بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

بازیابی پیشرفته‌ی اطلاعات

نیم‌سال دوم تحصیلی ۰۰-۰۱

تمرین پنجم: تحلیل لینک
نوشته‌های مرتبط با سلامت/بیو

محمدمهدی ابوترابی ۹۸۱۰۵۵۵۷
یاسمن زلفی موصلو ۹۸۱۰۵۷۹۵
فاطمه عسگری ۹۸۱۰۵۹۲۱

مقدمه

در این تمرین هدف پیاده‌سازی الگوریتم‌های تحلیل لینک از جمله PageRank و HITS بر روی پروپوزال انتخابی است. در نهایت تحلیلی که درباره‌ی نتیجه به آن رسیده‌ایم نیز مکتوب شده است.
در داک تمرین دو سناریوی متفاوت برای این پروپوزال تعریف شده که در این تمرین از سناریوی دوم استفاده شده است. این سناریو از این قرار است که جملات متون (سندها) در حکم گره در نظر گرفته می‌شوند. در صورتی که دو جمله بیش از یک حد مشخص دارای واژگان تکراری بودند، با هم متصل می‌شوند. سپس با الگوریتم‌های تحلیل لینک، مهم‌ترین یا محوری‌ترین جملات متون داده شده به عنوان خروجی داده می‌شود.

ساختار کلی پروژه

در همان پروژه‌ی پیشین که تا اینجا تمام تمارین بر روی انجام شده است، در پوشه‌ی notebooks یک نوت‌بوک link_analysis اضافه شده که همه‌چیز در آن پیاده‌سازی شده است.
در قسمت load documents مجموعه داده‌ی دارای رده‌ای که در تمرین قبل آماده شده بود لود شده است. سپس با استفاده از نرمالایزر، لمتایزر، stop wordها، توکنایزر جملات و توکنایزر کلمات، و همچنین پیش‌پردازش‌های لازم دیگر برای نیم فاصله‌ها و... مراحل اولیه‌ی پیش‌پردازش بر روی سندها انجام شده است. در آخر برای هر سند لیستی از لیست توکن‌های جملات به دست می‌آید.
(توجه شود که در اینجا ابتدا همه‌ی این مراحل روی سند یازدهم انجام شده تا به کمک آن گام به گام موارد مورد نیاز اعمال شود.)

در قسمت بعد ماتریس شباهت ساخته می‌شود؛ بدین صورت که اگر n تعداد جمله‌ها باشد، یک ماتریس صفر n در n ساخته و به ماتریس شباهت یک سری وزن نسبت داده می‌شود؛ یعنی هر دو جمله‌ی i و j با هم مقایسه می‌شوند تا تعداد کلمات مشترکشان پس از پیش‌پردازش‌ها به دست آید. (یک treshold فرض می‌شود و بر اساس همان، میزان شباهت حساب می‌شود. اگر تعداد کلمات مشترک بین دو جمله‌ی i و j کمتر از treshold باشد در خانه‌ی (i, j) صفر ذخیره می‌شود و اگر i=j باشد نیز صفر ذخیره می‌شود چراکه شباهت جمله با خودش خیلی زیاد است و عملا به حساب آوردنش بی‌معنی است. در غیر این صورت تعداد کلمات مشابه به توان 1.2 رسیده و تقسیم بر تعداد کلمات جمله‌ی i می‌شود. حاصل در خانه‌ی (i, j) ذخیره می‌شود.
ساده‌ترین کاری که می‌شد کرد این بود که اگر کمتر از treshold بود صفر دهیم و در غیر این صورت تعداد اشتراک کلمات را بگذاریم. مشکل این روش این است که جملات طولانی به دلیل داشتن تعداد کلمه‌ی زیاد با اکثر جملات اشتراک زیادی خواهند داشت و اینطور به نظر می‌رسد که این جملات خیلی مهم هستند. (در نتایج اولیه که دریافت شد جملاتی که به عنوان مرتبطترین جملات انتخاب شده بودند عمدتا جملات طولانی بودند) در روش انتخاب شده، این مشکل با تقسیم مقدار اشتراک بر طول جمله مرتفع شده است. اما یک کار دیگری هم که انجام شده به توان 1.2 رساندن مقدار اشتراک است؛ چراکه اگر این کار را نکنیم یک سری جملات کوتاه ۵-۴ کلمه‌ای که اشتراکشان با جملات دیگر خیلی زیاد است، این بار بایاس می‌شوند و به عنوان جملات مهم انتخاب می‌شوند. (مقدار 1.2 با آزمایش و خطای مقادیر مختلف انتخاب شده است چراکه نتیجه‌ی قابل قبول‌تری را حاصل می‌کند.)

