دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده ریاضی و علوم کامپیوتر

گزارش ششم

ارائه یک سیستم توصیه‌گر برای انتخاب کتاب

نگارش

سروش آریانا

استاد

دکتر مهدی قطعی

اردیبهشت ۱۴۰۰

مقدمه

سیستم‌های توصیه‌گر، همانطور که از نامشان مشخص است، سیستم‌ها و یا تکنیک‌هایی برای پیشنهاد دادن یک محصول، سرویس و یا یک موجودیت است. سیستم‌های توصیه‌گر به انواع مختلفی تقسیم بندی می‌شوند اما بصورت کلی سه نوع مختلف از سیستم‌های توصیه‌گر وجود دارد:

- فیلترکردن مشارکتی (Collaborative Filtering)
- سیستم‌های مبتنی بر محتوا (Content-based systems)
- توصیه‌گرهای مبتنی بر دانش (Knowledge-based recommenders)

سیستم‌های مبتنی بر Collaborative Filtering، یکی از پر طرفدارترین سیستم‌های توصیه‌گر در صنعت است که با استفاده از داده‌های قبلی موجود از کاربران و یافتن کاربران مشابه، سعی در ارائه پیشنهادات می‌دهد. نمونه استفاده موفق از این سیستم در کمپانی آمازون قابل مشاهده است. اما بزرگترین مشکل استفاده از این سیستم‌ها نیاز به داده بسیار بزرگ برای ایجاد یک سیستم موفق است. از این مشکل با عنوان Cold Start Problem یاد می‌شود.

از طرفی سیستم‌های Content-based نیاز به داده‌های بسیار بزرگ برای شروع ندارند. نمونه پیاده‌سازی این مدل سیستم در سایت نتفلیکس قابل مشاهده است. این سیستم‌ها سعی می‌کنند با استفاده از پروفایل کاربر و متا دیتاها به کاربر یک محصول را پیشنهاد کنند. با وجود اینکه این سیستم‌ها به داده کمتری برای شروع نیاز دارند اما در عمل خروجی سیستم‌های Collaborative filtering بسیار خلاقانه‌تر و بهتر است.

سیستم‌های Knowledge-based اصولا برای محصولاتی به‌کار می‌رود که اطلاعات زیادی از فروش آن‌ها در دسترس نیست. در چنین مواردی باید سیستمی ساخت که با گرفتن مشخصات و اولویت‌های کاربر از ورودی، سعی در پیدا کردن پیشنهادات بر اساس آن محدودیت‌ها بکند. این سیستم‌ها نیز همانند سیستم‌های Content-based از کمبود خلاقیت در ارائه پیشنهادات رنج می‌برند.

هدف از انجام این پروژه پیاده‌سازی یک سیستم Knowledge-based برای انتخاب کتاب است. برای ساخت این سیستم از دیتاست good-books10k و پایتون و کتابخانه‌های مربوطه استفاده شده است. همچنین کدهای مربوطه در کولب قابل دریافت و مشاهده است که در کنار فایل pdf گزارش ارسال شده است.

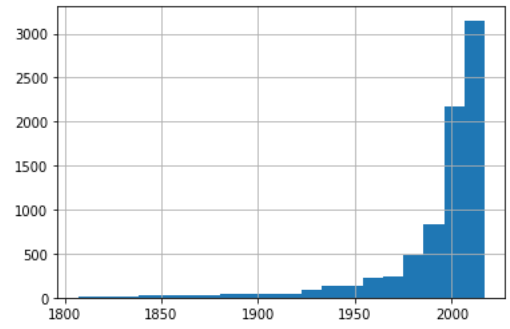
در این سیستم سعی داریم با دریافت ژانر، بازه انتشار کتاب و زبان مورد نظر به کاربر کتاب‌های مناسب را پیشنهاد کنیم. در ادامه به بررسی دیتاست و جزئیات سیستم پرداخته می‌شود.

