گزارش سوم

استاد راهنما: جناب استاد دکتر معینی
دانشجو: شهاب قدسی

بازه فعالیت: فروردین و اردیبهشت ۱۴۰۰

پس از تایید نهایی طرح پیشنهادی بررسی دقیق تر مقاله های اصلی شروع گردید. تلخیص یکی از این مقاله ها که توسط اکوقلو و همکاران[1] انجام گرفته به شرح زیر است:

- در این مقاله شبکه نظرات کاربران به شکل $G_s=(V,E)$ بیان شده است . G_s نشانگر گراف علامتداری است که در آن V مجموعه کاربران و محصولات (گره ها به دو بخش کاربران و محصولات افراز می شوند) و E مجموعه نظرات است.
- هریال در این شبکه به صورت $e(u_i,p_j,s)\in E$ تحت عنوان نظر کاربر u_i در مورد محصول p_j با علامت s می باشد.($s\in\{+,-\}$ ، نظر می تواند مثبت یا منفی باشد)
- به هر یک از موجودیت های کاربر، محصول و نظر برچسب های زیر تعلق می گیرد:
  - برچسب کاربر می تواند honest یا fraud باشد.
  - برچسب محصول می تواند good یا bad باشد.
  - برچسب نظر می تواند real یا fake باشد.
- به هر یک از برچسب ها یک احتمال نسبت می دهیم.
- اگر عناصری که از قبل برچسب خورده اند را با x نشان دهیم، می خواهیم انتساب y را به تمامی عناصر برچسب نخورده به نحوی نسبت دهیم که مقدار $P(Y|X)$ بیشینه شود.
- هدف، پیدا کردن انتساب y ای است که مقدار $P(Y|X)$ را بیشینه کند.

بعد از مرور و بررسی این مقاله پیاده سازی این روش بر روی مجموعه داده های آمازون[2] در چهارچوب اسپارک در دستور کار قرار گرفت. با توجه به اینکه اسپارک ابزار های خاصی برای کار با گراف ها طراحی کرده است، دو ابزار GraphX و GraphFrames را مورد بررسی قرار گرفت که در شکل زیرمشخصات کلی هریک نشان داده شده است.

1 L. Akoglu, R. Chandy and C. Faloutsos (2013), Opinion Fraud Detection in Online Reviews by Network Effects

2 https://nijianmo.github.io/amazon/index.html

	GraphFrames	GraphX
Core APIs	Scala, Java, Python	Scala only
Programming Abstraction	DataFrames	RDDs
Use Cases	Algorithms, Queries, Motif Finding	Algorithms
VertexIds	Any type (in Catalyst)	Long
Vertex/edge attributes	Any number of DataFrame columns	Any type (VD,ED)
Return Types	GraphFrames/DataFrames	Graph [VD,ED] or ...

جدول ۱: مقایسه GraphX و GraphFrames[3]

با توجه به اینکه GraphX رابط زبان پایتون ندارد و ابزار قدیمی تری است، GraphFrames به عنوان ابزار مورد استفاده انتخاب گردید. در این فرآیند اقدام به نصب GraphFrames بر روی اسپارک گردید که متاسفانه repository ای که کد های این ابزار بر روی آن قرار داشت به فعالیت خود خاتمه داده بود و دیگر قابلیت دسترسی نداشت. با جستجوی فراوان در گیت هاب نویسندگان و انجمن های مربوطه سرانجام در اواسط اردیبهشت ماه یک repository جدید توسط نویسندگان این ابزار ایجاد گردید و من موفق به نصب آن شدم.