پس از ساخته شدن ماتریس مشابهت، سطرهای آن l1 normalize شده است. سپس با استفاده از networkx گراف مربوط به ماتریس مجاورت ساخته شده است.

در قسمت بعد بر روی گراف به دست آمده الگوریتم page rank پیاده شده است.

تابع get_top_n با استفاده از خروجی الگوریتم page rank بهترین n جمله را برمی‌گرداند. به طور مثال خروجی آن به این صورت است:

```
برای درمان افسردگی چه بخوریم و چه نخوریم؟
https://namnak.com/مواد-غذایی-درمان-افسردگی/p23484.
---------------
غذاهای دیگر که با افسردگی مبارزه می‌کنند، عبارتند از: دانه‌های گیاه آفتابگردان و چیا آووکادو پرتقال ماست گوشت ماهی چرب لوبیا آب تره غلات کامل درمان
افسردگی با مواد غذایی، غذاهایی که نباید بخورید: متاسفانه زمانی که از افسردگی رنج می‌برید، ممکن است به غذاهای ناسالم علاقه داشته باشید.
---------------
مواد مغذی به فعالیت سیستم عصبی شما کمک می‌کنند و از اضطراب و افسردگی جلوگیری می‌کنند.
---------------
غذاهایی که خلق و خوی شما را بهبود میبخشد و افسردگی را درمان می‌کنند در ادامه این بخش از سلامت نمناک به توصیه‌های خوراکی برای درمان افسردگی و رفع
اضطراب می‌پردازیم: غذاهای حاوی مواد مغذی که شما را شاد می‌کنند، می‌توانند به رژیم غذایی دیگری که برای افسردگی نیاز دارید، اضافه شوند.
---------------
سبزیجات با برگ‌های سبز تیره: یک مواد مغذی فوق العاده هستند، که به جلوگیری از التهاب کمک می‌کند.
---------------
گردو: غنی از امگا ۳ و اسید چرب است.
---------------
```

در قسمت بعد الگوریتم hits پیاده‌سازی شده است که hubs و authorities را خروجی می‌دهد. نمونه‌های حاصل از آن‌ها به ترتیب عبارتند از:

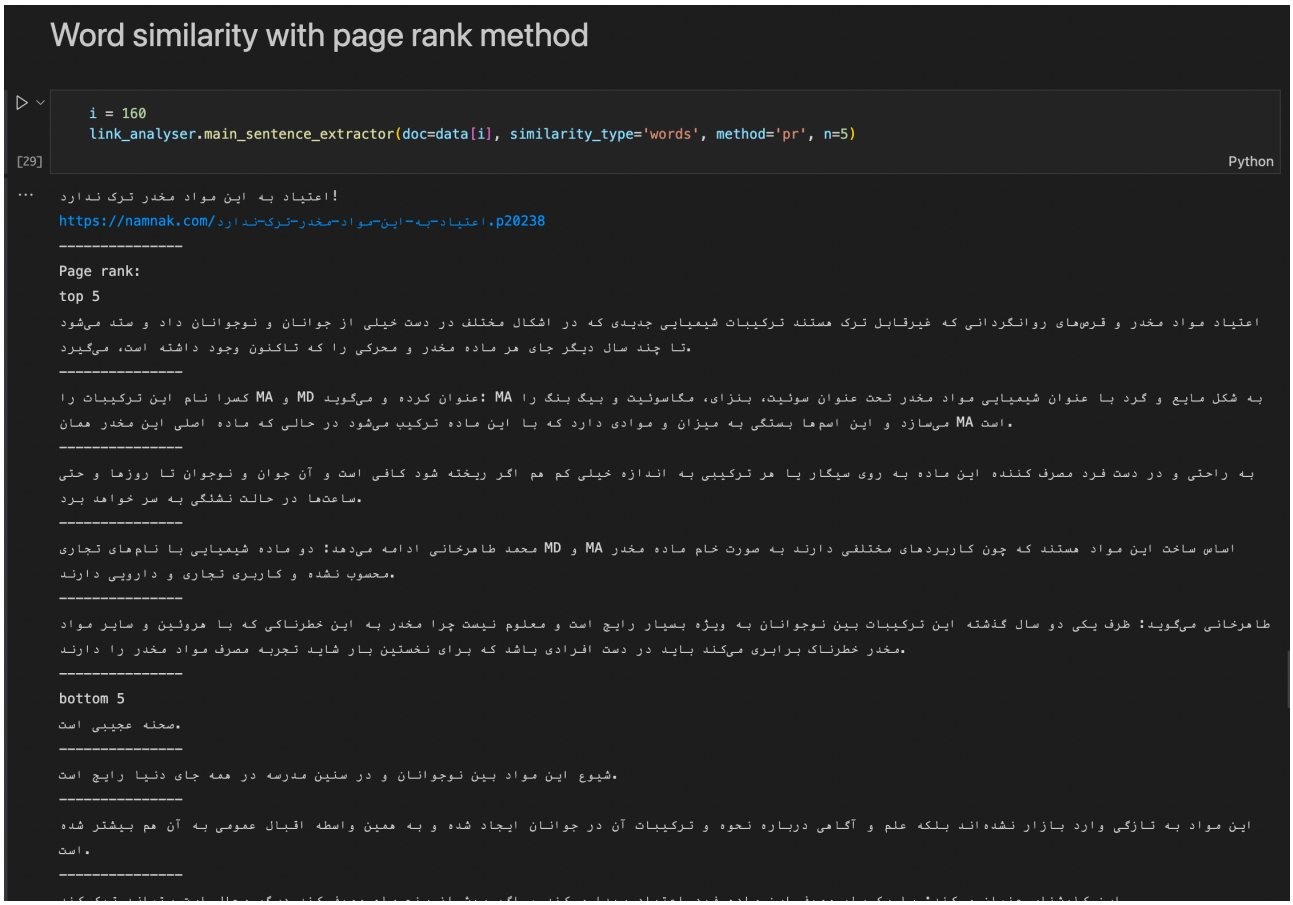
```
برای درمان افسردگی چه بخوریم و چه نخوریم؟
https://namnak.com/مواد-غذایی-درمان-افسردگی/p23484.
---------------
غذاهای دیگر که با افسردگی مبارزه می‌کنند، عبارتند از: دانه‌های گیاه آفتابگردان و چیا آووکادو پرتقال ماست گوشت ماهی چرب لوبیا آب تره غلات کامل درمان
افسردگی با مواد غذایی، غذاهایی که نباید بخورید: متاسفانه زمانی که از افسردگی رنج می‌برید، ممکن است به غذاهای ناسالم علاقه داشته باشید.
---------------
مانند: سوسیس قند و شکر و مشتقات آن‌ها نوشابه و سایر نوشیدنی‌های گازدار غذاهای سرخ شده مقدار بیش از حد کافئین شیرینی‌های تجاری پنیرهای چرب غلات غنی شده
علاوه بر اینها، استفاده از روش‌های دیگر برای مبارزه با افسردگی بسیار مهم است.
---------------
غذایی که برای مبارزه و درمان با افسردگی باید هر روز مصرف کنیم: ۱.
---------------
غذاهایی که خلق و خوی شما را بهبود میبخشد و افسردگی را درمان می‌کنند در ادامه این بخش از سلامت نمناک به توصیه‌های خوراکی برای درمان افسردگی و رفع
اضطراب می‌پردازیم: غذاهای حاوی مواد مغذی که شما را شاد می‌کنند، می‌توانند به رژیم غذایی دیگری که برای افسردگی نیاز دارید، اضافه شوند.
---------------
درمان افسردگی با مواد غذایی و تغذیه مناسب افسردگی یک مشکل سلامتی عمومی است، که بر میزان زیادی از جمعیت جهان تأثیر گذاشته است، با حالت غم انگیزی عمیق
مشخص میشود، می‌تواند احساسات ناامیدی، از دست دادن علاقه به زندگی را افزایش دهد.
---------------
```

