

به نام خدا



دانشگاه تهران
دانشکدگان فنی
دانشکده مهندسی برق و کامپیوتر



درس بازیابی هوشمند اطلاعات

تمرین ۴

استاد درس: خانم دکتر آزاده شاکری

سرپرست دستیاران آموزشی: سمانه پیمانی راد

طراح تمرین: امیرحسین صفریان

آذر ماه ۱۴۰۴

فهرست

۴	مقدمه
۵	سوالات عملی (۶۰ نمره).....
۶	سوال ۱: استخراج word association
۶	مجموعه‌داده
۶	پیش‌پردازش و آماده‌سازی پیکره
۶	استخراج روابط Paradigmatic با pseudo-document
۷	استخراج روابط Syntagmatic Mutual Information
۸	سوال ۲: کار با Pretrained Word Embedding
۸	بارگذاری مدل GloVe
۸	همسایه‌های معنایی در فضای Embedding و تحلیل با PCA
۹	ساخت Document Embedding برای Cranfield و Query Embedding
۱۰	سوال ۳: پیاده‌سازی مدل‌های Learning to Rank
۱۰	مجموعه‌داده
۱۰	ساخت Feature Vector برای هر زوج (Query, Document)
۱۰	پیاده‌سازی معیارهای MAP و Precision.NDCG.DCG
۱۱	آموزش و تحلیل مدل Pointwise LTR
۱۱	آموزش و تحلیل مدل Pairwise LTR
۱۱	آموزش و تحلیل مدل Listwise LTR
۱۲	سوالات تئوری (۴۰ نمره).....
۱۴	سوال ۱ – روابط Word Association
۱۴	سوال ۲ – تحلیل Mutual Information
۱۴	سوال ۳ – ماهیت فضای Word Embedding
۱۴	سوال ۴ – مقایسه Skip-gram و CBOW
۱۵	سوال ۵ – رویکردهای Pointwise و Pairwise LTR در Listwise
۱۵	سوال ۶ – سازگاری LambdaRank با معیارهای رتبه‌بندی
۱۶	سوال ۷ – شباهت واژگان با ویژگی‌های Interpretable
۱۷	ملاحظات (حتیا مطالعه شود)
۱۸	استفاده مسئولانه از هوش مصنوعی

۱۸	۱. هدف و اصول کلی
۱۸	۲. استفاده مجاز از LLMها
۱۹	۳. استفاده غیرمجاز از LLMها
۱۹	۴. مستندسازی
۱۹	۵. آمادگی ارائه شفاهی
۱۹	۶. پیامدهای تخلفات
۱۹	۷. موارد تکمیلی
۲۰	۸. اظهارنامه

اهداف تمرین

- درک عملی از انواع روابط واژگانی و مشاهده اینکه چگونه می‌توان آن‌ها را از داده واقعی استخراج کرد.
- آشنایی با ایده نمایش برداری واژه‌ها (word embeddings) و بررسی تفاوت آن با ویژگی‌های معنایی قابل تفسیر.
- ساخت نمایش برداری برای کوئری و سند و استفاده از شباهت‌های برداری برای تحلیل و بازیابی اسناد.
- آشنایی با مفهوم ویژگی‌سازی در IR و اینکه چگونه ویژگی‌های مختلف (آماری، معنایی، مبتنی بر embedding) بر کیفیت رتبه‌بندی تأثیر می‌گذارند.
- درک رویکردهای مختلف Learning to Rank و تفاوت نگاه Pairwise ، Pointwise و .Ranking به مسئله Listwise .
- توانایی ارزیابی کیفیت سیستم‌های بازیابی با استفاده از معیارهای استانداردی مثل NDCG و MAP .
- پیوند دادن مباحث نظری درس با تجربه عملی بر روی یک مجموعه داده واقعی.

