Sharif University of Technology

Masoud Tahmasbi Fard

| | |

Student ID: 402200275

EE577: Massive Data Analytics

Assignment #4- Random Walk algorithms

February 4, 2024

Table of Contents

1 Task Description

In this assignment, we implement a recommender system using random walks over the bipartite graph of Persian Twitter users and tweets. The dataset is the same previously analyzed tweet corpus from Assignment 3.

The approach constructs a bipartite user-tweet interaction graph where edges link users to the tweets they have engaged with. We then apply Pixie, an algorithm which performs personalized PageRank via random walks on the graph to uncover each user's most influential neighbors. These neighboring users with strong connections are deemed similar users.

To validate performance, for recommended tweets derived from similar users' engagements, we retrieve and display their natively associated tweets from the dataset. Comparing recommended tweets against their original contextual tweets allows qualitative evaluation of recommendation quality.

In summary, this system generates tweet recommendations by first modeling the user-tweet dataset as a bipartite network amenable to random walk analysis. It then exploits the Pixie algorithm to discern tweet preference patterns among similar users via graph walks, recommending tweets that those closely-linked users engaged with. Retrieving associated tweets provides ground truth context to gauge recommendation accuracy.

2 Algorithm Description

2.1 Random Walks

Random walks on graphs provide a technique to derive similarities between nodes based on network structure and connections. The intuition is that nodes linked by short walk lengths likely share related attributes or connectivity patterns.

Formally, a random walk starts at some initial node and iteratively moves to a randomly selected neighbor. This repeats, traversing the graph in a randomized fashion. Tracking walk sequences reveals nodes frequently visited together, having high transition probabilities between each other.

For similarity detection, numerous random walks originating from each node are analyzed to determine node affinity. The probability distribution over other nodes converging from the walks indicates global structural equivalencies.

Specifically, nodes reached faster and more frequently from the origin node are deemed more similar, as walkers consistently traverse between them. Transition likelihoods correlate with similarities in position and role within the graph.

2.2 Pixie Algorithm

Pixie is an algorithm introduced by Pinterest for item recommendation using random walks on user-item bipartite networks.

In the bipartite graph, the two node sets are users and items. An edge connects a user to an item if that user has interacted with (e.g. liked or saved) the item. Random walks originating from each user node transition between similar user nodes with higher probability based on their item overlap. The consistent random walk patterns reveal user affinity even where no explicit edge exists. Figure 1 illustrates an example of a bipartite graph, generated from user-item interactions

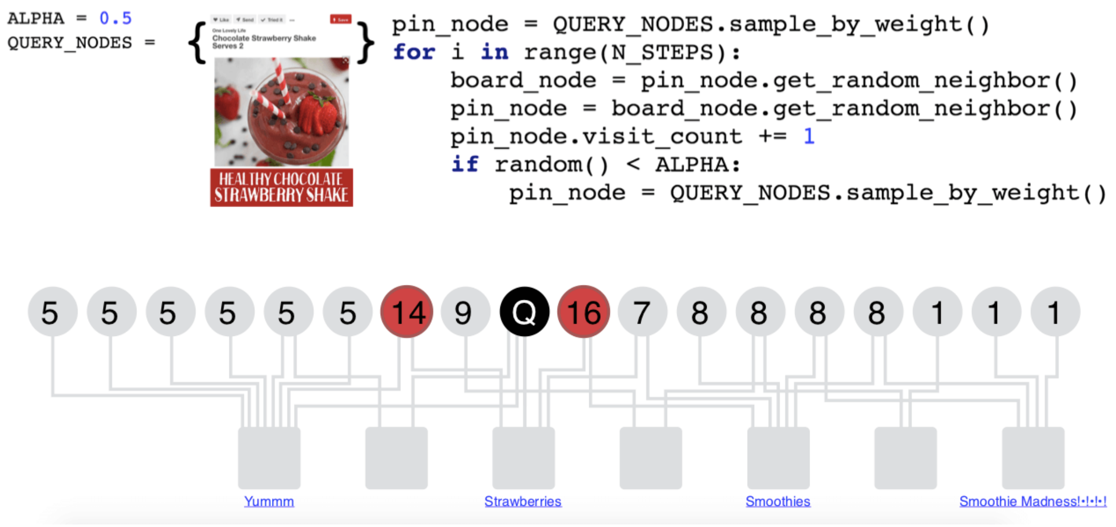

Figure 1: Pixie Algorithm

In our Twitter case, we construct a bipartite graph with user nodes on one side and tweet nodes on the other. An edge links a user to a tweet they engaged with, whether by replying, quoting or retweeting.

