کاوش داده‌گان انبوه

دکتر سامان هراتی زاده
دانشگاه تهران - دانشکده علوم و فنون
ترم اول سال تحصیلی – زمستان ۱۳۹۹

دستیاران آموزشی
مهدی صادقی (mahdisadeghi@ut.ac.ir)
شیوا پارسا راد (shiva.parsarad@ut.ac.ir)
ایمان برازنده (barazandeh.iman@ut.ac.ir)

مجموعه تمرین چهارم : پیج‌رَنک، مثل قالی کرمان!

مهلت تحویل این تمرین، جمعه ۱۹ دی ماه، ساعت ۲۳:۵۹ می‌باشد.

هدف از این تمرین آشنایی شما با ابزارهای رتبه‌بندی گره‌های یک شبکه یا گراف است. یکی از این ابزارها، الگوریتم پیج‌رَنک نام دارد و یکی از اولین الگوریتم‌های موتور جستجوی گوگل برای رتبه‌بندی نتایج پرس‌وجوهای کاربران بود و گهگاه رِوایت‌هایی منتشر می‌شود مبنی بر اینکه همچنان نیز در کنار دیگر الگوریتم‌های این شرکت مورد استفاده قرار می‌گیرد.

پیج‌رَنک و بطور کلی الگوریتم‌های مبتنی بر قدم‌زن تصادفی[1] مثل DeepWalk، Node2Vec و TransE و دیگر الگوریتم‌ها در تحقیقات آکادمیک نیز از الگوریتم‌های محبوب برای تحلیل گراف هستند، و همواره بطور مستقیم یا غیر مستقیم در توسعه بسیاری از الگوریتم‌های هوشمند مبتنی بر گراف در قالب کاربردهای یادگیری ماشین بکار گرفته می‌شوند.

۱-(پژوهش) تحقیقی از یک منبع معتبر (مجلات پایگاه‌هایی مثل .ACM, IEEE, ScienceDirect, Springer, etc) پیدا کنید که از پیج‌رَنک در کنار کار اصلی خود بهره گرفته باشد. بطور خلاصه در یک صفحه، این کاربرد را توضیح دهید.

[1] Random Walk

۲-(طراحی و پیاده سازی الگوریتم)

هدف این تمرین پیاده سازی یک موتور جستجوی ساده با استفاده از دو الگوریتم PageRank و Hub and Authorities است. دیتاستی که در اختیار شما قرار گرفته شامل لیستی از وب‌سایت‌ها، محتوا و موضوع آن است. **برنامه شما باید یک کلمه ورودی از کاربر گرفته و بر اساس رتبه‌بندی انجام شده وب سایت‌های برتر را نمایش لیست کند.**

مجموعه داده به صورت یک فایل JSON است. فرمت هر وب‌سایت در این دیتاست به شکل زیر است:

id: شناسه یکتای وب‌سایت

Content: محتوای وب‌سایت

Links: لیستی از شناسه‌ها که این وب‌سایت به آنها لینک داده است

Category: موضوع وب‌سایت

الف) گراف/شبکه ارجاعات را بسازید و با الگوریتم پیج‌رَنک تحلیل کنید.

ب) گراف/شبکه ارجاعات را بسازید و با الگوریتم پیج‌رَنک موضوعی تحلیل کنید. (در ورودی به جز کلمه کلیدی جستجو، یک موضوع نیز بگیرد)

ج) گراف/شبکه ارجاعات را بسازید و با الگوریتم "Hubs & Authorities" نیز تحلیل کنید.

د) نرم افزار گِفی (Gephi) برای تحلیل گراف/شبکه را نصب کرده و گراف/شبکه ارجاعات را رسم و تحلیل کنید. هر چیزی در مورد این گراف قابل توجه است گزارش کنید.

۳-ماتریس وِب M را با ابعاد $n \times n$ (n تعداد صفحات وب است) در نظر بگیرید. هر ورودی m_{ij} در سطر i و ستون j برابر با صفر است، مگر اینکه یالی از گره (صفحه) j به گره i وجود داشته باشد. در این صورت مقدار m_{ij} برابر با $1/k$ است و k تعداد یالهای خروجی گره j است. پس اگر گره j دارای $k > 0$ یال خروجی باشد، ستون j دارای k مقدار $1/k$ خواهد بود و بقیه درایه‌ها نیز صفر هستند. اگر گره j یک گره بُن‌بست باشد (یال خروجی نداشته باشد) همه درایه‌های ستون j صفر خواهد بود.

فرض کنید $r = [r_1, r_2, \dots, r_n]^T$ بردار پیج‌رَنک باشد. به این معنی که r_i تخمینی از پیج‌رَنک گره i است. فرض کنید $w(r)$ حاصل جمع مولفه‌های بردار r باشد، یعنی $w(r) = \sum_{i=1}^{n} r_i$.

در هر تکرار از الگوریتم، تخمین بعدی از بردار پیج‌رَنک یعنی r' به صورت $r' = Mr$ محاسبه می‌شود. به عبارت دیگر برای هر i خواهیم داشت $r'_i = \sum_{j=1}^{n} M_{ij} r_j$. همچنین فرض می‌کنیم که $w(r')$ برابر با مجموع مولفه‌های r' باشد، یعنی $w(r') = \sum_{i=1}^{n} r'_i$.

الف) فرض کنید در گراف وِب هیچ بُن‌بستی وجود ندارد. نشان دهید $w(r') = w(r)$.

ب) فرض کنید در گراف وِب هیچ بُن‌بستی وجود ندارد اما ما از یک احتمال teleportation برابر با $1 - \beta$ استفاده می‌کنیم به این معنی که با این احتمال به یک گره تصادفی خواهیم رفت $(0 < \beta < 1)$. بنابراین تخمین بعدی از r_i برابر خواهد بود با $r'_i = \beta(\sum_{j=1}^{n} M_{ij} r_j) + (1 - \beta)/n$. تحت چه شرایطی $w(r') = w(r)$ خواهد بود؟ ثابت کنید[۲].

[2] Credit: Stanford University, MMDS, Jure Leskovec.

نکات تکمیلی

- از زبان برنامه سازی پایتون کنید.
- کدها با نامگذاری مناسب در پوشه SOURCES قرار گیرد.
- گزارش کامل تمرین را با ساختار مناسب، بدون ابهام و به زبان ساده نوشته و با فرمت PDF در پوشه دیگری قرار دهید. برای نگارش گزارش‌ها نیز امتیاز مثبت و منفی در نظر گرفته شده است، پس برای نگارش گزارش خود وقت مناسبی اختصاص دهید.
- پاسخ تمرین خود را به ایمیل barazandeh.iman@ut.ac.ir ارسال نمایید. هر دو پوشه را در یک فایل ZIP با نام MMDS_STID_FullName_HW# مثلا MMDS_830123451_ImanBarazandeh_HW1 قرار داده و در ایمیلی با همین عنوان نیز ارسال کنید.
- در صورت مشخص شدن هر نوع تقلب انفرادی یا مشترک نمره تمرین صفر در نظر گرفته می‌شود.
- به ازای هر روز تاخیر ۲۵٪ از امتیاز تمرین از دست خواهید داد.
- هر گونه ابهام را با دستیاران آموزشی مطرح کنید.