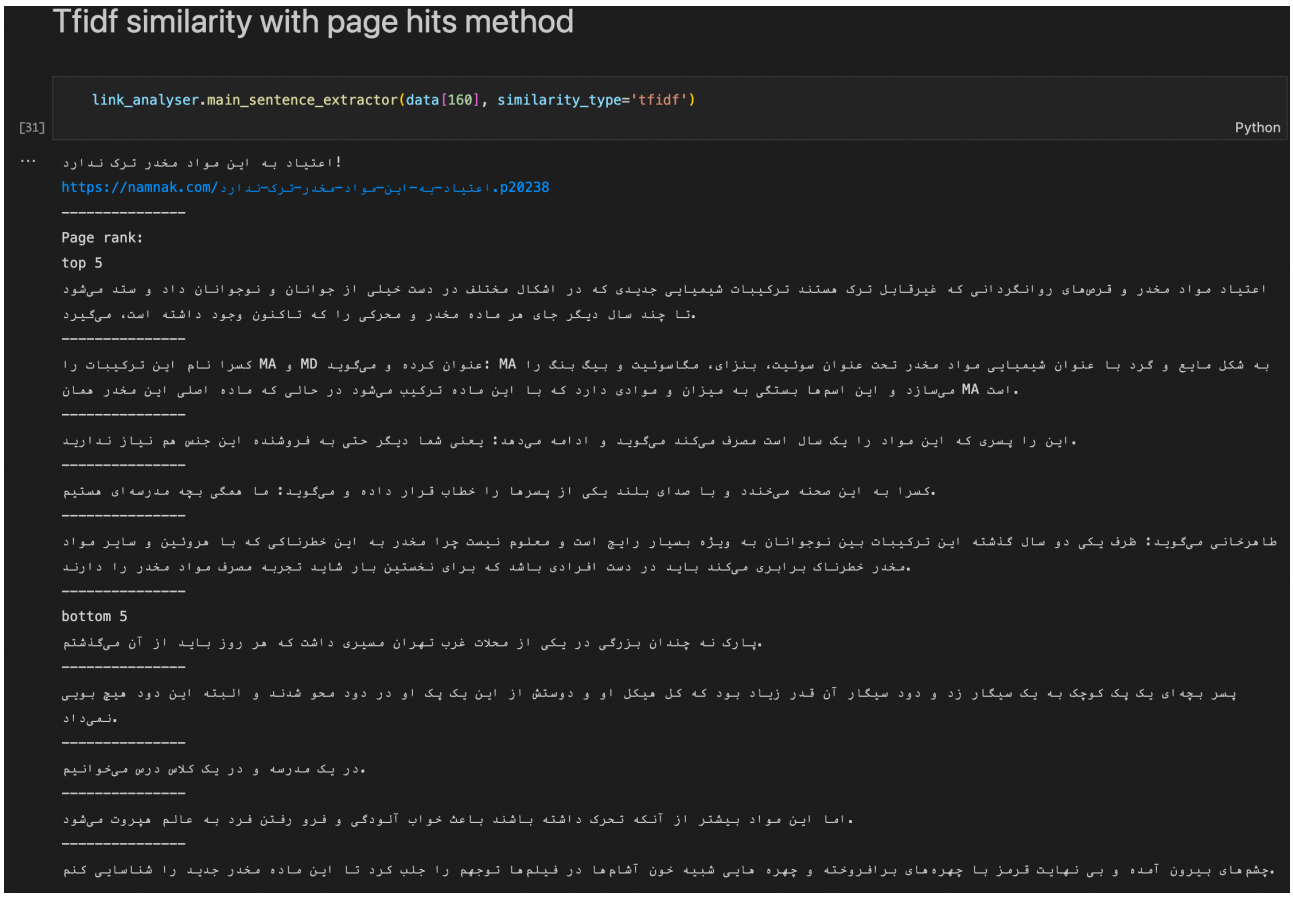
```
برای درمان افسردگی چه بخوریم و چه نخوریم؟
https://namnak.com/مواد-غذایی-درمان-افسردگی/p23484.
---------------
غذاهای دیگر که با افسردگی مبارزه می‌کنند، عبارتند از: دانه‌های گیاه آفتابگردان و چیا آووکادو پرتقال ماست گوشت ماهی چرب لوبیا آب تره غلات کامل درمان
افسردگی با مواد غذایی، غذاهایی که نباید بخورید: متاسفانه زمانی که از افسردگی رنج می‌برید، ممکن است به غذاهای ناسالم علاقه داشته باشید.
---------------
غذایی که برای مبارزه و درمان با افسردگی باید هر روز مصرف کنیم: ۱.
---------------
مانند: سوسیس قند و شکر و مشتقات آن‌ها نوشابه و سایر نوشیدنی‌های گازدار غذاهای سرخ شده مقدار بیش از حد کافئین شیرینی‌های تجاری پنیرهای چرب غلات غنی شده
علاوه بر اینها، استفاده از روش‌های دیگر برای مبارزه با افسردگی بسیار مهم است.
---------------
غذاهایی که خلق و خوی شما را بهبود میبخشد و افسردگی را درمان می‌کنند در ادامه این بخش از سلامت نمناک به توصیه‌های خوراکی برای درمان افسردگی و رفع
اضطراب می‌پردازیم: غذاهای حاوی مواد مغذی که شما را شاد می‌کنند، می‌توانند به رژیم غذایی دیگری که برای افسردگی نیاز دارید، اضافه شوند.
---------------
درمان افسردگی با مواد غذایی و تغذیه مناسب افسردگی یک مشکل سلامتی عمومی است، که بر میزان زیادی از جمعیت جهان تأثیر گذاشته است، با حالت غم انگیزی عمیق
مشخص میشود، می‌تواند احساسات ناامیدی، از دست دادن علاقه به زندگی را افزایش دهد.
---------------
```

