مینا جمشیدی
۴۰۱۱۲۶۲۱۳۴
مبینا محمدی
۴۰۱۱۲۶۲۰۹۸

مستندات پروژه NLP

۱ فاز اول: استخراج ویژگی‌ها

۱.۱ چالش‌های داده‌های خام

- وجود نویزهایی مانند نشانی‌های وب، ایمیل، علائم نگارشی و تکرار حروف
- شامل بودن کلمات پرتکرار (stopwords) که اطلاعات مفیدی منتقل نمی‌کنند
- وجود کلمات با فرمت‌های گرامری متفاوت (مانند running، run، ran) که باید به یک ریشه نگاشته شوند
- وجود حروف بزرگ/کوچک که می‌توانند مدل را دچار سردرگمی کنند (مثلاً `Apple` با `apple` متفاوت تفسیر شود)

۲.۱ راهکارها و منطق انتخاب‌شده

دلایل هر تکنیک استفاده شده:

- **تبدیل به حروف کوچک :(lowercase)** برای یکسان‌سازی نوشتار کلمات و جلوگیری از تکرار غیر ضروری ویژگی‌ها.
- **حذف لینک‌ها و ایمیل‌ها:** این اطلاعات معمولاً حاوی معنای خاصی برای دسته‌بندی محصول نیستند و حذف آن‌ها به ساده‌سازی متن کمک می‌کند.
- **حذف اعداد:** اعداد در این نوع مسئله‌ها معمولاً کاربرد خاصی ندارند (مثلاً شماره مدل یا سایز)، بنابراین حذف شدند مگر اینکه در تحلیل آینده به صورت جدا استفاده شوند.
- **حذف علائم نگارشی و :HTML** این کار برای کاهش نویز و تمرکز بر محتوای متنی اصلی انجام شد.
- **کاهش کشیدگی کلمات:** کلماتی مثل `soooo` به `soo` تبدیل شدند تا مدل دچار تکرار بی‌مورد نشود.
- **حذف کلمات توقف :(Stopwords)** کلماتی مثل `in is, the,` اطلاعات خاصی منتقل نمی‌کنند و حضور آن‌ها باعث می‌شود مدل یادگیری سخت‌تری داشته باشد.

- **Lemmatization:** برای تبدیل کلمات به شکل پایه‌شان (مثل `running` به `run`) از Lemma-tizer استفاده شد. این کار به مدل کمک می‌کند مفهوم اصلی کلمه را فارغ از فرم گرامری‌اش یاد بگیرد.
- **توکن‌سازی:** برای پردازش بهتر، متن به کلمات جداگانه شکسته شد تا بتوان روی هر کلمه عملیات انجام داد.

در پایان هر نمونه متنی به یک نسخه تمیز، یکنواخت و کاهش‌یافته از نظر اطلاعات زائد تبدیل شد. این داده‌ها اکنون آماده‌اند تا در مراحل بعدی برای برداری‌سازی (با TF-IDF و Word۲Vec) استفاده قرار گیرند.

	text	clean_text
35848	Kandy Men's Regular Fit Blazer Blue This produ...	kandy men regular fit blazer blue product made...
13005	HealthSense Chef-Mate KS 50 Digital Kitchen Sc...	healthsense chefmate digital kitchen scale grey
22719	Concept of Physics (2018-2019) Session (Set of...	concept physic session set volume
18453	Lista Stainless Steel Multi Functional Hammer ...	lista stainless steel multi functional hammer ...
20867	Gardening in Urban India update	gardening urban india update

۲ فاز دوم: بردارسازی متون با استفاده از TF-IDF

در این بخش، الگوریتم TF-IDF بدون استفاده از هیچ کتابخانه‌ای پیاده‌سازی شد.

۱.۲ چالش‌ها و راهکارها

- **محاسبه DF:** یکی از چالش‌ها، محاسبه دقیق Frequency Document برای هر واژه بود، به طوری که شمارش تکرار یک واژه فقط یک‌بار در هر سند لحاظ شود. برای حل این مسئله، از set استفاده شد تا کلمات تکراری در هر سند فقط یک‌بار شمرده شوند.
- **جلوگیری از تقسیم بر صفر:** در محاسبه IDF، برای جلوگیری از تقسیم بر صفر در فرمول $\log(N/df)$، یک واحد به df اضافه شد، یعنی از $\log(N/(1+df))$ استفاده شد.
- **ابعاد ماتریس:** با توجه به تعداد واژگان یکتا (واژگان نهایی)، لازم بود که ابتدا دیکشنری word2idx تعریف شود تا اندیس هر واژه در ماتریس TF-IDF مشخص شود. سپس یک ماتریس $N \times V$ (تعداد اسناد $\times$ تعداد واژگان) ساخته شد.

