مقدمه:

به دیتاستی دسترسی داریم که ۷۰۰۰۰ کامنت کاربران اسنپ فود را منتشر کرده است. این کامنت ها توسط دو لیبل "مثبت" و "منفی" تفکیک شده اند که تعداد مربوط به هر کدام ۳۵۰۰۰ عدد است.

توضیحات:

قبل از هر چیز نیاز است دیتا را آپلود کنیم. CSV

اولین هدف استفاده از لایبرری های هضم و سایکت لرن برای توکنایز کردن و نرمالایز کردن داده ها است.

سلول اول : استفاده از لایبرری هضم برای نرمالایز کردن

هدف نرمالسازی دیتافریم کامنت داده های متنی است. کتابخانه هضم یک کتابخانه برای پردازش زبان فارسی در پایتون است.

سلول دوم : توکنایز کردن

پس از نرمالسازی متنها با استفاده از این متد متن را به واحد های کوچک تر که به عنوان توکن ها شناخته می شوند تقسیم بندی میکنیم.

سلول سوم : استمینگ

این کار کلمات را به پایه ای که به آن اشاره دارد تبدیل میکند. هدف اصلی حذف پسوندها و پیشوندهای مختلف کلمات است.

استمینگ در برخی موارد میتواند کلمات را به ریشه های نامعتبر تبدیل کند.

سلول چهارم : حذف کلمات توقف

این کار در جهت کمک به کاهش حجم داده عملکرد سریعتر الگوریتم انجام میشود.

TF-IDF Vectorize

یک تکنیک نمایش عددی است که در پردازش زبان طبیعی استفاده می شود. هدف تبدیل مجموعه ای از متن های سندی به یک نمایش ماتریسی است. نتیجه این کار موجب میشود تا مدل های یادگیری ماشین بتوانند متن را فهمیده و پردازش کنند.

ترم فرکانس برابر با تعداد بارهایی است که یک ترم در یک سند تکرار شده تقسیم بر کل تعداد ترم های آن سند.

برعکس فرکانس سند اهمیت ترم ترم در کل مجموعه را اندازه میگیرد و این برابر لگاریتم تعداد کل بر تعداد سندهای شامل آ« ترم است.

Count Vectorize

در اینجا جملات را به بردارهای تعدادی تبدیل کرد ایم. برای نمایش برداری از جملات استفاده کردیم که در آن تعداد تکرار هر کلمه در هر جمله محاسبه می شود.

FastText Embeddings

برای هر جمله بردارهای متنظر را جمع آوری می کند. به کمک بردار میانگین میتوان به لیست اضافه کرد. در واقع برای هر جمله از دیتاست، برداری با ابعاد مشخص بر اساس معنای کلمات در جمله بدست می آید.

Word2Vec

به کمک توابع جملات را به توکن تبدیل می کنیم. برای هر جمله بردارهای متناظر با هر کلمه را جمع آوری میکنیم. اگر کلمه در مدل موجود باشد، بردار متناظر آنرا اضافه میکند و سپس لیست را به دیتافریم پانداس تبدیل میکنیم.

قبل از استفاده از مدل‌های رندم فارست، لجستیک رگرشن و نایو بیز برای کلاسیفیکیشن، باید داده‌های فارسی را پیش‌پردازش کنیم. چندین مرحله از پیش‌پردازش می‌تواند شامل مواردی باشد که انجام دادیم:

۱. توکن‌سازی : تقسیم جملات به توکن‌ها یا کلمات مجزا. برای زبان فارسی می‌توان از کتابخانه‌هایی مانند هضم استفاده کرد که ابزارهای متناظر با این زبان را فراهم می کند.
۲. حذف کلمات اضافی : حذف کلمات مشترک و کم ارزش مانند حروف اضافه و حروف ربط از متن.
۳. استم‌برداری : تبدیل کلمات به شکل اصلی یا ریشه آن‌ها.

نتایج:

TF-IDF Vectorize vs. Forest Classifier

.

روش رندم فارست یک الگوریتم یادگیری ماشینی است که بر اساس ساخت چندین درخت تصمیم تصادفی عمل می کند و عموماً در مسائل دسته‌بندی و رگرسیون استفاده می‌شود. این روش با ترکیب نتایج درخت‌ها به یک مدل کلی می‌رسد.

مستقیماً بر روی ویژگی‌های متنی تأثیر می گذارد، در حالی که روش رندم فارست با استفاده از اهمیت ویژگی‌ها، بررسی می کند که هر ویژگی به چه اندازه در تصمیم گیری نقش دارد.

قابلیت تعمیم‌پذیری برای متون جدید و ناشناخته را دارد، در حالی که روش رندم فارست نیاز به آموزش قبلی دارد و برای داده‌های جدید باید مراحل پیش‌پردازش و استخراج ویژگی‌ها را تکرار کنید.

Count Vectorize vs. Forest Classifier

اولی یک روش برای تبدیل متن به بردارهای عددی است. این روش تعداد تکرار کلمات را در متن محاسبه می کند و آنها را بعنوان ویژگی های عددی استفاده میکند. اما دومی یک الگوریتم دسته بندی است که بر اساس ساخت چندین درخت تصمیم عمل میکند.

.

Logistic Regression vs. TF-IDF Vectorize

دومی یک الگوریتم است که میتوانیم برای دسته بندی مسائل منطقی استفاده کنیم.