نکات قابل توجه در هنگام پاسخ به سؤالات:

- تمامی کدهای نوشته شده باید قابلیت اجرای مجدد داشته باشند. به تفسیرهایی که بدون آزمایش و صرفاً به صورت فرضی بیان گردند نمره‌ای تعلق نمی‌گیرد.
- نمره اصلی تمرین مربوط به کیفیت و درستی تفسیرهای ارائه شده است.
- بهتر است از نمودارها و کشف نمونه‌های مرتبط از اسناد و پرس‌وجوها برای افزایش کیفیت تفسیرها استفاده گردد.
- بدیهی است که حجم تمرین معیار نمره‌ی شما نیست، بلکه صحت انجام آزمایش‌ها و کیفیت تفسیرهای شما مهم است.
- توصیه می‌شود انجام تمرین را به روزهای آخر موکول نکنید! اجrai آزمایش‌ها و پیدا کردن پارامتر بهینه زمان بر بوده و تفسیر هر کدام نیازمند تحلیل و بررسی است.

سوالات عملی (۶۰ نمره)

این تمرین شامل سه بخش است:

در بخش اول، روی پیکره 20 newsgroups روابط واژگانی از نوع paradigmatic

و syntagmatic را استخراج می‌کنید.

در بخش دوم، از یک pretrained word embedding برای تحلیل همسایه‌های معنایی و همچنین

ساخت ویژگی‌های document و query برای embedding استفاده می‌کنید.

در بخش سوم، از مجموعه 1400cranfield مدل‌های Learning to Rank برای پیاده‌سازی در سه

رویکرد Listwise و Pairwise Pointwise استفاده می‌کنید و عملکرد آن‌ها را با معیارهای

رتبه‌بندی مانند NDCG و MAP مقایسه و تحلیل می‌کنید.

پاسخ سوالات این بخش را در همان نوتبوک تمرین خود (*IIR-CA4-Code.ipynb*) بنویسید.

سوال ۱: استخراج WORD ASSOCIATION ها

مجموعه داده

در این بخش از پیکره 20 newsgroups استفاده می کنید. این مجموعه شامل حدود ۲۰ هزار متن از گروههای خبری مختلف (sci, rec, comp و ...) است. نسخه آماده این مجموعه به صورت خام یا از طریق تابع fetch_20newsgroups در اختیار شما قرار می گیرد.

پیش پردازش و آماده سازی پیکره

ابتدا باید داده های 20 newsgroups را بارگذاری و پیش پردازش کنید. همه متن ها را lowercase کنید، punctuation ساده (مثل نقطه، ویرگول، علامت سؤال و ...) را حذف کنید و متن را به کلمات token کنید. سپس یا هر متن کامل، یا هر جمله را یک segment در نظر بگیرید.

در پایان این زیربخش، در گزارش بنویسید:

- چه تصمیم هایی برای پیش پردازش گرفتید (حذف stopword، حداقل فراوانی واژه، ...)
- تعداد segments ها و اندازه نهايی vocabulary چقدر شده است.
- و در یک پاراگراف کوتاه تحلیل کنید که اين نوع پیش پردازش چه مزایا و محدودیت هایی برای کارهای بعدی شما دارد.

استخراج روابط Paradigmatic با pseudo-document

در این زیربخش باید روابط paradigmatic را برای تعدادی واژه محاسبه کنید. ابتدا یک مجموعه نسبتاً بزرگتر از کلمات هدف انتخاب کنید (مثلاً ۱۵-۲۰ واژه نسبتاً پر تکرار و معنادار). برای هر واژه، همه بافت هایی را که آن واژه در آن ظاهر می شود جمع آوری کنید: هر بار که واژه هدف را در یک segment

می‌بینید، کلمات موجود در پنجره‌ای با اندازه k (مثلًاً ۲ یا ۵) در اطراف آن واژه را در نظر می‌گیرید. تمام این کلمات با هم یک context pseudo-document برای آن واژه می‌سازند.

برای هر pseudo-document یک بردار با وزن دهی شبیه BM25 (یا بردار با وزن دهی شبیه TF-IDF) بسازید و سپس cosine similarity بین بردارهای pseudo-document واژه‌های مختلف را محاسبه کنید. به این ترتیب ای از شbahet paradigmatic notion ای از شbahet paradigmatic notion بین واژگان به دست می‌آورید.