همزمان با تلاش برای نصب GraphFrames به دنبال مجموعه داده های آمازون رفتم. خلاصه ای از ویژگی های این مجموعه داده را در زیر آورده شده است:

- این مجموعه به بخش های متنوعی مانند کتاب، لباس، آهنگ، نرم افزار، فیلم، وسایل دیجیتال و ... تقسیم بندی شده است که شرح دقیق تر آن در جدول ۲ آمده است.
- سه نوع از این مجموعه داده موجود است:
  - raw review data با حجم ۳۴ گیگابایت: تمام ۲۳۳.۱ میلیون نظر
  - ratings only با حجم ۶.۷ گیگابایت: همانند مورد بالا با این تفاوت که فراداده و متن نظرات را شامل نمی شود.
  - ۵-core با حجم ۱۴.۳ گیگابایت: زیرمجموعه ای از داده ها که در آن همه کاربران و محصولات حداقل ۵ نظر را دارند.(۷۵.۲۶ میلیون نظر)

3 https://www.oreilly.com/library/view/learning-apache-spark/9781785885136/ch07s10.html

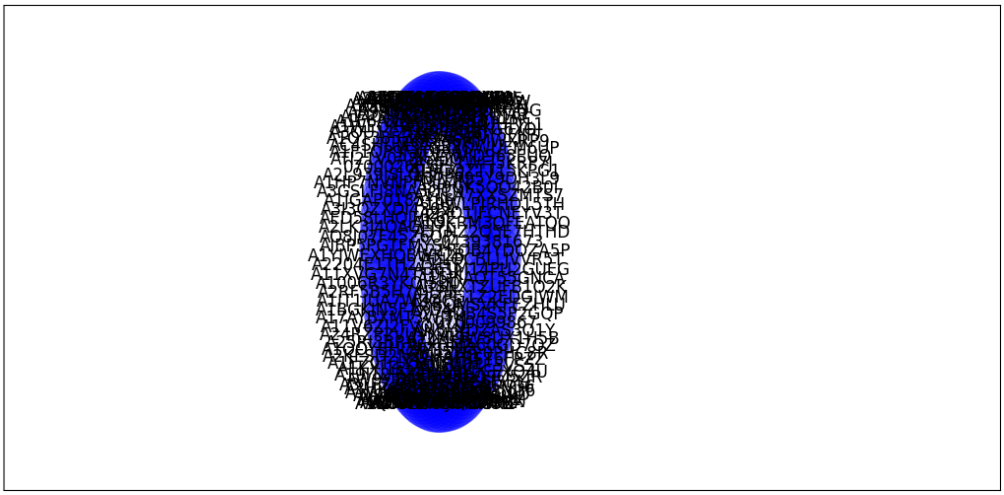
Amazon Fashion	reviews (883,636 reviews)	metadata (186,637 products)
All Beauty	reviews (371,345 reviews)	metadata (32,992 products)
Appliances	reviews (602,777 reviews)	metadata (30,459 products)
Arts, Crafts and Sewing	reviews (2,875,917 reviews)	metadata (303,426 products)
Automotive	reviews (7,990,166 reviews)	metadata (932,019 products)
Books	reviews (51,311,621 reviews)	metadata (2,935,525 products)
CDs and Vinyl	reviews (4,543,369 reviews)	metadata (544,442 products)
Cell Phones and Accessories	reviews (10,063,255 reviews)	metadata (590,269 products)
Clothing Shoes and Jewelry	reviews (32,292,099 reviews)	metadata (2,685,059 products)
Digital Music	reviews (1,584,082 reviews)	metadata (465,392 products)
Electronics	reviews (20,994,353 reviews)	metadata (786,868 products)
Gift Cards	reviews (147,194 reviews)	metadata (1,548 products)
Grocery and Gourmet Food	reviews (5,074,160 reviews)	metadata (287,209 products)
Home and Kitchen	reviews (21,928,568 reviews)	metadata (1,301,225 products)
Industrial and Scientific	reviews (1,758,333 reviews)	metadata (167,524 products)
Kindle Store	reviews (5,722,988 reviews)	metadata (493,859 products)
Luxury Beauty	reviews (574,628 reviews)	metadata (12,308 products)
Magazine Subscriptions	reviews (89,689 reviews)	metadata (3,493 products)
Movies and TV	reviews (8,765,568 reviews)	metadata (203,970 products)
Musical Instruments	reviews (1,512,530 reviews)	metadata (120,400 products)
Office Products	reviews (5,581,313 reviews)	metadata (315,644 products)
Patio, Lawn and Garden	reviews (5,236,058 reviews)	metadata (279,697 products)
Pet Supplies	reviews (6,542,483 reviews)	metadata (206,141 products)
Prime Pantry	reviews (471,614 reviews)	metadata (10,815 products)
Software	reviews (459,436 reviews)	metadata (26,815 products)
Sports and Outdoors	reviews (12,980,837 reviews)	metadata (962,876 products)
Tools and Home Improvement	reviews (9,015,203 reviews)	metadata (571,982 products)
Toys and Games	reviews (8,201,231 reviews)	metadata (634,414 products)
Video Games	reviews (2,565,349 reviews)	metadata (84,893 products)

