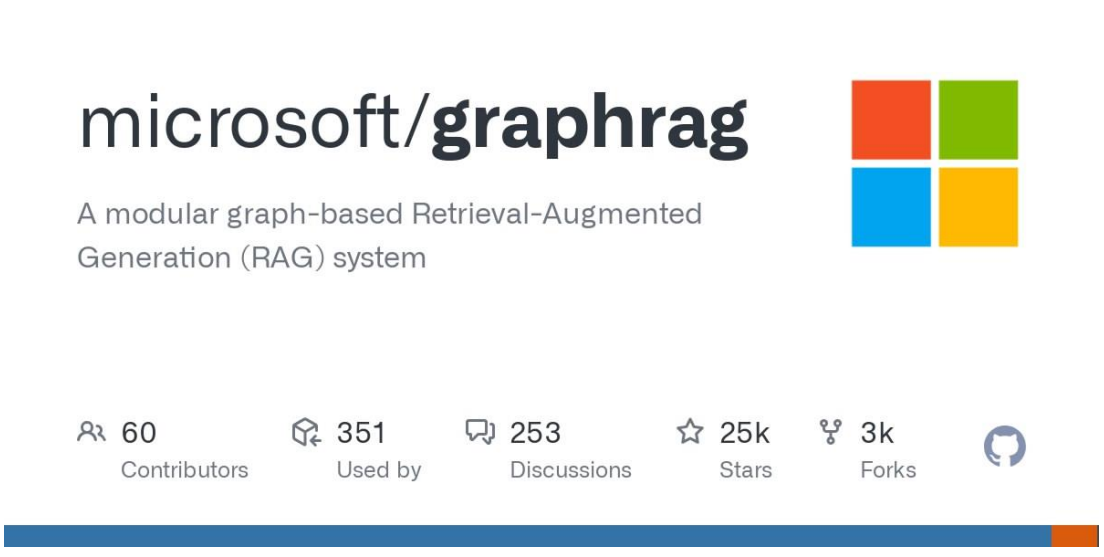
{daedge,trinhha,newmancheng,joshbradley,achao,moapurva,
steventruitt,dasham,robertness,jolarso}@microsoft.com

[†]These authors contributed equally to this work

DOI:2404.16130.viXra/10.48550v2
[Submitted on 24 Apr 2024 (v1), last revised 19 Feb 2025 (this version, v2)]

مرجع

Semantic Scholar

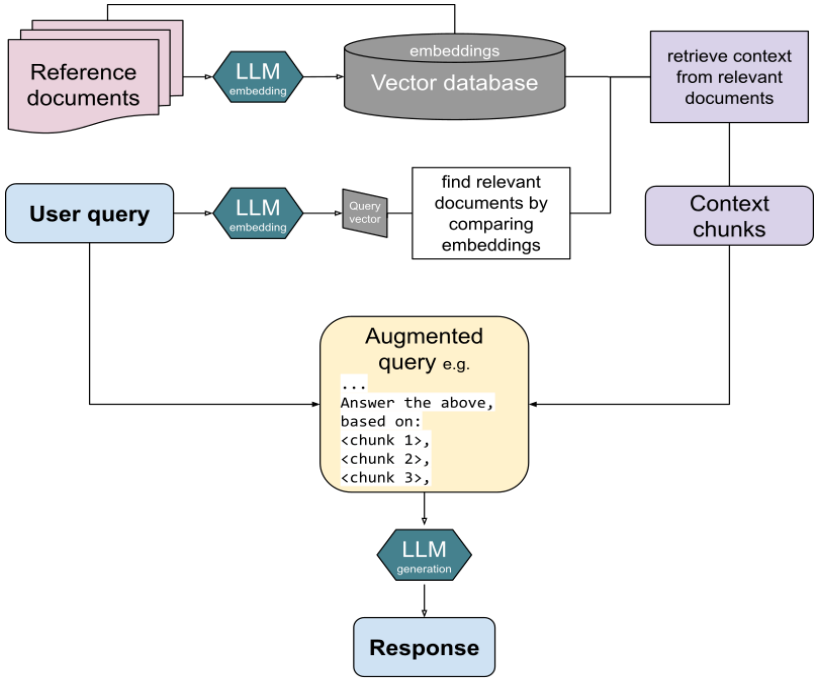


https://github.com/microsoft/graphrag

۱. مقدمه

❑ محدودیت RAG برداری (vector RAG)

- ❑ به مدل‌های زبانی بزرگ این امکان را می‌دهد تا به سؤالات در مورد اسناد خارجی/خصوصی که قبلاً بر روی آن تعلیم ندیده‌اند پاسخ دهند.
- ❑ برای سوالاتی که می‌توانند با اطلاعات محلی (local) در یک مجموعه داده پاسخ داده شوند، عملکرد خوبی دارد.
- ❑ با این حال، نمی‌تواند به سؤالاتی که نیاز به درک سراسری (global) از کل مجموعه داده دارند پاسخ دهند.