۱- دیتاست

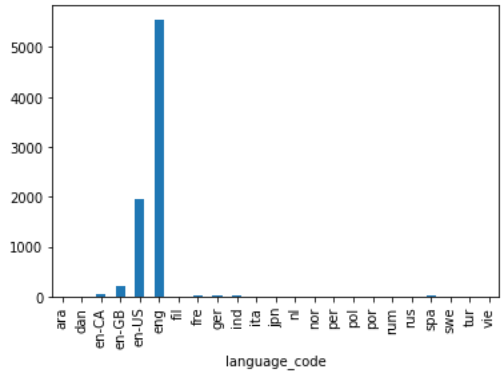
برای ساخت این سیستم نیاز به یک دیتاست مربوط به متا دیتاهای کتاب‌های مختلف نیاز داشتیم. انتخاب ما برای ساخت این سیستم دیتاست goodbooks-10k می‌باشد. این دیتاست از سایت goodreads جمع‌آوری شده است و شامل اطلاعات مربوط به ۱۰ هزار کتاب مختلف و نظرات کاربران است. یکی از تفاوت‌های مهم این دیتاست با دیتاست‌های مشابه مانند دیتاست آمازون این است که این دیتاست اطلاعات بیشتری از جمله تگ‌های کتاب‌ها نگه‌داری می‌کند که در روند معرفی کتاب بسیار کاربردی‌اند.

این دیتاست شامل کتاب‌هایی از سال ۱۷۵۰ قبل از میلاد تا سال ۲۰۱۷ است. کتاب‌های این دیتاست از زبان‌های مختلف از جمله انگلیسی، فرانسوی، آلمانی، ایتالیایی، فارسی، عربی و ... است. اما همانطور که در شکل ۱ و شکل ۲ قابل مشاهده است بیشتر این کتاب‌ها، کتاب‌های انگلیسی زبان و نسبتا جدید هستند.

همچنین میانگین امتیازات همه کتاب‌ها ۳,۹ و میانگین تعداد رای برای هر کتاب بیش از ۶۱۰۰۰ رای است.



شکل ۱ هیستوگرام سال انتشار کتاب‌ها



شکل ۲ نمودار توزیع کتاب‌ها در زبان‌های مختلف

این دیتاست شامل چندین فایل مختلف است که به بررسی آن‌ها می‌پردازیم:

- books.csv: این جدول شامل تمامی اطلاعات و متا دیتاهای مربوط به کتاب‌ها از جمله اطلاعاتی مانند عنوان کتاب، سال انتشار، نویسنده، شناسه کتاب، زبان کتاب، میانگین امتیازهای کاربران و ... است.

```
[15] books.head(5)
```

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publication_year	original_title	title	langua
0	1	2767052	2767052	2792775	272	439023483	9.780439e+12	Suzanne Collins	2008.0	The Hunger Games	The Hunger Games (The Hunger Games, #1)	
1	2	3	3	4640799	491	439554934	9.780440e+12	J.K. Rowling, Mary GrandPré	1997.0	Harry Potter and the Philosopher's Stone	Harry Potter and the Sorcerer's Stone (Harry P...	
2	3	41865	41865	3212258	226	316015849	9.780316e+12	Stephenie Meyer	2005.0	Twilight	Twilight (Twilight, #1)	
3	4	2657	2657	3275794	487	61120081	9.780061e+12	Harper Lee	1960.0	To Kill a Mockingbird	To Kill a Mockingbird	
4	5	4671	4671	245494	1356	743273567	9.780743e+12	F. Scott Fitzgerald	1925.0	The Great Gatsby	The Great Gatsby	

شکل ۳ جدول کتاب‌ها

- ratings.csv: در این فایل اطلاعات مربوط به تعدادی از نظرات کاربران وجود دارد. بصورت خاص در این پروژه نیازی به اطلاعات این فایل نداریم.
- tags.csv: در این جدول اطلاعات مربوط به تگ‌هایی است که کاربران به هر کتاب‌ها مقداردهی کرده‌اند. این تگ‌ها شامل اطلاعاتی مانند ژانر و اطلاعات دیگر است. رابطه بین جدول تگ‌ها و کتاب‌ها یک رابطه many to many است یعنی هر کتاب ممکن است چندین تگ داشته باشد و همچنین هر تگ ممکن است به چندین کتاب متصل باشد.

```
[20] tags.loc[5000:5005]
```

1 to 6 of 6 entries Filter

index	tag_id	tag_name
5000	5000	book-movie
5001	5001	book-movie-guide
5002	5002	book-of-ember
5003	5003	book-of-mormon
5004	5004	book-of-the-art
5005	5005	book-of-the-month

Show 25 per page

شکل ۴ جدول تگ‌ها

- book_tags.csv: این جدول شامل اطلاعات مربوط به Join کردن دو جدول کتاب‌ها و تگ‌ها است
- to_read.csv: این جدول شامل کتاب‌هایی است که کاربران برای خواندن علامت‌گذاری کردند. با توجه حوزه انجام این پروژه نیازی به استفاده از اطلاعات این فایل نداریم.

