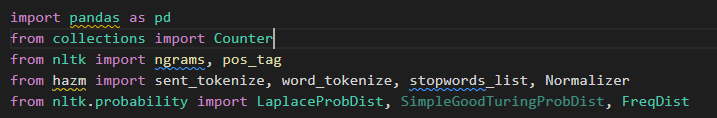
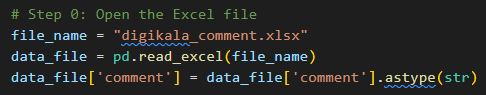
ایمپورت کردن کتاب خانه های لازم:



مرحله صفر:

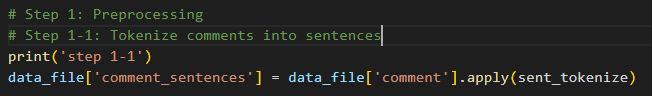
باز کردن فایل و تبدیل ستون کامنت به str



مرحله 1-1:

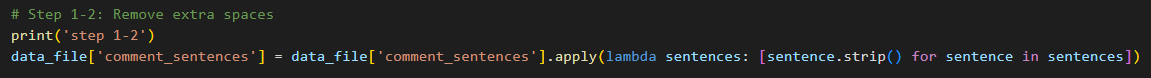
توکن کردن کامنت ها:

این مرحله با استفاده از hazm انجام شده است.



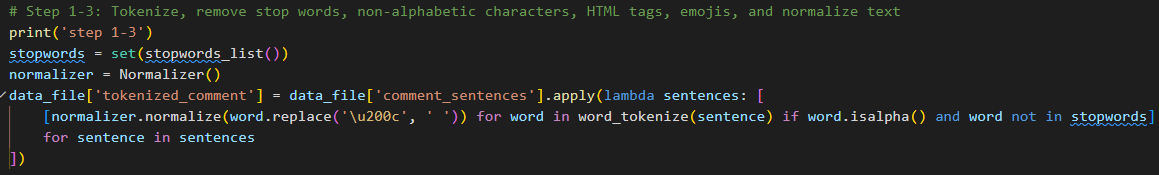
مرحله 1-2:

حذف فضاهای خالی اضافی با استفاده از strip



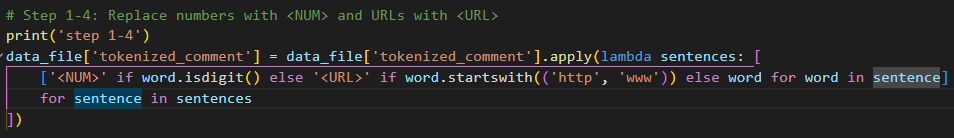
مرحله 1-3:

حذف استاپ ورد، علایم نگارشی و سایر موارد اضافی:



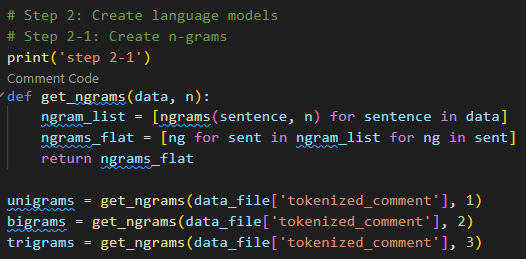
مرحله 1-4:

تبدیل اعداد به آدرس ها به جایگزین گفته شده:

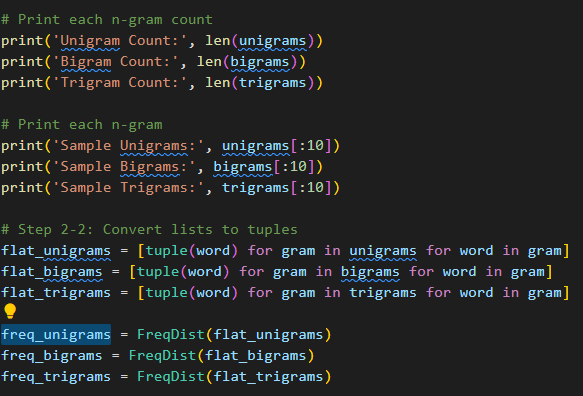


مراحل 2-1 2-2 2-3

ساخت n-gram:



چاپ موارد مربوط به این مرحله:



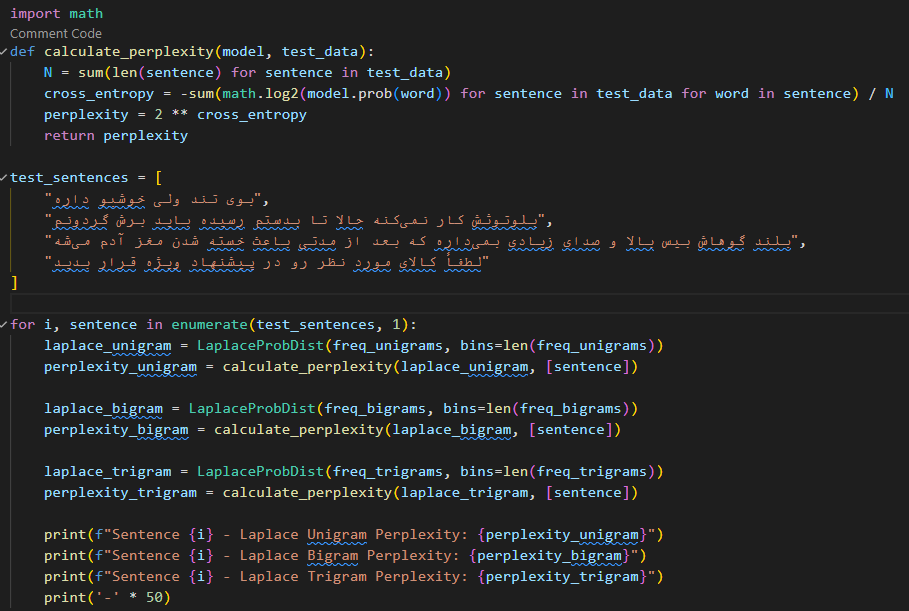
نتیجه چاپ:



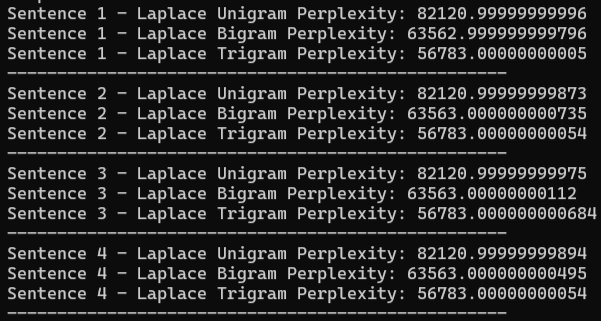
روش Laplace smoothing برای n-gram ها به خوبی عمل نمی‌ کند زیرا در این روش به هر کلمه یک تعداد ثابت از تعداد تکرارها اضافه می ‌شود تا از احتمال صفر برای کلماتی که در داده‌ ها دیده نشده ‌اند، جلوگیری شود. این اضافه کردن تعداد ثابت به تمام کلمات ممکن است باعث افزایش تعداد تکرارها و در نتیجه افزایش احتمالات شود. این امر می ‌تواند منجر به افزایش دقت نا همگن در تخمین احتمالات شود.

به طور مثال، اگر یک کلمه در داده ‌ها دیده نشده باشد، روش Laplace smoothing باعث افزایش احتمال آن کلمه می‌ شود، حتی اگر دیگر کلمات در متن با آن هماهنگ نباشند. این مشکل به خصوص برای n-gram هایی با تعداد زیادی از واژگان ممکن خودش را نشان می ‌دهد.

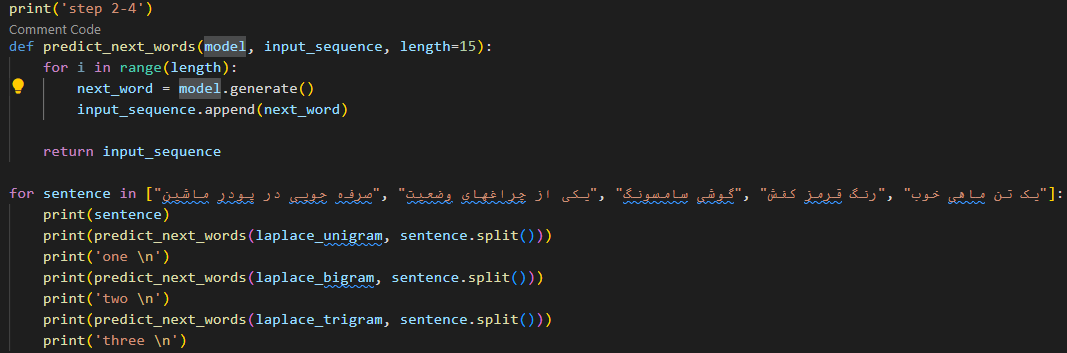
محاسبه perplexity:



گزارش این مرحله:



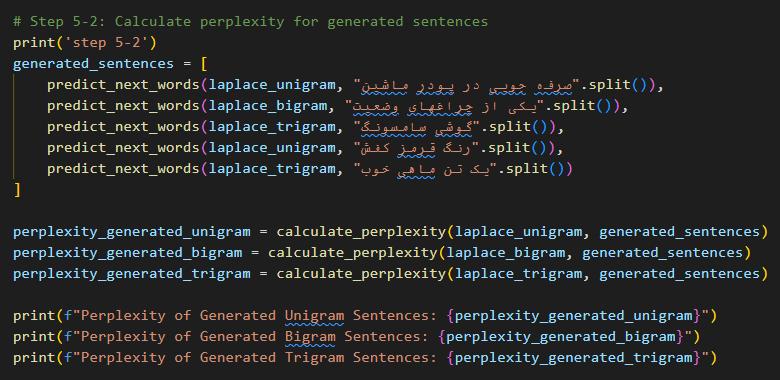
مرحله 2-4:



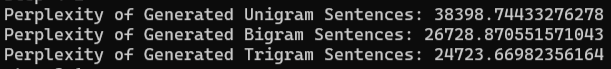
خروجی این قسمت یک مقدار طولانی بود قسمتی از خروجی نمایش داده شده:



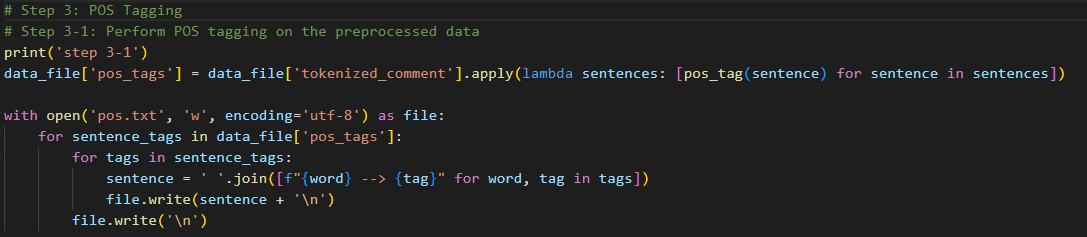
مرحله 2-5



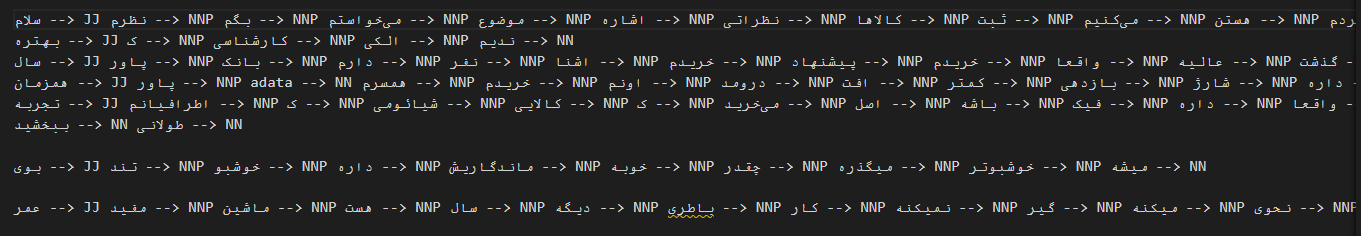
خروجی این مرحله:



مرحله 3-1:

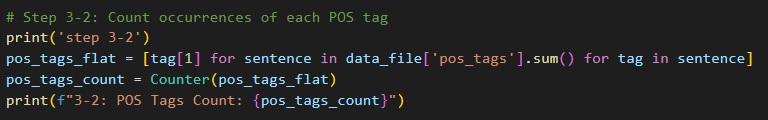


خروجی کار(فقط بخشی از خروجی در اسکرین شات است، فایل خروجی نیز ارسال شده):



مرحله 3-2:

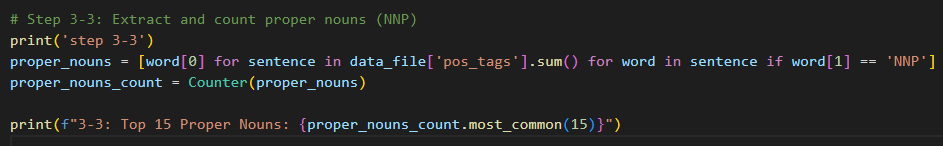
تعداد رخداد هر تگ:



خروجی:



مرحله 3-3:



خروجی:



چون کلمات فارسی هستن یکم خوانش پایین اومده این تو کنسول vscode



سوال 1

- مرحله یک: تولید کاندید ها (Candidate Generation): در این مرحله، با استفاده از مدل‌ های مختلف، یک لیست از کوئری‌ های پیشنهادی برای یک پیشوند (prefix) خاص تولید می ‌شود.