Random walks from a user node then hop between other user nodes that have interacted with common tweets. This models tweet preference similarity between the walked users - users consistently co-visited on short walks have demonstrated aligned tweet selections.

Finally, the tweets walked users engaged with that the source user did not are recommended, having been implicitly endorsed through the shared random traversal.

So in essence, Pixie transfers tweet endorsements from similar users discovered through randomized bipartite walk sequences in order to produce personalized recommendations while overcoming sparsity.

2.3 Implementation

To implement the Pixie algorithm, we first construct a bipartite graph representation consisting of user nodes on one side and tweet nodes on the other. Two RDDs encode the graph - one mapping users to their engaged tweet IDs, and another mapping tweets to their engaging user IDs.

Due to resource constraints and decreasing similarity from distant nodes, we extract a localized sub-graph around each target user for analysis by traversing nearby neighbors out to a depth of 2 hops empirically.

Random walks over the sub-graph originate at the target user node then transition to a randomly chosen tweet node from the target's engagements. Next, a random user is picked from that tweet's list who is then the origin for the next step.

We implement two walk modes - one returns to the source user on any dead-end, while the other enables teleportation back to source with some probability. Here we apply the former, as dead-ends occurred frequently enough to match typical teleport probabilities.

The most visited nodes indicate users most similar to the target. We retrieve both the target's and most similar users' tweets to enable qualitative evaluation of recommendation accuracy - if recommended tweets align with the target's tweet topics, the walk-based affinity estimates are demonstrated to translate to suggestion quality.

3 Results

Here is an edited version of the Results section intro:

This section presents recommended tweets generated for 5 users via bipartite Pixie walks. The first 3 users were previously analyzed in Assignment 3 for benchmarking. The final 2 users were randomly selected to demonstrate wider applicability. For each sampled user, their tweeted content provides ground truth context while the Pixie-recommended tweets based on graph walk user affinities showcase the system's ability to suggest relevant, personalized content matching users' interests. Analyzing recommendations across diverse users validates the implemented approach and provides qualitative insights into performance.

3.1 User #1

This user was previously analyzed using a collaborative filtering approach in Assignment 3. Figure 2 compares the similar users recommended by the random walk method versus collaborative filtering. As expected given the different underlying algorithms, there is some overlap as well as variance between the recommendation lists. Critically, the common users identified provides mutual validation between the two approaches.

Further analysis of the target user's tweets versus the tweets from a sample of top recommended similar users reveals strong topical alignment. As evidenced in the accompanying example figures, the recommended users share common interests and tweeting behavior with the target. This provides qualitative confirmation that the random walk algorithm properly identifies users posting about analogous subjects who would be expected to engage with similar content. Seeing analogous tweet topics verifies that the graph walk-based user affinity estimates appropriately capture preference similarities that translate to semantically relevant tweet suggestions.

Overall, inspection of recommendations for previously seen users builds additional confidence in the recommender while allowing interesting comparative analysis against the collaborative filtering method. The meaningful user overlaps combined with tweet topic matches support the validity of employing Pixie random walks on user-tweet bipartite networks to uncover personalized tweet recommendations.

```
[('1654179648761102337', 420),
 ('1710302619371982848', 308),
 ('1487097029990002688', 205),
 ('1648401324608827392', 166),
 ('1593669637102866432', 124),
 ('1226113547542773762', 122),
 ('1696453872074539008', 87),
 ('1253707166919397377', 85),
 ('1724483603600486400', 81),
 ('1632351425383522312', 80),
 ('1714031594598490113', 78),
 ('1398658879827886085', 73),
 ('1710903076574359552', 71),
 ('1221434210927464448', 70),
 ('1305617097339400192', 68),
 ('1654409094076268546', 61),
 ('1549397580869664775', 56),
 ('1643233712916578309', 54),
 ('1698191107786518528', 53)]
```

