بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

بازیابی پیشرفته اطلاعات

نیم‌سال دوم ۰۱ـ۰۰

استاد: احسان‌الدین عسگری

تمرین سری پنجم — مهلت تحویل: ۳۱ خرداد

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در تمرین‌هایی که چند ترک دارند، فقط یک نفر از هر گروه در سامانه CW باید ترک مورد نظر گروه را انتخاب کند. امکان تغییر ترک تا قبل از زمان ددلاین انتخاب ترک وجود دارد. البته ذکر این نکته ضروری است که هر ترک محدودیتی برای تعداد افرادی که آن را انتخاب می‌کنند، دارد. بنابرین در اسرع وقت برای انتخاب ترک اقدام کنید.
- در طول ترم امکان ارسال با تاخیر تمرین‌ها بدون کسر نمره تا سقف ۱۲ روز وجود دارد. محل بارگزاری جواب تمرین‌ها بعد از ۳ روز بسته خواهد شد و پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند شد. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۵ درصد از نمره تمرین کسر خواهد شد.
- توجه داشته‌باشید که نوت‌بوکهای شما باید قابلیت بازاجرای ۱۰۰ درصد داشته باشند و در صورت نیاز به نصب یک کتابخانه یا دسترسی به یک فایل، مراحل نصب و دانلود (از یک محل عمومی) در نوت‌بوک وجود داشته باشد.
- تمامی فایل‌های مرتبط به پروژه که حجم کمی دارند باید به شکل فایل زیپ در سامانه CW اپلود شوند. اگر حجم یک فایل زیاد بود (مانند فایل ذخیره شده یک مدل در صورتیکه بیش از ۲۰۰ مگابایت باشد)، تنها همان فایل را در یک محل عمومی، مثل گوگل درایو آپلود بفرمایید و لینک دانلود را در نوت‌بوک و مستندات قرار دهید.
- در پروژه‌های گروهی کافی است که فقط یکی از اعضای گروه پروژه را آپلود کند. اما حتما در گزارش کار نام همه اعضای گروه همراه با شماره دانشجویی آن‌ها آورده شود.
- بخشی از نمره شما به گزارش کار شما اختصاص دارد. در گزارش کار لازم نیست خط به خط کاری را که کرده‌اید توضیح دهید. بلکه باید به شکل کلی ایده‌تان برای حل مساله را شرح دهید. لازم است چند نمونه از خروجی‌های مساله را در گزارش بیاورید و براساس آن رفتار برنامه‌تان را تحلیل کنید. همچنین اگر پارامتری در صورت مساله خواسته شده (مانند دقت، صحت و ...) که در گزارش آورده شود شما باید آن را حساب کنید و در گزارش خود بیاورید.
- در صورت وجود هرگونه ابهام یا مشکل، در کوئرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به دستیاران آموزشی خودداری کنید.

توضیحات کلی

در این تمرین نیز هریک از گروه‌ها بر روی پروژه‌ی خود به صوت جداگانه کار خواهند کرد در این تمرین در بسیاری از بخش‌ها می‌توانید از حاصل کار عزیزان ترم گذشته که با زحمات تدریس‌یاران درس در قالب کتابخانه parsi.io ایجاد شده بهره ببرید. به امید خدا در ترم‌های آینده حاصل تجمیع زحمات شما عزیزان در قالب محصولات متن‌باز (البته با ذکر نام خودتان) در اختیار دیگر دانشجویان و بلکه جامعه ایرانی قرار می‌گیرد تا در اثر این تلاش‌ها محصولاتی ارزشمند برای پردازش متن‌های فارسی و بلکه زبان‌های ایرانی و فراتر از آن داشته باشیم. می‌توانید به این کتابخانه از طریق این لینک دسترسی داشته باشید

مدل‌های مورد استفاده برای تحلیل لینک

در این تمرین شما بر روی پروپوزال خود الگوریتم‌های تحلیل لینک را پیاده خواهید کرد. برای تحلیل لینک از هر دو الگوریتم PageRank و HITS و استفاده شود. همچنین نیاز به پیاده‌سازی آن‌ها نیست و می‌توانید از کتابخانه‌های آماده استفاده کنید. دربخش ارزیابی در این تمرین نداریم و شما باید تحلیل خود را از خروجی به صورت کتبی در گزارش‌تان بنویسید.

