به نام و یاد او

بهار ۱۴۰۱
وقت امتحان ۱۲۰ دقیقه
دانشکده مهندسی کامپیوتر
امتحان درس بازیابی پیشرفته اطلاعات

قوانین آزمون

- همراه داشتن سه برگه A4 (پشت و رو یا یک رو) از اطلاعات مورد نیاز مجاز است.
- استفاده از تلفن همراه و کامپیوتر در امتحان مجاز نیست.

بخش الف) سوالات مربوط به مفاهیم (۱۲ نمره از ۲۰)

۱۲ مفهوم از ۱۴ مفهوم زیر را به اختصار در حد یک سطر (حداکثر حدودا ۲۰ کلمه) توضیح دهید. اگر به بیش از ۱۲ مورد اشاره شود، فقط ۱۲ مورد اول بررسی می‌شوند. لطفا فقط موارد اصلی را بنویسید تا وقتتان برای سوالات بعد گرفته نشود.

۱. تعریف زبان
۲. تعریف مدل زبانی
۳. فرق lemmatisation و stemming را بیان فرمایید.
۴. منطق کاربردی بودن tf-idf را در بازیابی توضیح دهید (نقش هر کدام از tf و idf به اختصار).
۵. یک مثال از انواع روابط واژگانی و یک مثال توجه به آن در بازیابی اطلاعات را بیان فرمایید.
۶. معناشناسی توزیعی[1] برای واحد کلمه را بیان کنید. یعنی در این نظریه معنای یک کلمه را چه چیز مشخص می‌کند؟
۷. یکی از دلایلی که مدل زبانی، مساله مناسبی برای یادگیری بازنمایی[2] واحدهای متنی است را بیان کنید.
۸. یک مزیت استفاده از روش‌های امبدینگ مدل زبانی بر روش نمایش برداری tf-idf کلمات را در سامانه بازیابی ذکر کنید.
۹. یکی از مزایای اصلی ترنسفرمرها را نسبت به مدل‌های زبانی BiLSTM ذکر فرمایید.
۱۰. یکی از دلایل اهمیت بهبود پرسمان کاربر در سامانه بازیابی را نام ببرید. یک از راههای آن را نیز فقط ذکر کنید.
۱۱. تفاوت اصلی logistic regression و naive Bayes در مدل‌سازی متن ورودی چیست؟
۱۲. یکی از دلایل استفاده از یادگیری رتبه‌بندی[3] در بازیابی را توضیح دهید.
۱۳. یکی از نقش‌های الگوریتم‌های تحلیل لینک مانند page rank در بازیابی چیست؟
۱۴. در کراولینگ پیوسته صفحات وب، دو ملاحظه که در اولویت‌بندی انتخاب آدرس مناسب است رعایت کنیم را بیان کنید.

بخش ب) حل مساله و محاسبه (۴ نمره از ۲۰)

تنها یکی از دو مساله بخش ب را انتخاب کنید و حل بفرمایید.

ب-۱) فاصله کمینه ویرایشی (جریمه تغییر حرف: ۲ و جریمه حذف یا اضافه: ۱)

- ماتریس برنامه‌نویسی پویا برای محاسبه امتیاز تطبیق دو رشته «سنت» و «صنعت» را تشکیل دهید و پر بفرمایید.
- تطبیق دو رشته را زیر هم با مشخص کردن محل فاصله و تغییر تشکیل دهید.
- اگر بدانید که پرسمان کاربر از سامانه دستیار صوتی آمده‌اند، چه طور جریمه تغییر حرف را تغییر می‌دهید؟ (به صورت خاص برای «س» و «ص» بیان بفرمایید)

---

[1] distributional semantics

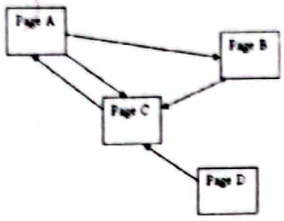
[2] representation

[3] learning to rank

به نام و یاد او

بهار ۱۴۰۱
وقت امتحان ۱۲۰ دقیقه
دانشکده مهندسی کامپیوتر
امتحان درس بازیابی پیشرفته اطلاعات

ب-۲) محاسبه Page-Rank



۱. ماتریس مجاورت احتمالی گراف صفحات زیر را تشکیل دهید (یال‌ها ضرایب یکسان دارند).
۲. ماتریس مجاورت احتمالی گراف صفحات زیر را با فرض حرکت تصادفی با نرخ ۰.۱ محاسبه فرمایید (teleporting rate =۰.۱).
۳. اگر یک محرک تصادفی در نقطه شروع در بردار احتمالی (۰.۲۵،۰.۲۵،۰.۲۵،۰.۲۵) باشد که از چپ به راست احتمال شروع از صفحات ABCD به ترتیب الفبای انگلیسی است. با توجه به ماتریس مجاورت خواسته قبلی (شماره ۲)، موارد زیر را محاسبه بفرمایید:

۳.۱) احتمال توقف آن در هر صفحه را پس از انجام یک حرکت محاسبه کنید.
۳.۲) احتمال توقف آن در هر صفحه را پس از انجام دو حرکت محاسبه کنید.
۳.۳) کدام صفحه از این صفحات بیشترین احتمال توقف را در تعداد بسیار زیاد حرکت دارد؟

بخش ج) تحلیل (۴ نمره از ۲۰)

فرض کنید از شما خواسته شده برای یک مجموعه شعر مربوط به قرن پنجم یک سامانه بازیابی طراحی و پیاده‌سازی کنید. این اشعار خروجی یک سامانه OCR (تشخیص متن از تصویر کتاب) از یک نسخه قدیمی هستند که خطاهایی در تشخیص کلمات آن نیز در آن وجود دارد. یکی از متخصصین تصحیح متون ادبی می‌خواهد با این داده کار کند و در آن جست‌وجوهای ظاهری و مفهومی انجام دهد.

۱) چه پیش پردازش‌هایی برای این داده در نظر می‌گیرید. چه مراحلی را متفاوت از پردازش عادی زبان فارسی در نظر می‌گیرید؟

۲) برای جست‌وجوهایی که کاربر به دنبال ظاهر رشته[۴] مشابه پرسمان خود است، از چه مدل‌هایی در فضای برداری استفاده می‌کنید؟ اگر کاربر به جست‌وجوهای مفهومی در فضای برداری نیاز داشته باشد چه طور؟ شباهت را چگونه بین پرسمان و اشعار موجود محاسبه می‌کنید.

۳) اگر بخواهید با کمک متخصصین تصحیح متون، برای سامانه نهایی یک ارزیابی روی روشهای بازنمایی بالا داشته باشید، چگونه داده ارزیابی را ایجاد می‌کنید؟ و از چه متریک‌هایی برای ارزیابی روش‌های بالا استفاده می‌نمایید؟

۴) اگر بخواهید با استفاده از یکی از ترنسفرمرهای زبان فارسی، یک طبقه‌بندی روی ۴ شاعر این قرن انجام دهید، این کار را در چه مراحلی انجام می‌دهید؟

۵) اگر مقدار داده از این ۴ شاعر یکسان نباشد، طبقه‌بندی را چگونه و با چه معیاری ارزیابی می‌کنید؟

۶) از شما خواسته شده یک خوشه‌بندی روی اشعار این شاعران، با استفاده از بردار امبدینگ ابیات انجام دهید و ببینید که خوشه‌بندی چه نسبتی با تقسیم بندی بر اساس شاعر دارد. از چه معیاری برای سنجش این خوشه‌بندی استفاده می‌کنید؟ این معیار را شرح دهید (یک مورد کافیست).

---

[۴] surface form