(a) Random Walks

```
[(1.6035674514745464, '1724483603600486400'),
 (1.6035674514745464, '1632351425383522312'),
 (1.6035674514745464, '1487097029990002688'),
 (1.341640786499738, '774470366676877312'),
 (1.341640786499738, '1725652340043100160'),
 (1.341640786499738, '1725280261300965378'),
 (1.341640786499738, '1721135891308007424'),
 (1.341640786499738, '1719916251404021760'),
 (1.341640786499738, '1719773177503682560'),
 (1.341640786499738, '1719696919692292096'),
 (1.341640786499738, '1718837061237530624'),
 (1.341640786499738, '1710302619371982848'),
 (1.341640786499738, '1708866964020514816'),
 (1.341640786499738, '1704606183858880513'),
 (1.341640786499738, '1704531218174251012'),
 (1.341640786499738, '1692155672807325699'),
 (1.341640786499738, '1689309272180482048'),
 (1.341640786499738, '1685928050708418560'),
 (1.341640786499738, '1675154446966312962'),
 (1.341640786499738, '1648401324608827392')]
```

(b) Collaborative Filtering

Figure 2: Similar users by RW and CF

```
[ ] target_user_tweets = parsed_rdd.flatMap(lambda x: target_user_tweets_parser(target_user,x))
    target_user_tweets.collect()

[('وصیت نامه یکی از  شهدای بزرگوار :  خدایا نشود که در زندگی دو دستی به درخت زندگی بچسبم و همچون میوه ابی پوسیده بیافتم خدایا تا ایمانی هست مرا برای خود قربانی کن\n#گمنام_مثل_مادر\n#لشکر_علویان https://t.co/CzBw8KoKmM',
  'generated'),
 ('زندگی پر از انتخاب هایی است که مسیر ها رو عوض می\u200cکنند، اینکه قهرمان زندگی چه شخصی باشه در عاقبت بخیری خیلی مهمه\n#گمنام_مثل_مادر\n#لشکر_علویان https://t.co/QGEkchpZl4',
  'generated'),
 ('حضور مردم قدر شناس دیار علویان در مراسم وداع با لاله\u200cها\n#وداع_با_لاله\u200cها\n#گمنام_مثل_مادر https://t.co/JWvgiokZzn',
  'generated'),
 ('دیار علویان و مردم همیشه در صحنه\n#وداع_با_لاله\u200cها\n#گمنام_مثل_مادر https://t.co/1NNoFp1AGe',
  'generated'),
 ('در پی اسم و...\n#گمنام_مثل_مادر\n#وداع_با_لاله\u200cها https://t.co/QgF5gIXki9',
  'generated'),
 ('اهل مقدسات علیه اینجا متأسفانه زنند،توهین می کنند و یه آب هم میزنن\u200cسنت قلم\u200cقبلا باور نمی کردم،ولی باچشمان خودم می بینم همونایکه دم از وحدت و برادری می زنند،همه توصیه وتأکیدانشون اطاعت نمی کنید؟\u200cانگش!نمی دونم فازشون چیه!شمایکه خودتون رو پیرو رهبری فرزانه مان میدونید،چجور از آن',
  'generated'),
 ('و اینک رهروان مکتب حاج قاسم در بابل ،فرزندان خوشنام\nخسته نباشی پهلوون\n\nمسافر من\u200cاومدی\u200cخوش\n\nشهید گمنام سلام💔وطن را در آغوش میکشد\n\n#گمنام_مثل_مادر\n#لشکر_علویان https://t.co/W8TdS4BqlM',
  'generated'),
 ('#گمنام_مثل_مادر\n #وداع_با_لاله\u200cها\n بوی مادر میدهد\n آمدنت چه آمدنیست که عجب هوای شهر\n شهید گمنام سلام',
  'generated'),
 ('🤲 خاطرات مردانی که پروای نام ندارند و در کهف گمنامی هستند\n.. چه زیباست\n#گمنام_مثل_مادر\n#لشکر_علویان https://t.co/Zj4jzA9tCG',
  'generated'),
 ('تو نشان از مادر داری گمنام بمان\nبخدا گمنام بودن زیباست\nگمنام بمان شهید عزیز\n#لشکر_علویان\n#گمنام_مثل_مادر',
  'generated'),
 ('دیار علویان این روزها مهمان دارد\nمهمانانی ویژه از سال\u200cهای حماسه و دفاع مقدس\nمهمانانی که #گمنام_مثل_مادر هستند\n\n#لشکر_علویان https://t.co/ioQCzwZBqC',
  'generated'),
 ('شهید گمنام مایه برکت برای نظام و باعث وحدت و همدلی میشود.\n#گمنام_مثل_مادر\n#لشکر_علویان',
  'generated'),
```

