فاز اول پروژه یادگیری ماشین

امیر سامان کریمی شرق، مرضیه کامکار، مهیار آلبالان

۱۰ فروردین ۱۴۰۳

۱ مقدمه

در سال‌های اخیر ما شاهد رشد مخاطبین و تولیدکنندگان محتوای پادکست فارسی بوده‌ایم به گونه‌ای که در حال حاضر بیش از ۹ هزار عنوان پادکست فارسی در پلتفرم‌های مختلف ثبت شده‌اند که در حدود ۸۰ درصد آن‌ها نیز فعال می‌باشد. در میان پلتفرم‌های پادگیر، اپلیکیشن castbox با اقبال بیشتری مواجه بوده به طوری که سهم آن از میزبانی پادکست‌های فارسی ۸۰ درصد است. در این پروژه ما قصد داریم با تمرکز بر پادکست‌های فارسی زبان منتشر شده در این اپلیکیشن، سیستم پیشنهاددهنده‌ای ارائه دهیم که با دانستن اطلاعات یک کاربر، پادکست‌هایی که در زمان‌های بعد دنبال می‌کند را پیش‌بینی کرده و به کاربر پیشنهاد دهد.

۲ داده

جمع آوری داده توسط گروه ما صورت نگرفته و از داده‌ی مربوط به پروژه درس علم شبکه، که در ترم پیش ارائه شد، استفاده کردیم. این داده توسط آقای سینا معمر از طریق API اپلیکیشن castbox به دست آمده است. داده‌ی خامی که به دست ما رسید شامل یک فایل all_podcasts_with_channel.json و یک پوشه followers است که هر دو در گوگل درایو آپلود شده‌اند.

در اپلیکیشن castbox هر کاربر به طور پیش فرض می‌تواند پادکستر باشد و ما کاربرانی را که حداقل یک پادکست به زبان فارسی داشته باشند را پادکستر و دیگر کاربران را شنونده یا همان کاربر عادی در نظر گرفتیم. در پوشه followers اطلاعات مربوط به دنبال کنندگان یک پادکستر آمده است که با استفاده از این اطلاعات ما دیتافریم followers_df را ساختیم که در یک ستون آن آیدی پادکستر و در ستونی دیگر آیدی کاربری که پادکستر را دنبال کرده و در آخرین ستون زمان دنبال کردن پادکستر آورده شده است. هنگام ساختن این دیتافریم ما پادکسترهایی که دنبال کننده‌ای ندارند را حذف کردیم و بدین صورت در کل ۳۳۲۵ پادکستر متمایز و ۴۳۱۶۴۴۶ کاربر یکتا داریم.

در فایل all_podcasts_with_channel اطلاعات مربوطه به پادکست‌های مختلف وجود دارد که در مجموع شامل ۱۱۶۹۰ پادکست است. نکته‌ی قابل توجه اینجاست که ما در مورد اینکه یک کاربر چه پادکستی را دنبال می‌کند نداریم و تنها ارتباط بین کاربران و پادکسترها برایمان مشخص است. یک پادکستر ممکن است چندین پادکست ارائه کند. با توجه به اینکه یک کاربر وقتی یکی از پادکست‌های پادکستری را دنبال می‌کند، دیگر پادکست‌های او نیز به احتمال زیاد در حوزه‌ی علاقه‌مندی‌اش است خطایی که این موضوع در پاسخ به این مسئله ایجاد می‌کند از نظرمان قابل صرف نظر کردن می‌آید.
در این بخش ما دو دیتافریم دیگر ساختیم که یکی از آن‌ها شامل اطلاعات تمامی پادکست‌ها نظیر آیدی پادکست، آیدی پادکستر، عنوان پادکست، تعداد دنبال کننده و ... است و دیتا فریم دیگر مربوط به پادکسترهایی است که در دیتافریم followers_df وجود دارند. اولین دیتافریم به صورت all_podcasts_df.csv و دومی به صورت my_podcaster_df.csv ذخیره شده‌اند. کد مربوط به این بخش در پوشه Data در گیت‌هاب قرار دارد.

