به نام خدا

پروژه دوم درس جبر خطی

مدرس: دکتر محمدسعید سرافراز

فردین عباسی

دانشجوی کارشناسی مهندسی برق دانشگاه تهران

fardinabbasi1381@gmail.com

عرفان عسگری

دانشجوی کارشناسی مهندسی برق دانشگاه تهران

erfanasgari21@gmail.com

نفیسه مقنی زاده

دانشجوی کارشناسی مهندسی برق دانشگاه تهران

nmoghanizadeh@gmail.com

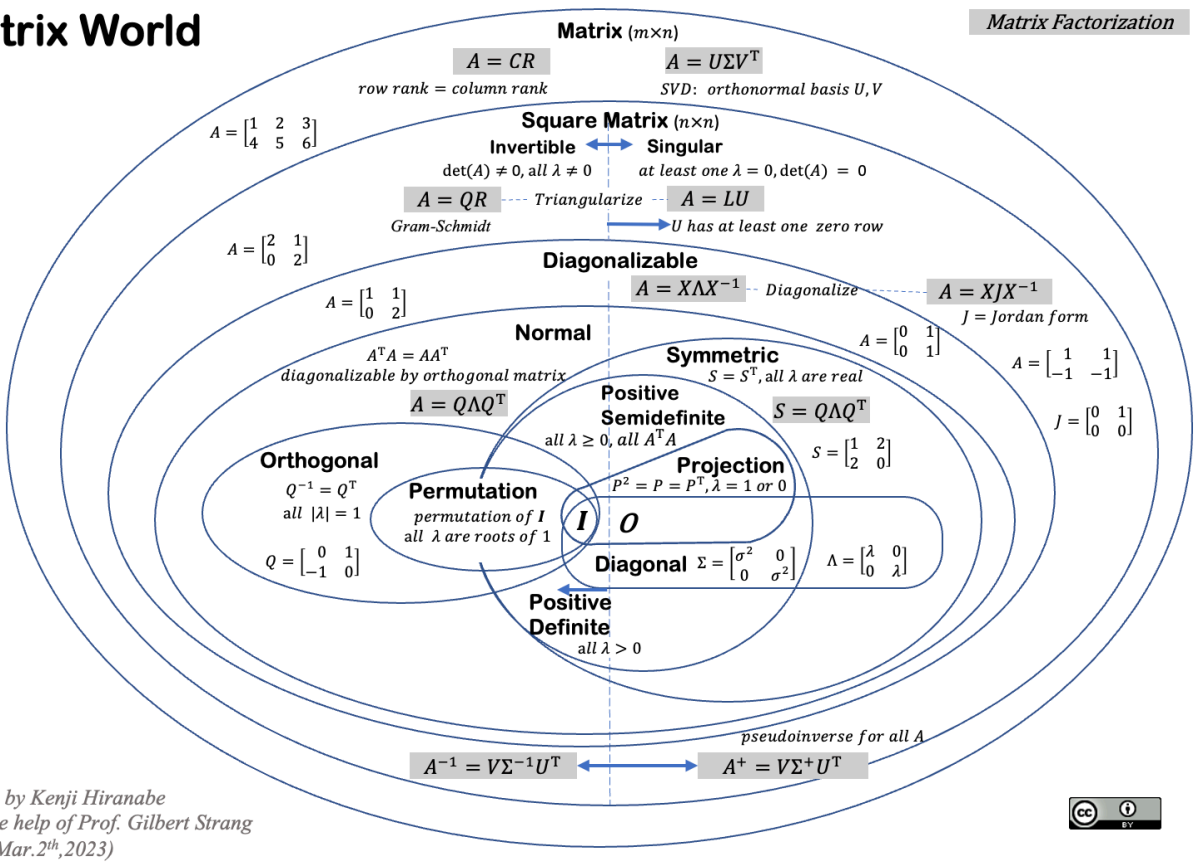
دانشکده مهندسی برق و کامپیوتر

دانشکدگان فنی

دانشگاه تهران

زمستان ۱۴۰۲

فهرست مطالب



فصل ۱

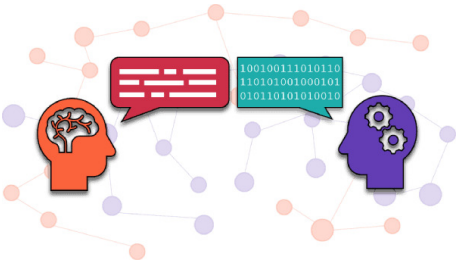
شرح پروژه

در این پروژه سعی شده است، به جنبه های مختلف کاربرد جبر خطی پرداخته شود. امید است با انجام این پروژه دانش قبلی شما تثبیت شود و با مفاهیم جدید در جبر خطی آشنا شوید.