Figure 3: Targer user tweets

```
[('1654179648761102337', 420),
 ('1710302619371982848', 308),
 ('1487097029990002688', 205),
 ('1648401324608827392', 166),
 ('1593669637102866432', 124),
 ('1226113547542773762', 122),
 ('1696453872074539008', 87),
 ('1253707166919397377', 85),
 ('1724483603600486400', 81),
 ('1632351425383522312', 80),
 ('1714031594598490113', 78),
 ('1398658879827886085', 73),
 ('1710903076574359552', 71),
 ('1221434210927464448', 70),
 ('1305617097339400192', 68),
 ('1654409094076268546', 61),
 ('1549397580869664775', 56),
 ('1643233712916578309', 54),
 ('1698191107786518528', 53)]
```

Figure 4: Similar users and corresponding similarity score

Figure 5: Recommended tweets 1

Figure 6: Recommended tweets 2

3.2 User #2

Figure 7: Targer user tweets

```
[('824950806411624448', 127),
 ('302678816', 116),
 ('1562505373285658631', 77),
 ('855095592', 68),
 ('60171166', 67),
 ('767126934', 67),
 ('1271067471760568322', 60),
 ('1556954568797392897', 57),
 ('1576644054699642880', 52),
 ('1639263584407416839', 49),
 ('17532913', 44),
 ('1676645415758733333', 38),
 ('597673178', 38),
 ('1010561817997922308', 37),
 ('840241125499777024', 37),
 ('1321668098076954624', 36),
 ('46853901', 36),
 ('1231450050389315585', 34),
 ('1567482915600142339', 34)]
```

Figure 8: Similar users and corresponding similarity score

```
[ ] recommended_tweets = parsed_rdd.flatMap(lambda x: tweet_parser(sample_tweets,x))
    recommended_tweets.collect()
```

```
[('یه نفر اومده اعتراف کرده یه دختر (احتمالا کراش طرف بوده) اونو آنفالو کرده بعد از اینکه فهمیده مقالش MIT، تو پیج اعترافات\n\n\u200cهم جای عجیبیه MIT دانشگاه ریجکت شده 😂\nحقیقتا این یکی خیلی دارک بود. مثلا طرف کات کنه با دوست پسرش بهش بگه مقالت ریجکت شده من دیگه نمیتونم باهات باشم\n\n#فقر_فساد_گرونی',),
 ('عالیه:)) https://t.co/747xf8YIx9',),
 ('الان میگه کارت رو بخوای شارژ کنی باید تتر بگیری. هر تتر بگیری 8 % بر میدارم\n. این کیف پول هم چقدر کارمزد میگیره. هر کاری میکنی یه درصد میگیره\n.کارمزد\n.هر تتر هم 23 لیر زده\n.یعنی شما 100 تتر بگیری\n92 تتر میده بهت\n2116 لیر\nمیشه',),
 ('@hkazazi بهش بزنی بگیری دیگه جای از رو لیر نمیکنه قبولم حتی اخه :/\nترن ارزون خیلی اصلا میزدم حواله من',),
 ('بابا اینقدر بلند اعلام اسکم نکن :( یکی از این پرایا تقسیم سود صد درصدی گذاشته.',),
 ('هر سری میام اینجا لذت میبرم\nپراگ اسکان زیباست https://t.co/VVVEl8NYGo',),
 ('@SeyedAllyn گذاشته خرید اولین برای فقط و ساعت 100 فقط . ها باشه اسکم نکنم فکر . بیشتر یاد تخفیف های سنگین ریحون میوفتم که جنب مخاطب کنه\n',),
 ('@hkazazi اره سید بیشتر اکانتای خریداری شده خود این پروموشنو میگم نه کلی',),
 ('وضعیت: https://t.co/5O26C5fBy7',),
 ('وای پسر این مدل کیچاها منو سرویس کردن!!!! https://t.co/FZ4Q8O9zYS',),
 ('فروخته ۲۴ میلیون. با تحویل یک ماهه/۳۵ روزه و تقریبا هیچ تحویل سر موعدی نداشته و کلی تحویل عقب افتاده داره، در اقدامی\u200cیک شرکتی که آیفون ۱۳ رو می\u200cجالب هم اینه برای هیچ نهاد نظارتی، مسئولی سوال پیش نمیاد چطوری طرف آیفون رو\n.کنه ۲۰میلیون تومان\u200cفروش می\u200cجالب قیمت رو کاهش داده و آیفون رو پیش\u200cیک\n.کنن\u200cنام رو بیشتر و بیشتر و بیشتر می\u200cیا چرا بعد از این همه تحویل عقب افتاده، با کاهش قیمت و افزایش تبلیغات، ثبت\n.فروشه\u200cارزون تر از خود اپل می\u200cشعبه\u200c(طبق گفته خودشون) شون هم پلمپ شد البته. نه به خاطر این فعالیت مشکوک. بلکه به خاطر حجاب.',),
 ('@Nikolaona برو خدا رو شکر کن. ما هی راه پله و چراغ راهنمایی و موتور تشخیص میدیم',),
 ('@hkazazi @Nikolaona البته خودت استادی :) شوخی و جدیشو نفهمیدم ولی راحت میتونی تغییر ریجن بدی کل گوگلتو،',),
 ('امروز چقدر تست گرفتیم',),
 ('شب بخیر',),
 ('اندر احوالات آیفون های ۲۰ میلیون تومنی و رویا فروشی به مردمی که به سخت ترین شکل ممکن پول در میارن اما به راحت ترین شکل ممکن سرشون رو واسه کلاه های گشاد اما پنهان آماده میکنن ...\nآزاده، حتی بدون تگ، بازنشر این ویدیو https://t.co/99k4fKIZNX',),
 ('اختلاس چای دبش که حاصل اقدام چند وزارتخانه و گمرک و بانک\n#خراش ساختهٔ شده به دست انسان = ۱.۵ میلیارد دلار\u200cهزینه ساخت برج خلیفه بلندترین آسمان خراش ساختهٔ مرکزی است = ۳.۴ میلیارد دلار https://t.co/wF7ohsZwG0',),
 ('ابتدا*\n\n🔥34 Billion of dollars! 🔥\n   تکمیلی: بزرگترین فساد کشف شده مربوط به پدیده شاندیز به ارزش روز، ۳۴ میلیارد دلار است👈\nحساب نرخ تورم دلار (سایت فدرال رزرو) ارزش روز محاسبه شد \u200cارزش دلاری همان سال محاسبه و سپس با توجه به کاهش ارزش خود دلار، با کمک ماشین https://t.co/LmfPm59Vki',),
 ('پس از ۱۳/۸روز بررسی سایت #قمار #سایسی_ماتکن چطور سه درگاه پرداخت کارت به کارت بانک @mahramian رسانه/مطالبه گر/میتونی از معاون محترم بانک مرکزی تمام سایتهای #قمار این خدمات رو داره و "حتی یک روز"\n\nیادداشت\u200cبی❌\n\u200cهشتاد\u200cهف❌\nهمراه کارت❌\n:دو روز پیش\nمرکزی داره و جواب هم بگیری؟ ۲/۳\nقطع نشده! https://t.co/7K9DlalN3Q',),
```