- **محدودیت RAG برداری (vector RAG)**

- نمونه‌ای از پرس‌وجوی سراسری (global):
  «موضوعات اصلی اشاره شده در این پیکرهٔ متنی کدامند؟»
- نیاز به استدلال در مورد «اتصالات» (connections) بین افراد، مکان‌ها و رویدادها دارد.
- این گونه سؤالات، خلاصه‌سازی متمرکز بر پرسش‌وجو هستند، نه وظایف بازیابی سنتی.
- **Query-focused summarization (QFS)**
- رویکردهای RAG برداری استدلال بر روی کل پیکره متنی پشتیبانی نمی‌کنند.
- همچنین، روش‌های قبلی QFS در پیکره‌های بزرگ مقیاس کارا نیستند.

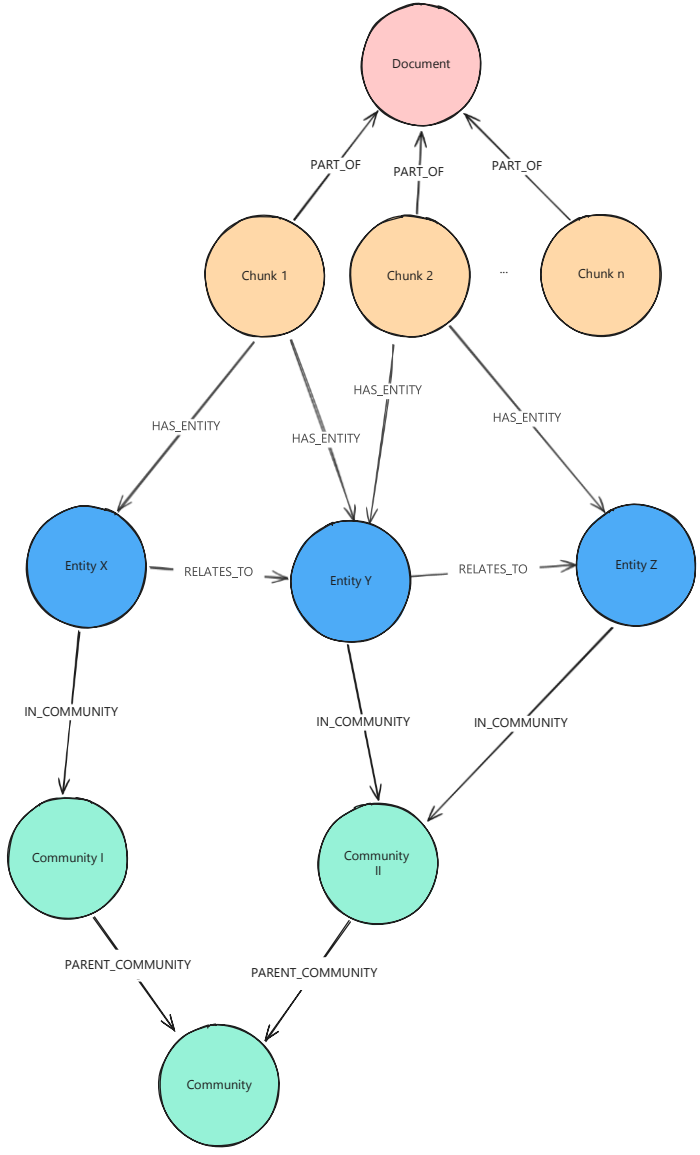
- ارائهٔ GraphRAG
  - یک رویکرد RAG مبتنی بر گراف.
  - امکان درک‌معنا (sensemaking) را در سطح کل یک مجموعه داده بزرگ متنی فراهم می‌کند.
  - استفاده ازLLM برای ساخت نمایهٔ گراف؛ در دو مرحله:

۱. ایجاد گراف دانش موجودیت از متن

- گره‌ها موجودیت‌های کلیدی و یال‌ها راوبط بین موجودیت‌ها هستند.

۲. قرار دادن موجودیت‌های مربوط به هم در جوامع (community) و ایجاد خلاصه برای هر جامعه

- این خلاصه‌ها، توصیف‌ها و بینش‌های کلی را در مورد مجموعه داده ارائه می‌دهند و برای پاسخ‌دهی به پرس‌وجوی به کاربر استفاده می‌شوند.


Document
PART_OF
PART_OF
PART_OF
Chunk 1
Chunk 2
...
Chunk n
HAS_ENTITY
HAS_ENTITY
HAS_ENTITY
HAS_ENTITY
Entity X
RELATES_TO
Entity Y
RELATES_TO
Entity Z
IN_COMMUNITY
IN_COMMUNITY
IN_COMMUNITY
Community I
Community II
PARENT_COMMUNITY
PARENT_COMMUNITY
Community

- روش GraphRAG و توانایی آن در انجام درک‌معنای (sensemaking) سراسری در کل یک مجموعه داده

- مقیاس پذیر هم به نسبت عمومی بودن سوالات کاربران و هم به نسبت حجم متن پیکره (بررسی شده برای پیکره به مقیاس ۱ میلیون توکن)

- توسعهٔ یک کاربرد نوآورانه از تکنیک LLM به عنوان داور؛ مناسب سؤالاتی که به مسائل و موضوعات گسترده‌ای اشاره دارند و حقیقت پایه‌ای برای آن‌ها وجود ندارد.

