پیشنهاد محصول در کنار متن مرتبط

فاطمه نجفی

۹ اسفند ۱۳۹۹

مقدمه

قراردادن آگهی مناسب در زمان مناسب در برابر کاربر مناسب، هدف کلیدی تبلیغات مربوط به متن است. در این گونه تبلیغات، آگهی در صفحاتی با موضوع مرتبط با آن قرار داده می‌شود؛ برای مثال تبلیغ مربوط به یک مدل خاص از عطر می‌تواند در کنار متنی مربوط به انواع مختلف عطر قرار بگیرد.

در ادامه فرایند طراحی سیستمی مناسب این کار را بررسی می‌کنیم. در این کار حالتی درنظر گرفته‌شده‌است که اطلاعاتی از کاربر نداریم و فقط باید با شباهت بین کالا و متن تصمیم بگیریم.

جمع‌آوری داده

برای جمع‌آوری داده، فقط از سایت دیجی‌کالا استفاده کردم. دلایل این کار:

۱. مجموعه داده یک‌پارچگی داشته باشد. (حداقل در این سطح)

۲. محصول‌های سایت‌هایی مثل تخفیفان به شهر کاربری که متن را می‌بیند بستگی دارد و در این مرحله اطلاعات کاربر را نداریم. هرچند حالتی وجود دارد که کاربر مقاله‌ای درمورد رستوران‌های یک شهر بخواند و قصد سفر به آن شهر را داشته باشد، نشان دادن تخفیف از یک رستوران در آن شهر می‌تواند مفید باشد، این کار پردازش جدایی می‌طلبد که در این سطح، زمان کافی نبود.

۳. زمان محدود برای انجام این پروژه.

برای جمع‌آوری اطلاعات مربوط به کالاهای متفاوت در دیجی‌کالا از کتاب‌خانه beautifulsoup4 استفاده کردم. متاسفانه سرعت پردازش این کتاب‌خانه بسیار پایین بود و چون به کمک Google colaboratory این برنامه را نوشتم و به دلیل محدودیت زمانی اجرا، مجبور بودم اطلاعات را در بازه‌های کوچک جمع و جدا ذخیره کنم تا در نهایت با چسباندن آن‌ها به هم به مجموعه‌داده‌ی خواسته شده برسم. این اطلاعات در نهایت در فایل digikala_product.csv ذخیره شده‌است. جمع‌آوری این مجموعه داده بیش از ۴۲ ساعت زمان برد و در نهایت اطلاعات ۱۸۸۲ محصول از دیجی‌کالا را به‌دست آوردم.

اطلاعاتی که از هر محصول ذخیره کردم:

- شماره‌ی محصول؛ از این شماره در لینک کالا استفاده شده‌است. ذخیره‌سازی این مورد مهم است چون برای دسترسی به لینک کالا به آن نیاز داریم. شماره‌ی محصول با عنوان dkp ذخیره می‌شود.

- نام کالا که با عنوان product name ذخیره می‌شود.
- نقد و بررسی محصول که با عنوان description ذخیره می‌شود. البته تعداد زیادی از محصولات نقد و بررسی ندارند.
- هر کالا ممکن است در چند دسته از محصولات قرار گیرد. تمام دسته‌هایی که شامل کالا می‌شوند را در لیستی با نام classes ذخیره می‌کنیم.
- در بین صفحاتی که برای معرفی محصول در دیجی‌کلا وجود داد، صفحاتی هستند که آن کالا دیگر در انبار موجود نیست. موجود بودن یا نبودن کالا در stocked نگه‌داری می‌شود. فایده‌ی نگهداری این است که اگر کالا موجود باشد همان کالا را می‌توانیم پیشنهاد دهیم و اگر موجود نباشد، از کالاهای دیگری که در آن دسته قرار دارند استفاده می‌کنیم.

سایر اطلاعات هر محصول که می‌توانست استفاده شود:

- امتیاز؛ هنگام پیشنهاد محصول، اگر چند محصول با توجه به معیاری که تعریف می‌شود، عدد برابری برای انتخاب شدن داشته‌باشند، می‌توان کالایی با امتیاز بالاتر را پیشنهاد داد.
- قیمت؛ مخصوصا اگر کالا تخفیف نیز داشته باشد، هنگام نمایش می‌تواند وسوسه برانگیز باشد.
- مشخصات فنی محصول؛ با توجه به این که مشخصات فنی محصولات در دسته‌های مختلف با یکدیگر متفاوتند، در این سطح از پروژه تصمیم گرفتم وارد کار نشوند. در حالی‌که ممکن است با درنظر گرفتن مشخصات فنی دقت مدل بالا برود.

