۱۴۰۰/۱۰/۲۲

# بازیابی هوشمند اطلاعات

## پروژه پایانی


مینا فریدی

۸۱۰۱۰۰۴۳۰

آریا وارسته‌نژاد

۸۱۰۱۰۰۴۹۸


## گام اول: پیش‌پردازش

ابتدا مراحل پیش پردازش را به این صورت انجام می‌دهیم که نخست متن سوالات و پاسخ‌های ورودی را با کمک کتابخانه nltk توکن‌توکن می‌کنیم. بعد از حذف استاپ وردها، کوچک‌سازی حروف و استم کردن با PorterStemmer را بازهم با استفاده از کتابخانه nltk انجام می‌دهیم در ادامه بخش‌های مورد نیاز از داده‌ها یعنی:

- RELQ_ID
- RelQBody
- RELC_ID
- RELC_RELEVANCE2RELQ
- RelCText

را از فایل‌های train_data.xml، dev_data.xml و test_data.xml جدا می‌کنیم. برای صفت RelCText به ازای مقدار Good عدد ۲، PotentiallyUseful عدد ۱ و به ازای Bad عدد ۰ در نظر می‌گیریم و به صورت مناسب به ترتیب در فایل‌های pa_pp_train_data.csv، pa_pp_dev_data.csv و pa_pp_test_data.csv ذخیره می‌کنیم.

| | | | | | |
|---|---|---|---|---|---|
| ... | ... | ... | ... | ... | ... |
| 1 | Q268_R16 | hi ti ql what bank use are use bank affili hom... | Q268_R16_C1 | bank use us talk taken credit card loan know | 0 |
| 2 | Q268_R16 | hi ti ql what bank use are use bank affili hom... | Q268_R16_C2 | in qatar like say best std | 0 |
| 3 | Q268_R16 | hi ti ql what bank use are use bank affili hom... | Q268_R16_C3 | i surpris see feedback qatar bank is seriou pr... | 0 |
| 4 | Q268_R16 | hi ti ql what bank use are use bank affili hom... | Q268_R16_C4 | well arman noth wrong bank i feel par uae bank... | 2 |
| 2436 | Q317_R23 | i experienc time water boil certain smell like... | Q317_R23_C6 | appar water hard could damag kidney but i hear... | 2 |
| 2437 | Q317_R23 | i experienc time water boil certain smell like... | Q317_R23_C7 | i ur cat so listen | 0 |
| 2438 | Q317_R23 | i experienc time water boil certain smell like... | Q317_R23_C8 | so use bottl water brush rins mouth wash dish ... | 0 |
| 2439 | Q317_R23 | i experienc time water boil certain smell like... | Q317_R23_C9 | duhh cours i wo drink water i avoid i paranoid... | 2 |
| 2440 | Q317_R23 | i experienc time water boil certain smell like... | Q317_R23_C10 | tri take disinfect glass drink stop tast it da... | 1 |

## گام دوم و سوم: استخراج ویژگی و بازیابی پاسخ به کمک شبکه پرسپترون چندلایه

پس از انجام مراحل پیش‌پردازش داده‌ها را برای مرحله استخراج ویژگی آماده‌سازی می‌کنیم. چون در صورت تمرین دسته‌بندی باینری خواسته شده که در آن کلاس‌های مورد تقاضا True به ازای پاسخ‌های Good و False به ازای پاسخ‌های Bad و PotentiallyUseful است، در این گام مقادیر label_array را برای ورودی ۲ به ۱ و برای ورودی‌های ۱ و ۰ را به ۰ تبدیل می‌کنیم.