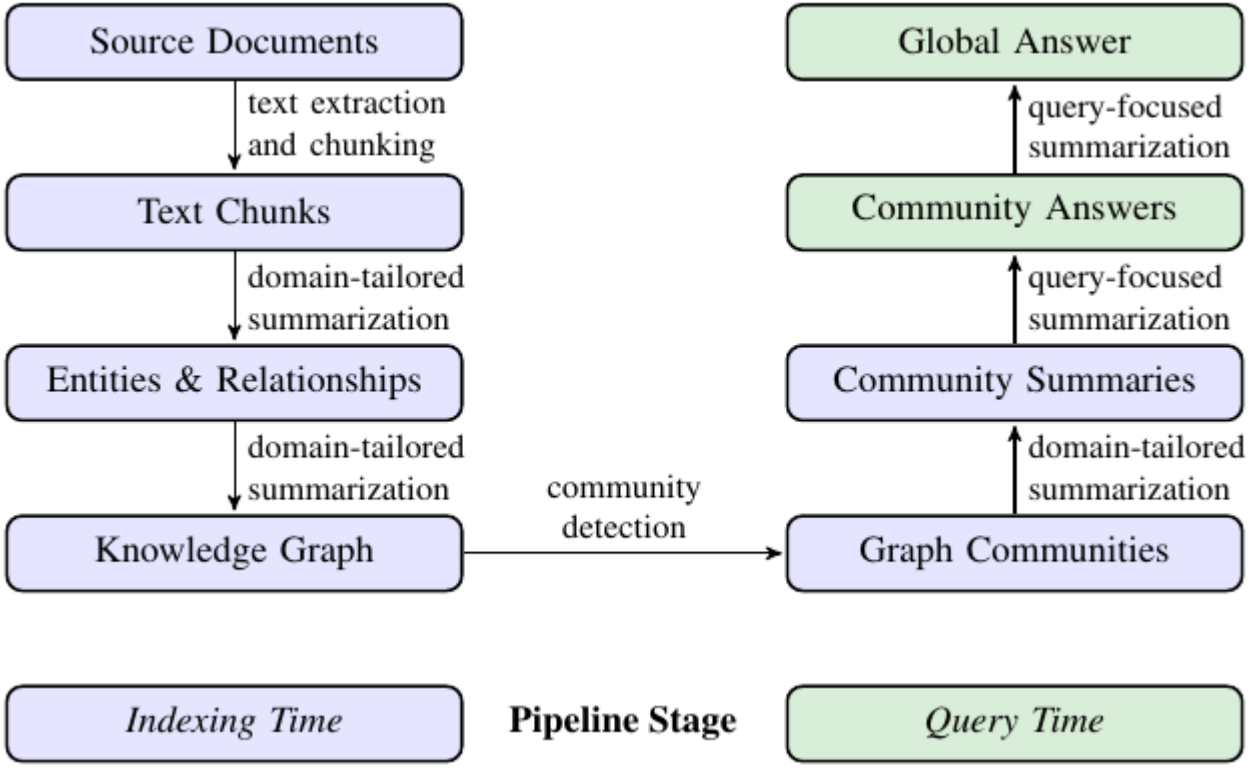
- GraphRAG، با استفاده از GPT-4 به عنوان LLM، در پاسخ‌دهی به سؤالات سراسری، به‌طور قابل توجهی از RAG برداری پیشی می‌گیرد.

- پیاده‌سازی روش پیشنهادی به صورت متن‌باز در گیت‌هاب و کتاب‌خانه‌های متن‌باز در دسترس است:

- https://github.com/microsoft/graphrag
- Langchain
- LammaIndex
- NebulaGraph
- Neo4j

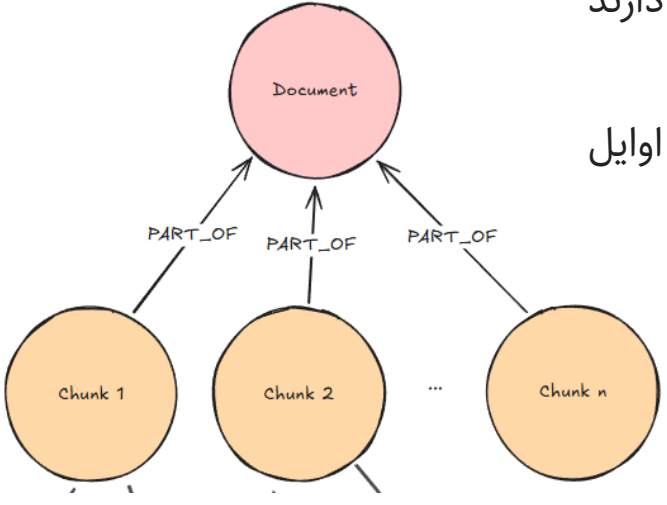
۲. جریان کاری GraphRAG

جریان داده در GraphRAG



۱. اسناد ⟸ قطعات متنی (text chunks)

