تمرین اول درس شبکه‌های اجتماعی

فایل dolar-gte20180828-lt20180928.csv.zip در فولدر HW1-dataset حاوی مطالبی است که در کانالهای عمومی تلگرام در یک بازه یک ماه از 20180828 تا 20180928 به زبان فارسی منتشر شده و حاوی کلمه "دلار" بوده است. هر سطر مشخصات یک مطلب منتشر شده را نشان میدهد که فیلدهای مرتبط با آن به صورت زیر در فایل نوشته شده است:

- Id: شماره مطلب در کانال (با شناسه مطلب اشتباه نشود). این مطلب، مطلب چندم کانال بوده است؟
- peer_id: شناسه منتشر کننده مطلب (ترکیب peer_id و Id میتواند شناسه مطلب باشد)
- peer_type: نوع منتشر کننده مطلب (کانال عمومی، کانال خصوصی، کاربر، ربات)
- peer_username: شناسه انگلیسی منتشر کننده مطلب (در صورت وجود)
- peer_title: عنوان منتشر کننده مطلب (در صورت وجود)
- peer_participants_count: تعداد اعضای کانال
- date: تاریخ انتشار مطلب
- در صورتی که مطلب از منتشر کننده دیگری فوروارد شده باشد در ستونهای بعدی فیلدهای اطلاعاتی مرجع با پیشوند _fwd ذکر میشود:
- fwd_id
- fwd_peer_id
- fwd_peer_type
- fwd_peer_username
- fwd_peer_title
- fwd_peer_participants_count
- fwd_date

به معنی ستونها توجه داشته باشید. مثلا Fwd_date تاریخی است که مطلب اولین بار در کانال مرجع منتشر شده و سپس در تاریخ date توسط کانال peer_id بازنشر شده است.

- Type: نوع مطلب (اعم از متنی، تصویری، ویدئویی و ...)
- Views: تعداد بازدیدهای مطلب در کانالهای مختلف. در صورتی که مطلب فورواردی باشد این فیلد نامعتبر است. براساس روش تلگرام، بازدید مطالب فقط یک شمارنده دارد و در صورتی که مطلب در کانالی فوروارد شود بازدید آن با مطلب اصلی یکی است و با هم آپدیت میشوند. در نتیجه تعداد بازدیدها به حساب مطلب اصلی گذاشته میشود)
- Text: متن فارسی مطلب
- Link: لینک مطلب در کانال، قابل استفاده در مرورگرهای وب برای دسترسی به اصل مطلب در تلگرام
- cid: فعلا از آن استفاده نشده و برای همه مطالب خالی است
- peer_about: توضیحات کانال منتشر کننده
- fwd_peer_about: توضیحات کانال مرجع

خواسته های تمرین:

ابتدا **شبکه بازنشر** مطلب بین **کانالهای عمومی** را استخراج کنید به گونه ای که:

a. هر کانال عمومی یک نود شبکه باشد (**توجه داشته باشید که کانالهای عمومی در فایل دادگان در ستون type برچسب channel دارند**).
b. بین کانالهایی که از هم مطلبی بازنشر کرده اند یک لینک ایجاد کنید.
c. مبدا لینک کانال بازنشر کننده و مقصد لینک کانال مرجع باشد.
d. وزن هر لینک تعداد دفعات ارجاع باشد.

روی این شبکه و با استفاده از یکی از نرم افزارهای Gephi، Cytoscape و یا NetworkX محاسبات زیر را انجام دهید و نتیجه آن را گزارش کنید:

۱- آمار شبکه:
   i. تعداد نودها،
   ii. تعداد یالها
   iii. چگالی شبکه (density) – معنی آن را در اینترنت پیدا کنید
۲- توزیع آماری درجه نودها روی نمودار log-log:
   i. توزیع آماری درجه ورودی
   ii. توزیع آماری درجه خروجی
   iii. توزیع آماری درجه کل نودها
۳- توزیع آماری درجه وزن دار نودها روی نمودار log-log:
   i. توزیع آماری درجه ورودی وزن دار
   ii. توزیع آماری درجه خروجی وزن دار
   iii. توزیع آماری قدرت نودها (strength) (تعریف قدرت نودها را از منابع اینترنتی پیدا کنید)
۴- لیست نودهای برتر (ذکر اسم انگلیسی کانال یعنی فیلد username کافی است):
   i. ۲۰ نود برتر از نظر تعداد نودهایی که به آن ارجاع داده اند.
   ii. ۲۰ نود برتر از نظر تعداد نودهایی که از آن نود مورد ارجاع واقع شده اند.
   iii. ۲۰ نود برتر از نظر تعداد دفعاتی که به آن نود ارجاع شده است.
   iv. ۲۰ نود برتر از نظر تعداد دفعاتی که آن نود ارجاع داده است.
۵- توزیع آماری وزن یالها
۶- توزیع Clustering coefficient نودها و متوسط آن
۷- مولفه های همبند (weakly connected components)
   i. تعداد مولفه‌های همبن
   ii. توزیع سایز آنها

iii. سایز بزرگترین مولفه و نسبت آن با سایز شبکه

۸- مولفه های قویا همبند (strongly connected components)

i. تعداد مولفه‌های همبن

ii. توزیع سایز آنها

iii. سایز بزرگترین مولفه و نسبت آن با سایز شبکه

iv. قطر بزرگترین مولفه قویا همبند

نمره اضافه (۵۰ نمره اضافه بر ۱۰۰ نمره تمرین):

۱- پردازش متن و استخراج قیمت دلار اعلام شده در کانالها در روزهای مختلف بر روی کانالهایی که بیش از ۳۰۰ پیام در این بازه یک ماهه منتشر کرده اند.

خروجی باید در قالب یک فایل csv با فیلدهای زیر تولید شود:

Peer_username, Date, Price

که این فیلدها به ترتیب عبارتند از:

- Peer_username: نام انگلیسی کانال اعلام کننده قیمت
- Date: زمان اعلام قیمت
- Price: قیمت "دلار تهران" اعلام شده در کانالها. توجه داشته باشید که دلارهای اعلامی انواع مختلفی همچون دلار هرات، دلار سلیمانیه، دلار تهران و ... دارند. حتی قیمت دلار در تهران نیز براساس خرید و فروش در خیابانها و محله های مختلف نامهایی همچون دلار سبزه میدان و دلار افشار و دلار منوچهری دارند. منظور این تمرین قیمت دلار در "هر نقطه از تهران" بوده است. در صورت اعلام چند نوع دلار تهران در یک مطلب، متوسط آنها را ملاک بگیرید.

۲- استخراج شبکه همبستگی قیمت در کانالها:

- حساب کنید قیمتهای اعلام شده در هر کانال i با کانال دیگر j چقدر همبستگی آماری (pearson correlation) دارد؟
- فقط در صورتی که همبستگی بالاتر از ۰,۷ است لینک را به شبکه همبستگی اضافه کنید.
- خروجی باید در قالب یک فایل csv با فیلدهای زیر تولید شود:

Peer_username1, Peer_username2, correlation