تحلیل لینک

در این بخش یکی از سناریوهای زیر یا سناریوی دل‌خواه خودتان (با هماهنگی با تدریس‌یار) را در نظر گرفته و بر روی آن، تحلیل لینک (نود) انجام دهید.

۱. سناریوی یکم

با استفاده از ابزارهای تشخیص موجودیت‌های نام‌دار، برای هر حکایت در بوستان (یا گلستان سعدی)، نام شخص‌ها و یا موقعیت‌های جغرافیایی (نام شهر و یا کشورها) را استخراج کنید. سپس گراف این موجودیت‌ها را بسازید. هنگام ساخت گراف، گره‌های گراف همان موجودیت‌ها و یال‌های گراف ارتباط میان آن موجودیت‌ها است. این ارتباط می‌تواند توسط خود شما و به روش‌های مختلف تعیین شود. برای نمونه، دو موجودیت به یکدیگر توسط یک یال پیوند می‌خورند در صورتی که آن دو موجودیت در یک حکایت مشترک (یا یک باب مشترک) از گلستان (یا بوستان) وجود داشته باشند. یا این‌که آن دو موجودیت در یک پنجره‌ی مشخص از هم اتفاق افتاده باشند (منظور آن است که مثلا در صورتی دو موجودیت A و B به هم پیوند بخورند که فاصله‌ی این دو موجودیت در متن گلستان، بیش از عدد K نباشد که K را به صورت تجربی خودتان به دست آورید. K همان اندازه‌ی پنجره نامیده می‌شود). سپس با استفاده از الگوریتم‌های تحلیل لینک (گره)، محوری‌ترین (یا به عبارتی مهم‌ترین) شخصیت موجود در گلستان را بیابید.

۲. سناریوی دوم

جمله‌های (یا پاراگراف‌های) بوستان (یا گلستان) در حکم گره در نظر گرفته شوند. در صورتی که دو جمله (یا پاراگراف) بیش از یک «حد مشخص» به یکدیگر شبیه بودند، باهم متصل شوند. سپس با الگوریتم تحلیل لینک، مهم‌ترین (یا محوری‌ترین) جمله‌ی گلستان سعدی را بیابید.

توضیحات کلی تمرین

هدف از این تمرین، آشنایی با مفاهیم تحلیل لینک می‌باشد. برای بخش تحلیل لینک، دو سناریو مختلف در نظر داریم که می‌توانید یکی از آن‌ها را انتخاب کرده و پیاده‌سازی کنید.

جمع‌آوری داده

با توجه به اینکه جنس مسئله در این تمرین با دو تمرین قبلی متفاوت است، باید یکبار دیگر داده‌های مقالات جمع‌آوری گردند. در این تمرین نیز شما لازم است که اطلاعات حداقل ۱۰۰۰ مقاله از یک فیلد خاص را جمع‌آوری (crawl) نمایید. اطلاعات مورد نیاز هر مقاله به صورت زیر می‌باشند:

- id مقاله
- عنوان مقاله
- سال انتشار
- نویسندگان مقاله
- چکیده مقاله
- تعداد استنادهای مقاله
- تعداد ارجاعات مقاله
- ارجاعات مقاله (۱۰ ارجاع اول کافی است)

می‌توانید تمرین‌های قبل از سایت Semantic Scholar برای جمع‌آوری داده‌ها استفاده نمایید.

تحلیل لینک

دو سناریو اصلی برای این بخش وجود دارد. یکی از آن‌ها را به دلخواه انتخاب کنید و به انجام آن بپردازید:

ارزش‌گذاری مقالات

در این بخش شما باید با استفاده از الگوریتم مناسب و ایجاد گراف ارجاعات، به ارزش‌گذاری مقالات بپردازید.

