به نام خدا

پروژه درس جبر خطی – ترم ۴۰۴۲

Indexing Semantic Latent

استاد : دکتر نیکان

اعضای گروه: علی کارخانه - مصطفی شیبانی – حسین نادری – مرتضی رحیمی

سوالات مفهومی:

۱. **از روش‌های ارزش‌گذاری کلمات در متون، روش‌های TF و IDF را توضیح دهید. چرا مطالعه مستقل این دو روش ممکن است گمراه‌کننده باشد؟**

تعریف TF: TF مخفف Term Frequency است؛ یعنی یک واژه چند بار در یک سند تکرار شده است.
اگر واژه‌ای فقط یک بار در سند آمده باشد ، آنگاه:

$$TF = 1$$

تعریف IDF: IDF مخفف Inverse Document Frequency است؛یعنی آن واژه در کل مجموعهٔ اسناد چقدر نادر یا رایج است.
فرمول رایج آن:

$$IDF(t) = \log\left(\frac{N}{df(t)}\right)$$

که در آن:

- N : تعداد کل اسناد
- $df(t)$: تعداد اسنادی که واژه t در آن‌ها آمده است

هرچه یک واژه در اسناد کمتری دیده شود، IDF بزرگ‌تر می‌شود.

چرا مطالعه مستقل TF و IDF ممکن است گمراه‌کننده باشد؟
چون هر کدام فقط یک **بعد** از اهمیت واژه را نشان می‌دهند و به‌تنهایی کافی نیستند.

- **اگر فقط به TF نگاه کنیم،** ممکن است واژه‌ای در یک سند خیلی زیاد تکرار شود، اما اصلاً واژهٔ مهمی نباشد. مثلا کلماتی مثل: (است ، در ، و) . این‌ها ممکن است TF بالایی داشته باشند، ولی چون در تقریباً همهٔ اسناد هستند، برای تشخیص موضوع سند مفید نیستند.
  پس TF بالا لزوماً به معنی مهم بودن واژه نیست.
- **اگر فقط به IDF نگاه کنیم،** ممکن است واژه‌ای در کل مجموعه نادر باشد، اما در سند موردنظر فقط یک‌بار و آن هم به‌صورت حاشیه‌ای آمده باشد. مثلاً نام یک مکان یا عدد خاص که فقط یک بار در متن آمده و نقش اصلی در موضوع سند ندارد.
- **پس IDF بالا هم به‌تنهایی لزوماً به معنی مهم بودن در آن سند خاص نیست.**

۲. **یکی از روش‌های تعیین آستانه مقادیر تکین برای کاهش مرتبه Truncated SVD رویکرد نقطه زانویی است. این رویکرد را شرح دهید.**

در روش Truncated SVD هدف این است که به جای نگه‌داشتن همه مقادیر تکین، فقط چند مقدار تکین مهم‌تر را حفظ کنیم تا مرتبه ماتریس کاهش یابد. اگر ماتریس داده را به صورت زیر تجزیه کنیم:

$$A = U\Sigma V^T$$

در ماتریس Σ، مقادیر تکین به‌صورت نزولی مرتب می‌شوند:

$$\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \cdots \geq \sigma_r$$

در Truncated SVD باید تصمیم بگیریم چند مقدار تکین اول را نگه داریم؛ یعنی مقدار k را انتخاب کنیم.

ایده اصلی نقطه زانویی:

در بسیاری از داده‌ها، مقادیر تکین ابتدایی بزرگ هستند و بخش اصلی اطلاعات را نگه می‌دارند. اما بعد از مدتی، مقدارهای تکین خیلی کوچک می‌شوند و کاهش آن‌ها آهسته‌تر می‌شود. اگر مقادیر تکین را روی نمودار رسم کنیم:

- محور افقی: شماره مؤلفه یا شماره مقدار تکین
- محور عمودی: اندازه مقدار تکین

معمولاً نمودار در ابتدا کاهش زیادی دارد و سپس تقریباً **صاف** می‌شود. نقطه‌ای که در آن نمودار از حالت افت سریع به حالت افت آهسته تغییر می‌کند، به نام **نقطه زانویی** (Elbow Point یا Knee Point) شناخته می‌شود.