برای تحلیل، از میان واژه‌های هدف خودتان ۵ واژه دلخواه انتخاب کنید. برای هر یک از این ۵ واژه، نزدیکترین واژه‌ها از نظر cosine similarity در فضای pseudo-document را (مثلًاً ۱۰ همسایه نزدیک) پیدا کنید و در گزارش، برای هر واژه یک جدول حاوی همسایه‌ها بیاورید. سپس برای هر یک از این ۵ واژه در چند جمله توضیح دهید که این همسایه‌ها چه رابطه‌ای با واژه اصلی دارند و آیا از نظر شما واقعًا «همنوع» یا «قابل جایگزینی در context» هستند یا نه.

در ادامه، روی بردارهای pseudo-document همان ۵ واژه و تعدادی از همسایه‌های آن‌ها، PCA را به ۲ بعد اعمال کنید و نقاط را روی یک نمودار دو بعدی رسم کنید (هر نقطه یک واژه، همراه با label آن). در گزارش توضیح دهید که آیا کلماتی که از نظر معنایی به هم نزدیک‌اند در این نمودار نیز نزدیک به هم قرار گرفته‌اند یا خیر، و اگر خوش‌هایی دیده می‌شود آن‌ها را توصیف کنید.

استخراج روابط Syntagmatic با Mutual Information

در این زیربخش باید روابط syntagmatic را اندازه‌گیری کنید. برای هر واژه مورد نظر، یک متغیر دودویی روی segment‌ها تعریف کنید: مقدار این متغیر ۱ است اگر واژه در آن segment ظاهر شود و ۰ در غیر این صورت. برای جفت‌هایی از واژه‌ها که هر دو حداقل در N segment حضور دارند (مثلًاً $N = 20$)، فراوانی حضور هر واژه و حضور همزمان هر دو را بشمارید و با یک smoothing کوچک، احتمال‌ها را تخمین بزنید و مقدار Mutual Information را محاسبه کنید.

پس از محاسبه MI برای تعداد زیادی جفت‌واژه، دوباره بر همان ۵ واژه انتخابی تمرکز کنید. برای هر یک از این ۵ واژه، واژگانی را که بالاترین MI را با آن دارند (مثلًاً ۱۰ همسایه syntagmatic برتر) استخراج کنید و در گزارش جدول آن‌ها را بیاورید. سپس برای هر واژه، تفاوت بین لیست همسایه‌های paradigmatic (زیربخش قبل) و syntagmatic را در چند جمله توضیح دهید.

سوال ۲: کار با PRETRAINED WORD EMBEDDING

بارگذاری مدل GloVe

در این زیربخش باید فایل GloVe 6B, 100d را بارگذاری کنید و با خواندن آن، یک map از واژه به بردار ۱۰۰ بعدی بسازید. در گزارش بنویسید که اندازه واژگان موجود در فایل چقدر است و تقریباً چه نسبتی از واژه‌های پیکره newsgroups20 شما در این واژگان پوشش داده می‌شوند.

همسايه‌های معنایي در فضای Embedding و تحليل با PCA

با استفاده از همان ۵ واژه انتخابی بخش اول (به شرط حضور در واژگان GloVe)، همسایه‌های آن‌ها را در فضای embedding پیدا کنید. برای هر واژه، cosine similarity آن را با سایر واژگان محاسبه کرده و نزدیک‌ترین کلمات (مثلاً ۱۰ همسایه) را استخراج کنید. در گزارش، برای هر واژه یک جدول از این همسایه‌ها نشان دهید و درباره نوع رابطه آن‌ها (متراff، مرتبط موضوعی، همآیند و...) و تفاوت‌ها با همسایه‌های syntagmatic و paradigmatic صحبت کنید.