رتبه‌بندی نویسندگان

برای رتبه‌بندی نویسندگان، مفهوم ارجاع نویسندگان به یکدیگر مطرح می‌شود. زمانی که نویسنده A در مقاله خود به مقاله P که نویسنده B جزو نویسندگان آن مقاله (P) می‌باشد، ارجاع دهد، می‌گوییم که نویسنده A به نویسنده B ارجاع داده است. با توجه به این رابطه، می‌توان گراف ارجاعات بین نویسندگان را ایجاد و سپس با استفاده از الگوریتم مناسب، نویسندگان را رتبه‌بندی کرد.

تحلیل لینک

در این بخش یکی از سناریوهای زیر یا سناریوی دل‌خواه خودتان (با هماهنگی با تدریس‌یار) را در نظر گرفته و بر روی آن، تحلیل لینک (نود) انجام دهید.

۱. با استفاده از ابزارهای تشخیص موجودیت‌های نام‌دار، برای هر مقاله‌ی خبری، نام شخص‌ها و یا موقعیت‌های جغرافیایی (نام شهر و یا کشورها) و یا هر دو را استخراج کنید. سپس گراف این موجودیت‌ها را بسازید. هنگام ساخت گراف، گره‌های گراف همان موجودیت‌ها و یال‌های گراف ارتباط میان آن موجودیت‌ها است. این ارتباط می‌تواند توسط خود شما و به روش‌های مختلف تعیین شود. برای نمونه، دو موجودیت به یکدیگر توسط یک یال پیوند می‌خورند در صورتی که هر دوی آن‌ها در یک مقاله‌ی خبری مشترک وجود داشته باشند. سپس با استفاده از الگوریتم تحلیل لینک (گره)، محوری‌ترین (یا به عبارتی مهم‌ترین) شخص یا موقعیت جغرافیایی در دیتاست خبری خود را بیابید.

۲. عناوین مقالات خبری یا خلاصه‌ی آن‌ها را به عنوان گره در نظر بگیرید. در صورتی که تعداد واژگان مشترک دو عنوان/خلاصه بیش از یک «تعداد مشخص» که خودتان تعیین می‌کنید باشد، آن دو عنوان به هم متصل شوند. سپس با الگوریتم تحلیل لینک، مهم‌ترین (یا محوری‌ترین) عنوان/خلاصه‌ی خبر را بیابید.

تحلیل لینک

در این بخش یکی از سناریوهای زیر یا سناریوی دلخواه خودتان (با هماهنگی با تدریس‌یار) را در نظر گرفته و بر روی آن، تحلیل لینک (نود) انجام دهید.

۱. شخصیت های شاهنامه

در این بخش لازم است تا با پیدا کردن شخصیت های مرتبط در شاهنامه، ابتدا گراف ارتباط را به دست بیاورید و سپس با استفاده از الگوریتم های page rank و HITS ، میزان اهمیت افراد را در شاهنامه به دست بیاورید.
برای این کار اگر لازم شد، می توانید این کار را به تفکیک داستان انجام دهید و برای هر داستان یک گراف مجزا تشکیل دهید.
لازم به ذکر است که می توانید ارتباط (لینک)افراد را، رخداد نام یا صفات آن ها در نزدیکی یکدیگر (یک یا چند بیت قبل یا بعد تر) در نظر بگیرید.

۲. شهرها ی شاهنامه

مانند قسمت قبلی (شخصیت های شاهنامه)، این بار همان کار ها را برای مکان ها و شهرهای شاهنامه انجام دهید و نتایج را توسط هر دو الگوریتم مذکور در قسمت قبل به دست آورید.

تحلیل لینک

در این بخش یکی از سناریوهای زیر یا سناریوی دلخواه خودتان (با هماهنگی با تدریس‌یار) را در نظر گرفته و بر روی آن، تحلیل لینک انجام دهید.