۲-۱- Data Cleaning

ابتدا نیاز است تا داده‌ها را برای ادامه ساخت سیستم آماده‌سازی کنیم. بنابرین نیاز است تا ستون‌های بی‌کاربرد در جدول books را حذف کنیم. ستون‌های کاربردی در این قسمت عبارتند از آیدی کتاب، نام نویسندگان، عنوان کتاب، سال انتشار، زبان کتاب، میانگین امتیاز و تعداد نظرات که بعدا برای ارائه پیشنهادات مناسب هستند.

پس از حذف ستون‌های بی‌کاربرد در قدم بعد جدول کتاب را با جدول تگ‌ها ترکیب (Join) می‌کنیم. خروجی این بخش جدولی است که به ازای هر کتاب یک ردیف در جدول داریم که به یک تگ متصل است.

```
[23] books.head()
```

1 to 5 of 5 entries Filter

index	goodreads_book_id	authors	original_publication_year	original_title	language_code	average_rating	ratings_count	tag_name
0	2767052	Suzanne Collins	2008.0	The Hunger Games	eng	4.34	4780653	to-read
1	2767052	Suzanne Collins	2008.0	The Hunger Games	eng	4.34	4780653	fantasy
2	2767052	Suzanne Collins	2008.0	The Hunger Games	eng	4.34	4780653	favorites
3	2767052	Suzanne Collins	2008.0	The Hunger Games	eng	4.34	4780653	currently-reading
4	2767052	Suzanne Collins	2008.0	The Hunger Games	eng	4.34	4780653	young-adult

شکل ۵ خروجی اتصال جدول کتاب‌ها و تگ‌ها

۲- پیاده‌سازی سیستم

پیاده‌سازی سیستم ما بصورت یک فانکشن پایتون است که ورودی‌های آن، محدودیت‌هایی است که کاربر برای انتخاب داده گذاشته است. در این قسمت ما از قابلیت‌های کتابخانه pandas برای اعمال کوئری‌های محدود کننده روی تمامی کتاب‌ها کردیم. اما برای نمایش خروجی، باید یک معیار خاص برای نمایش ترتیبی کتاب‌ها به کاربر داشته باشیم.

معیار رتبه‌بندی

ساده‌ترین المان قابل استفاده به عنوان معیار در این بخش، استفاده از میانگین امتیازات کاربران به هر کتاب است. اما اساسی این معیار این است که محبوبیت آن کتاب را در نظر نمی‌گیرد. مثلا کتابی که توسط نفر امتیاز ۴,۵ گرفته است در این معیار بالاتر از کتابی که توسط ۱۰۰۰۰۰ نفر امتیاز ۴ گرفته است.

این اتفاق از نظر ما نامطلوب است، به این دلیل که هرچه تعداد رای دهندگان بیشتر باشد، امتیاز رای بیشتر بازتاب کننده کیفیت واقعی کتاب است. اما با تعداد کم رای نمی‌توان بصورت قطعی راجع به کیفیت آن کتاب قضاوت کرد. بنابرین ما به معیاری نیاز داریم که علاوه بر امتیاز میانگین، تعداد رای دهندگان را هم در رتبه‌بندی اثر دهد.

$$Weighted\ Rating = \left(\frac{v}{v+m} \times R\right) + \left(\frac{m}{v+m} \times C\right)$$

فرمول بالا، فرمول استفاده شده برای محاسبه وزن‌دار امتیاز برای هر کتاب است که به تابع محاسبه امتیاز میانگین IMDb معروف است. در این فرمول v همان تعداد نظرات کاربران برای هر کتاب، m حداقل تعداد نظرات برای نمایش در لیست پیشنهادات، R میانگین امتیاز کتاب مورد نظر و C میانگین امتیازات تمامی کتاب‌های داخل دیتاست است.

برای انتخاب m (حداقل تعداد نظرات برای نمایش در لیست پیشنهادات)، روشی که ما استفاده کردیم، انتخاب آن بصورت نسبی از داخل دیتاست بود. به اینصورت که با تنظیم پارامتر percentile روی عدد ۰٫۸، m برابر با تعداد نظرات کتابی می‌شود که از ۸۰ درصد کل کتاب‌ها بیشتر نظر دارد. با افزایش پارامتر percentile، خروجی محدود به کتاب‌هایی می‌شود که محبوبیت بیشتری دارند.