انتخاب آستانه با نقطه زانویی:

در این روش، مقدار k را برابر با شماره نقطه زانویی انتخاب می‌کنیم.
پس مقادیر تکین زیر نگه داشته می‌شوند:

$$\sigma_1, \sigma_2, \ldots, \sigma_k$$

و بقیه مقادیر تکین حذف می‌شوند:

$$\sigma_{k+1}, \sigma_{k+2}, \ldots$$

مثال:
فرض کنید مقادیر تکین به شکل زیر باشند:

$$100,\ 80,\ 60,\ 40,\ 20,\ 5,\ 4,\ 3.8,\ 3.5,\ 3.3$$

در اینجا تا مقدار پنجم کاهش زیاد است، اما بعد از آن مقادیر خیلی آهسته کم می‌شوند. پس نقطه زانویی حدوداً در $k = 5$ است؛ یعنی ۵ مقدار تکین اول را نگه می‌داریم و بقیه را حذف می‌کنیم.

دلیل استفاده از نقطه زانو:

- **حفظ اطلاعات مهم:** مقادیر تکین بزرگ‌تر معمولاً بیانگر ساختار اصلی داده هستند.
- **حذف نویز و کاهش پیچیدگی:** مقادیر تکین کوچک‌تر معمولاً شامل جزئیات کم‌اهمیت یا نویز هستند.

پس با انتخاب نقطهٔ زانویی، سعی می‌کنیم تا جایی مؤلفه‌ها را نگه داریم که هنوز اطلاعات قابل‌توجهی اضافه می‌کنند.

۳. **نحوه محاسبه خطای بازسازی در SVD Truncated را بیان کنید.**

در **Truncated SVD** ماتریس اصلی A را با یک ماتریس تقریبیِ کم‌مرتبه بازسازی می‌کنیم .اگر تجزیهٔ SVD ماتریس A به‌صورت زیر باشد:

$$A = U\Sigma V^T$$

آنگاه در نسخه برش‌خورده، فقط k مقدار تکین اول را نگه می‌داریم:

$$A_k = U_k \Sigma_k V_k^T$$

که در آن:

- U_k شامل k بردار ویژه چپ اول است.
- Σ_k شامل k مقدار تکین بزرگ‌تر است.
- V_k شامل k بردار ویژه راست اول است.

خطای بازسازی برابر است با اختلاف بین ماتریس اصلی و ماتریس بازسازی‌شده:

$$E = A - A_k$$

برای اندازه‌گیری بزرگی این خطا، معمولاً از یک نرم استفاده می‌شود.
نرم فروبنیوس: رایج‌ترین معیار

$$\|A - A_k\|_F$$

که به این خطا، **خطای فروبنیوس** می‌گویند.
این نرم از رابطه زیر به دست می‌آید:

$$\|A - A_k\|_F = \sqrt{\sum_{i,j} (A_{ij} - (A_k)_{ij})^2}$$

یعنی: ریشه مجموع مربعات اختلاف درایه‌های ماتریس اصلی و بازسازی‌شده.
نکته: در Truncated SVD خطای بازسازی فروبنیوس را می‌توان مستقیماً از مقادیر تکین حذف‌شده محاسبه کرد:

$$\|A - A_k\|_F = \sqrt{\sigma_{k+1}^2 + \sigma_{k+2}^2 + \cdots + \sigma_r^2}$$

که r رتبهٔ اصلی ماتریس است، یعنی:
خطای بازسازی برابر است با ریشهٔ مجموع مربعات مقادیر تکینی که حذف شده‌اند.
نرم طیفی: گاهی از نرم طیفی هم استفاده می‌شود:

$$\|A - A_k\|_2 = \sigma_{k+1}$$

یعنی در این معیار، خطای بازسازی برابر با **اولین مقدار تکین حذف‌شده** است.

تفسیر

- اگر k را بزرگ‌تر بگیریم، مقادیر تکین بیشتری نگه داشته می‌شود.
- در نتیجه، خطای بازسازی کوچک‌تر می‌شود.
- اگر همهٔ مقادیر تکین را نگه داریم، آنگاه

$$A_k = A$$

و خطا صفر می‌شود.

مثال

فرض کنید مقادیر تکین یک ماتریس چنین باشد:

$$\sigma_1 = 10, \qquad \sigma_2 = 7, \qquad \sigma_3 = 4, \qquad \sigma_4 = 2$$

اگر فقط دو مقدار اول را نگه داریم ($k = 2$)، خطای بازسازی فروبنیوس برابر است با:

$$\|A - A_k\|_F = \sqrt{4^2 + 2^2} = \sqrt{16 + 4} = \sqrt{20}$$

و خطای طیفی برابر است با:

$$\|A - A_k\|_2 = 4$$