عکس از: The Art of Linear Algebra

Latent Semantic Indexing

پردازش زبان طبیعی [۱] علمی است که به واسطهٔ آن ماشین‌ها زبان انسانی را درک می کنند و یا حتی به بازتولید آن می‌پردازند.



این علم به دو شاخه درک زبان طبیعی [۲] و تولید زبان طبیعی [۳] تقسیم می شود. درک زبان طبیعی سعی می‌کند ماشین‌ها را قادر سازد تا مفهوم [۴] جملات را درک کنند. در حالی که تولید زبان طبیعی بر ساختارمند کردن داده‌های زبانی تمرکز کرده تا جملات تولیدی برحسب آن ساختار، شبیه به جملات تولیدی انسان ها به نظر برسد.

Markov Chains!

یکی از روش‌های جالب تولید متن، استفاده از زنجیره‌های مارکوف است. در این روش، زنجیره ای از کلمات بر حسب احتمال ظهور توالی هر دو کلمه تشکیل می شود. پیشنهاد می‌شود برای اطلاعات بیشتر این کتابچه را مطالعه نمایید.

پردازش زبان طبیعی در حوزه‌های گوناگون مرتبط با زبان کاربرد دارد. برخی از این کاربردها شامل ترجمه ماشین، تحلیل احساس [۵] ، بازیابی اطلاعات و... است.

در این پروژه به بررسی یکی از کاربرد های آن در استخراج مفهوم متون می پردازیم.

---

۱ Natural Language Processing
۲ Natural Language Understanding
۳ Natural Language Generation
۴ Semantic
۵ Sentiment Analysis

مقدمه

ماشین ها کلمات را مانند انسان‌ها تعبیر نمی کنند، از این رو کار با داده های زبانی چالش های خاص خود را به همراه دارد. انسان‌ها با خواندن هر کلمه معنای آن را با کلمات نزدیک و مرتبط تداعی می کنند، در صورتی که ماشین به صورت پیش فرض هیچ درکی از کلمات ندارد.

در مقابل کامپیوترها درک خیلی خوبی از اعداد دارند؛ در نتیجه اگر بتوانیم کلمات را به صورت بردارهای عددی با اطلاعات غنی تعبیه [۶] کنیم، ماشین ها توانایی درک آن را خواهند داشت.

کوله کلمات [۷] رویکردی ابتدایی برای تعبیه متون است. این رویکرد فارغ از جایگاه کلمات، تعداد تکرار لیستی از کلمات **مشخص** را به عنوان بردار عددی از متن تعبیه می کند. اگر ماتریس کوله کلمات را $BW = (f_{i,j})$ بنامیم، $f_{i,j}$ تعداد تکرار کلمه j ام در متن i ام است.

Embedding Projector

کوله کلمات علی رغم سادگی، در درک زبان طبیعی رویکرد کارایی است. رویکرد های پیچیده تری با عملکرد بهتری، توسعه یافته است که می توانید عملکرد برخی از آن‌ها را در این لینک آزمایش کنید.

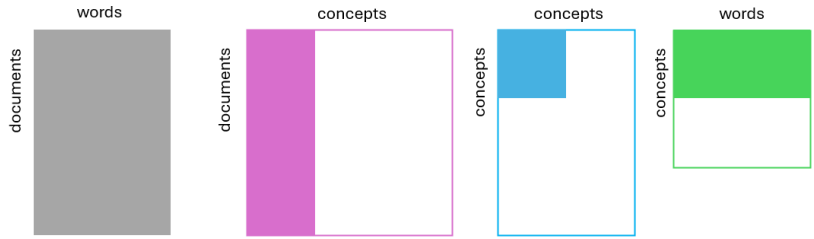
در تعبیه متون باید تلاش کنیم حداکثر اطلاعات مفید را استخراج کنیم و از استخراج اطلاعات کم ارزش پرهیز کنیم. در انتخاب کلمات کوله کلمات نیز باید به کلمات کلیدی که در مفهوم متن تاثیر گذار

[۶] Embed
[۷] Bag of Words

هستند توجه کنیم؛ به عنوان مثال حروف ربط و یا به طور کلی کلمات پرتکرار مانند افعال اسنادی از اهمیت کمتری برخوردار هستند.