سپس همسایه‌های همان ۵ واژه و تعدادی از همسایه‌هایشان را با PCA به ۲ بعد کاهش دهید و آن‌ها را روی نمودار دو بعدی رسم کنید. نام هر واژه را کنار نقطه مربوط به آن قرار دهید. در گزارش، به صورت کيفی تحليل کنید که آیا کلماتی که از نظر شما از نظر معنا به هم نزدیک‌اند در اين فضای embedding دو بعدی نيز به هم نزدیک شده‌اند، آیا خوش‌های قابل توجهی می‌بینید و اين خوش‌ها چه تفاوتی با نمودارهای به دست آمده در بخش ۱ دارند.

ساخت Cranfield Document Embedding و Query Embedding و تحلیل

در این زیربخش embedding ها را وارد دنیای Cranfield می کنید. ابتدا اسناد و کوثری های Cranfield را مشابه بخش ۱ پیش پردازش کنید (lowercase، حذف punctuation و tokenization) برای هر document، واژه هایی را که در واژگان GloVe هستند انتخاب کنید و document embedding را به صورت میانگین بردارهای آن ها بسازید. برای هر query نیز همین کار را انجام دهید و query embedding بسازید.

برای هر زوج (query, document) یک ویژگی embedding تعريف کنید که برابر است با cosine similarity بین document embedding و query embedding. این مقدار را برای همه زوج های دارای برچسب ربط در train/val/test محاسبه و ذخیره کنید.

در گزارش، چند query نمونه انتخاب کنید و برای هر query یکی دو document مرتبط و یکی دو document نامرتبط پیدا کنید. برای این اسناد مقدار embedding similarity را در جدول بنویسید و در چند جمله تحلیل کنید که آیا این similarity مطابق انتظار شماست یا خیر؛ یعنی آیا معمولاً برای اسناد مرتبط بالاتر است و اگر موارد خلاف مشاهده کردید، حدس بزنید چرا embedding ممکن است این طور رفتار کرده باشد.

سوال ۳: پیاده‌سازی مدل‌های LEARNING TO RANK

مجموعه داده

در این بخش از مجموعه کلاسیک Cranfield 1400 استفاده می‌کنید. این مجموعه شامل حدود ۱۴۰۰ چکیده مقاله در حوزه آیرودینامیک، بیش از ۲۰۰ query و relevance judgment مشخص بین queryها و اسناد است. نسخه این داده همراه با یک تقسیم‌بندی مشخص به train و validation (براساس Query ID) در اختیارتان قرار داده می‌شود.

می‌توانید از این آدرس به این مجموعه داده دسترسی پیدا کنید:

https://ir.dcs.gla.ac.uk/resources/test_collections/cran.tar.gz

ساخت Feature Vector برای هر زوج (Query, Document)

در این زیربخش باید برای هر زوج (query, document) در مجموعه‌های train و validation و test یک بردار ویژگی بسازید. حداقل ویژگی‌هایی که باید استفاده کنید عبارت‌اند از: cosine similarity، TF/IDF، تعداد يا نسبت کلمات مشترک، طول document (یا \log آن)، طول query و embedding-based similarity که در بخش ۲.۴ تعریف کرده‌اید.

اگر مایل بودید می‌توانید ویژگی‌های دیگری (مثلًاً میانگین IDF واژه‌های query، ، یا BM25 یا ویژگی‌های مبتنی بر Word Association) نیز اضافه کنید. در گزارش، یک جدول کوچک تهیه کنید که نام هر feature را بیاورد و در یک جمله توضیح دهد آن feature چه چیزی را اندازه می‌گیرد و چرا ممکن است برای رتبه‌بندی مفید باشد.

پیاده‌سازی معیارهای MAP و Precision، NDCG، DCG

برای محاسبه معیارهای ارزیابی نظیر Precision@k، NDCG@k، DCG@k و MAP می‌توانید از فایل evaluation.py که در تمرین‌های قبلی در اختیار شما قرار داده شده است کمک بگیرید.