اغلب روش‌های یادگیری ماشین بر روی داده‌های عددی قابل اجرا هستند و برای استفاده و اجرای آن‌ها روی داده‌های متنی نیاز به تبدیل متون به مجموعه اعداد است. پس هدف رویکردهای مختلف تبدیل متن به بردارهای

عددی، استخراج و انتخاب مجموعه‌ای از ویژگی‌های مناسب از متون زبان طبیعی است. لذا انتخاب ویژگی مرحله‌ای بسیار مهم در فعالیت ما به شمار می‌رود، زیرا در این مرحله باید واژه‌های کلیدی انتخاب شوند تا به عنوان بهترین نمایش‌دهنده برای سند متنی مورد استفاده قرار بگیرند. اگر تعداد واژه‌های کلیدی انتخاب شده کم باشد صحت و کارایی سیستم تحت تاثیر قرار می‌گیرد و کاهش می‌یابد و در مقابل اگر تعداد واژه‌های کلیدی انتخاب شده زیاد باشد باعث کاهش کارایی سیستم در بعد زمان خواهد شد و سرعت آموزش در فاز آموزش پایین می‌آید.

از جمله روش‌های مطرح در استخراج ویژگی برای داده‌های متنی، روش‌های زیر هستند:

۱) روش فرکانس سند
۲) روش وزن‌دهی منطقی
۳) روش فرکانس کلمه
۴) روش فرکانس معکوس سند
۵) روش فرکانس کلمه-فرکانس معکوس سند
۶) روش بهره اطلاعاتی
۷) روش اطلاعات متقابل
۸) BM25
۹) روش CHI
۱۰) روش ضریب همبستگی

در انجام این پروژه به روش‌های ۱ الی ۸ را به دلیل آشنایی با آن‌ها در طول دوره کلاس پس از آموزش دادن با دادگان آموزشی، بر روی مجموعه dev_data مورد تحلیل و ارزیابی قرار داده و سعی کردیم با تنظیم پارامترها نتایج را بهبود بدهیم. در این میان به دلیل کسب نتایج بهتر و نیز بهینه بودن زمان آموزش و سایر موارد روش‌های ۱ و ۵ را به عنوان روش‌های استخراج ویژگی در نظر گرفتیم.

پس از انجام استخراج ویژگی و بدست آوردن خصیصه‌های متن، می‌توان از روش‌های انتخاب ویژگی استفاده کرد. مطرح‌ترین روش‌ها برای انجام این کار عبارت اند از:

۱) **آستانه واریانس:** یک رویکرد پایه ساده برای انتخاب ویژگی است. همه ویژگی‌هایی را که واریانس آن‌ها آستانه‌ای را برآورده نمی‌کند حذف می‌کند. به‌طور پیش‌فرض، تمام ویژگی‌های واریانس صفر، یعنی ویژگی‌هایی که در همه نمونه‌ها مقدار یکسانی دارند، حذف می‌کند.
۲) **انتخاب ویژگی تک‌متغیره:** این روش با انتخاب بهترین ویژگی ها بر اساس آزمون های آماری تک متغیره کار می‌کند. مثلاً روش SelectKBest، K ویژگی با بهترین معیار آماری را نگه داشته و از باقی ویژگی‌ها صرف نظر می‌کند که در اینجا مقدار K می‌تواند به وسیله ما مشخص شود. یا روش SelectPercentile

که درصد مشخصی از برترین ویژگی‌ها از نظر معیارهای آماری را نگهداری کرده و سایر ویژگی‌ها را حذف می‌کند.

در این پروژه پس از استخراج ۵۰۰۰۰ ویژگی با روش‌های ذکر شده بالا، برای انتخاب ویژگی از روش SelectKBest با توابع آماری chi2، f_classif، f_regression و mutual_info_regression برای انتخاب ویژگی‌های مجموعه‌های آموزش، توسعه و تست استفاده کرده و مدل شبکه پرسپترون چند لایه را با ویژگی‌های بدست آمده از این روش‌ها از مجموعه آموزش، آموزش داده و مدل آموزش داده شده را بر روی مجموعه dev_data با پارامترهای مختلف اجرا کرده و سعی کردیم تا بهترین پیکربندی برای هر مدل که معیار score را بیشینه می‌کند را با تحلیل و تغییر دادن پیکربندی‌ها بدست آوریم. در جدول زیر معیار score برای منتخب برخی از روش‌های گوناگون که تست شده قابل ملاحظه است.

