هوالحکیم

مبانی بازیابی اطلاعات و جستجو وب

نیم سال دوم ۰۲-۰۱

دانشکده مهندسی کامپیوتر

تمرین سری سوم — مهلت تحویل : ۱۴۰۲/۰۴/۰۲

نکات:

۱- پاسخ تمرینات در قالب فایل پی دی اف صرفا در lms قرار گیرد.
۲- بخش تئوری تمرین توسط هر نفر جداگانه تحویل داده می شود، در lms دو تمرین جداگانه تئوری و عملی تعریف شده است.
۳- بخش عملی تمرین در گروه های حداکثر ۲ نفره انجام میشود. (گروه ها تا انتهای ترم یکسان باقی می ماند)
۴- تمرین عملی در زمانی که بعدا اعلام خواهد شد ، توسط حل تمرین به صورت حضوری تحویل گرفته می شود.

الف) بخش تئوری

۱- در تمرین قصد داریم از الگوریتم means-k و فاصله اقلیدسی برای خوشه بندی داده های زیر به سه خوشه استفاده کنیم.

A1=(2,10) A2=(2,5) A3=(8,4) A4=(5,8) A5=(7,5) A6=(6,4) A7=(1,2) A8=(4,9)

ماتریس فاصله براساس فاصله اقلیدسی را برای داده‌های فوق حساب کنید.

فرض کنید مراکز خوشه ابتدایی A ۱و A ۴و A ۷هستند. الگوریتم means_k را برای یک تکرار epoch اجرا کنید و سپس به سوالات زیر پاسخ دهید.

الف) خوشه های جدید را با نام بردن اعضای آنها مشخص کنید.
ب) مراکز خوشه های جدید را مشخص کنید.

۲- مجموعه داده زیر را درنظر بگیرید. هر سطر جدول نشان دهنده یک رکورد داده است که سه ویژگی ورودی و یک ویژگی کلاس (وضعیت) دارد. وضعیت میتواند شاغل (yes) یا بیکار (no) باشد. برای رکورد دادهای جدید x = (کارشناسی، تهران، بله) مقدار تخمینی وضعیت و مقدار احتمال زیر، مطایق روش بیز ساده (Bayes Naïve) ، برابر چه مقادیری است؟

P(y = yes | x)

وضعیت	سابقه کار	شهر	تحصیلات
شاغل	خیر	تهران	کارشناسی
شاغل	بله	شاهرود	ارشد
بیکار	خیر	مشهد	ارشد

بیکار	بله	مشهد	ارشد
شاغل	خیر	تهران	کارشناسی
بیکار	خیر	شاهرود	ارشد
شاغل	بله	تهران	ارشد
شاغل	خیر	مشهد	کارشناسی

۳. یک شبکه دلخواه با حداقل چهار نود را مثال بزنید و یک مرحله اجرای الگوریتم page rank را در آن توضیح دهید. به بیان دیگر ماتریس احتمال انتقال را برای شبکه مثالی خود بنویسید، سپس بردار page rank اولیه را نیز نوشته و در یک مرحله آن را به روز کنید. (۳,۵)؟

ب) بخش عملی

قسمت اول

در این پروژه یک موتور جستجو برای یکی از سایت‌های معروف فیلیمو یا استک اور فلو ایجاد می کنیم. شما باید یک خزنده بنویسید که از صفحه اصلی سایت آغاز به کار کند و ارجاعاتی که در همین دامنه هستند را دنبال کند و حداقل ۱۰۰ صفحه را جمع آوری کند (اگر این صفحات مرتبط با سه بخش این سایت باشند مثال در فیلیمو سه ژانر مشخص و یا در است اور فلو سه زبان مشخص در بخش دوم تمرین کار شما تسهیل خواهد شد). بخش های مختلف یک صفحه مانند عناوین و متن و... را جمع آوری کنید. برای پردازش صفحات Html می توانید از کتابخانه های مناسب موجود استفاده کنید. سپس محتوای جمع شده را با استفاده از Lucene یا برنامه خودتان که در پروژه قبل با آن کار کرده اید شاخص گذاری کنید. با وارد نمودن تعدادی پرس و جوی دلخواه عملکرد مناسب جستجو را نشان دهید .به صورت اختیاری بر اساس بخشهای مختلف صفحه مکانیزم رتبه بندی ایجاد کنید. به عنوان مثال اگر در سندی کلمه در عنوان متن ظاهر شده بود، این سند در رتبه بالاتری نسبت به سندی که کلمه را در متن خود دارد قرار بگیرد. با وارد نمودن تعدادی پرس و جوی دلخواه عملکرد مناسب جستجو را نشان دهید .نحوه پیاده سازی خزشگر و موتور جستجو، به همراه انواع پرس و جوها که قابلیت برنامه شما را نشان میدهند را در گزارش نهایی ذکر کنید

قسمت دوم

صفحات مرتبط با حداقل سه بخش مختلف سایت مثل سه زبان یا سه ژانر را خزش کنید. از هر بخش حداقل ۲۵ صفحه را خزش کنید .سپس برای صفحات هر بخش برچسب همان بخش را در نظر گرفته و یک رده بند بیز ساده را برای رده بندی صفحات اجرا کنید. به این منظور می توانید از کتابخانه های هر زبانی که استفاده میکنید بهره ببرید یا از ابزارهای آماده مثل وکا استفاده کنید. شرح کار و دقت رده بند را گزارش کنید.