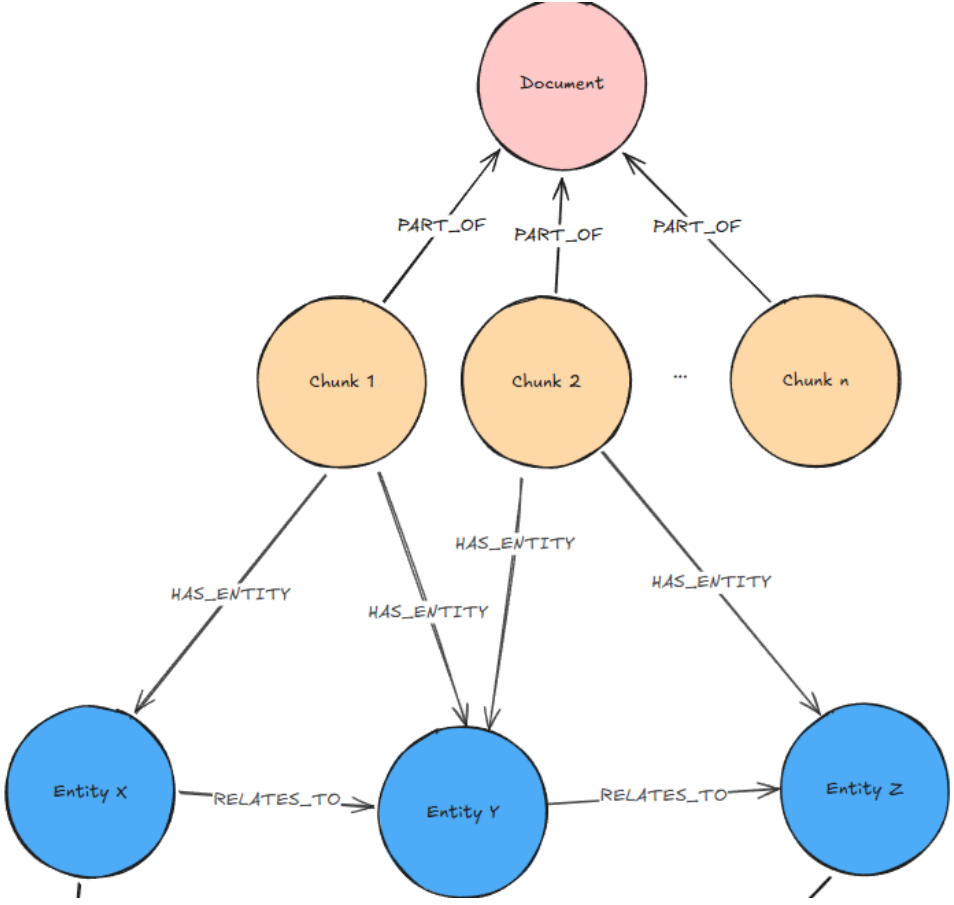
- اسناد موجود در مجموعه داده به قطعات متنی تقسیم می‌شوند.
- در مراحل بعد LLM هر قطعه را پردازش می‌کند.
- انتخاب طول قطعه‌ها یک تصمیم طراحی اساسی است.
- قطعات متنی بلندتر نیاز به فراخوانی‌های LLM کمتری دارند (کاهش هزینه).
- اما با افزایش طول قطعه، میزان یادآوری اطلاعاتی که در اوایل قطعه ظاهر می‌شوند کاهش می‌یابد.



۲. قطعات متنی ⟸ موجودیت‌ها و روابط

- استخراج موجودیت‌های مهم و روابط بین آن‌ها از هر قطعه؛ توسط LLM با استفاده از مهندسی اعلان (prompt)
- توضیحات کوتاهی نیز برای موجودیت‌ها و روابط تولید می‌شود.
- اعلان پیش‌فرض، کلاس‌های مرسوم موجودیت‌های نام‌دار مانند افراد، مکان‌ها و سازمان‌ها را استخراج می‌کند و در دامنه‌های عمومی قابل استفاده‌ست.
- در دامنه‌هایی با دانش تخصصی (مانند علم، پزشکی و حقوق)، می‌توان اعلان‌ها را با نمونه‌های few-shot تخصصی برای آن حوزه سفارشی‌سازی کرد.
- علاوه بر این، بیانیه‌های واقعی درباره موجودیت‌ها («ادعاها» (claims)) با استفاده از LLM استخراج می‌شود.

۲. قطعات متنی ⟸ موجودیت‌ها و روابط



۳. قطعات متنی ⟸ موجودیت‌ها و روابط

NeoChip’s (NC) shares surged in their first week of trading on the NewTech Exchange. However, market analysts caution that the chipmaker’s public debut may not reflect trends for other technology IPOs. NeoChip, previously a private entity, was acquired by Quantum Systems in 2016. The innovative semiconductor firm specializes in low-power processors for wearables and IoT devices.