۲.۲ نتایج

نتیجه داکیومنت اول نشان داد که واژگانی مانند kandy، velvet و buttoned دارای مقادیر TF-IDF بالاتری نسبت به کلمات عمومی‌تر مانند product یا made هستند. این نشان می‌دهد که TF-IDF به درستی کلمات متمایزکننده را نسبت به کلمات پرتکرار تشخیص داده است.

```
kandy men regular fit blazer blue product made velvet finished attractive blue color feature plain solid pattern long sle
kandy → tf-idf: 7.824046010856292
men → tf-idf: 2.8134107167600364
regular → tf-idf: 3.519980917652122
fit → tf-idf: 2.501036031717884
blazer → tf-idf: 5.626821433520073
blue → tf-idf: 2.9603651297166995
product → tf-idf: 1.7070507413011007
made → tf-idf: 1.789761466565382
velvet → tf-idf: 5.8781358618009785
finished → tf-idf: 4.06284589516273
attractive → tf-idf: 3.519980917652122
blue → tf-idf: 2.9603651297166995
color → tf-idf: 1.9589953886039688
feature → tf-idf: 2.0826466726287842
plain → tf-idf: 4.406319327242926
solid → tf-idf: 3.649658740960655
pattern → tf-idf: 3.4357888264317746
long → tf-idf: 2.4557362724882204
sleeve → tf-idf: 3.789805372703897
buttoned → tf-idf: 6.907755278982137
closure → tf-idf: 5.051457288616511
targeted → tf-idf: 5.8781358618009785
towards → tf-idf: 5.149897361429764
men → tf-idf: 2.8134107167600364
furthermore → tf-idf: 5.115995809754082
recommended → tf-idf: 4.173387769562553
kept → tf-idf: 5.221356325411908
away → tf-idf: 3.9220733412816475
extreme → tf-idf: 5.051457288616511
heat → tf-idf: 3.533586569707901
fire → tf-idf: 5.339139361068292
corrosive → tf-idf: 5.8781358618009785
liquid → tf-idf: 4.474141923581687
avoid → tf-idf: 4.268697949366879
form → tf-idf: 3.763603000309873
damage → tf-idf: 4.06284589516273
```

۳ فاز دوم: بردارسازی متون با استفاده از Word۲Vec

از مدل Word2Vec کتابخانه gensim استفاده شد.

۱.۳ چالش‌ها و تصمیمات

- **توکن‌سازی دقیق:** برای آموزش Word۲Vec نیاز به توکن‌سازی دقیق داشتیم. در ابتدا از split() استفاده شد، اما برای نتایج بهتر از word_tokenize از کتابخانه nltk استفاده شد.
- **تنظیمات مدل:** برای آموزش مدل، پارامترهایی مانند vector_size=100، window=5 و min_count=2 انتخاب شد. این تنظیمات با توجه به اندازه داده‌ها و برای جلوگیری از نویز ناشی از کلمات بسیار کم‌تکرار انجام شد.

- **کلمات نادیده گرفته‌شده:** برخی کلمات به دلیل min_count وارد مدل نمی‌شوند، بنابراین در تحلیل نهایی ممکن است برخی واژه‌ها بردار نداشته باشند.

۲.۳ نتایج

با بررسی بردار کلمه painting مشاهده شد که شباهت زیادی با واژگان مفهومی مانند art، drawing، و craft دارد. این نشان می‌دهد که مدل Word۲Vec توانسته ارتباط معنایی بین کلمات را تا حد مناسبی بیاموزد. همچنین خروجی بردار عددی کلمه painting به‌خوبی ساختار توزیعی مدل را نشان می‌دهد.

```
Vector for 'painting':
[-0.5508629   0.36387342 -0.36982006 -1.056777   -0.22129601 -0.11388729
 -1.4613373  -0.73587257 -1.2004316   2.6038702   0.2659391  -0.6010603
  1.0231673  -0.92522097 -0.79301536 -1.2032902   1.4779582   2.5695066
 -0.15849325  1.3117727  -0.2737219  -0.61334205 -0.5723068   0.25922787
  0.81314796  0.60878175 -0.35792628 -0.6534402   0.7085712  -1.3418529
 -0.53413486 -0.42737135  0.68361074 -0.19693144 -0.2622345  -0.19879907
 -2.0610995  -0.6067788  -0.26163772 -0.28893083  0.2548341  -1.48044
 -2.6918893   0.86935455 -1.7898625   0.23239349 -1.2008233  -1.0039829
  1.537418   -1.3866178  -1.1436353  -0.8453917  -1.0944076   0.17852592
  0.9740102   0.05182031 -0.6212762  -1.4812379  -1.5863237  -0.35590577
  0.42921606 -1.2403116   0.48564795 -1.1095368   0.5501702  -1.0542358
 -0.35811844  1.2013218  -1.0504943   1.2214376  -0.88091445  0.7849959
  0.2985102  -0.42124787  0.01212086 -0.2500986   0.52507186 -0.41478604
  0.80750847  1.1138465  -0.0455182  -1.648858   -1.3382894   1.0623164
 -0.88975775 -0.12223998  0.5895954   2.7580974  -0.5656033  -0.42690265
 -0.60894865  3.018483    0.7489087   0.8160427   0.4866975   2.2931328
 -1.6081651   0.40424594 -0.9389486   0.14996298]

Similarity between 'painting' and 'art': 0.6907795071601868
Most similar words to 'painting':
art: 0.6907795071601868
watercolour: 0.6684693098068237
craft: 0.6585366129875183
drawing: 0.6484317779541016
stencil: 0.6409387588500977
```