آموزش و تحلیل مدل Pointwise LTR

در این مرحله باید یک مدل Pointwise Learning to Rank بسازید. هر زوج (query, document) یک نمونه آموزشی مستقل است؛ feature vector آن را در زیربخش ۳.۲ ساخته‌اید و label آن همان درجه ربط است. می‌توانید برای این بخش از مدل‌ها یا کتابخانه‌های آماده یادگیری ماشین (مانند پیاده‌سازی‌های موجود در scikit-learn) استفاده کنید.

از یک مدل مناسب (مثلًا Decision Tree یا LinearRegression) استفاده کنید و آن را روی train آموزش دهید. سپس روی test، برای هر query نمره predicated برای اسناد را محاسبه کرده و براساس این نمره آن‌ها را رتبه‌بندی کنید. با استفاده از کد زیربخش ۳.۳ NDCG@10 و P@10 و MAP را روی query‌ها محاسبه کنید.

در گزارش، علاوه بر نوشتمن مقادیر کلی NDCG@10 و P@10 و MAP، چند query را انتخاب کنید که مدل روی آن‌ها خوب یا بد عمل کرده است. برای هر مورد، بخشی از لیست رتبه‌بندی‌شده اسناد را نشان دهید و در چند جمله تحلیل کنید که چرا مدل ممکن است روی آن query خاص موفق یا ناموفق بوده باشد (مثلًا به خاطر کوتاهی/بلندی query، فقر یا غنای feature‌ها، یا رفتار embedding).

آموزش و تحلیل مدل Pairwise LTR

در این زیربخش یک مدل Pairwise Learning to Rank آموزش می‌دهید. برای هر query در مجموعه train، جفت‌سندهایی را که یکی از دیگری مرتبط‌تر است استخراج کرده و از آن‌ها برای آموزش یک مدل pairwise (مانند SVM Ranking) استفاده کنید.

در این بخش نیز می‌توانید از پیاده‌سازی‌های آماده موجود در کتابخانه‌های یادگیری ماشین استفاده کنید. پس از آموزش، روی test مانند مدل pointwise برای هر سند یک score تعريف کنید (مثلًا با استفاده از $w^T x$) و اسناد را بر اساس این score رتبه‌بندی کنید و NDCG@10 و P@10 و MAP را محاسبه کنید. در گزارش، نتایج مدل pairwise را در کنار مدل pointwise در یک جدول قرار دهید و تحلیل کنید که کدام model بهتر عمل کرده است.

آموزش و تحلیل مدل Listwise LTR

در این بخش باید با استفاده از یک کتابخانه آماده (مانند XGBoost یا LightGBM) یک مدل Listwise Learning to Rank بازسازی کرد. از فرمت مناسب کتابخانه برای group کردن داده بر اساس Query ID استفاده کنید و از objective رتبه‌بندی (نظیر LambdaMART) بهره بگیرید. می‌توانید با استفاده از داده validation چند پارامتر ساده مثل تعداد درخت‌ها، حداکثر عمق و learning rate را تنظیم کنید.

پس از آموزش، روی test، برای هر query score اسناد را به دست بیاورید و P@10 و NDCG@10 و MAP را محاسبه کنید. در گزارش،

- ۱) توضیح دهید چه کتابخانه و چه objective ای استفاده کرده‌اید،
- ۲) خروجی نهایی 10 NDCG@10 و P@10 و MAP مدل pairwise و pointwise را کنار listwise قرار دهید،
- ۳) برای چند query نمونه لیست‌های رتبه‌بندی سه مدل را مقایسه کنید و در چند جمله تحلیل کنید که کدام روش و به چه دلیل عملکرد بهتری داشته است.

سوالات تئوری (۴۰ نمره)

باتوجه به مباحث تدریس شده سوالات زیر را حل کنید. پاسخ سوالات این بخش باید در قالب مربوط به سوالات تئوری (IIR-CA4-Theory) بنویسید.

نکات سوالات تئوری:

- صرفا به جواب نهایی نمره تعلق نمی‌گیرد و راه حل نوشته شده بخش اصلی نمره‌ی شماست.
- می‌توانید جهت اطمینان از محاسبات خود از کدنویسی استفاده کنید، اما باید راه حل به طور کلی به صورت ریاضی در گزارش تمرين آمده باشد و صرفا قراردادن کد راه حل باعث کسر نمره شما می‌شود.