Figure 9: Recommended tweets 1

```
[ ] recommended_tweets = parsed_rdd.flatMap(lambda x: tweet_parser(sample_tweets,x))
    recommended_tweets.collect()

[('دوستان پلاس ولی استیشن از کجا بگیرم قیمتش خوب باشه؟',),
 ('دوستان من بدنبال کار بصورت پروژه ای هستم(پایتون)\nممنون میشم ریتوییت کنید🙏 \n#ریتوییت',),
 ('البته الان اسپاتیفای پلن ۳ ماهه پرمیوم رایگان هم داره.\nمیتونید کارت فیک هم بزنید و اگه قبول شد استفادش کنید.\n\nهمین الان برای دوستم با کارت فیک گرفتم :)))) https://t.co/QEw09pML8u',),
 ('به نظرتون جواب همچین متنی چیه؟ https://t.co/Hy74QC5m9G',),
 ('@nimbit__ من ازشون زیاد خرید داشتم و برای من کاملا اوکی بوده تا الان. قیمت هاشونم کم بوده به نسبت بقیه جاها\nولیییی\nاین دلیل خوبی نیست. من هیچی تضمین نمیکنم 😬',),
 ('@Nikolaona نه فقط میخواستم بدونم تا حالا ازشون خرید داشتی یا نه:(\nبازم ممنون\nباید ترکیه vpn فقط با مستر کارت ترکیه هم بزنیم دیگه؟\nجون ریجن اسپاتیفای هم باید عوض بشه فک کنم',),
 ('@nimbit__ نه نیاز نیس\nموقع زدن شماره کارت از پایین کشور رو بکن ترکیه',),
 ('@Nikolaona بعد یه سوال دیگه\nبعد بخوام با همین ریجن گوگل پلی عوض کنم میشه؟\nمثلا همین مستر کارت رو اد کنم گوگل پلی ریجنم بشه ترکیه(تا حالا نکردم)',),
 ('از من به شما نصیحت! \nماشین قراره بخری؟ خفه شو\nگوشی قراره بخری؟ خفه شو\nخونه قراره بخری؟ خفه شو\nهرکاری که میخوای بکنی فقط خفه شو\nشاید الان بدنون بیاد از این حرف ولی بعدش نتیجه رو میفهمین!',),
 ('@nimbit__ کوچیکه داداش تو تلویزیون بگیر',),
 ('به عنوان یه آدم قدیمی این آهنگ قدیمی رو خیلی دوست دارم:\nhttps://t.co/Y35UrznmiC',),
 ('@realbardia فک کردم فقط منم سی سی کج گوش میدم:)))))\nولی Are you man enough بهترینشه',),
 ('آقا کسی اینجا هست یه اکانت تلگرام که شماره غیر ایرانی داره داشته باشه؟\nکارم خیلی واجبه یکاری هست که فقط رو اون شماره ها میشه انجام داد...',),
 ('زیرنویس عالی https://t.co/T8iZNddaGq',)]
```

