پروژه سوم درس بازیابی پیشرفته اطلاعات

ارزیابی سیستم‌های بازیابی با استفاده از یک مجموعه آزمون فارسی

معرفی مجموعه آزمون "محک"

از مجموعه‌های آزمون (Test Collections) برای ارزیابی سیستم‌های بازیابی اطلاعات استفاده می‌شود. یک مجموعه تست استاندارد شامل تعدادی مستند، تعدادی پرسش و لیست مستندات مرتبط با هر پرسش است. در حال حاضر تعداد زیادی مجموعه آزمون استاندارد برای زبان انگلیسی تهیه شده است که مهمترین آنها TREC می باشد. Time، Reuters و CF نمونه‌های کوچکتری (از لحاظ تعداد مستندات و پرسش‌ها) از این مجموعه‌ها هستند. با توجه به این نکته که کیفیت نتایج سیستم‌های بازیابی اطلاعات تا حد زیادی به وابسته به زبان استفاده شده در نگارش مستندات می‌باشد لازم است برای ارزیابی موتورهای جستجوی مورد استفاده در یک زبان خاص از مجموعه آزمون مخصوص همان زبان استفاده شود.

متاسفانه در حال حاضر هیچ مجموعه آزمونی -که به صورت رایگان در دسترس باشد- برای زبان فارسی معرفی نشده است. به همین دلیل در آزمایشگاه وب معنایی، سعی شده است تا به عنوان بخشی از یک پروژه تحقیقاتی چنین مجموعه‌ایی تهیه شود. این مجموعه که "محک" نام دارد حاوی تعدادی مستند (حدود ۵۰۰۰ عدد) و پرسش (حدود ۴۰۰ عدد) می‌باشد که از سایت خبری ایسنا بارگذاری شده است. پرسش‌های مذکور با مطالعه مجموعه مستندات و پس از دسته بندی آنها، استخراج گردیده است، بنابراین برای هر پرسش، لیست مستندات مرتبط با آن معلوم می‌باشد.

شرح پروژه

هدف از این پروژه پیاده‌سازی یک موتور جستجو و ارزیابی کارایی آن با استفاده از "محک" می‌باشد. از این مجموعه آزمون فقط بخش مستندات و پرسش‌ها در سایت درس قرار داده خواهد شد. هر گروه باید با استفاده از موتور پیاده‌سازی شده خود، لیست همه مستندات مرتبط (به ترتیب میزان شباهت) برای هرکدام از پرسش‌های داده شده را پیدا نموده و در قالب یک فایل Excel تحویل دهد. نمره این پروژه ۵ درصد از نمره نهایی این درس را تشکیل می‌دهد و یک نمره (از بیست نمره نهایی) نیز به عنوان نمره تشویقی در نظر گرفته شده است. این نمره به گروه‌هایی تعلق می‌گیرد که با مطالعه بخشی از مستندات (مثلا ۵۰۰ عدد) میزان شباهت هر مستند را با هرکدام از پرسش‌ها تعیین مشخص کنند (به صورت دستی).

نکات مهم

- نحوه انجام این پروژه به صورت گروهی می‌باشد (حداکثر ۲ نفر)
- می‌توانید از موتور بازمتنِ Lucene (http://lucene.apache.org) در انجام این پروژه استفاده و آنرا برای زبان فارسی سازگار نمایید. اما برای حصول کارایی بالاتر توصیه می‌شود یک موتور جستجو را به طور کامل پیاده‌سازی نمایید
- موعد نهایی تحویل جدول نتایج **۸۴/۹/۳۰** می‌باشد. این مهلت به هیچ‌عنوان تمدید نخواهد شد
- گروه‌هایی که علاقه‌مند به کسب نمره تشویقی می‌باشند برای تعیین شماره مستنداتی که باید مطالعه نمایند با حل تمرین درس تماس بگیرند
- برای پرسیدن سوالات خود از بخش Discussion Area در وب سایت درس استفاده کنید