Latent Semantic Indexing یا به اختصار **LSI**، یک الگوریتم برای استخراج رابطه پنهان معنایی میان واژه‌های یک متن است. این الگوریتم کاربردهای متعددی در بازیابی اطلاعات [۸] به ویژه در بهینه سازی موتور جست و جو [۹] دارد.

در این روش ماتریس عددی حاصل از تعبیه کوله کلمات را با استفاده از تجزیه مقادیر تکین [۱۰] به ماتریس هایی تجزیه می کنیم که هر یک بیانگر روابط دوتایی مفهوم با کلمات و متون است. با این کار فضای کلمات و متون را به فضایی از مفاهیم نگاشت می کنیم که پایه های این فضا، بردارهای تکین ماتریس متون-کلمات است.



$$A_{d\times n} = U_{d\times d}\ \Sigma_{d\times n}\ V^{T}_{n\times n} \quad \text{Full SVD}$$

$$A_{d\times n} \approx \widetilde{U}_{d\times c}\ \widetilde{\Sigma}_{c\times c}\ \widetilde{V}^{T}_{c\times n} \quad \text{Truncated SVD}$$

با تجزیه ماتریس متون-کلمات، ۳ ماتریس زیر حاصل می شود.

- متون – مفاهیم : این ماتریس بیانگر شدت هر مفهوم در هر متن است.
- مفاهیم – مفاهیم : این ماتریس **قطری** بیانگر شدت هر مفهوم در تمامی متون است.
- مفاهیم – کلمات : این ماتریس بیانگر شدت هر مفهوم در هر کلمه است.

[۸] Information Retrieval
[۹] Search Engineering Optimization
[۱۰] Singular Value Decomposition

پس از انجام این تجزیه، هر سطر از ماتریس متون-مفاهیم یک نمایش جدید برای آن متن بر حسب مفاهیم است که به آن نمایش فضای نهان [۱۱] گفته می‌شود. هر ستون از ماتریس مفاهیم-کلمات نیز یک نمایش جدید برای آن کلمه در فضای نهان و همان بردار تعبیه [۱۲] کلمه می‌باشد.

توجه کنید پس از تجزیه، تعدادی از درایه‌های ماتریس مفاهیم-مفاهیم که از یک آستانه کمتر می‌باشند را صفر در نظر می‌گیریم. این مفاهیم چون در تمامی متون کم تکرار بوده‌اند، ماشین درک درستی از آن‌ها ندارد و درنتیجه نمی توان درباره وجود یا عدم وجود آنها در هر کلمه یا متن ابراز نظر کرد.

در تصویر قبل نیز به همین دلیل، ماتریس مفاهیم-مفاهیم به جای $d \times n$ ،دارای ابعاد کاهش یافته $c \times c$ می باشد.

به این روش، تجزیه مقادیر تکین برش خورده یا Truncated SVD می‌گویند. این روش یکی از مشتقات SVD است که در آن با کاهش مرتبه ماتریس مقادیر تکین و به تبع، ماتریس‌های بردارهای تکین سمت راست و چپ، **تخمین** مرتبه پایینی از ماتریس اصلی خواهیم داشت.

[۱۱] Latent Space Representation
[۱۲] Embedding Vector

سوالات مفهومی

۱. از روش های ارزش گذاری کلمات در متون، روش های TF [۱۳] و IDF [۱۴] را توضیح دهید. چرا مطالعه مستقل این دو ارزش ممکن است گمراه کننده باشد؟

۲. یکی از روش های تعیین آستانه [۱۵] مقادیر تکین برای کاهش مرتبه در Truncated SVD، رویکرد نقطه زانویی [۱۶] است. این رویکرد را شرح دهید.

۳. نحوه محاسبه خطای بازسازی [۱۷] در Truncated SVD را بیان کنید.

۴. دو معیار شباهت بردارهای عددی، Cosine Similarity و Euclidean Distance را ضمن ارائه روابط ریاضی، توضیح دهید. بیشینه و کمینه این دو معیار را تفسیر کنید.