Figure 10: Recommended tweets 2

3.3 User #3

```
[ ] target_user_tweets = parsed_rdd.flatMap(lambda x: target_user_tweets_parser(target_user,x))
    target_user_tweets.collect()

[('تو » دیگر«\nحتی «تو» نیستی\n«شما» شده اید\nو این دنیای مودب جای «ما» نیست\nما دیوانه شده ها \n#افشین_خماند',
  'generated'),
 ('...تمام سیاره ها را گشتم تا به جایی رسیدم که خالی از هر کسی بود و آنگاه دلم آرام گریست\n...از آن روز به بعد شدم باغبان باغچه ی کوچکم با یک گل\nگلم قرمز ، بنفش و اکنون آبی را در خود دارد...',
  'generated'),
 ('...امروز خواستم از عشق بگم، دیدم عشقی نیست! خواستم از عاشقونه ها حرف بزنم دیدم چیزی نیست، یادم اومد از دل تنگ بنویسم که فهمیدم!!\nنه! هیچی، چیزی نفهمیدم، فقط خوابیدم. همین...',
  'generated'),
 ('اگه دستت بی وزن شد، حرکتش دادی و نتونستی لمسش کنی، باختی\nاگه چشات صدات کرد، سرتو بلند کردی ندیدیش، باختی\nاگه گوشات هوشیار شد، تیزش کردی و صداش رو نشنیدی، باختی\nاگه دلت تنگ شد، موبایلت رو ورداشتی، زنگش زدی، جوابت رو نداد، باختی\nو امشب من بازنده ترینم...',
  'generated'),
 ('یه روز یکی از آدمای اطرافم که دنیاش خیلی رنگیه بهم گفت: چقدر دوسم داری؟!\n...خیلی زود جواب دادم دوس داشتن که اندازه نداره رفیق\nگفت یعنی جی؟! گفتم دوس داشتن یا هست یا نیست، اندازه نداره.\n...گفت باز نفهمیدم چی میگی!!! گفتم هر وقت کسی رو دوس داشتی متوجه میشی. یهو داد زد و گفت آهان',
  'generated'),
 ('از بوی تنت فهمیدم\nشاید دیر رسیدم خیلی دیر \nاما مگر قانون این نبود که هر آنچه دیر آید می آید\nعاقبت روزی به خانه ی ما خواهد رسید؟_...',
  'generated')]
```

Figure 11: Targer user tweets

```
[('1574904470835798041', 1193)]
```

Figure 12: Similar user and corresponding similarity score

```
[ ] recommended_tweets = parsed_rdd.flatMap(lambda x: tweet_parser(sample_tweets,x))
    recommended_tweets.collect()

    [('تو\nآنجایی\nو\n "آنجا" \nنمی داند\nکه\nچقدر\nخوشبخت\nاست!',),
     ('…کنار کسی که دوستش داری شروع میشه\nبخیر ترین صبح قطعا اونی که',),
     ('…آنجا که تو نیستی، کجا خواهد بود',),
     ('…و خدا همه چیز را درست میکند\n، گمان میکنی پایان است',),
     ('…عاقبت روزی به خانه ی ما خواهد رسید؟\nاما مگر قانون این نبود که هر آنچه دیر می آید\n شاید خیلی دیر رسیدیم خیلی دیر\nاز بوی تنت فهمیدم',),
     ('…شهر حسودی بکند\nبغلم کن همه جا',),
     ('…در هستی پنهان من\nای هست تو پنهان شده\nای روی تو ایمان من\nای دیدن تو دین من',),
     ('به او می گویم حالت چطور است؟ و به تو فکر\n به آن زن تلفن میکنم و درعین حال به تو فکر میکنم\n به خانه برمی گردم وبه تو فکر میکنم\nکار می کنم و به تو فکر میکنم
    ...به او لبخند می زنم ،چشمانم را می بندم و به تو فکر می کنم\n کجا برویم ؟وبه تو فکر میکنم\nمیکنم',)]
```