| روش استخراج ویژگی | ایتریشن | تابع انتخاب ویژگی | K | Score |
|---|---|---|---|---|
| Count | ۲۰ | - | - | ۰.۵۴۳۱ |
| Count | ۵۰ | - | - | ۰.۵۳۳۶ |
| Tf-Idf | ۲۰ | - | - | ۰.۵۸۱۲ |
| Tf-Idf | ۵۰ | - | - | ۰.۵۵۱۸ |
| Count +Tf-Idf | ۲۰ | - | - | ۰.۶۱۳۵ |
| Count +Tf-Idf | ۵۰ | - | - | ۰.۵۹۲۴ |
| Count +Tf-Idf | ۲۰ | chi2 | ۵۰ | ۰.۷۰۴۵ |
| Count +Tf-Idf | ۲۰ | chi2 | ۱۰۰ | ۰.۶۸۳۶ |
| Count +Tf-Idf | ۲۰ | chi2 | ۲۰۰ | ۰.۵۷۶۲ |
| Count +Tf-Idf | ۲۰ | chi2 | ۳۰۰ | ۰.۴۹۸۷ |
| Count +Tf-Idf | ۲۰ | chi2 | ۴۰۰ | ۰.۵۷۴۵ |
| Count +Tf-Idf | ۲۰ | chi2 | ۵۰۰ | ۰.۵۲ |
| Count +Tf-Idf | ۲۰ | chi2 | ۶۰۰ | ۰.۴۶۶۳ |
| Count +Tf-Idf | ۵۰ | chi2 | ۵۰ | ۰.۶۸۶۸ |
| Count +Tf-Idf | ۵۰ | chi2 | ۱۰۰ | ۰.۶۵۴۹ |
| Count +Tf-Idf | ۵۰ | chi2 | ۲۰۰ | ۰.۵۶۷۲ |
| Count +Tf-Idf | ۵۰ | chi2 | ۳۰۰ | ۰.۴۷۸۶ |
| Count +Tf-Idf | ۵۰ | chi2 | ۴۰۰ | ۰.۵۴۹۵ |
| Count +Tf-Idf | ۵۰ | chi2 | ۵۰۰ | ۰.۵۰۴۵ |
| Count +Tf-Idf | ۵۰ | chi2 | ۶۰۰ | ۰.۴۵۶۵ |
| Count +Tf-Idf | ۲۰ | f_classif | ۵۰ | ۰.۶۹۳ |
| Count +Tf-Idf | ۲۰ | f_classif | ۱۰۰ | ۰.۶۸۶۴ |
| Count +Tf-Idf | ۲۰ | f_classif | ۲۰۰ | ۰.۶۲۵ |
| Count +Tf-Idf | ۲۰ | f_classif | ۳۰۰ | ۰.۶۳۹۳ |
| Count +Tf-Idf | ۲۰ | f_classif | ۴۰۰ | ۰.۵۷۶۲ |
| Count +Tf-Idf | ۲۰ | f_classif | ۵۰۰ | ۰.۵۰۴۵ |

| | | | | |
|---|---|---|---|---|
| Count +Tf-Idf | ۲۰ | f_classif | ۶۰۰ | ۰.۵۵۶۵ |
| Count +Tf-Idf | ۵۰ | f_classif | ۵۰ | ۰.۶۸۲۳ |
| Count +Tf-Idf | ۵۰ | f_classif | ۱۰۰ | ۰.۶۸۲۳ |
| Count +Tf-Idf | ۵۰ | f_classif | ۲۰۰ | ۰.۵۷۳۳ |
| Count +Tf-Idf | ۵۰ | f_classif | ۳۰۰ | ۰.۶۱۸۴ |
| Count +Tf-Idf | ۵۰ | f_classif | ۴۰۰ | ۰.۵۶۷۲ |
| Count +Tf-Idf | ۵۰ | f_classif | ۵۰۰ | ۰.۵۰۴۵ |
| Count +Tf-Idf | ۵۰ | f_classif | ۶۰۰ | ۰.۵۵۵۷ |

همانطور که ملاحظه می‌شود استفاده از روش استخراج ویژگی Count +Tf-Idf منجر به بدست آمدن نتایج بهتری می‌شود، لذا از این روش برای رتبه بندی استفاده می‌کنیم.