۵. ابتدا درباره استاندارد کردن [۱۸] بردارهای عددی جست و جو کنید. سپس ضمن ارائه رابطه ریاضی، درباره لزوم استاندارد کردن ماتریس داده‌ها، پیش از تجزیه آن بحث کنید.

۶. در برخورد با ماتریس های بزرگ، یکی از روش های تجزیه کارآمد، Randomized SVD است. ضمن ارائه شبه کد این الگوریتم، نحوه عملکرد و کاربردهای آن را تشریح کنید.

---

[۱۳] Term Frequency
[۱۴] Inverse Document Frequency
[۱۵] Threshold
[۱۶] Elbow Point
[۱۷] Reconstruction Error
[۱۸] Standardization

پیاده سازی

یک دادگان [۱۹] شامل ۲۲۲۵ خبر کوتاه متنی به زبان انگلیسی و با موضوعات مختلف در اختیارتان قرار گرفته است. در این بخش با پیاده سازی تکنیک LSI آن را تحلیل می کنیم.

۷. فایل dataset.csv شامل متن و برچسب موضوعی خبرها می‌باشد. این فایل را بخوانید و متن‌ها را برای تحلیل کلمه‌ای پیش پردازش کنید. این پیش پردازش شامل حذف علائم نگارشی و کوچک کردن تمامی حروف می‌باشد.

۸. از بین تمامی متون ۳۰ کلمه پر تکرار را بیابید و نمودار ستونی فراوانی آنها را رسم کنید. آیا دانستن تعداد این کلمات در یک خبر، کمکی به درک موضوع آن خبر می‌کند؟

۹. نمودار ابر کلمات [۲۰] کلی متون را رسم کنید. تفسیر این نمودار و معیار انتخاب کلمات برگزیده در آن را تحقیق کنید و به اختصار شرح دهید.

۱۰. تعدادی کلمه از بین این متون با همان روش استفاده شده در نمودار ابر کلمات استخراج شده است و در فایل words.csv به ترتیب حروف الفبا در اختیارتان قرار گرفته است. کوله کلمات را با استفاده از این کلمات روی متون اجرا کنید. درواقع باید تعداد تکرار هر کلمه در هر متن را بشمارید و ماتریس کوله کلمات را تشکیل دهید. ابعاد این ماتریس را گزارش کنید.

۲۰۰۰ ردیف اول داده‌ها را از باقی ردیف‌ها جدا کنید. ۲۲۵ ردیف باقی مانده را به عنوان داده‌های آزمون [۲۱] کنار بگذارید و تا سوال آخر از آنها استفاده نکنید. موارد خواسته شده در سوالات پیش رو را تنها روی ۲۰۰۰ ردیف اول اعمال کنید.

۱۱. بردارهای کوله کلمات تشکیل شده را استاندارد کنید و سپس تجزیه SVD را روی ماتریس بدست آمده اعمال کنید. ابعاد ماتریس های بدست آمده را گزارش کنید.

۱۲. با رسم نمودار لازم، یک آستانهٔ مناسب برش مقادیر تکین برای کاهش مرتبه پیشنهاد دهید. بر این اساس، با برش ماتریس های موجود، Truncated SVD، را اجرا کنید و خطای بازسازی را گزارش کنید.

---

[۱۹] Dataset
[۲۰] Word Cloud
[۲۱] Test set

۱۳. تابع Randomized SVD را **پیاده سازی** کنید. این تابع ماتریس و رتبه فضای مقصد را دریافت می کند و بر اساس شبه کدی که ارائه کردید، ماتریس ورودی را تجزیه کرده و ماتریس های مقادیر تکین و بردارهای تکین سمت راست و چپ را خروجی می دهد.

۱۴. تابع Randomized SVD را با رتبه متناظر با آستانه پیشنهادی ، بر روی ماتریس کوله کلمات اجرا کنید و خطای بازسازی آن را با Truncated SVD مقایسه کنید. اگر دادگان ما شامل تمام مجموعه متون موجود در اینترنت بود، کدام روش تجزیه را پیشنهاد می‌کردید؟

۱۵. برای هرکدام از مولفه‌های حاصل از Truncated SVD ، ۵ کلمه‌ای که بیشترین اندازه را در آن راستا دارند بیابید و نمایش دهید. از روی کلمات بدست آمده سعی کنید، معنای نهانی که برخی از مولفه‌ها در خود دارند را حدس بزنید.