Figure 13: Recommended tweets

3.4 User #4

```
[ ] target_user = users_rdd.takeSample(False,1,487)[0]
    target_user

    ('1343961231166734336',
     ['1736503973559906442',
      '1736422396721062257',
      '1736384901635539042',
      '1736407359461704039',
      '1736455332614877486',
      '1736307951638323471',
      '1736442052454310094'])
```

Figure 14: Targer user 4

```
[ ] target_user_tweets = parsed_rdd.flatMap(lambda x: tweet_parser(target_user_tweets_ids,x))
    target_user_tweets.collect()

    [('انتقام_سخت#\nشهدای_فاطمیه#\n.الله در میان اشک و آه بدرقه کردند\u200cمردم گیلان شهید حسن باباتیا را در روز شهادت حضرت زهرا سلام
    https://t.co/AXAcw9R9xc',),
     ('شرم بر تو\n!!!!!!هستند ، اومدن نمایش ترازیک و عذاب آور ببیند IR خانم زینب اسماعیلی خبرنگار سابق روزنامه جاسوس پرور شرق ، این جمعیت میلیونی که میبینی مردم ایران
    👌 https://t.co/JiUnVKOXzK',),
     ('علیها در حسینیه امام خمینی(ره). ۱۴۰۲/۰۹/۲۶\u200cالله\u200cزهرا سلام\u200cاکنون؛ تصویری از رهبر انقلاب در ابتدای مراسم عزاداری شام شهادت حضرت فاطمه\u200cهم
    https://t.co/Z3AZqJsY5f',),
     ('به کجا چنین شتابان. ، سردار از #آیزنهاور پرسید 😁😁 https://t.co/m24Eb8AyGW',),
     ('برادران\n فاطمه سلام الله علیها بنیانگذار این مکتب است\n… آن مظلوم باید مقتدر باشد نه ذلیل\n:منتها یک شرط دارد\n.هیچ چیز مانند مظلومیت، رسواکننده ظلم و ظالم نیست
    غزه_تقاوم#\n مظلوم_مقتدر#\n\n.خواهرانمان در غزه خواه بدانند یا ندانند پیرو این مکتب هستند\u200cو',),
     ('شوند، زیرا\u200cتبدیل می\u200cها\u200cترین انسان\u200cترین و خطرناک\u200cاند به متوهم\u200cکسانی که چند کتاب محدود خوانده\nاند!یا زیاد بخوانید یا اصلا نخوانید
    ...کنند\u200cتعصب شدیدی روی دانش اندکشان پیدا می\n کارل سیگن',),
     ('علیها در حسینیه امام خمینی(ره). ۱۴۰۲/۰۹/۲۶\u200cالله\u200cهای متفاوتی از مراسم عزاداری شام شهادت حضرت فاطمه زهرا سلام\u200cقاب
    https://t.co/IWPzZrVEXS',)]
```

Figure 15: Targer user tweets

```
[('952453705', 755),
 ('37050268', 370),
 ('1591547142136627202', 336),
 ('1303599606551531520', 327),
 ('1343961231166734336', 43),
 ('1575260422516662272', 25),
 ('1705256827213041669', 24),
 ('1593256090908532736', 23),
 ('959142611025059842', 22),
 ('1272999611607719942', 21),
 ('1712456826292428800', 21),
 ('1501786831725240321', 19),
 ('1089593176808148997', 18),
 ('1545831035208093696', 18),
 ('1021430809474076674', 17),
 ('1269654925', 17),
 ('1420411453975273474', 17),
 ('1579439373636046848', 17),
 ('1267421104861261826', 16)]
```

