

Data Structures and algorithms

Instructor : Dr. Amin Eskandari

Head-Ta's : Hamid Namjoomanesh , Amir Hossein Hemati , Ali Mojahed

Ta's : Amir Mohammad Asadjoo , Padina razmjooi , Armin Janfaza , Alireza Heydari , Mahdi Torabi ,
Maryam Ghorbani , Fatemeh Keshavarz , MohammadReza Fasli , Abolfazl Tadrissi

Guest TA: Fatemeh Amiri

دو نکته قبل از شروع کوییز:

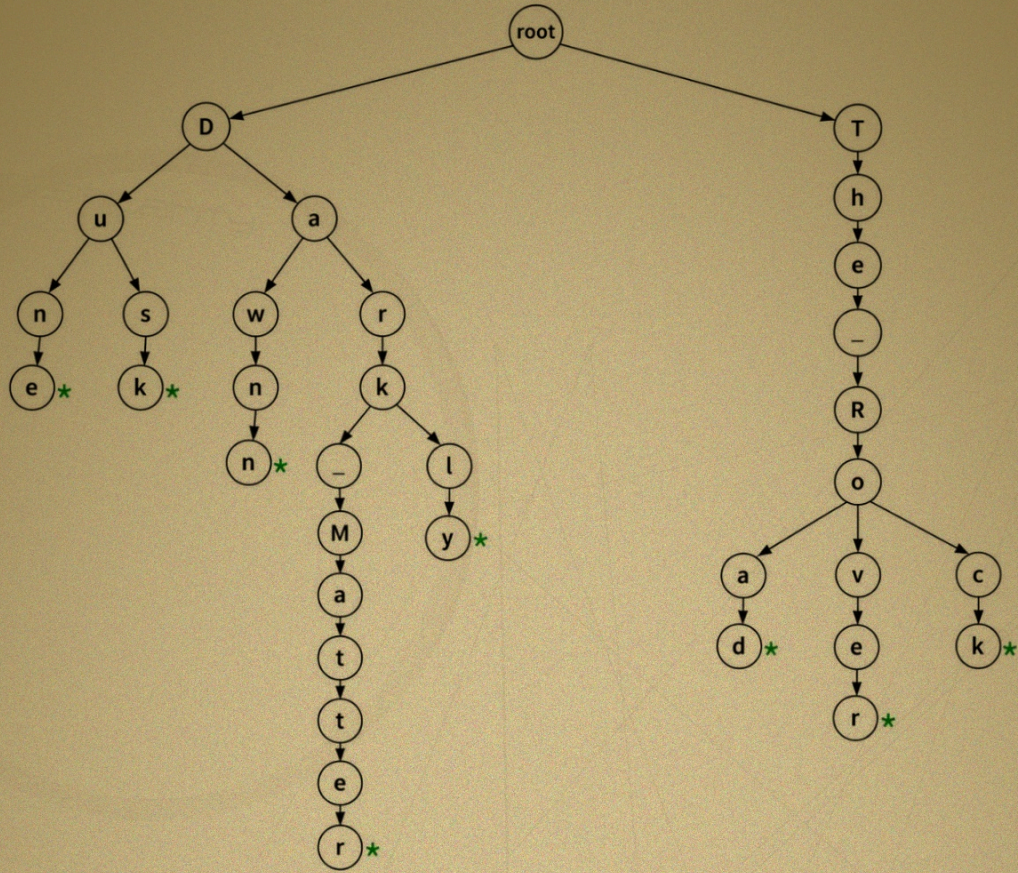
1. در این آزمون، برخی پارامترها بر اساس شماره دانشجویی شما تعریف می‌شوند. جایگذاری اشتباه عدد، باعث صفر شدن آن بخش می‌شود.
2. پاسخ‌های بخش‌های مختلف ممکن است به یکدیگر وابسته باشند. از بخش قبلی خودتان استفاده کنید.

Goodreads یک پلتفرم اجتماعی برای کتاب‌خوان‌هاست. کاربران می‌توانند کتاب بخوانند، امتیاز بدهند، نظر بنویسند، و یکدیگر را دنبال کنند.