۳- نمونه خروجی سیستم

همانطور که در خروجی ساده شکل ۶ دیده می‌شود، امتیاز بالاتر و تعداد رای بیشتر باعث بالاتر بودن در لیست خروجی کتاب‌ها شده است. از طرفی با مشاهده خروجی‌های شکل ۷ و شکل ۸ می‌توان اهمیت پارامتر کمترین تعداد نظرات در خروجی سیستم را مشاهده کرد. هنگامی که این پارامتر روی عدد ۰٫۸ تنظیم شده است، کتاب‌هایی که تعداد رای بیشتری دارند، در ترتیب ما در رتبه بالاتری قرار گرفته‌اند.

Try Recommender System

genre: fantasy

from_year: 0

to_year: 2017

language: French

percentile: 0.8

list_size: 5

1 to 2 of 2 entries Filter

index	goodreads_book_id	authors	original_publication_year	original_title	language_code	average_rating	ratings_count	tag_name	score
54501	10629	Stephen King, Marie Milpois	1983.0	Christine	fre	3.72	151160	fantasy	3.7985689088796546
76701	11570	Stephen King, William Olivier Desmond	2001.0	Dreamcatcher	fre	3.59	115855	fantasy	3.735836611725015

شکل ۶ نمونه خروجی اول

Try Recommender System

genre: horror

from_year: 1998

to_year: 2017

language: English

percentile: 0.5

list_size: 5

1 to 5 of 5 entries Filter

index	goodreads_book_id	authors	original_publication_year	original_title	language_code	average_rating	ratings_count	tag_name	score
146773	6585201	Jim Butcher	2010.0	Changes	eng	4.54	66402	horror	4.395095965673409
150870	12216302	Jim Butcher	2012.0	Cold Days	en-US	4.51	57779	horror	4.358608167655082
55242	13615	Tsugumi Ohba, Takeshi Obata	2004.0	デスノート #1 (Desu Nōto) Taikutsu (退屈)	eng	4.42	139501	horror	4.3574884243528125
188168	19486421	Jim Butcher	2014.0	Skin Game	eng	4.55	41512	horror	4.346357492873004
123973	17683	Jim Butcher	2005.0	Dead Beat	eng	4.43	78123	horror	4.3275590897193075

شکل ۷ نمونه خروجی دوم

Try Recommender System

genre: horror

from_year: 1998

to_year: 2017

language: English

percentile: 0.8

list_size: 5

1 to 5 of 5 entries Filter

index	goodreads_book_id	authors	original_publication_year	original_title	language_code	average_rating	ratings_count	tag_name	score
55242	13615	Tsugumi Ohba, Takeshi Obata	2004.0	デスノート #1 (Desu Nōto) Taikutsu (退屈)	eng	4.42	139501	horror	4.295014023419884
146773	6585201	Jim Butcher	2010.0	Changes	eng	4.54	66402	horror	4.28561402259883
123973	17683	Jim Butcher	2005.0	Dead Beat	eng	4.43	78123	horror	4.24434661695362
109547	13154150	Hajime Isayama, Sheldon Drzka	2010.0	進撃の巨人 1	eng	4.42	82565	horror	4.2443396549618235
36179	2282133	Richelle Mead	2008.0	Frostbite	eng	4.3	256745	horror	4.243663004183874

شکل ۸ نمونه خروجی سوم

منابع و مراجع

[1] "zygmuntz / goodbooks-10k," 1 October 2018. [Online]. Available: https://github.com/zygmuntz/goodbooks-10k.

[2] "Weighted Average Ratings," [Online]. Available: https://help.imdb.com/article/imdb/track-movies-tv/weighted-average-ratings/GWT2DSBYVT2F25SK#.

[3] "Understanding the IMDb weighted rating function for usage on my own website," [Online]. Available: https://math.stackexchange.com/questions/169032/understanding-the-imdb-weighted-rating-function-for-usage-on-my-own-website.

[4] "pandas," [Online]. Available: https://pandas.pydata.org.

[5] "Plotting with matplotlib," [Online]. Available: https://pandas.pydata.org/pandas-docs/version/0.13/visualization.html.