پس از این با استفاده از روش‌های بهتر در مدل بدست آمده، عملیات رتبه‌بندی را برای مجموعه test_data.xml انجام می‌دهیم. خروجی این رتبه‌بندی با فرمت اشاره شده در صورت پروژه در فایل IR-Final-Prj-Step-3.csv تنظیم شده‌است.

| | Question Number | Answer Number | Predict Rank | Predict Proba | Answer class |
|---|---|---|---|---|---|
| 0 | Q388_R14 | Q388_R14_C5 | 1 | 0.958196 | 1 |
| 1 | Q388_R14 | Q388_R14_C7 | 2 | 0.950809 | 1 |
| 2 | Q388_R14 | Q388_R14_C4 | 3 | 0.865108 | 1 |
| 3 | Q388_R14 | Q388_R14_C3 | 4 | 0.864406 | 1 |
| 4 | Q388_R14 | Q388_R14_C9 | 5 | 0.863566 | 1 |
| ... | ... | ... | ... | ... | ... |
| 2925 | Q475_R9 | Q475_R9_C10 | 6 | 0.789277 | 1 |
| 2926 | Q475_R9 | Q475_R9_C7 | 7 | 0.710056 | 1 |
| 2927 | Q475_R9 | Q475_R9_C5 | 8 | 0.697191 | 1 |
| 2928 | Q475_R9 | Q475_R9_C4 | 9 | 0.643386 | 1 |
| 2929 | Q475_R9 | Q475_R9_C1 | 10 | 0.629225 | 0 |

2930 rows × 5 columns

## گام چهارم: بازنمایی طیفی کلمات

GloVe مخفف Global Vectors برای نمایش کلمه است. این یک الگوریتم یادگیری بدون نظارت است که توسط محققان دانشگاه استنفورد ایجاد شده است که هدف آن ایجاد جاسازی کلمات با جمع‌آوری ماتریس‌های co-occurrence کلمه سراسری از یک پیکره معین است.

ایده اصلی پشت جاسازی کلمه GloVe این است که رابطه بین کلمات را به کمک آمار بدست آوریم. برخلاف ماتریس رخداد، ماتریس co-occurrence می‌گوید که چند بار یک جفت کلمه خاص با هم اتفاق می‌افتد. هر مقدار در ماتریس co-occurrence نشان دهنده یک جفت کلمه است که با هم اتفاق می‌افتند.

برای پیاده‌سازی این قسمت از کتابخانه zeugma.embeddings استفاده کردیم. ابتدا یک transformer را با استفاده از تابع EmbeddingTransformer که پارامتر 'glove' به آن داده شده است می‌سازیم و سپس داده‌ی ورودی را به آن می‌دهیم که با این بازنمایی ویژگی‌های آن استخراج شود. در ادامه نیز خروجی این نمایش را به همراه داده‌ی های برچسب آموزشی به شبکه MLP می‌دهیم.

دلیل انتخاب MLP برای آموزش این است که MLP قابلیت کار با داده‌ی حجیم را دارا بوده و سرعت آموزش آن نیز مناسب امکانات در دسترس ما است و همچنین توانایی اداره ویژگی‌های متعدد را به عنوان ورودی دارا بوده و نیز در حجم داده‌ی کمتر دقت آن تغییر نمی‌کند لذا با استفاده از شبکه MLP عملیات رتبه‌بندی را برای مجموعه test_data.xml انجام می‌دهیم. خروجی این رتبه‌بندی با فرمت اشاره شده در صورت پروژه در فایل IR-Final-Prj-Step-4.csv قرار داده شده‌است.

| | Question Number | Answer Number | Predict Rank | Predict Proba | Answer class |
|---|---|---|---|---|---|
| 0 | Q388_R14 | Q388_R14_C5 | 1 | 0.999996 | 1 |
| 1 | Q388_R14 | Q388_R14_C7 | 2 | 0.999914 | 1 |
| 2 | Q388_R14 | Q388_R14_C9 | 3 | 0.999584 | 1 |
| 3 | Q388_R14 | Q388_R14_C3 | 4 | 0.994282 | 1 |
| 4 | Q388_R14 | Q388_R14_C4 | 5 | 0.991382 | 1 |
| ... | ... | ... | ... | ... | ... |
| 2925 | Q475_R9 | Q475_R9_C3 | 6 | 0.994994 | 1 |
| 2926 | Q475_R9 | Q475_R9_C7 | 7 | 0.994703 | 1 |
| 2927 | Q475_R9 | Q475_R9_C9 | 8 | 0.989387 | 1 |
| 2928 | Q475_R9 | Q475_R9_C4 | 9 | 0.982252 | 1 |
| 2929 | Q475_R9 | Q475_R9_C1 | 10 | 0.658059 | 0 |