سؤال ۱ – روابط WORD ASSOCIATION

الف) تفاوت دو نوع رابطه *syntagmatic* و *paradigmatic* را توضیح دهید.

ب) هر کدام چه نوع شباهتی را در زبان مدل می‌کنند؟

پ) برای هر کدام یک مثال واقعی بزنید.

سؤال ۲ – تحلیل MUTUAL INFORMATION

الف) توضیح دهید چرا *Mutual Information* برای تشخیص واژه‌هایی که در یک موضوع مشترک مشارکت می‌کنند مناسب است، حتی اگر شباهت معنایی نداشته باشند.

ب) یک مثال از واژه‌هایی بدهید که *MI* بالایی دارند ولی ارتباط معنایی مستقیمی ندارند.

سؤال ۳ – ماهیت فضای WORD EMBEDDING

توضیح دهید که *word embedding* ها چه نوع اطلاعاتی را در بردارهای خود رمزگذاری می‌کنند. سپس مفهوم «خطی بودن» در فضای *embedding* را توضیح دهید و بیان کنید چرا رابطه‌هایی مثل:

$$\text{king} - \text{man} + \text{woman} \approx \text{queen}$$

در فضای *embedding* مشاهده می‌شود.

سؤال ۴ – مقایسه SKIP-GRAM و CBOW

در مدل‌های *Word2Vec*، تفاوت‌های دو روش *Skip-gram* و *CBOW* را توضیح دهید.

در پاسخ خود مشخص کنید:

- ورودی هر مدل چیست؟
- خروجی هر مدل چیست؟
- کدام مدل برای کلمات پر تکرار بهتر است و چرا؟
- کدام مدل برای کلمات کم تکرار عمل می‌کند و چرا؟

سؤال ۵ – رویکردهای LTR در LISTWISE و PAIRWISE و POINTWISE

در یادگیری به روش رتبه‌بندی (*Learning to Rank*)، سه رویکرد زیر را مقایسه کنید:

- *Pointwise*
- *Pairwise*
- *Listwise*

برای هر کدام بیان کنید:

- چه چیزی را مدل می‌کند (*label* چیست)
- تابع هدف چه فرمی دارد؟
- چه مزایا و محدودیت‌هایی دارد؟
- کدام رویکرد با معیارهایی مثل $NDCG$ سازگارتر است و چرا؟

سؤال ۶ – سازگاری LAMBNDARANK با معیارهای رتبه‌بندی

در الگوریتم *LambdaRank* توضیح دهید.

- مفهوم λ -gradient چیست؟
- چرا تغییر رتبه سندهای بالای لیست نسبت به سندهای پایین‌تر gradient بزرگ‌تری ایجاد می‌کند؟
- چگونه این روش به طور مؤثر با معیارهایی مانند $NDCG$ سازگار می‌شود؟

سؤال ۷ – شباهت واژگان با ویژگی‌های INTERPRETABLE

در جدول زیر تعدادی واژه و چند ویژگی معناشناختی *interpretable* آورده شده‌اند (۰ و ۱ نشان‌دهنده نبودن/بودن ویژگی است):

Word	<i>is_animal</i>	<i>has_tail</i>	<i>can_fly</i>	<i>is_fruit</i>	<i>is_vehicle</i>	<i>is_food</i>
<i>dog</i>	۱	۱	۰	۰	۰	۰
<i>cat</i>	۱	۱	۰	۰	۰	۰
<i>bird</i>	۱	۰	۱	۰	۰	۰
<i>apple</i>	۰	۰	۰	۱	۰	۱
<i>banana</i>	۰	۰	۰	۱	۰	۱
<i>car</i>	۰	۰	۰	۰	۱	۰

الف) برای هر واژه، بردار ویژگی آن را یادداشت کنید.