- موجودیت‌ها، روابط و توضیحات استخراج شده

- NeoChip, with description “NeoChip is a publicly traded company specializing in low-power processors for wearables and IoT devices”.
- Quantum Systems, with description “Quantum Systems is a firm that previously owned NeoChip”.
- A relationship between NeoChip and Quantum Systems, with description “Quantum Sytems owned NeoChip from 2016 until NeoChip became publicly traded”.

۳. قطعات متنی ⟸ موجودیت‌ها و روابط

NeoChip’s (NC) shares surged in their first week of trading on the NewTech Exchange. However, market analysts caution that the chipmaker’s public debut may not reflect trends for other technology IPOs. NeoChip, previously a private entity, was acquired by Quantum Systems in 2016. The innovative semiconductor firm specializes in low-power processors for wearables and IoT devices.

❑ ادعاهای استخراج شده

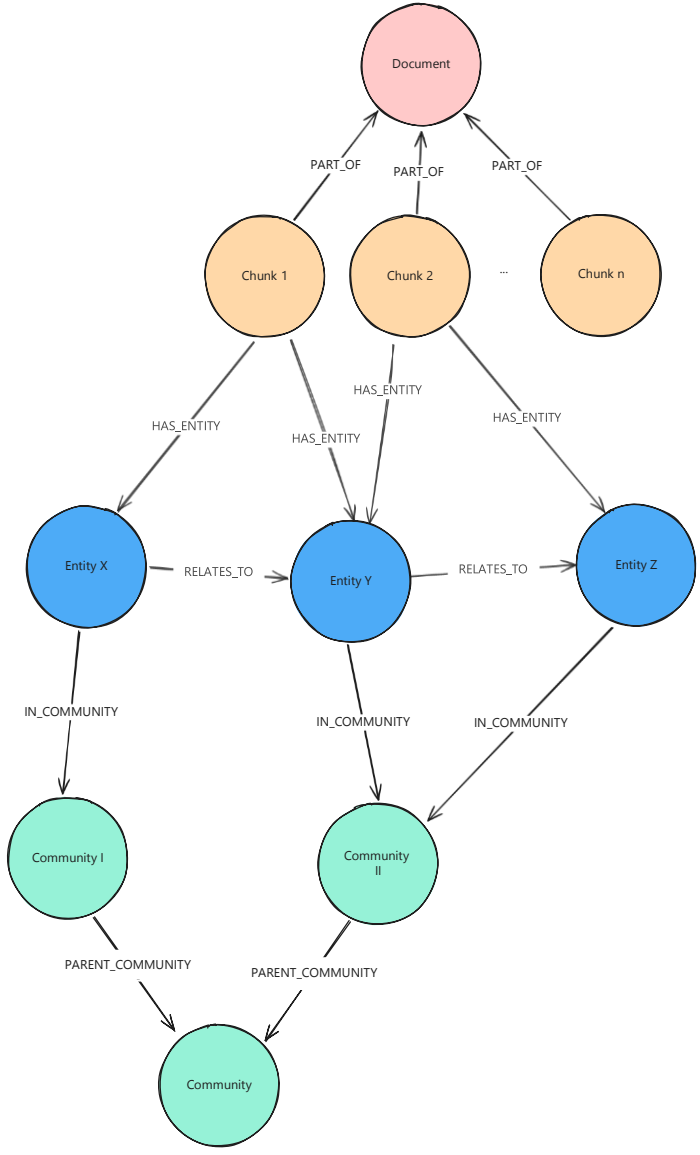
- ❑ NeoChip’s shares surged during their first week of trading on the NewTech Exchange.
- ❑ NeoChip debuted as a publicly listed company on the NewTech Exchange.
- ❑ Quantum Systems acquired NeoChip in 2016 and held ownership until NeoChip went public.

3. گراف دانش ⟸ جوامع گرافی (graph communities)

- تقسیم گراف به جوامعی از گره‌های به‌هم مرتبط
- از الگوریتم شناسایی جامعهٔ Leiden، به‌صورت سلسله‌مراتبی استفاده شده است.
- به‌صورت بازگشتی زیرجامعه‌ها را در هر جامعه شناسایی‌شده تشخیص می‌دهیم تا به جوامع برگ برسیم که دیگر نمی‌توانند به زیر جامعه‌ها تقسیم شوند.
- هر سطح از این سلسله‌مراتب یک تقسیم‌بندی جامعه ارائه می‌دهد
- گره‌های جوامع به‌طور متقابل انحصاری (mutually exclusive) و به‌طور جمعی جامع (collectively exhaustive) پوشش می‌دهد
- امکان خلاصه‌سازی کلی به روش تقسیم و غلبه را فراهم می‌کند.