۳ تمیز کردن داده

همانطور که در بالا اشاره شد، ما تنها می‌دانیم یک کاربر چه پادکستری را دنبال کرده است در نتیجه تنها اطلاعات پادکست‌هایی برایمان اهمیت دارد که پادکستر پشت آن‌ها در دیتافریم my_podcaster_df وجود دارد.
پس در گام اول از دیتافریم all_podcasts_df تنها پادکست‌هایی را نگه داشتیم که پادکسترهای مد نظرمان آن‌ها را تولید کرده‌اند. با گذاشتن این فیلتر تعداد پادکست‌ها از ۱۱۶۹۰ به ۴۳۱۵ پادکست رسید. اما با دقت در نتیجه متوجه شدیم بعضی پادکست‌ها چندین بار تکرار شده‌اند. یعنی چند پادکست با عنوان، تعداد قسمت‌ها و پادکستر یکسان، با cid (cid در واقع آیدی پادکست است) و اطلاعاتی متفاوت تکرار شده‌اند.
برای اینکه در بین پادکست‌های تکراری کدام را نگه داریم چند روش را امتحان کردیم اما در نهایت به این نتیجه رسیدیم که در میان پادکست‌های تکراری، اطلاعات مربوط به پادکستی که بیشترین دنبال کننده را دارد معتبرتر است. پس در نهایت تصمیم گرفتیم از میان پادکست‌هایی که عنوان و پادکستر مشابه دارند پادکستی که بیشترین دنبال کننده را دارد نگه داریم و بدین صورت از ۴۳۱۵ پادکست به ۳۸۸۰ پادکست رسیدیم.
این فایل تحت عنوان my_podcasts_df.csv ذخیره شده و هم در گیت‌هاب و هم در گوگل درایو بارگذاری شده است. کد مربوط به این بخش در پوشه Data_Cleaning در گیت‌هاب قرار داده شده است.

۴ پیش پردازش داده

در این بخش ما نیاز داریم مسئله را واضح‌تر بیان کنیم تا به شناخت بهتری در مورد X و Y در این مسئله برسیم. سوالی که ما تلاش بر پاسخ دادن آن داریم این است که با داشتن اطلاعات یک کاربر احتمال اینکه در بازه‌ی زمانی مشخصی، به عنوان مثال یک سال یا شش ماه بعد، پادکست‌های دیگر را دنبال کند چقدر است. پس در نهایت ما لیستی از احتمال دنبال کردن پادکست‌ها توسط یک کاربر مشخص را به عنوان Y پیش‌بینی شده توسط مدل خواهیم داشت.
سوالی که در ابتدا باید به آن پاسخ بدهیم این است که چه اطلاعاتی در مورد یک کاربر به منظور پاسخ دادن به مسئله‌مان سودمند است؟ یا به طور دقیقتر X را چطور تعریف کنیم؟
اولین چیزی که برایمان اهمیت دارد این است که کاربر به چه حوزه‌هایی علاقه‌مند است. مثلا ممکن است یک کاربر بیشتر پادکست‌هایی در زمینه اقتصادی یا جامعه‌شناسی و یا ... را دنبال کند. در داده‌ی مربوط به all_podcasts_with_channel برای هر پادکست ما اطلاعاتی به عنوان categories داریم. که مقادیر مربوط به این لیست می‌تواند به سوالمان در مورد سلیقه کاربر پاسخ دهد.
در گام اول تعداد کل گتکوری‌های متمایز و تعداد تکرارشان را به دست آوردیم. در بین پادکست‌های مد نظرمان در کل ۱۱۶ کتگوری متمایز وجود دارد. برای اعتبارسنجی کتگوری‌ها، پادکست‌هایی با کتگوری مشخص را نشان دادیم و مشاهده کردیم همخوانی نسبتا خوبی با یکدیگر دارند. اما چیزی که در اصل به دنبال آن هستیم این است که یک کاربر چه کتگوری‌هایی را دنبال می‌کند. تصمیم گرفتیم که پنج کتگوری پرتکرار که توسط یک کاربر دنبال می‌شود را نگه داریم. با بررسی تعداد تکرار کتگوری‌ها، پنج کتگوری پر تکرار در میان تمامی پادکست‌ها را از داده‌ی مربوط به کاربر حذف کردیم زیرا تعداد تکرار آن‌ها به قدری بالاست که نمی‌تواند دسته‌بندی دقیقی را ارائه کند و اکثر پادکست‌ها زیر مجموعه آن‌ها قرار می‌گیرند. پس در نهایت می‌خواهیم برای یک کاربر پنج کتگوری پر تکراری که دنبال می‌کند به جز پنج کتگوری اول پر تکرار در پادکست‌ها را گزارش کنیم.
از طریق followes_df ما می‌دانیم هر کاربر کدام پادکسترها را دنبال می‌کند. در ابتدا تصمیم گرفتیم کاربرانی که کمتر از پنج پادکستر دنبال می‌کنند را از دیتافریم حذف کنیم. دلیل این کار در واقع این بود که برای چنین کاربرانی احتمال آنکه پنج کتگوری متمایز غیر از کتگوری‌های پر تکرار وجود داشته باشد کم است. با این کار تعداد کاربران به حدود یک ملیون و هفتصد هزار کاهش پیدا کرد. سپس از طریق تابعی با داشتن user_id تمامی کتگوری‌هایی که یک کاربر دنبال می‌کند، کتگوری‌هایی که غیر از ۵ کتگوری پر تکرار دنبال می‌کند و تعداد هر کدام را به دست آوردیم.
با انجام این کار برای تمامی کاربران، همچنان حدود چهار هزار کاربر وجود داشت که کمتر از ۵ کتگوری را دنبال می‌کنند، البته با صرف نظر کردن از کتگوری‌های پر تکرار که اطلاعات زیادی به مسئله اضافه نمی‌کنند. با توجه به اینکه تعداد این کاربران در مقایسه با تعداد کل، یعنی حدود یک میلیون و هفتصد کاربر، ناچیز است تصمیم بر حذف این کاربران گرفتیم و در نهایت