ب) برای جفت‌های زیر *cosine similarity* را محاسبه کنید:

۱. *cat, dog*

۲. *bird, dog*

۳. *banana, apple*

۴. *apple, dog*

۵. *car, apple*

پ) نتایج به دست‌آمده را تحلیل کنید:

- کدام جفت‌ها بیشترین شباهت را دارند؟

- چه نوع شباهتی در این فضای *interpretable* اندازه‌گیری شده است؟

- این شباهت‌ها چگونه با شباهتی که از *word embedding* انتظار داریم تفاوت دارد؟

- به نظر شما کدام نوع نمایش برای مدل‌کردن *paradigmatic* مناسب‌تر است و کدام برای *syntagmatic*؟

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان IIR-CA4-StudentID تحویل داده شود.

- خوانایی و دقیق بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرين‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد. **دقیق** کنید که حتما گزارشات خود را در قالب ارائه شده برای تحویل تکالیف که در سامانه برای شما بارگذاری شده است ارسال بفرمایید.
- کدهای نوشته شده برای هر بخش را با نام مناسب مشخص کرده و به همراه گزارش تکلیف ارسال کنید. همه‌ی کدهای پیوست گزارش بایستی قابلیت اجرای مجدد داشته باشند. در صورتی که برای اجرا مجدد آن‌ها نیاز به تنظیمات خاصی می‌باشد بایستی تنظیمات مورد نیاز را نیز در گزارش خود ذکر کنید. **دقیق** کنید که تمامی کدها باید توسط شما اجرا شده باشند و نتایج اجرا در فایل کدهای ارسالی مشخص باشد. به کدهایی که نتایج اجرای آن‌ها در فایل ارسالی مشخص نباشد نمره‌ای تعلق نمی‌گیرد.
- تمرين تا یک هفته بعد از مهلت تعیین شده با تأخیر تحویل گرفته می‌شود. دقیق کنید که شما جماعت برای تمام تکالیف، ۱۴ روز زمان تحویل بدون جریمه دارید که تنها از ۷ روز آن برای هر تمرين می‌توانید استفاده کنید، در صورتی که این ۱۴ روز به اتمام رسیده باشد، به ازای هر روز تأخیر در ارسال تمرين، ۵ درصد جریمه می‌شود.
- توجه کنید این تمرين باید به صورت تک نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرين نیز ممنوع است). در صورت مشاهده تشابه به همه افراد مشارکت کننده، نمره ۵۰- تعلق می‌گیرد و به استاد نیز گزارش می‌گردد.

در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

<mailto:safdarian2000@gmail.com>

استفاده مسئولانه از هوش مصنوعی

۱. هدف و اصول کلی

هدف

- ترویج استفاده اخلاقی و مسئولانه از LLM‌ها (مانند Deepseek، ChatGPT) به عنوان ابزار کمکی
 - اطمینان از مشارکت فعال دانشجویان در تکالیف و درک راه حل‌های آن‌ها
 - حفظ صداقت علمی در عین بهره‌گیری از ابزارهای مدرن هوش مصنوعی
- #### اصول کلی
- تمرين باید نتیجه تلاش و زحمت شخصی شما باشد.
 - باید به تمام بخش‌های تمرين، اعم از پیاده‌سازی و تحلیل نتایج مسلط باشید.
 - تمامی کدها باید توسط خود شما اجرا شده و نتایج قابل مشاهده باشند.
 - تمام مراحل انجام تمرين باید مستند و قابل پیگیری باشد.
 - هرگونه نتیجه‌گیری و تحلیل باید بر اساس درک شخصی شما باشد.
- LLM‌ها ممکن است پاسخ‌های نادرست یا قدیمی تولید کنند، اولویت با مطالب و کارگاه‌های درس است.

موارد ذکر شده در ادامه این سند، به عنوان راهنمایی بیشتر برای انجام تمرين آورده شده‌اند. با این حال، مسئولیت تطبیق کار با اصول کلی فوق بر عهده شماست. توجه داشته باشید که ممکن است مواردی در ادامه ذکر نشده باشند که با اصول کلی ذکر شده در تضاد باشند. در چنین مواردی به تشخیص دستیار آموزشی و دستیار مسئول، شما موظف به پاسخ‌گویی در قبال تمرين خود هستید. عدم رعایت هر یک از اصول فوق می‌تواند منجر به کسر نمره یا عدم پذیرش تمرين شود.