3. گراف دانش ⟸ جوامع گرافی (graph communities)

- تقسیم گراف به جوامعی از گره‌های به‌هم مرتبط
- از الگوریتم شناسایی جامعهٔ Leiden، به‌صورت سلسله‌مراتبی استفاده شده است.
- به‌صورت بازگشتی زیرجامعه‌ها را در هر جامعه شناسایی‌شده تشخیص می‌دهیم تا به جوامع برگ برسیم که دیگر نمی‌توانند به زیر جامعه‌ها تقسیم شوند.
- هر سطح از این سلسله‌مراتب یک تقسیم‌بندی جامعه ارائه می‌دهد
- گره‌های جوامع به‌طور متقابل انحصاری (mutually exclusive) و به‌طور جمعی جامع (collectively exhaustive) پوشش می‌دهد
- امکان خلاصه‌سازی کلی به روش تقسیم و غلبه را فراهم می‌کند.


Document
PART_OF
PART_OF
PART_OF
Chunk 1
Chunk 2
...
Chunk n
HAS_ENTITY
HAS_ENTITY
HAS_ENTITY
HAS_ENTITY
Entity X
RELATES_TO
Entity Y
RELATES_TO
Entity Z
IN_COMMUNITY
IN_COMMUNITY
IN_COMMUNITY
Community I
Community II
PARENT_COMMUNITY
PARENT_COMMUNITY
Community

3. جوامع گرافی ⟸ خلاصهٔ‌های جوامع

- ایجاد خلاصه‌های گزارش-مانند (report-like) از هر جامعه در سلسله‌مراتب جوامع
- با استفاده از روشی که برای مجموعه داده‌های بسیار بزرگ نیز مقیاس پذیر باشد.
- این خلاصه‌ها به‌طور مستقل به‌عنوان روشی برای درک ساختار و معنای کلی مجموعه داده مفید هستند، حتی بدون حضور یک پرس‌وجوی خاص.
- خلاصه‌های گره‌ها، لبه‌ها و ادعاها (claim) را ترکیب می‌کند.
- خلاصه‌ها به‌صورت سلسله‌مراتبی ساخته می‌شوند، از سطح برگ تا جوامع بالاتر.

3. خلاصه‌های جوامع ⟸ پاسخ‌های جوامع ⟸ پاسخ سراسری

- از خلاصه‌های جوامع برای پاسخ‌دهی به پرس‌وجو در یک فرآیند چندمرحله‌ای استفاده می‌شود:

۱. آماده کردن خلاصه‌های جامعه

- خلاصه‌های جامعه به‌صورت تصادفی مرتب شده و به تکه‌هایی با اندازه توکن از پیش تعیین‌شده تقسیم می‌شوند.

۲. نگاشت پاسخ‌های جامعه

- از LLM خواسته می‌شود یک نمره بین ۰ تا ۱۰۰ تولید کند که نشان‌دهنده میزان مفید بودن پاسخ تولید شده در پاسخ به پرسش هدف باشد.
- پاسخ‌هایی که نمره ۰ دارند، دور ریخته می‌شوند.

۳. کاهش به یک پاسخ سراسری

- پاسخ‌های میانی جامعه بر اساس نمره مفید بودن به‌صورت نزولی مرتب شده و به یک پنجره متنی جدید اضافه می‌شوند تا به حد مجاز توکن برسند.
- این متن نهایی برای تولید پاسخ سراسری که به کاربر بازگردانده می‌شود، استفاده می‌شود.

تولید پرس‌وجوهای درک سراسری

- **هدف**
  - ❑ ارزیابی عملکرد سیستم‌های RAG در درک سطح بالا از یک مجموعه داده کامل
  - ❑ معیارهای موجود برای ارزیابی RAG برداری (بازیابی حقایق صریح) طراحی شده‌اند.
- **رویکرد**
  - ❑ استفاده از از یک خط لولهٔ مبتنی LLM برای تولید خودکار پرسش‌های سراسری از مجموعه داده‌ای خاص که نیازی به بازیابی حقایق خاص و سطح پایین ندارند.
- **مراحل**
  - ❑ از LLM خواسته می‌شود تا شخصیت‌های کاربران فرضی سیستم RAG را تولید کند.
  - ❑ سپس برای هر کاربر فرضی، از LLM خواسته می‌شود تا وظایفی را مشخص کند که این کاربر از سیستم RAG برای انجام آن‌ها استفاده خواهد کرد.
  - ❑ در نهایت، برای هر ترکیب کاربر و وظیفه، از LLM خواسته می‌شود تا پرسش‌هایی تولید کند که نیاز به درک کل مجموعه داده دارند.

Algorithm 1: Prompting Procedure for Question Generation

1: **Input:** Description of a corpus, number of users K, number of tasks per user N, number of questions per (user, task) combination M.
2: **Output:** A set of $K * N * M$ high-level questions requiring global understanding of the corpus.
3: **procedure** GENERATEQUESTIONS
4: Based on the corpus description, prompt the LLM to:

1. Describe personas of K potential users of the dataset.
2. For each user, identify N tasks relevant to the user.
3. Specific to each user & task pair, generate M high-level questions that:
   - Require understanding of the entire corpus.
   - Do not require retrieval of specific low-level facts.