دیتا فریمی ساختیم که یک ستون آن آیدی کاربر و پنج ستون دیگر به ترتیب پنج کتگوری اصلی‌ای است که توسط کاربر دنبال می‌شود. این فایل تحت عنوان user_categories_df ذخیره شده است و در گوگل درایو قرار دارد. و در نهایت با توجه به اینکه تعداد زیادی از کاربران را حذف کردیم، از دیتافریم followers، دیتافریمی تنها شامل کاربرانی که نگه داشتیم ساختیم که تحت عنوان my_followers_df در گوگل درایو گذاشته شده است. تعداد کاربران از حدود چهار میلیون به ۱۶۸۶۶۹۳ رسید. کد مربوط به این بخش در پوشه Data_Preprocessing و در فایل categoty.ipynb قابل مشاهده است.

کاری که تا اینجا انجام دادیم در نهایت منجر به پیشنهاد بر اساس محتوا می‌شود. اما چیزی که از اول علاقه‌مند به پیاده‌سازی آن بودیم این بود که چطور بر اساس مشابهت کاربران این پیشنهاد صورت بگیرد. اولین چیزی که به ذهنمان رسید با توجه به مقالاتی که دیدیم این بود که ماتریسی داشته باشیم که سطرهای آن کاربران و ستون‌های آن پادکسترها باشد و از طریق Cosine similarity کاربران مشابه را پیدا کرده و بر اساس این مشابهت پادکست جدید پیشنهاد دهیم. اما حجم این ماتریس زیاد است و نسبت به حجم اشغال شده اطلاعات کمی را در خود ذخیره می‌کند. به همین دلیل با مشورت‌هایی که انجام دادیم تصمیم گرفتیم از طریق شبکه‌ی اصلی که یک شبکه‌ی دو بخشی است، به طوری که یک بخش آن پادکسترها و بخش دیگر آن کاربران هستند، شبکه‌ی جدیدی بسازیم که بین پادکسترها یال وجود دارد. برای اینکه تشخیص دهیم بین دو پادکستر یال وجود دارد یا نه قصد داشتیم از تعداد دنبال کنندگان مشترک دو پادکستر استفاده کنیم. برای این کار به حد بالا و پایینی برای دنبال کنندگان یک پادکستر نیاز داریم. حد بالا به این دلیل که بدون در نظر گرفتن آن پادکسترهایی که تعداد زیادی دنبال کننده دارند همسایه تمامی پادکسترها خواهند شد و حد پایین برای مشخص کردن کمینه تعداد دنبال‌کننده مشترک دو پادکستر به منظور پیوند زدن بینشان.
ما کد مربوط به هر دو رویکرد را زدیم. یعنی هم می‌توانیم ماتریسی داشته باشیم که سطر و ستون‌های آن کاربران و پادکسترها باشند و هم دنبال کنندگان مشترک دو پادکستر را به دست آوردیم. اما برای مشخص کردن شرط پیوند زدن بین دو پادکستر نتوانستیم به نتیجه برسیم. با توجه به اینکه تابع توزیع درجه راس در این شبکه توانی است، اگر حد پایین را کمی بیشتر در نظر بگیریم تعداد زیادی از پادکسترها حذف می‌شوند و اگر حد پایین را کم در نظر بگیریم پادکسترها با درجه راس بیشتر همسایه تمامی پادکسترها خواهند شد. کد مربوط به این بخش در گیت‌هاب و در پوشه Data_Preprocessing و در فایل podcasts network.ipynb قرار دارد. همچنین ماتریس common_followers که در حقیقت خانه i و j آن بیانگر این است که پادکست i و پادکست j چند دنبال کننده مشترک دارند و همچنین آرایه podcaster_ids_array که درایه iام آن آیدی پادکستر iام است را ذخیره کردیم تا بعدا اگر برای شرط پیوند زدن بینشان به نتیجه رسیدیم روی این ماتریس شرط را اعمال کنیم.