۲. استفاده مجاز از LLM‌ها

شما می‌توانید از LLM‌ها برای موارد زیر استفاده کنید:

- روشن‌سازی مفاهیم (مثال: "خوشبندی DBSCAN چگونه کار می‌کند؟")
- کمک در اشکال‌زدایی (مثال: شناسایی خطاهای گرامری یا منطقی در کد)
- ایده‌پردازی رویکردها (مثال: "روش‌های مدیریت داده‌های missing را پیشنهاد دهید")

الزامات استفاده مجاز:

- ثبت تعاملات اصلی: (به بخش ۴ مراجعه کنید).
- درک راه حل: باید قادر به توضیح هر خط کد یا منطق استفاده شده باشید.

۳. استفاده غیرمجاز از LLMها

اقدامات ممنوع شامل:

- کپی-پیست مستقیم خروجی‌های LLM بدون تغییر
- استفاده از LLM‌ها برای حل اصلی مسائل (مثال: "این سؤال تکلیف را برای من حل کن")
- گرفتن کد از سایر دانشجویان به هر شکل غیر مجاز است، تغییر و پارافریز کردن کد دیگران توسط LLM نیز قابل قبول نیست.
- هرگونه استفاده که منجر به عدم احاطه شما به موضوع تمرین شود.

۴. مستندسازی

ارجاع به مشارکت‌های LLM: افروzen پانویس یا توضیح (مثال: کد با رعایت قوانین به کمک ChatGPT نوشته شده است).

- نیازی به اشتراک گذاری پرامپت‌ها و سابقه چت نیست.
- مستندسازی تک تک تعاملات با هوش مصنوعی هدف این بخش نیست. اشاره کوتاه و کلی در بخش‌های مورد استفاده کافی است. در نظر داشته باشید که مستندسازی به معنای رفع مسئولیت نبوده و باید اصول کلی را رعایت کنید.

۵. آمادگی ارائه شفاهی

آماده دفاع از کار خود باشید: در صورت درخواست دستیار تمرین در بازه زمانی اعلام شده برای ارائه شفاهی، باید:

- رویکرد، کد یا نتایج خود را توضیح دهید.
 - درک مفاهیم کلیدی را نشان دهید (مثلاً چرا یک الگوریتم خاص انتخاب شده است)
- عدم توضیح کافی کار شما ممکن است منجر به جریمه شود (بخش ۶)

۶. پیامدهای تخلفات

- تخلفات جزئی (مثل مستندسازی ناقص): کاهش نمره
- تخلفات عمده (مثل کپی-پیست بدون تغییر): نمره ۵۰- در تکلیف
- تخلفات مکرر: نمره ۵۰- در تکلیف و گزارش به استاد

۷. موارد تکمیلی

- از LLM‌ها به عنوان معلم استفاده کنید، نه پاسخنامه تمرین‌ها: اولویت را به مهارت‌های حل مسئله خود بدهید.

- خروجی‌ها را متقابلاً تأیید کنید: پیشنهادات LLM را با کتاب مرجع درس، اسلایدها و کارگاهها مقایسه کنید.
- از دستیاران آموزشی کمک بگیرید: اگر پاسخ LLM یا نحوه استفاده شما را گیج می‌کند، در ساعت متعارف از دستیاران آموزشی کمک بگیرید.

۸. اظهارنامه

این عبارت را در تکلیف ارسالی خود قرار دهید:

"تأیید می‌کنم که از LLM‌ها مطابق با دستورالعمل‌های بارگذاری شده در سامانه Elearn درس به طور مسئولانه استفاده کرده‌ام. تمام اجزای کار خود را درک می‌کنم و آماده بحث شفاهی درباره آنها هستم."