5: Collect the generated questions to produce $K * N * M$ test questions for the dataset.
6: **end procedure**

در این ارزیابی $K = M = N = ۵$
۱۲۵ سوال برای هر مجموعه داده

Dataset	Example activity framing and generation of global sensemaking questions
Podcast transcripts	*User*: A tech journalist looking for insights and trends in the tech industry *Task*: Understanding how tech leaders view the role of policy and regulation *Questions*: 1. Which episodes deal primarily with tech policy and government regulation? 2. How do guests perceive the impact of privacy laws on technology development? 3. Do any guests discuss the balance between innovation and ethical considerations? 4. What are the suggested changes to current policies mentioned by the guests? 5. Are collaborations between tech companies and governments discussed and how?
News articles	*User*: Educator incorporating current affairs into curricula *Task*: Teaching about health and wellness *Questions*: 1. What current topics in health can be integrated into health education curricula? 2. How do news articles address the concepts of preventive medicine and wellness? 3. Are there examples of health articles that contradict each other, and if so, why? 4. What insights can be gleaned about public health priorities based on news coverage? 5. How can educators use the dataset to highlight the importance of health literacy?

- عدم وجود پاسخ استاندارد، نیازمند مقایسه مستقیم رویکردها با استفاده از یک ارزیاب مبتنی LLM است.
- چهار معیار هدف طراحی شده است که ویژگی‌های مطلوب برای فعالیت‌های درک سراسری را در بر می‌گیرد.

۱. **جامعیت** (Comprehensiveness): چقدر جزئیات در پاسخ ارائه می‌شود تا تمام جنبه‌ها و جزئیات پرسش را پوشش دهد؟

۲. **تنوع** (Diversity): چقدر پاسخ در ارائهٔ دیدگاه‌ها و بینش‌های مختلف دربارهٔ سؤال متنوع و غنی است؟

۳. **توانمندسازی** (Empowerment): چقدر پاسخ به خواننده کمک می‌کند تا موضوع را درک کند و قضاوت‌های آگاهانه‌ای انجام دهد؟

۴. **مستقیم بودن** (Directness): چقدر پاسخ به طور خاص و واضح به سؤال مطرح شده توجه دارد؟

- سؤال و پاسخ تولید شده از دو سیستم رقیب به LLM ارائه می‌شود و از آن خواسته می‌شود تا دو پاسخ را بر اساس معیار مشخص شده مقایسه کند.
- LLMیک برنده یا در صورتی که پاسخ‌ها به‌طور بنیادی مشابه باشند، نتیجه را مساوی اعلام می‌کند.
- برای در نظر گرفتن تصادفی بودن ذاتی LLM، هر مقایسه چندین بار تکرار می‌شود نتایج میانگین گرفته می‌شود.

برای ارزیابی نتایج جامعیت و تنوع داوری LLM ⟸ معیارهای مبتنی ادعا (claim-based)

- ❑ استخراج ادعاها با استفاده از Claimify (*Metropolitansky and Larson, 2025*)
- ❑ ادعا: یک بیانیه واقعی قابل تأیید
- ❑ جامعیت: میانگین تعداد ادعاهای منحصر به فرد در پاسخ‌های تولید شده
- ❑ تنوع: خوشه‌بندی ادعاها برای هر پاسخ و محاسبه میانگین تعداد خوشه‌ها

“California and New York implemented incentives for renewable energy adoption, highlighting the broader importance of sustainability in policy decisions”

- ❑ **ادعاهای استخراج شده:**

(1) California implemented incentives for renewable energy adoption
(2) New York implemented incentives for renewable energy adoption.

۳. تحلیل

- استفاده از GPT4 به‌عنوان LLM
- اندازه پنجره زمینه ثابت ۸۰۰۰ توکن برای استفاده از LLM در زمان پرسش
- **روش‌های مقایسه شده:**
  - GraphRAG در چهار سطح مختلف جامعهٔ گراف (C0-C3)
  - خلاصه‌سازی متن (TS): خلاصه‌سازی به‌طور مستقیم بر روی متون منبع اعمال می‌شود، بدون گراف.
  - RAGبرداری (SS): جستجوی معنایی که تکه‌های متنی مشابه پرس‌وجو را بازیابی می‌کند.

- دو مجموعه دادهٔ در محدودهٔ یک میلیون توکن
- متن‌های پادکست: حدود ۱ میلیون توکن (۱۶۶۹ تکه ۶۰۰ توکنی).
  - گراف تولید شده ۸۵۶۴ گره و ۲۰۶۹۱ رابطه
- مقالات خبری: حدود ۱.۷ میلیون توکن (۳۱۹۷ تکه ۶۰۰ توکنی).
  - گراف تولید شده ۱۵۷۵۴ گره و ۱۹۵۲۰ رابطه

	Podcast Transcripts					News Articles				
	C0	**C1**	**C2**	**C3**	**TS**	**C0**	**C1**	**C2**	**C3**	**TS**
Units	34	367	969	1310	1669	55	555	1797	2142	3197
Tokens	26657	225756	565720	746100	1014611	39770	352641	980898	1140266	1707694
% Max	2.6	22.2	55.8	73.5	100	2.3	20.7	57.4	66.8	100