در قسمت check similarity matrix symmetricity، بررسی و تایید شده است که ماتریس متقارن نیست. با این کار اطمینان حاصل می‌شود که خروجی hubs و authorities حتی اگر کاملا یکسان باشد به دلیل متقارن بودن ماتریس نیست. دلیل مشابهت زیاد این دو خروجی در ادامه بیان شده است.

در قسمت Link analysis using tfidf vectorizer تحلیل لینک با استفاده از tfidf انجام شده است بدین معنا که به جای آنکه معیار شباهت تعداد کلمات مشابه باشد، بردار tfidf جملات به دست آمده و با هم شباهت گرفته شده است. پس از ساخته شدن ماتریس مشابهت، گراف ساخته شده و page rank و hits روی آن پیاده شده است:

در آخر یک کلاس LinkAnalyzer پیاده‌سازی شده است. این کلاس یک تابع به نام main_sentence_extractor دارد که یک سند ورودی گرفته و با آرگومان ورودی similarity_type که می‌تواند words یا tfidf باشد به یکی از دو روش توضیح داده شده ماتریس مشابهت را می‌سازد و با آرگومان method نیز به روش page rank یا hits تحلیل لینک را انجام می‌دهد. این تابع n جمله‌ی top و n جمله‌ی bottom را نشان می‌دهد .
در زیر دو نمونه از خروجی‌ها قابل مشاهده است:





تحلیل نتایج

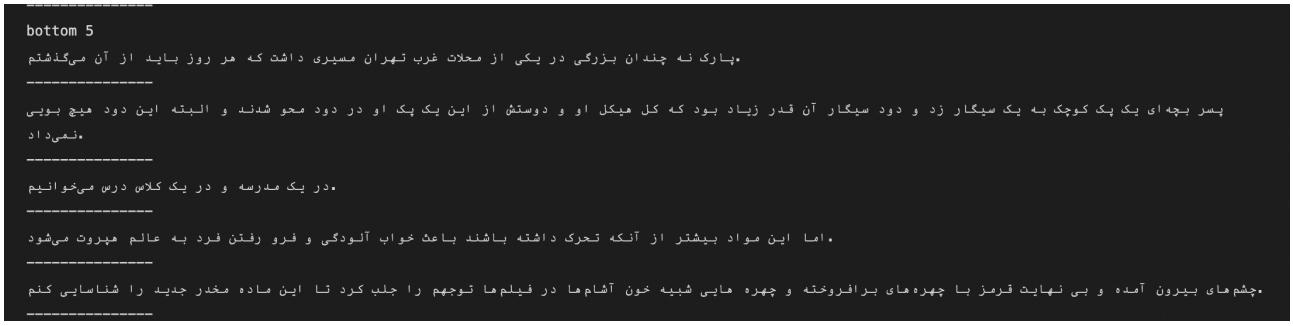
به طور کلی شانس حضور جملات طولانی در میان جملات top بیشتر است که منطقی هم هست چرا که از سویی جمله‌ی مهم جمله‌ای است که اطلاعات بیشتری در آن باشد. در جملات bottom نیز معمولا جملات کوتاه که اطلاعات خاصی ندارند حاضر شده‌اند.