۱۶. با استفاده از معیار های Cosine Similarity و Euclidean Distance شباهت جفت کلمات زیر در فضای نهان را بسنجید و تفسیر خود از نتایج را بیان کنید.

(mobile, technology) (director, film) (win, won)

(play, game) (play, law) (government, music)

۱۷. با توجه به اینکه اکنون برای هر متن نیز یک بردار در فضای نهان تشکیل شده است که ترکیب مفاهیم موجود در متن را نشان می‌دهد، معیار Cosine Similarity متنی که شماره آن با ۳ رقم آخر شماره دانشجویی شما یکسان است را برای تک تک کلمات داده شده محاسبه کنید و به صورت نمودار ستونی عریضی نمایش دهید. همچنین تعداد تکرار هر یک از کلمات در این متن را نیز زیر آن نمایش دهید و این دو نمودار را مقایسه کنید.

۱۸. چنانچه بخواهیم در یک پایگاه داده متنی، کلمهٔ technology را جستجو کنیم، استفاده از فضای نهان چگونه می‌تواند در مرتبط تشخیص دادن متنی که در آن از کلمهٔ technology استفاده نشده اما کلمات mobile و digital در آن وجود دارد تاثیر گذار باشد؟ جستجو در فضای نهان به جای فضای کوله کلمات چه مزیتی از نظر هزینه محاسباتی دارد؟

۱۹. ستون دوم مجموعه داده‌هایی که در اختیارتان قرار گرفته است برچسب موضوعی متون را نشان می‌دهد که از ۰ تا ۴ کد گذاری شده‌اند.

[0: Politics, 1: Sport, 2: Technology, 3: Entertainment, 4: Business]

برای متون متعلق به هر دسته بندی، بردار فضای نهان میانگین را بدست آورید و آنها را روی یک نقشه

گرمایی [۲۲] نمایش دهید. روشی پیشنهاد دهید که بتوان به وسیله نمایش فضای نهان، متون را برچسب گذاری کرد و دسته بندی آنها را تشخیص داد.

روش پیشنهادی باید با در نظر گرفتن این حقیقت باشد که ما مفهوم هرکدام از مولفه های فضای نهان را به صورت معین نمی‌دانیم و با وجود اینکه در سوال ۱۵ سعی کردید نسبت آنها شهود پیدا کنید، این امر در کاربرد واقعی مقیاس پذیر نیست.

۲۰. روش برچسب‌گذاری پیشنهادی‌تان را پیاده سازی کرده و دقت آن را روی **داده‌های آزمون** بسنجید. همچنین دقت برچسب گذاری هریک از دسته‌های مختلف داده‌ها را به صورت مجزا بدست آوردید.

$$Accuracy = \frac{CorrectPredictions}{TotalPredictions}$$

[۲۲] Heat Map

فصل ۲

نکات کلی

لطفا قبل از بارگذاری به موارد زیر توجه نمایید، در صورت عدم رعایت هر یک از موارد زیر پروژه شما تصحیح نخواهد شد.

عکس از: istockphoto

- هیچگونه شباهتی در انجام پروژه بین افراد مختلف پذیرفته نمی شود. در صورت کشف هر گونه تقلب مطابق قوانین درس با افراد خاطی برخورد خواهد شد.
- گزارش شما مهم ترین معیار ارزیابی خواهد بود؛ در نتیجه لطفا زمان کافی را برای رسم جامع نمودارها و گزارش نتایج خواسته شده اختصاص دهید.
- کد پروژه باید به زبان های Python و یا MATLAB نوشته شود. کد Python حتما به فرمت ipynb. و کد MATLAB حتما در محیط MATLAB Live Editor تحویل داده شود.
- کد پروژه به همراه گزارش آن به صورت یک فایل تجمیع و با نام

  P2_<LastName>_<StdID>.zip ارسال گردد.
- استفاده از مراجع با ارجاع به آنها بلامانع می باشد.
- مهلت تحویل پروژه تا ۱۴۰۳/۰۴/۲۰ می باشد و امکان ارسال با تاخیر وجود ندارد!
- پس از مطالعه کامل پروژه، در صورت ابهام پیشنهاد می شود از طریق تالار گفتگوی صفحه درس و یا گروه مجازی درس سوال خود را مطرح کنید، در غیر اینصورت با دستیاران آموزشی پروژه، عرفان عسگری یا فردین عباسی در ارتباط باشید.