پروژه اختیاری درس طراحی الگوریتم: پیدا کردن جامعه توییتری و حلقه های متصل

سید صالح اعتمادی ـ مریم سادات هاشمی

نیمسال دوم سال تحصیلی ۹۷ـ۹۸

هدف از این پروژه کار با داده واقعی و معنا دار در قالب گراف و الگوریتم های مربوط به آن میباشد. داده مورد استفاده برای این پروژه از توییتر میباشد. هدف اول ما پیدا کردن همه اعضاء دانشکده کامپیوتر در توییتر میباشد. هدف بعدی پیدا کردن گروه های متصل دانشکده است.

۱ پیدا کردن جامعه توییتری

میتوانید توییتر را به صورت یک گراف مدل کنید که در این گراف هر اکانت توییتر یک گره است. ولی تعریف یال جهت دار بر عهده شماست. یال را باید طوری تعریف کنیم که هدف نهایی ما برآورده شود. مثلا در مرحله اول که پیدا کردن اعضاء توییتر دانشکده است ممکن است تعریف یال رابطه follow کردن باشد.

- خوب بااین تعریف به نظر شما اگر الگوریتم DFS را ازاکانت شما اجرا کنیم نتیجه چه اکانت هایی پیدا میشوند[۱]؟
- یک نظریه هست که هر دو نفری در دنیا با حداکثر هفت واسط همدیگر را میشناسنند. نظر شما چیست؟ فکر میکنید در توییتر هم این درست باشد؟
- آیا میتوانید DFS را بگونه ای تغییر دهید که یک عدد n به عنوان حداکثر فاصله دریافت کند و تنها نودهایی را که حداکثر فاصله n را دارند پیدا کند؟
- خوب حالا اگر این الگوریتم تغییر یافته DFS را با $n = 2$ از اکانت خودتان اجرا کنید، چه اکانت هایی پیدا میشوند؟
- آیا کل اکانت های دانشکده پیدا میشوند؟ آیا به جز اکانت های دانشکده اکانت های دیگری هم پیدا میشوند؟
- قطعا به جز اکانت های دانشکده، اکانت های دیگری هم پیدا میشوند. به نظر شما چطور میتوان این اکانت های دیگر را پیدا و فیلتر کرد، بطوری که اکثر اکانت های باقیمانده مربوط به دانشکده باشند؟

[۱] خوب است به این سوالات فکر کنید. علاوه بر فکر کردن، به توییتر مراجعه کنید و یک مقداری بررسی کنید ببینید بر اساس داده های موجود چه جوابی به نظرتون میرسه. از اون بهتر اینکه بعد از درست کردن account developer برای توییتر داده های لازم را برای جواب دادن به بعضی از سوالات زیر پیدا کنید و داده ها را با روش های مختلف مرتب کنید، ببینید چه جوابی به نظرتون میرسه.

- شاید راه حل خوبی به نظرتون برسه. شاید هم نرسه. یه راه حل میتونی این باشه که یه اکانت مربوط به مثلا انجمن علمی دانشکده یا چیز مشابهی را که اکثر follower هاش مربوط به دانشکده باشند را پیدا کنیم و از آنجا شروع کنیم. مثلا DevCampIUST یک اکانت اینطوری است. خوب اگر فرض کنیم که همه follower های این اکانت مربوط به دانشکده هستند، چه راه حلی به نظرتون میرسه؟
- مثلا به نظرتون اگر follower های این اکانت شصت نفر باشند (S_1) و ما همه follower های اینها را پیدا کنیم و بشوند ۶۰۰۰ اکانت (S_2). آیا از این اطلاعات میشه چیزی را پیدا کرد؟
- فکر میکنید اگر تمام اعضاء S_2 را بر اساس اینکه چند تا از اعضاء S_1 را دنبال میکنند به صورت نزولی مرتب کنیم ابتدای لیست با چه اکانت هایی شروع میشود؟ این اکانت ها را پیدا کنید. آیا مربوط به دانشکده هستند؟ اگر نیستند، آیا راهی برای فیلتر کردن آنها به نظر شما میرسد؟
- به نظر شما اگر دو اکانت در دانشکده یک اکانت را دنبال کنند، آن اکانت در کدام یک از این زیر گروه ها میگنجد:
  - دوست/آشنا/همکلاسی مشترک در دانشکده
  - دوست/آشنا مشترک خارج از دانشکده
  - Celebrity
  - ...