۱. یک سناریو برای انجام این بخش می‌تواند ارزش‌گذاری صفحات و ایجاد گراف ارجاعات باشد.

۲. جمله‌های (یا پاراگراف‌های) هر صفحه در حکم گره در نظر گرفته شده و در صورتی که دو جمله (یا پاراگراف) بیش از یک «حد مشخص» دارای واژگان تکراری بودند، بهم متصل شوند. سپس با الگوریتم تحلیل لینک، مهم‌ترین جمله‌ی کل منبع برای یک پرس‌وجو را بیابید.

۳. با استفاده از ابزارهای تشخیص موجودیت‌های نام‌دار، برای هر صفحه، نام شخص‌ها و یا موقعیت‌های جغرافیایی یا هر موجودیت مورد نظر خودتان را استخراج کنید. سپس گراف این موجودیت‌ها را بسازید. هنگام ساخت گراف، گره‌های گراف همان موجودیت‌ها و یال‌های گراف ارتباط میان آن موجودیت‌ها است. این ارتباط می‌تواند توسط خود شما و به روش‌های مختلف تعیین شود. برای نمونه، دو موجودیت به یکدیگر توسط یک یال پیوند می‌خورند در صورتی که آن دو موجودیت در یک صفحه‌ی مشترک وجود داشته باشند. سپس با استفاده از الگوریتم تحلیل لینک (گره)، مهم‌ترین موجودیت انتخابی در منبع خود (برای مثال شخص) را بیابید.

۴. یک سناریو می‌تواند این باشد که برای یک موجودیت یا برچسب در تمام مجموعه صفحات خود، به رتبه‌بندی آن موجودیت‌ها در مستند خود بپردازید.

تحلیل لینک

در این بخش یکی از سناریوهای زیر یا سناریوی دل‌خواه خودتان (با هماهنگی با تدریس‌یار) را در نظر گرفته و بر روی آن، تحلیل لینک انجام دهید.

۱. به کمک الگوریتم‌های گفته شده، ریاضی‌دانان را بر اساس معیاری مشخص، مانند تعداد شاگردان یا نوادگان (descendants) ارزش‌گذاری کرده و گراف ارجاعات بر حسب روابط استادی و شاگردی را تشکیل دهید.

۲. با استفاده از ابزارهای تشخیص موجودیت‌های نام‌دار برای هر ریاضی‌دان، نام اشخاص و یا موقعیت‌های جغرافیایی یا هر موجودیت دیگر را استخراج کنید و سپس گراف این موجودیت‌ها را بسازید. هنگام ساخت گراف، گره‌های گراف را همان موجودیت‌ها و یال‌های گراف را ارتباط میان آن موجودیت‌ها در نظر بگیرید. این ارتباط می‌تواند توسط خود شما و به روش‌های مختلف تعیین شود. برای نمونه، دو موجودیت به یکدیگر توسط یک یال پیوند می‌خورند در صورتی که آن دو موجودیت در صفحه‌ی یک ریاضی‌دان وجود داشته باشند. سپس با استفاده از الگوریتم تحلیل لینک (گره)، مهم‌ترین موجودیت انتخابی در منبع خود (برای مثال شخص) را بیابید.

۳. برای یک موجودیت، مثلاً یک کشور، در تمام مجموعه صفحات ریاضی‌دانان به رتبه‌بندی آن موجودیت‌ها بپردازید.

۴. هر ریاضی‌دان را به عنوان یک موجودیت و یک گره در گراف در نظر بگیرید. با بررسی خلاصه‌ی بیوگرافی ریاضی‌دانان، در صورت شباهت بیشتر از حد مشخصی بین دو بیوگرافی (مثلاً داشتن تعداد معینی کلمات یکسان) یالی بین آن دو گره رسم نمایید.