Podcast transcripts

	SS	TS	C0	C1	C2	C3
SS	50	17	28	25	22	21
TS	83	50	50	48	43	44
C0	72	50	50	53	50	49
C1	75	52	47	50	52	50
C2	**78**	**57**	**50**	**48**	**50**	**52**
C3	79	56	51	50	48	50

Comprehensiveness

	SS	TS	C0	C1	C2	C3
SS	50	18	23	25	19	19
TS	82	50	50	50	43	46
C0	77	50	50	50	46	44
C1	75	50	50	50	44	45
C2	81	57	54	56	50	48
C3	**81**	**54**	**56**	**55**	**52**	**50**

Diversity

	SS	TS	C0	C1	C2	C3
SS	50	42	57	52	49	51
TS	**58**	**50**	**59**	**55**	**52**	**51**
C0	43	41	50	49	47	48
C1	48	45	51	50	49	50
C2	51	48	53	51	50	51
C3	49	49	52	50	49	50

Empowerment

	SS	TS	C0	C1	C2	C3
SS	**50**	**56**	**65**	**60**	**60**	**60**
TS	44	50	55	52	51	52
C0	35	45	50	47	48	48
C1	40	48	53	50	50	50
C2	40	49	52	50	50	50
C3	40	48	52	50	50	50

Directness

News articles

	SS	TS	C0	C1	C2	C3
SS	50	20	28	25	21	21
TS	80	50	44	41	38	36
C0	72	56	50	52	54	52
C1	**75**	**59**	**48**	**50**	**58**	**55**
C2	79	62	46	42	50	59
C3	79	64	48	45	41	50

Comprehensiveness

	SS	TS	C0	C1	C2	C3
SS	50	33	38	35	29	31
TS	67	50	53	45	44	40
C0	62	47	50	40	41	41
C1	65	55	60	50	50	50
C2	**71**	**56**	**59**	**50**	**50**	**51**
C3	69	60	59	50	49	50

Diversity

	SS	TS	C0	C1	C2	C3
SS	50	47	57	49	50	50
TS	53	50	58	50	50	48
C0	43	42	50	42	45	44
C1	**51**	**50**	**58**	**50**	**52**	**51**
C2	50	50	55	48	50	50
C3	50	52	56	49	50	50

Empowerment

	SS	TS	C0	C1	C2	C3
SS	**50**	**54**	**59**	**55**	**55**	**54**
TS	46	50	55	53	52	52
C0	41	45	50	48	48	47
C1	45	47	52	50	49	49
C2	45	48	52	51	50	49
C3	46	48	53	51	51	50

Directness

- رویکردهای سراسری (GraphRAG C0-C3، TS) به طور قابل توجهی در هر دو زمینهٔ جامعیت و تنوع نسبت به رویکرد RAG برداری (SS) در هر دو مجموعه داده عملکرد بهتری داشتند.
- نرخ‌های پیروزی جامعیت در مقایسه با SS: ۸۳٪-۷۲٪ (پادکست)، ۸۰٪-۷۲٪ (اخبار).
- نرخ‌های پیروزی تنوع در مقایسه با SS: ۸۲٪-۷۵٪ (پادکست)، ۷۱٪-۶۲٪ (اخبار).
- RAG برداری (SS)، بیشترین پاسخ‌های مستقیم را تولید کرد.

۴. کارهای آینده و نتیجه‌گیری

- ارزیابی بر روی دو مجموعه خاص متنی (~۱ میلیون توکن هر کدام) متمرکز بود.
- کار بیشتری بر روی تعمیم به حوزه‌های مختلف مورد نیاز است.
- نمایه‌گذاری گراف، حاشیه‌نویسی‌های متنی غنی و ساختار جامعه‌ای سلسله‌مراتبی امکانات زیادی برای بهبود و سازگاری ارائه می‌دهند.
- از جمله بررسی طرح‌های هیبریدی RAG که ترکیبی از تطابق مبتنی بر نهفته‌سازی (embedding) و تولید گزارش‌های جامعه هستند.

- GraphRAG یک رویکرد جدید RAG است که ترکیبی از تولید گراف دانش و QFS برای درک معنا بر روی کل مجموعه‌های متنی می‌باشد.
- ارزیابی‌های اولیه نشان‌دهنده بهبودهای قابل توجهی نسبت به RAG برداری در زمینه جامعیت و تنوع پاسخ‌ها است.
- GraphRAG در مقایسه با یک رویکرد خلاصه‌سازی متنی سراسری و بدون گراف، نتایج بهتری را نشان می‌دهد.
- برای سناریوهایی تعداد پرسش‌های سراسری زیاد است، خلاصه‌های جامعه در سطح ریشه (C0) یک شاخص داده کارآمد ارائه می‌دهند که از RAG برداری برتر است و در مقایسه با سایر روش‌های سراسری، هزینه توکن بسیار کمتری دارد.

سپاس