  آیا میتوان محدودیتی روی این رابطه گذاشت که فرد سوم با احتمال خوبی داخل دانشکده باشد؟ مثلا اگر دو اکانت خودشان همدیگر رد دنبال نکنند ولی هر دو یک نفر را دنبال کنند، آیا آن فرد با احتمال بیشتری داخل دانشکده است؟ آیا اگر این رابطه را بصورت متناوب برای تمام اعضاء پیدا شده (تا حالا) در دانشکده اجرا کنیم و بعد تکرار کنیم تا اینکه فرد جدیدی پیدا نشود، بیشتر اعضاء دانشکده با احتمال خوبی پیدا میشوند؟

این روند فکری ـ آزمایشی را در جهت پیدا کردن تمام اعضاء دانشکده ادامه دهید تا به نتیجه ای که به نظرتون قابل قبول است برسید. قابل قبول بودن دو جنبه دارد.

۱. میزان پوشش/coverage: اینکه مثلا از ده نفر رندوم داخل دانشکده که اکانت توییترشون را میشناسید چندتاشون با این روش پیدا شد. مثلا اگر هشت تا شون پیدا شد، یعنی میزان پوشش تون تقریبا هشتاد درصد است. آیا اصلا پیداشون نکردین، یا اینکه پیداشون کردین ولی با روشی که برای فیلتر کردن داشتین، حذ شدند. اگر راهی به نظرتون میرسه که پوشش را کاملتر بکنید انجام دهید. علت پیدا نکردن این اکانت ها را تحلیل کنید.

۲. میزان دقت/precision: از اینهایی که پیدا کردین، چند درصدشون واقعا توی دانشکده هستند. مثلا اگر ده تا رندوم از توی این لیست انتخاب کنید، و بعد از بررسی متوجه بشین که سه تاشون توی دانشکده نیستند، دقت شما هفتاد درصد است. باز باید فکر کنید ببینید راهی به نظرتون میرسه که اینها را از لیست حذف کنید یا نه.

۲ پیدا کردن حلقه های متصل

خوب حالا یه تعداد اکانت دارین که با تقریب خوبی داخل دانشکده هستند. هدف شما این است که حلقه های دوستی/آشنایی/ارتباط را پیدا کنید. برای این کار میتوانید از اکانت خودتان شروع کنید. ببینید

شما و گروهی که با آنها در ارتباط هستید را چگونه میتوانید پیدا کنید. مثلا یک راهش این است که اگر من در هزار توییت آخر خودم بیش از n بار یک نفر را mention کردم یک یال جهت دار از گره من به گره او اضافه کنم. با توجه به شناختی که از توییتر دارید میتوانید ایده های مختلف را امتحان کنید و ببینید کدام بهتر کار میکند. برای شروع میتوانید اطلاعات کامل همه اکانت ها و ۱۰۰۰ تا توییت آخر هر کدام را بگیرید و شروع به امتحان کردن ایده های مختلف بکنید. بعد از پیدا کردن تمام یال ها SCC ها را پیدا کنید. بعد اکانت ها را نگاه کنید، به نظرتون آیا گروه های معناداری پیدا شده اند؟ مثلا ورودی های سال های مختلف،...؟

۳ جمع آوری داده

برای شروع باید یک account developer درست کنید. برای درست کردن این اکانت لطفا وقت کافی بگذارید، ترجیحا از وی پی ان با آی پی آمریکا اکانت را باز کنید، توضیحات کامل با انگلیسی خوب بنویسید و کشور را هم آمریکا انتخاب کنید تا اکانت تون بدون دردسر درست بشود. بد از کد ضمیمه شده برای گرفتن اطلاعات اولیه استفاده کنید.

برای شروع کار قبل از اینک account developer درست کنید، داده اولیه ای ضمیمه شده که تمام follower های DevCampIUST و تمام follower های آنها ضمیمه شده. همچنین کدی برای پردازش اولیه داده موجود است.

نکته ای که باید دقت کنید این است که توییتر میزان دفعاتی که از آنها اطلاعات دریافت میکنید را محدود کرده و کار دانلود داده از آنها وقت گیر است. این است که کدتان را باید بگونه ای بنویسید که هر اطلاعاتی که دانلود میکنید با ساختاری مناسب در فایل و پوشه ها نگهداری کنید که لازم نباشد هر بار همه اطلاعات را از ابتدا دانلود کنید.

اگر به مشکلی برخوردید حتما از استاد یا حل تمرین ها کمک بگیرید.

موفق باشید!