به صورت کلی با آزمایش روش‌های مختلف، از برای ساختن ماتریس مشابهت گرفته تا انتخاب آلفا برای page rank حالت tfidf، مقادیر و روش‌های بهینه‌تر استفاده شده‌اند تا جملاتی که اطلاعات مرتبط زیادی در بر دارند و مهم هستند خروجی داده شود.. از مثال‌های زده‌شده هم تا حدودی مشخص است که بهترین جملات خروجی داده شده‌اند.

به طور مثال در خروجی‌ای که برای سند با عنوان «اعتیاد به این مواد مخدر ترک ندارد» داده شده است ۵ جمله‌ی مرتبط به شکل زیر است:

Page rank:
top 5
اعتیاد مواد مخدر و قرص‌های روانگردانی که غیرقابل ترک هستند ترکیبات شیمیایی جدیدی که در اشکال مختلف در دست خیلی از جوانان و نوجوانان داد و ستد می‌شود تا چند سال دیگر جای هر ماده مخدر و محرکی را که تاکنون وجود داشته است، می‌گیرد.
به شکل مایع و گرد با عنوان شیمیایی مواد مخدر تحت عنوان سونیت، بنزای، مگاسونیت و بیگ بنگ را MA :عنوان کرده و می‌گوید MD و MA کسرا نام این ترکیبات را می‌سازد و این اسم‌ها بستگی به میزان و موادی دارد که با این ماده ترکیب می‌شود در حالی که ماده اصلی این مخدر همان MA است.
این را پسری که این مواد را یک سال است مصرف می‌کند می‌گوید و ادامه می‌دهد: یعنی شما دیگر حتی به فروشنده این جنس هم نیاز ندارید.
کسرا به این صحنه می‌خندد و با صدای بلند یکی از پسرها را خطاب قرار داده و می‌گوید: ما همگی بچه مدرسه‌ای هستیم.
طاهرخانی می‌گوید: ظرف یکی دو سال گذشته این ترکیبات بین نوجوانان به ویژه بسیار رایج است و معلوم نیست چرا مخدر به این خطرناکی که با هروئین و سایر مواد مخدر خطرناک برابری می‌کند باید در دست افرادی باشد که برای نخستین بار شاید تجربه مصرف مواد مخدر را دارند.

جملاتی که به عنوان جملات مرتبط انتخاب شده‌اند همگی درباره‌ی اعتیاد و ترک مواد مخدر است که کاملا به متن سند مرتبط است.
در حالیکه پنج جمله‌ی نامرتبط به شکل زیر است:



نامرتبطترین جمله که «پارک نه چندان بزرگی در یکی از محلات غرب تهران مسیری داشت که هر روز باید از آن می‌گذشتم.» است واقعا هیچ ارتباطی به متن سند ندارد و جمله‌ی بسیار نامرتبطی است با آنکه آنقدر هم کوتاه نیست.
داخل نوتبوک مربوط به این تمرین مثال‌های بیش‌تری برای بررسی قرار داده شده است که می‌توانید به آن مراجعه کنید.
توجه: در بخش hubs و authorities توضیح دادیم که برای یکسان نشدن خروجی hubs و authorities ماتریس باید نامتقارن می‌بود که حتی این مسئله بررسی و تایید شد. اما دلیل اینکه خروجی این دو مخصوصا در جملات top یکسان به نظر می‌رسد این است که ماتریس شباهت بر اساس تعداد اشتراک کلمات میان جملات مختلف تعریف شده است و وقتی جمله‌ی a با b اشتراک زیادی دارد، منطقا b نیز با a اشتراک زیادی خواهد داشت. حال همانطور که می‌دانیم hubs و authorities برای ارجاع تعریف می‌شوند یعنی یکی بیانگر جملاتی است که خیلی به آن ارجاع داده می‌شود و دیگری بیانگر جملاتی است که زیاد به بقیه ارجاع می‌دهند. از آنجایی که در تعریف ماتریس و گراف ساخته شده، این ارجاع همان اشتراک است و اشتراک هم دو سویه است، خروجی hubs و authorities نسبتا مشابه شده است.