Figure 16: Similar users and corresponding similarity score

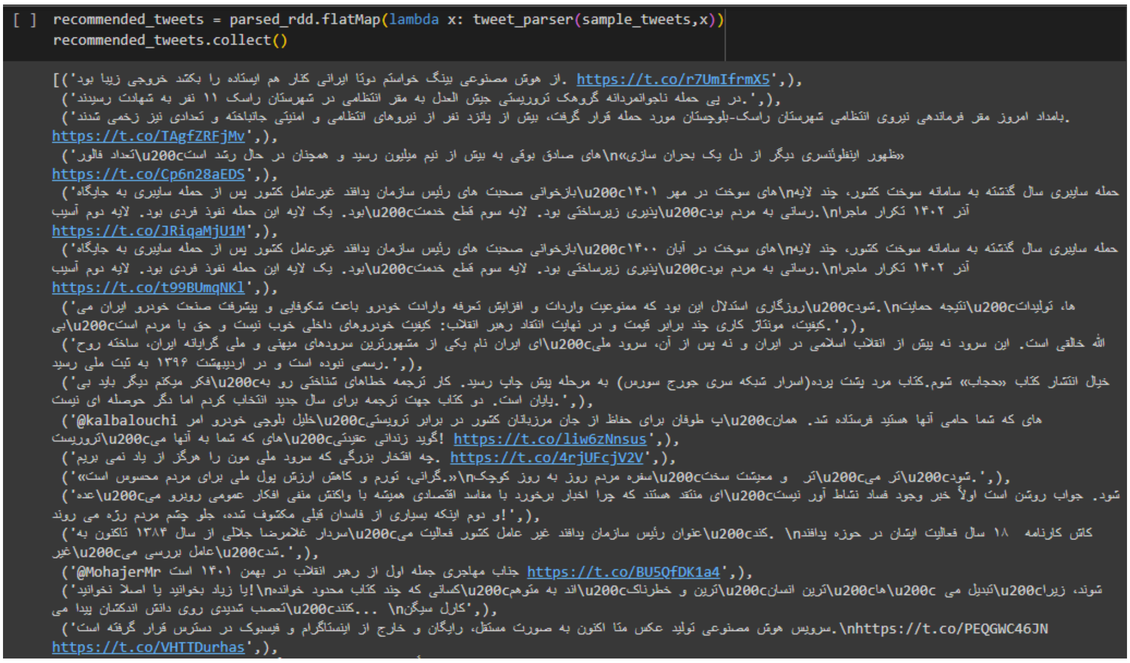

Figure 17: Recommended tweets 1

Figure 18: Recommended tweets 2

4 Discussion

Is the implemented model online or offline?

The current implementation operates in an offline manner. Some key points:

- The full bipartite graph is preprocessed into localized target user-centric subgraphs upfront before walking
- There is no incremental updates to the graph representation
- The random walk computations originate from static snapshots of the subgraphs

This indicates new tweets or users would require recomputation of the subgraphs and re-execution of the random walks to refresh recommendations. The upfront bipartite graph partitioning and materialization makes directly incorporating dynamic new nodes challenging.

However, there is potential to evolve this approach to an online system:

- Subgraphs could be generated in a streaming fashion using sliding windows

- Random walks could run continuously on subgraph snapshots
- Edge updates could trigger targeted partial recomputation

  In summary, the current system follows an offline batch analysis paradigm over static snapshots. But an online workflow could be achieved by adapting the preprocessing and incremental updating the graph analytics. This would allow dynamically adjusting recommendations in real-time as new edges arrive without full rebuilds.

Comparison between Random Walk and Collaborative Filter

Advantages of random walks:

- Captures implicit relationships beyond explicit engagement edges
- Handles data sparsity well, can connect disjoint user clusters
- Computationally efficient to approximate recommendations
- Naturally incorporates bipartite structure between users and tweets
- Visualizable walks aid debugging and insight

Disadvantages of random walks:

- Preprocessing overhead to construct bipartite graph
- Static batches lack easy online update process
- Walk hyperparameter tuning is challenging
- Hard to provide interpretable explanations

Advantages of collaborative filtering:

- Intuitive and transparent user-user correlations
- Incremental Updates with new data inserts easier

- Widely analyzed approach with known strengths
- No required modeling of tweet relationships
- Explains recommendations via user alignments

Disadvantages of collaborative filtering:

- Limited by sparsity of engagement edges
- User correlations can be less robust
- Suffers from cold start problem
- Hard to go beyond direct connections
- User biases and outliers degrade performance

Random walk's graph diffusion captures subtle signals across disjoint data, overcoming sparsity issues. However collaboration filtering is more interpretable and naturally incrementally. Hybrid approaches combining explicit correlations augmented with structural walk affinities may balance the tradeoffs. The best approach depends on the dataset connectivity, domain biases, and product requirements. But both move beyond content features with user behavioral modeling.