تحلیل لینک

در این بخش یکی از سناریوهای زیر یا سناریو دلخواه خودتان (با هماهنگی تدریس‌یار) را در نظر گرفته و بر روی آن، تحلیل لینک انجام دهید. لازم به ذکر است که موارد زیر با در نظر گرفتن توییتر به عنوان شبکه اجتماعی مطرح شده‌اند و در صورتی که شبکه اجتماعی انتخابی شما شبکه دیگری است، باید سناریو خود را با هماهنگی تدریس‌یار انتخاب کنید.

۱. با استفاده از ارجاعات میان کاربران که به صورت‌های مختلف نظیر retweet، qoute tweet و ... قابل بررسی است، گراف مربوط به کاربران و ارجاعات آنها را تشکیل دهید. هنگام ساخت گراف، گره‌های گراف را همان کاربران در نظر بگیرید. یال‌های گراف نیز نمایانگر ارتباط میان دو کاربر هستند که به همان صورت retweet، qoute tweet و ... ایجاد می‌شوند. در نهایت با استفاده از دو الگوریتم ذکر شده برای تحلیل لینک، محوری‌ترین کاربران (کاربرانی که ارجاع بیشتری داشته‌اند) را بیابید.

۲. با استفاده از ابزارهای تشخیص موجودیت‌های نام‌دار، برای هر توییت یا مجموعه توییت‌های یک کاربر، نام شخص‌ها و یا موقعیت‌های جغرافیایی یا هر موجودیت مورد نظر خودتان را استخراج کنید. سپس گراف این موجودیت‌ها را بسازید. هنگام ساخت گراف، گره‌های گراف همان موجودیت‌ها و یال‌های گراف ارتباط میان آن موجودیت‌ها است. این ارتباط می‌تواند توسط خود شما و به روش‌های مختلف تعیین شود. برای نمونه، دو موجودیت به یکدیگر توسط یک یال پیوند می‌خورند در صورتی که آن دو موجودیت در یک توییت یا مجموعه توییت‌های یک کاربر وجود داشته باشند. سپس با استفاده از الگوریتم تحلیل لینک (گره)، مهم‌ترین موجودیت انتخابی در منبع خود (برای مثال شخص) را بیابید.

تحلیل لینک

در این بخش یکی از سناریوهای زیر یا **سناریوی دل‌خواه خودتان** (با هماهنگی با تدریس‌یار) را در نظر گرفته و بر روی آن، تحلیل لینک (نود) انجام دهید.

۱. با استفاده از ابزارهای تشخیص موجودیت‌های نام‌دار، برای هر بیماری یا عبارت مرتبط با پزشکی و سلامت در (مثلا در متون معرفی شده) عبارات کلیدی (مشابه تگ‌ها) را استخراج کنید. سپس گراف این موجودیت‌ها را بسازید. هنگام ساخت گراف، گره‌های گراف همان موجودیت‌ها و یال‌های گراف ارتباط میان آن موجودیت‌ها است. این ارتباط می‌تواند توسط خود شما و به روش‌های مختلف تعیین شود. برای نمونه، دو موجودیت به یکدیگر توسط یک یال پیوند می‌خورند در صورتی که آن دو موجودیت در یک دسته‌بندی مشترک (یا برچسب مشترک) از متون وجود داشته باشند. سپس با استفاده از الگوریتم‌های تحلیل لینک (گره)، محوری‌ترین (یا به عبارتی مهم‌ترین) موضوع موجود در نوشته‌ها را بیابید.

۲. پاراگراف‌های متون (مثلاً دکتر سلام یا نم‌ناک) در حکم گره در نظر گرفته شوند. در صورتی که دو جمله (یا پاراگراف) بیش از یک »حد مشخص» دارای واژگان تکراری بودند، با هم متصل شوند. سپس با الگوریتم‌های تحلیل لینک، مهم‌ترین (یا محوری‌ترین) جملات متون داده شده را بیابید.

تحلیل لینک

در این بخش یکی از سناریوهای زیر یا سناریوی دل‌خواه خودتان (با هماهنگی با تدریس‌یار) را در نظر گرفته و بر روی آن، تحلیل لینک (نود) انجام دهید.