## گام پنجم (امتیازی) BERT finetuning

BERT [1] روشی برای بازنمایی زبان به صورت از پیش آموزش داده‌شده است که به کمک آن برنامه نویسان می‌توانند به طور رایگان مدل‌های از پیش آموزش داده شده را استفاده کنند. این مدل‌ها می‌توانند برای استخراج ویژگی‌های موثرتر یا برای دقیق کردن تنظیمات (fine tuning) مختص کاربردهایی مثل دسته بندی، شناسایی موجودیت و پاسخ‌دهی به پرسش استفاده شوند.

از مزایای استفاده از BERT این است که زمان بسیار کمتری برای آموزش مدل fine tuned نیاز است، زیرا لایه‌های پایینی شبکه عصبی قبلا آموزش دیده شده‌اند و تنها بایستی که به تدریج با استفاده از خروجی آن لایه‌ها که به عنوان ویژگی به شبکه‌ی ما داده می‌شوند تنظیم شوند. علاوه بر این، به دلیل وزن‌های از پیش آموزش‌دیده‌شده، این روش به ما اجازه می‌دهد تا تسک خود را روی مجموعه داده‌ای بسیار کوچک‌تر از آنچه در مدلی که از ابتدا ساخته شده است، تنظیم کنیم. یک اشکال عمده مدل های NLP که از ابتدا ساخته شده‌اند این است که ما اغلب به یک مجموعه داده بسیار بزرگ نیاز داریم تا بتوانیم شبکه خود را با دقت معقول آموزش دهیم، به این معنی که زمان و انرژی زیادی باید برای ایجاد مجموعه داده صرف شود.

لذا با استفاده از BERT عملیات بازرتبه‌بندی را برای مجموعه test_data.xml انجام می‌دهیم. خروجی این رتبه‌بندی پاسخ‌های مرتبط با هر پریش در قالب خواسته شده در صورت پروژه در فایل IR-Final-Prj-Step-5.csv قرار دارد.

| | Question Number | Answer Number | Predict Rank | Predict Proba | Answer class |
|---|---|---|---|---|---|
| 0 | Q388_R14 | Q388_R14_C9 | 1 | 1.000000 | 1 |
| 1 | Q388_R14 | Q388_R14_C8 | 2 | 0.999986 | 1 |
| 2 | Q388_R14 | Q388_R14_C7 | 3 | 0.999938 | 0 |
| 3 | Q388_R14 | Q388_R14_C5 | 4 | 0.996208 | 1 |
| 4 | Q388_R14 | Q388_R14_C4 | 5 | 0.994334 | 0 |
| ... | ... | ... | ... | ... | ... |
| 2925 | Q475_R9 | Q475_R9_C10 | 6 | 0.999454 | 1 |
| 2926 | Q475_R9 | Q475_R9_C3 | 7 | 0.999219 | 1 |
| 2927 | Q475_R9 | Q475_R9_C9 | 8 | 0.999201 | 1 |
| 2928 | Q475_R9 | Q475_R9_C5 | 9 | 0.997918 | 1 |
| 2929 | Q475_R9 | Q475_R9_C1 | 10 | 0.996837 | 1 |

2930 rows × 5 columns

[1] Bidirectional Encoder Representations from Transformers

**) برای اجرای کدها بایستی که فایل‌های train_data.xml، dev_data.xml و test_data.xml که همان فایل‌های اولیه ارائه شده در تمرین است و نیز فایل‌های pa_pp_train_data.csv، pa_pp_dev_data.csv و pa_pp_test_data.csv در کنار فایل نوت بوک ژوپیتر قرار داشته باشند.