(1) نوار جستجو

آریا، مهندس سیستم‌های جستجو در Goodreads، متوجه شد جستجوی خطی بسیار کند است. او از یک Trie استفاده کرد که کتاب‌ها را ذخیره کرده و نتایج را بر اساس تعداد دفعات سرچ مرتب می‌کند. (ساختار Trie و تعداد دفعات جستجو در جدول زیر آمده است)



کتاب	تعداد سرچ
Dune	560
Dawn	890
Darkly	125
Dark Matter	410
Dusk	230
The Road	850
The Rock	310

الف) کاربری پیشوندی از نام کتابی را جستجو می‌کند. این پیشوند بر اساس شماره دانشجویی شما تعیین می‌شود:

- اگر رقم آخر شماره دانشجویی شما زوج است: پیشوند "Da" را جستجو کنید.
- اگر رقم آخر شماره دانشجویی شما فرد است: پیشوند "The R" را جستجو کنید.

مسیر پیمایش را از ریشه گام به گام توضیح دهید.

چه کتاب‌هایی پیشنهاد می‌شوند؟

پیچیدگی زمانی این عملیات را بر حسب p (طول پیشوند) و k (تعداد نتایج) بیان کنید.

ب) دو کاربر با پیشوندهای متفاوت جستجو می‌کنند:

کاربر اول: "Dar" را سرچ می‌کند و 2 نتیجه می‌گیرد.

کاربر دوم: "The R" را سرچ می‌کند و 3 نتیجه می‌گیرد.

سوال: آیا پیچیدگی زمانی این دو عملیات با هم فرق دارد؟ چرا؟ (بر اساس تعریف ارائه‌شده در کلاس ما برای تحلیل *Worst-Case* در *Trie* پاسخ دهید، نه بر اساس شهود عمومی خود!)

ج) مرتب‌سازی نتایج بر اساس تعداد سرچ:

آریا باید نتایج را بر اساس تعداد دفعات سرچ مرتب کند تا محبوب‌ترین کتاب‌ها اول ظاهر شوند. الگوریتم Merge Sort را روی لیست کتاب‌های بدست‌آمده از بخش (الف) خودتان اجرا کنید. درخت بازگشت را رسم کنید و در هر مرحله وضعیت لیست را نشان دهید (چون پاسخ الف به شماره دانشجویی شما بستگی دارد، هر دانشجو درخت متفاوتی رسم می‌کند، پس فکر تقلب با همدیگه و AI نباشید 😊)

2) برترین های ماه

مدیر تیم: باید ویژگی جدیدی اضافه کنیم که در پایان هر ماه، 5 کتاب برتر اون ماه رو نمایش بده. ولی میلیون ها کتاب در دیتابیسمون خروار شده!

سارا (مهندس ارشد): "بهتره بجای مرتب کردن همه، فقط ۵ تا برتر رو پیدا کنیم"

کتاب‌ها و امتیازاتشان:

کتاب	امتیاز
Dune	4.8
The Road	4.2
Dawn	3.9
Darkly	4.5
The Rock	3.7
Dark Matter	4.1
Dusk	4.6
1984	4.7
Flowers for Algernon	4.9

الف) به سارا برای انتخاب تنها 5 کتاب برتر، چه ساختار داده‌ای پیشنهاد می‌دهید؟ نحوه ساخت آن را روی لیست بالا گام به گام نشان دهید و در هر مرحله وضعیت ساختار را رسم کنید.

ب) پیچیدگی زمانی الگوریتم خود را بر حسب n بیان کنید و توجیه کنید.

ج) حالا که ساختار شما در بخش الف) ساخته شد، فرض کنید یک کتاب جدید با نام "Simulated Reality" و امتیاز S به دیتابیس اضافه می‌شود، که در آن اینگونه

$$S = (\text{SumOfDigits}(\text{StudentID}) \% 5) + 1$$

است. بدون بازسازی کل ساختار، نشان دهید این کتاب جدید چگونه وارد ساختار شما می‌شود. آیا در ۵ کتاب برتر قرار می‌گیرد یا خیر؟

3) شبکه اجتماعی

Goodreads به کاربران اجازه می‌دهد یکدیگر را دنبال کنند. این رابطه یک‌طرفه است. شبکه اجتماعی فعلی به صورت زیر است:

شبکه فعلی:

کاربر	فالو می‌کنه
Armin	Mina, Sara
Mina	Armin, Danial
Sara	Danial, Raha
Danial	Raha, Ryan
Raha	Mina, Armin
Ryan	Sara, Mina
Nasim	Danial