۱. برای هر دستور غذایی موجودیت‌های نام‌دار، مانند مواد اولیه بکاررفته و یا شهر و کشوری که مربوط به آن غذاست و به نوعی در متن یا عنوان دستور غذایی ذکر شده را استخراج کند و سپس گراف این موجودیت‌ها را بسازید. هنگام ساخت این گراف، گره‌های گراف را همان موجودیت‌ها در نظر گرفته و یال‌های گراف را ارتباط میان آن موجودیت‌ها در نظر بگیرید؛ توجه کنید که این ارتباط می‌تواند توسط خود شما تعیین شود. برای نمونه هم می‌توانید ارتباط را به این شکل تعریف کنید: دو موجودیت به یکدیگر وصل می‌شوند در صورتی که هر دو در یک دستور غذایی آمده باشند.
در نهایت با استفاده از الگوریتم تحلیل لینک(گره)، مهم‌ترین موجودیت انتخابی در منبع خود را بیابید.

۲. جملات یا پاراگراف‌های هر دستور غذایی را بعنوان گره درنظر بگیرید و در صورتی که دو جمله یا پاراگراف بیش از یک "حد مشخص" دارای واژگان تکراری بودند، به یکدیگر متصل شوند. سپس با الگوریتم تحلیل لینک، مهم‌ترین جمله یا پاراگراف کل دستورات غذایی موجود را بیابید.

تحلیل لینک

در این بخش یکی از سناریوهای زیر یا سناریوی دلخواه خودتان (با هماهنگی با تدریس‌یار) را در نظر گرفته و بر روی آن، تحلیل لینک انجام دهید.

۱. هر آیه در یک سوره در حکم گره در نظر گرفته شده و در صورتی که دو آیه بیش از یک «حد مشخص» دارای واژگان تکراری بوده و یا با هر روش دیگری به هم شبیه بودند به یکدیگر متصل شوند. سپس با الگوریتم تحلیل لینک، محوری‌ترین آیه‌ها را بیابید.

۲. با استفاده از ابزارهای تشخیص موجودیت‌های نام‌دار، برای هرسوره، نام شخص‌ها و یا موقعیت‌های جغرافیایی یا هر موجودیت مورد نظر خودتان را استخراج کنید. سپس گراف این موجودیت‌ها را بسازید. هنگام ساخت گراف، گره‌های گراف همان موجودیت‌ها و یال‌های گراف ارتباط میان آن موجودیت‌ها است. این ارتباط می‌تواند توسط خود شما و به روش‌های مختلف تعیین شود. برای نمونه، دو شخص زمانی به یکدیگر توسط یک یال پیوند می‌خورند که بین آن دو شخص صفات مشترک وجود داشته باشند. سپس با استفاده از الگوریتم تحلیل لینک (گره)، مهم‌ترین موجودیت انتخابی در منبع خود (برای مثال شخص) را بیابید.

۳. در قرآن آیات و پندهای زیادی مختص به یک دسته خاص از افراد داده شده است. در این قسمت هر پند و آیه را بر اساس مخاطب آن به یکدیگر متصل کنید و از در نهایت مهم‌ترین پندها را از میان تمام آیات بیابید .

تحلیل لینک

در این بخش یکی از سناریوهای زیر یا سناریوی دل‌خواه خودتان (با هماهنگی با تدریس‌یار) را در نظر گرفته و بر روی آن، تحلیل لینک انجام دهید.

۱. می‌توانید مخزن‌های کد را بر حسب نیازمندی‌هیاشان به هم وصل کنید. دو صفحه در صورتی بهم وصل می‌شوند که دارای دستکم یک نیازمندی مشترک باشند.

۲. دو صفحه‌ی گیت در صورتی به یکدیگر در گراف وصل می‌شوند که متن Readme های‌شان بیش از یک مقدار مشخص بهم شبیه باشد