به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس بازیابی هوشمند اطلاعات

تمرین ۳

آذر ماه ۱۴۰۰

*فهرست

شرح دادگان

- در سوالات بخش ۱ مجموعه داده‌ای که استفاده خواهیم کرد (فایل wikitext.txt)، شامل ۲۰۰۰۰ سند متنی جمع آوری شده از سایت Wikipedia می باشد. در اینجا هر سطر از این فایل به عنوان یک سند در نظر گرفته می شود.

- در سوالات بخش ۲ از یک مجموعه داده شامل دو فایل با فرمت txt استفاده می‌شود. فایل spider10k.txt شامل ۱۰۰۰۰ صفحه وب و ۷۸۳۲۳ لینک و فایل spider800k.txt شامل ۸۷۵۷۱۳ صفحه وب و ۵۱۰۵۰۳۹ لینک می باشد. در هر سطر از این دادگان دو ستون از id صفحات وجود دارد که بیانگر وجود لینک از صفحه‌ی ستون اول به صفحه‌ی ستون دوم می باشد.

پیش نیاز ها

به منظور پاسخ به تمامی سوالات، حتما از زبان **پایتون** استفاده کنید.

برای پاسخ به سوالات بخش ۱، استفاده از کتابخانه‌ی **MRJob** توصیه می‌شود و برای پاسخ به سوال دوم بخش ۲، استفاده از کتابخانه‌ی **NetworkX الزامی** است.

بخش ۱ – نگاشت/کاهش (Map/Reduce)

در این بخش قصد داریم با استفاده از مدل مبتنی بر map/reduce اقدام به پیدا کردن لیستی از ۱۰۰ جفت کلمه با بالاترین میزان فراوانی نسبی در دادگان موجود در فایل wikitext.txt نماییم.

روش محاسبه میزان فراوانی نسبی کلمه‌ی A نسبت به کلمه‌ی B به صورت زیر است:

$$f(A|B) = \frac{count(A,B)}{count(B)} = \frac{count(A,B)}{\sum_{A'} count(A',B)}$$

که در اینجا $count(A,B)$ تعداد تکرار کلمه‌ی A و B به طور همزمان در یک سند می‌باشد. همچنین $count(B)$ تعداد تکرار کلمه‌ی B در همه سندها می‌باشد.

در روش map/reduce با توجه به نحوه استفاده از کلیدها[1] و مقادیر[2] آن ها رویکردهای متفاوتی را می‌توان پیاده سازی کرد. هدف از این مسئله، پیاده سازی و مقایسه‌ی دو رویکرد Stripes و Pairs در استفاده از map/reduce میباشد. تفاوت این دو روش در نحوه انتشار جفت کلید-مقدارهای مختلف می باشد که این موضوع در زمان اجرای کلی الگوریتم تاثیرگذار می باشد.

برای مطالعه‌ی بیشتر و آشنایی با رویکردهای Stripes و Pairs می توانید از دو منبع زیر استفاده نمایید:

- Data-Intensive Text Processing with MapReduce
- MapReduce Algorithm Design

سوال ۱ – با استفاده از رویکرد Stripes، روش map/reduce را برای محاسبه‌ی یک لیست نزولی از ۱۰۰ جفت کلمه با بالاترین میزان فراوانی نسبی پیاده سازی کنید و خروجی را در فایل stripes.csv ذخیره نمایید. نحوه‌ی عملکرد این رویکرد را در فایل گزارش خود شرح دهید.

سوال ۲ – با استفاده از رویکرد Pairs، روش map/reduce را برای محاسبه‌ی یک لیست نزولی از ۱۰۰ جفت کلمه با بالاترین میزان فراوانی نسبی پیاده سازی کنید و خروجی را در فایل pairs.csv ذخیره نمایید. نحوه عملکرد این رویکرد را در فایل گزارش خود شرح دهید.

سوال ۳ – عملکرد دو رویکرد را از نظر حافظه مصرفی و زمان اجرا مقایسه کنید و جمع بندی خود را از این دو رویکرد با ذکر دلیل بیان کنید.

[1] Keys
[2] Values

بخش ۲ – PageRank

سوال ۱- پیدا کردن همه بن‌بست‌ها[1]: یک نود را بن بست می گوییم اگر یا هیچ یال خروجی نداشته باشد و یا همه‌ی یال‌های خروجی آن به یک نود بن‌بست متصل باشند. برای مثال این گراف را در نظر بگیرید: $A -> B -> C -> D$. در اینجا، همه‌ی نودهای موجود در این گراف از نوع نود بن‌بست می باشند. نود D بن‌بست است چون یال خروجی ندارد، نود C بن‌بست است چون تنها یک یال خروجی دارد که آن هم به یک نود بن‌بست متصل است و به همین ترتیب.

در این بخش همه‌ی نودهای بن‌بست دادگان را برای هر دو فایل ورودی داده شده به دست آورید و تحت عنوان فایل های de_10k.csv و de_800k.csv ذخیره و ارسال نمایید.

توجه:

- استفاده از کتابخانه‌های آماده جهت محاسبه‌ی بن‌بست **مجاز نمی باشد.**
- حداکثر زمان مورد نیاز جهت یافتن بن بست های فایل 800k برابر۳ دقیقه می باشد.

سوال ۲- الگوریتم PageRank را برای هر دو فایل ورودی اجرا کنید و نتیجه را در دو فایل جداگانه pr_10k.csv و pr_800K.csv ذخیره نمایید. مقادیر فایل ها باید به صورت دو ستون PageRank و page id باشند.

توجه: در این سوال، می توانید از کتابخانه networkx با مقادیر پیش فرض و با تعداد تکرار ۱۰ استفاده کنید.

سوال ۳- با استفاده از خروجی حاصل از سوال اول و حذف نود های بن بست از دادگان، الگوریتم PageRank را دوباره اجرا کنید. عملکرد الگوریتم را از نظر صحت نتایج و زمان اجرا، بررسی و نتایج خود را شرح دهید. دلایل خود را برای توجیه نتایج به دست آمده بیان کنید. (**پیاده سازی الزامی است**).

[1] dead end

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان IR_CA3_StudentID تحویل داده شود.

- این فایل فشرده، بایستی حاوی یک فایل با فرمت PDF (گزارش تایپ شده) و یک پوشه به نام Codes باشد که کدهای نوشته شده را به تفکیک هر بخش شامل شود.
- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.
- گزارش به صورت تایپ شده در قالبPDF شامل شرح آزمایش‌های انجام شده، پارامترهای آزمایش، نتایج و تحلیل‌ها باشد.
- مهلت تحویل تمرین به هیچ عنوان تمدید نخواهد شد. تمرین تا یک هفته بعد از مهلت تعیین شده با جریمه تحویل گرفته می‌شود که جریمه تاخیر تحویل تمرین تا یک هفته ۳۰درصد است.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mailto:alihomayouni@ut.ac.ir

مهلت تحویل بدون جریمه: **۲۶ آذرماه ۱۴۰۰**

مهلت تحویل با تاخیر، با جریمه ۳۰ درصد: ۳ **دی ماه ۱۴۰۰**