ماهان، مهندس بخش گراف‌ها، می‌خواهد در جلسه‌ای برای مدیران ارشد، نحوه کارکرد بخش شبکه اجتماعی را ارائه دهد. برای پاسخ به سوالات مدیران، به او کمک کنید:

الف) کوتاه‌ترین زنجیره فالو از Nasim به Armin چیست؟ الگوریتم BFS را گام به گام اجرا کنید و در هر مرحله صف را نشان دهید. (طبق قرارداد کلاس ما، در صورت برابری اولویت در صف BFS، همیشه گرهی که از نظر ترتیب حروف الفبای انگلیسی زودتر می‌آید را زودتر از صف خارج کنید)

ب) یک پست توسط Armin منتشر می‌شود و همزمان یک پست توسط Nasim منتشر می‌شود. چه کاربرانی هر دو پست را می‌بینند؟ چطور این مجموعه را پیدا می‌کنید؟

ج) کارآموزی ادعا میکند این شبکه به گونه ایست که آگه یک پست به اندازه کافی viral شود، بالاخره به همه کاربران می رسد. آیا این ادعا درست است؟ چرا؟

4) مسیر خواندن (Reading Path)

تینا ، متقاضی استخدام بخش مدیریت محتوا در Goodreads ، در مصاحبه فنی خود با سوال زیر روبرو میشود:

Goodreads یک فیچر جدید می خواهد:

"کاربران یک مسیر منطقی برای خواندن کتاب ها داشته باشند. مثلاً قبل از هوش مصنوعی، ریاضیات و آمار را بخوانند"

پیش نیازها:

کتاب	پیش نیاز
Physics Fundamentals	ندارد
Mathematics	ندارد
Philosophy of Science	ندارد
Quantum Mechanics	Physics Fundamentals, Mathematics
Relativity	Physics Fundamentals, Mathematics
Philosophy of Mind	Philosophy of Science
Artificial Intelligence	Mathematics, Philosophy of Mind

کتاب	پیش‌نیاز
Astrophysics	Quantum Mechanics, Relativity
Consciousness and Machines	Artificial Intelligence, Philosophy of Mind
Theory of Everything	Astrophysics, Consciousness and Machines

الف) یک ترتیب معتبر مطالعه برای تمام کتاب‌ها پیدا کنید. الگوریتم Topological Sort را گام به گام اجرا کنید و در هر مرحله in-degree هر node و وضعیت صف را نشان دهید. (مشابه سوال ۳، در صفِ گره‌هایی که in-degree صفر دارند، همیشه گرهی که نام کتابش در الفبای انگلیسی اول است را انتخاب کنید)

ب) دو متقاضی این مسئله را حل کرده‌اند و به دو ترتیب متفاوت رسیده‌اند. هر دو ادعا می‌کنند جوابشان درست است. آیا ممکن است هر دو درست باشند؟ چرا؟

ج) کاربر می‌تواند چند کتاب را همزمان بخواند، به شرطی که پیش‌نیازهایشان را خوانده باشد. اگر هر کتاب یک هفته وقت ببرد، حداقل چند هفته طول می‌کشد تا کاربر نظریه همه چیز را بخواند؟ چطور این عدد را پیدا می‌کنید؟

5) سیستم پیشنهاد

انتشاراتی‌ها برای افزایش فروش محصولات خود، پیشنهاد اضافه کردن قابلیت به برنامه را ارایه می‌دهند:

"کسانی که این کتاب را خواندند، اینها را هم دوست داشتند" را پیاده کنند.

جدول امتیازات کاربران به کتاب‌ها صورت زیر است:

ریاضیات	آگاهی و ماشین	نسبیت	فلسفه ذهن	هوش مصنوعی	کیهان‌شناسی	نظریه همه چیز	
2	4	5	5	2	3	51	Armin
5	3	2	2	5	5	4	Mina
1	5	5	2	3	4	5	Sara
4	2	3	4	5	5	2	Danial
3	5	4	5	4	2	3	Raha
5	2	3	5	2	4	2	Ryan
4	4	2	3	1	5	4	Nasim

الف) کاربر جدیدی کتاب «نظریه همه چیز» را با امتیاز T ارزیابی کرده است.