تجزیه و تحلیل داده‌ها [1]

اطلاعاتی از مجموعه‌داده:

- مجموعه‌داده جمع‌آوری شده شامل ۱۸۸۲ محصول از سایت دیجی‌کالا است.
- با بررسی نام دسته‌های آن متوجه می‌شویم که از تمام ۹ دسته اصلی کالاهای آن، کالا داریم.
- اگر دسته‌بندی فرعی (یک مرحله بعد از اصلی) را درنظر بگیریم، کالاها در ۵۳ دسته‌ی متفاوت وجود دارند.
- از ۱۸۸۲ محصول، ۹۵۲ محصول در انبارها هستند و ۹۳۰ محصول موجود نیستند.
- از ۱۸۸۲ محصول، ۸۲۶ محصول متنی تحت عنوان نقد و بررسی دارند و سایر محصولات این متن را ندارند. به همین دلیل تصمیم گرفتم از این ویژگی استفاده نکنم.

[1] Exploratory Data Analysis(EDA)

بررسی چند روش معمول برای یک سیستم پیشنهاددهنده

- بر اساس رفتار کاربر یا کاربران. در این دسته از الگوریتم‌ها، محصول بر اساس سابقه‌ی رفتاری کاربران، پیشنهادات داده می‌شوند. مانند مدل Collaborative filtering ، Content-based filtering و غیره.

- بدون توجه به رفتار کاربر. برای کار خواسته شده، الگوریتم‌های دسته‌ی قبل پاسخ‌گو نیستند. تنها اطلاعاتی که داریم، تعدادی محصول است و باید برای هر متن تشخیص دهیم کدام محصول یا محصولات برای تبلیغ کنار متن مناسب است. یکی از راه‌ها برای انجام این کار محاسبه‌ی شباهت بین متن و محصولات و در نهایت انتخاب محصولات با بیشترین شباهت است.

ارزیابی

- user studies . در مقیاس‌های کوچک می‌توان از این روش استفاده کرد. در این حالت افرادی به صورت دستی پیشنهادها را بررسی می‌کنند.

- online evaluations یا A/B tests . در این روش باید به ازای یک متن، تبلیغ‌های مختلف را نمایش داد و با توجه به بازخوردی که از کاربران گرفته می‌شود، می‌توانیم میزان مناسب بودن پیشنهاد را متوجه شویم.

- offline evaluations . این نوع ارزیابی به کمک تعریف متریک انجام می‌شود. که از انواع آن می‌توان Mean Squered Error (MSE)، Root Mean Squared Error (RMSE) و Normalized Discounted Cumulative Gain (nDCG) را نام برد. البته باید درنظر بگیریم که ما حالت ایده‌آل برای هر پیشنهاد را لازم داریم.

کارهایی که کردم

جمع‌آوری داده

به کمک توابعی که در فایل crawler_functions.py قرار دارند و با اجرای ده‌ها بار تابع run_crawler تعدادی فایل pickle ساختم و با چسباندن آن‌ها به هم، آن را به Pandas DataFrame تبدیل کردم. در نهایت این دیتافریم را در فایل csv ذخیره کردم. متوجه شدم گاهی فایل csv راحت خوانده نمی‌شود؛ پس برای راحتی کار و در مواقع اضطراری، داده را در فایل pickle هم ذخیره کردم که استفاده از آن نیز آسان است.

نمایش داده

برای کار با اطلاعات محصول، لازم بود متن‌ها و کلمات به صورت برداری و با عدد نمایش داده شوند. ترجیح دادم برای این کار از روش Term Frequency استفاده کنم. این روش نسبت به Bag of Words بهتر است چون تعداد را نیز در نظر می‌گیرد و از طرفی پیچیدگی‌های tf-idf را ندارد. استفاده از روش‌های تعبیه کلمات نیز مفید بود ولی با توجه به وقت کم ترجیح دادم از روش ساده‌تری استفاده کنم. کار دیگری که می‌توانستم انجام دهم آموزش مدل BERT به کمک مجموعه‌داده‌های دریافت شده بود اما باز هم مشکل کمبود زمان داشتم.

الگوریتم

چون کار خواسته شده در دسته‌ی دوم سیستم‌های پیشنهاددهنده قرار می‌گیرد، از معیارهای اندازه‌گیری شباهت استفاده کردم. به این منظور از دو معیار شباهت Jaccard و Cosine استفاده کردم. روش کار به این صورت است که برای هر متن دریافتی، شباهت متن با تمام محصولات محاسبه می‌شود و ۱۰ محصول با بیشترین شباهت به عنوان خروجی داده می‌شود.

معیار ارزیابی

برای ارزیابی معیار nDCG را انتخاب کردم اما مشکل بزرگ آن نبود مجموعه‌داده مناسب برای آزمایش است. که جمع‌آوری آن زمان زیادی می‌طلبد. در ادامه لازم است متن‌هایی را درنظر بگیریم و با توجه به موضوع متن، از دسته‌ی مرتبط، محصولات مرتبط را به ترتیب مشخص کنیم.

ایده‌های دیگر

- استفاده از متن نقد محصول به عنوان داده‌ی آموزشی .
- استفاده از مجموعه تعبیه کلمات از پیش آماده شده.
- آموزش مدل BERT روی دیتاست‌های متنی داده شده .