- مرحله دو: رتبه‌ بندی کاندید ها (Candidate Ranking): کاندید ها بر اساس اهمیت و ارتباط با کوئری ورودی رتبه ‌بندی می‌ شوند. این مرحله با استفاده از مدل ‌های یادگیری عمیق انجام می‌ شود.

سوال 2

- حافظه محدود: توانایی مدل در به‌ خاطر سپاری کاندید ها و اطلاعات مرتبط با کاربر.

- کیفیت بالای پیش ‌بینی: توانایی مدل در پیش‌ بینی کاندید های با کیفیت بالا برای ورودی‌ های مختلف.

- کارآیی و سرعت: اجرای سریع و کارآیی بالا در محیط‌ های واقعی.

- بهبود بازیابی (Recall): افزایش تعداد کوئری‌ های مناسب در لیست پیشنهادی.

- بهبود دقت (Precision): افزایش ارتباط کاندیداها با نیاز واقعی کاربر.

- مدیریت بازه زمانی (Latency): افزایش سرعت و کاهش تاخیر در تولید پیشنهاد ها.

- مدیریت منابع (Resource Management): بهره‌ گیری بهینه از منابع محاسباتی برای اجرای بهتر سیستم.

سوال 3

- توانایی در مدل ‌سازی ارتباطات پیچیده: مدل‌ های زبان عصبی قابلیت مدل ‌سازی ارتباطات پیچیده بین کلمات را دارا هستند.

- قابلیت یادگیری از داده: مدل‌ های زبان عصبی با توجه به حجم بالای داده ‌ها، قابلیت یادگیری و به‌ روز رسانی بهتری دارند.

- انعطاف‌ پذیری در تطبیق با داده جدید: مدل‌ های زبان عصبی قابلیت تطبیق با داده ‌های جدید را دارا هستند و می ‌توانند به تغییرات در زبان و استفاده‌ کنندگان پاسخ دهند.

- مدل سازی همه ‌جانبه (End-to-End Modeling): قابلیت یادگیری ارتباطات پیچیده درون کاندیداها و کوئری ورودی.

- قابلیت افزودن ویژگی‌ های اضافی (Incorporating Additional Features): قابلیت اضافه کردن ویژگی‌ های شخصی ‌سازی مانند شناسه کاربر در مدل.

- مدیریت همسانی دنباله (Sequence Coherence Modeling): توانایی مدل‌ سازی همسانی دنباله ‌های کلمات.

سوال 4

رویکرد MCG یا Maximum Context Generation که در این مقاله مطرح شده است، در فاز تولید کاندید ها بر اساس الگوریتم گریدی (Greedy Matching) بر روی کلمات پیشوندی ورودی عمل می‌ کند. این الگوریتم از آخرین چندین کلمه پیشوندی به صورت گریدی (از طولانی ‌ترین به کوتاه ترین) برای پیدا کردن مطلوب‌ ترین کاندید ها برای تکمیل خودکار کوئری استفاده می‌ کند.

سوال 5

Unnormalized Language Model (LSTMEmb): یک مدل زبان عصبی بر اساس شبکه عصبی با ساختار LSTM برای محاسبه امتیاز رتبه ‌بندی.

Frequency-Based Models (MPC, LWG, MCG): مدل‌ های مبتنی بر فراوانی که بر اساس تکرار و فراوانی واژگان محاسبه می ‌شوند.

Hybrid Models (NN+Frequency): مدل‌ های ترکیبی که از اطلاعات فراوانی و مدل‌ های عصبی برای رتبه ‌بندی استفاده می ‌کنند.

Convolutional Latent Semantic Model (CLSM): یک مدل عصبی مبتنی بر شبکه‌ های عصبی کانولوشنی برای محاسبه امتیاز رتبه ‌بندی.