جدول ۲: تقسیم بندی مجموعه داده آمازون[۴]

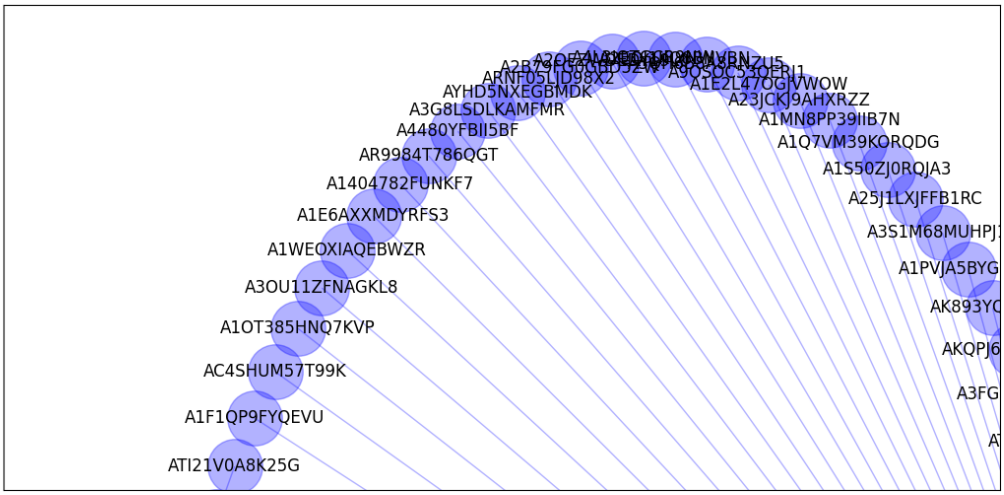
نکته: حجم فایل های ذکر شده، حجم فایل فشرده شده می باشد.

طبقه بندی video Games (یکی از کم حجم ترین طبقه بندی ها با حجم ۵۲۲ مگابایت) برای شروع کار دانلود گردید. چکیده ای از ویژگی ها و شکل این مجموعه داده که توسط **اینجانب** تولید شده در شکل های زیر آمده است.

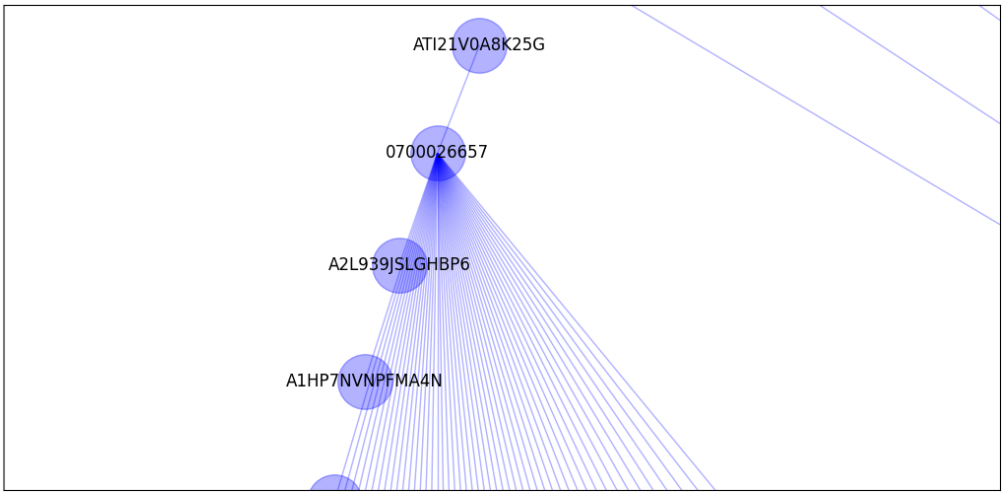
4 https://nijianmo.github.io/amazon/index.html