$$T = (\text{Last2Digit}(\text{StudentID}) \bmod 3) + 2$$

سیستم فرض می‌کند کاربرانی که به این کتاب امتیازی در بازه

$$T - 0.5 \leq \text{Rating} \leq T + 0.5$$

داده‌اند، سلیقه‌ای مشابه او دارند. بر اساس امتیازهای این کاربران مشابه به سایر کتاب‌ها، کتاب‌ها را به ترتیب اولویت برای کاربر جدید پیشنهاد دهید. **صرفاً نوشتن لیست نهایی کافی نیست و تمامی مراحل محاسبه، میانگین‌گیری و رتبه‌بندی باید نشان داده شوند.**

(دقت کنید! در ردیف اول جدول (کاربر Armin)، امتیاز کتاب "نظریه همه چیز" به اشتباه ۵۱ درج شده است (به ازای ۵). به عنوان یک مهندس داده، ابتدا این خطای آشکار را اصلاح کنید، سپس محاسبات خود را انجام دهید. اگر خطا را اصلاح نکنید، میانگین‌های شما کاملاً اشتباه از آب درخواهند آمد!..)

ب) Goodreads واقعی چندین میلیون کاربر کتاب دارد. آیا همین رویکرد بخش الف در این مقیاس کار می‌کند؟ پیچیدگی زمانی‌اش چیست؟

ج) تیم مهندسی راه‌حلی پیشنهاد می‌کند که بدون مقایسه مستقیم همه کاربرها، کاربرهای مشابه را پیدا می‌کند، اما نتیجه‌اش تضمین دقیق نیست. این trade-off را توضیح دهید. آیا برای یک سیستم پیشنهاد کتاب این trade-off قابل قبول است؟

⚡ آیا اصولاً می‌توان الگوریتمی طراحی کرد که در یک سیستم بسیار بزرگ، دقیق‌ترین موجودیت مشابه را همیشه پیدا کند؟ اگر نه، چرا؟ اگر بله، چه محدودیت‌هایی دارد؟

6) نگهداری سیستم

دیتابیس کتاب‌های Goodreads بسیار بزرگ است و دائماً در حال تغییر است:

- کتاب جدید اضافه می‌شود
- امتیاز کتاب‌ها تغییر می‌کند
- رتبه‌بندی‌ها به‌روز می‌شوند

سیستم باید همیشه بتواند ۵ کتاب برتر را نمایش دهد.

اعضای تیم نگهداری با چالش‌هایی روبرو هستند :

الف) اگر هر بار کل داده‌ها دوباره مرتب شوند چه مشکلی ایجاد می‌شود؟

ب) چه نوع ساختار داده‌ای می‌تواند این مسئله را بهتر حل کند؟

ج) چرا این روش از مرتب‌سازی کامل مناسب‌تر است؟

یک کاربر جدید وارد سیستم می‌شود و هیچ فعالیتی ندارد:

- هیچ کتابی نخوانده

- هیچ امتیازی نداده

د) چگونه می‌توان به چنین کاربری کتاب پیشنهاد داد؟

و) چرا روش‌های مبتنی بر شباهت کاربران در این حالت کار نمی‌کنند؟

ه) یک روش ساده و عملی برای شروع پیشنهاد بدهید

7) تقلب-تعصب یاب

تحلیلگران سیستم متوجه رفتار غیر عادی برخی کاربران در سیستم امتیازدهی میشوند:

- به همه کتاب‌های یک نویسنده خاص امتیاز بسیار بالا می‌دهند (نویسنده مورد

علاقه‌شون، حتی اگر همه کتاب‌هاش رو نخونده باشن!)

- یا به یک نویسنده خاص امتیاز غیرعادلانه پایین می‌دهند

الف) این رفتار چه مشکلی در سیستم امتیازدهی ایجاد می‌کند؟

ب) چرا میانگین ساده امتیازها قابل اعتماد نیست؟

(جایزه)

ج) یکی از راه‌حل‌ها استفاده از "میانگین وزنی" یا "میانگین بیزی (Bayesian)" (Average) است. اما یک پارادوکس وجود دارد: اگر یک کتاب فقط ۱ رأی بسیار بالا (مثلاً ۵ ستاره) داشته باشد، میانگین بیزی آن را به میانگین کل سایت نزدیک می‌کند و رتبه‌اش پایین می‌آید. از طرفی، یک کتاب با ۱۰۰۰ رأی میانگین ۴.۸، رتبه بالایی می‌گیرد. آیا این رفتار الگوریتم عادلانه است؟ تله کجاست و چگونه باید این الگوریتم را اصلاح کرد تا هم از تقلب جلوگیری شود و هم به کتاب‌های جدید فرصت رشد داده شود؟

The End..