شکل ۱: نمای کلی گراف



شکل ۲: گره های کاربران و یال نظرات آن ها

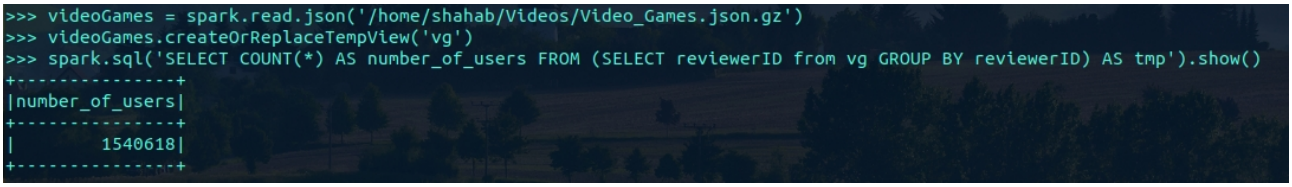


شکل ۳: یک گره محصول و یال های نظرات مرتبط با آن

```
>>> spark.sql('SELECT COUNT(asin) as product_degree from vg GROUP BY asin ORDER BY COUNT(asin) DESC').show()
+--------------+
|product_degree|
+--------------+
|          7630|
|          6462|
|          5135|
|          4359|
|          3962|
|          3960|
|          3930|
|          3634|
|          3520|
|          3345|
|          3167|
|          3162|
|          2990|
|          2959|
|          2952|
|          2928|
|          2912|
|          2905|
|          2879|
|          2859|
+--------------+
only showing top 20 rows
```

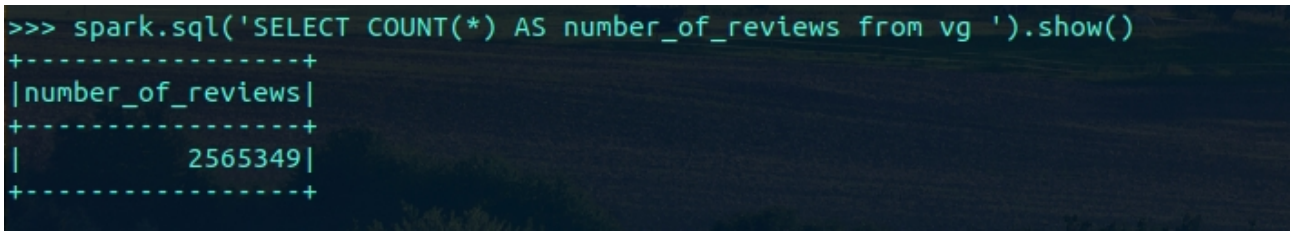
شکل ۴: درجه گره های محصولات به صورت نزولی

شکل ۵ : درجه گره های کاربران به صورت نزولی



شکل ۶ : تعداد کاربران

شکل ۷: تعداد محصولات



شکل ۸ : تعداد نظرات

اکنون درصدد طراحی پلتفرم مناسب برای اجرای الگوریتم طرح شده در مقاله اکوقلو بر روی مجموعه داده فوق به عنوان نقطه آغازی برای دستیابی به الگوریتم و روش ابداعی مورد نظر در پایان نامه هستم.