بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

بازیابی پیشرفته اطلاعات

نیم‌سال دوم ۰۱ـ۰۰

استاد: احسان‌الدین عسگری

پروژه‌ی پایانی

مهلت تحویل: ۳۱ تیر

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در تمرین‌هایی که چند ترک دارند، فقط یک نفر از هر گروه در سامانه CW باید ترک مورد نظر گروه را انتخاب کند. امکان تغییر ترک تا قبل از زمان ددلاین انتخاب ترک وجود دارد. البته ذکر این نکته ضروری است که هر ترک محدودیتی برای تعداد افرادی که آن را انتخاب می‌کنند، دارد. بنابرین در اسرع وقت برای انتخاب ترک اقدام کنید.
- در طول ترم امکان ارسال با تاخیر تمرین‌ها بدون کسر نمره تا سقف ۱۲ روز وجود دارد. محل بارگزاری جواب تمرین‌ها بعد از ۳ روز بسته خواهد شد و پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند شد. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۵ درصد از نمره تمرین کسر خواهد شد.
- توجه داشته‌باشید که نوت‌بوکهای شما باید قابلیت بازاجرای ۱۰۰ درصد داشته باشند و در صورت نیاز به نصب یک کتابخانه یا دسترسی به یک فایل، مراحل نصب و دانلود (از یک محل عمومی) در نوت‌بوک وجود داشته باشد.
- تمامی فایل‌های مرتبط به پروژه که حجم کمی دارند باید به شکل فایل زیپ در سامانه CW اپلود شوند. اگر حجم یک فایل زیاد بود (مانند فایل ذخیره شده یک مدل در صورتیکه بیش از ۲۰۰ مگابایت باشد)، تنها همان فایل را در یک محل عمومی، مثل گوگل درایو آپلود بفرمایید و لینک دانلود را در نوت‌بوک و مستندات قرار دهید.
- در پروژه‌های گروهی کافی است که فقط یکی از اعضای گروه پروژه را آپلود کند. اما حتما در گزارش کار نام همه اعضای گروه همراه با شماره دانشجویی آن‌ها آورده شود.
- بخشی از نمره شما به گزارش کار شما اختصاص دارد. در گزارش کار لازم نیست خط به خط کاری را که کرده‌اید توضیح دهید. بلکه باید به شکل کلی ایده‌تان برای حل مساله را شرح دهید. لازم است چند نمونه از خروجی‌های مساله را در گزارش بیاورید و براساس آن رفتار برنامه‌تان را تحلیل کنید. همچنین اگر پارامتری در صورت مساله خواسته شده (مانند دقت، صحت و ...) که در گزارش آورده شود شما باید آن را حساب کنید و در گزارش خود بیاورید.
- در صورت وجود هرگونه ابهام یا مشکل، در کوئرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به دستیاران آموزشی خودداری کنید.

توضیحات

در پروژه‌ی پایانی درس هریک از گروه‌هایی که پروپوزال‌شان مشترک است می‌توانند با یکدیگر ترکیب بشوند. تا روز یک‌شنبه ۱۹ تیرماه وقت دارید تا در کوئرا گروه‌هایی که می‌خواهند باهم ترکیب شوند این درخواست را به همراه شماره و اعضای گروه اعلام کنند. در صورتی که تمایل به ترکیب گروه نیز ندارید باید این درخواست خود را در کوئرا اعلام کنید. همچنین دقت شود که حداکثر اندازه‌ی گروه‌های ترکیب‌شده باید ۶ دانشجو باشد.

مواردی در پروژه‌ی پایانی درس وجود دارد که به صورت کلی در همه‌ی پروپوزال‌ها مشترک است که به شرح زیر است:

۱. دو تسک دسته‌بندی و خوشه‌بندی در پروژه‌ی پایانی باید انجام شود و نتایج آن به عنوان خروجی گزارش شود. معیارهایی که باید به عنوان خروجی برای این دو تسک گزارش شوند دقیقا همان معیارهایی است که در تمرین سری چهارم نیز وجود داشتند. در صورتی که گروه‌ها در تمرین سری چهارم این دو تسک را انجام داده‌اند می‌توانند از همان مدل‌های گذشته استفاده کنند فقط باید دقت کنند که حتما یک بهبود هرچقدر کوچک نیز بر روی مدل‌های دسته‌بند و خوشه‌بند موجود در تمرین سری چهارم خود اعمال کنند. این بهبود می‌تواند در زمینه‌های مختلف باشد. برای نمونه ممکن است با اعمال برخی از کارهای پیش‌پردازشی بتوانید به یک بهبود برسید. ممکن است با استفاده از مدل‌های جدید و یا تغییر پارامترهای همان مدل‌های پیشین به بهبود برسید و یا هر تکنیک دیگری... به بیان دیگر در پروژه‌ی پایانی باید گروه شما دارای یک سامانه‌ی خوشه‌بند و یک سامانه‌ی دسته‌بند باشد که نسبت به تمرین‌های سری چهارم بهتر شده باشد. در صورتی که کد خوشه‌بند یا دسته‌بند شما پیش‌تر دارای مشکل یا باگ‌هایی بوده یا تدریس‌یار مربوطه ایرادهایی به آن وارد کرده است باید در پروژه‌ی پایانی این ایرادها برطرف شده باشند.

۲. در پروژه‌ی پایانی باید بتوانید گسترش پرس‌وجو را انجام دهید. روش‌های مختلفی برای این کار می‌تواند صورت بگیرد که استفاده از یک روش برای این بخش کفایت می‌کند. برای نمونه می‌توان از مدل‌های زبانی برای پیشنهاد کلمات استفاده کرد. روش دیگر استفاده از الگوریتم Rocchio است.

۳. استفاده از یکی از موتور جستجوهای رایج مانند Search Elastic و یا Milvus در پروژه‌ی پایانی ضروری است. باید بتوانید داده‌های خود را در موتور جستجو ایندکس کرده و سپس بر روی آن پرس‌وجو بزنید و خروجی را گزارش کنید. دستکم ۱۰ پرس‌وجو به همراه خروجی آن را گزارش کنید. همچنین مقایسه کنید که اگر این ۱۰ پرس‌وجو توسط سامانه‌هایی که در تمرین سری سوم گسترش داده‌اید یعنی بولین، tfidf، ترنسفورمرها و fasttext انجام می‌شد، خروجی به چه شکل بود؟ آیا خروجی‌های موتور جستجو یعنی milvus و یا elastic search بهتر بوده است؟
**یک سند راهنما برای کار با موتورهای جستجو در کنار این توضیحات قرار داده می‌شود تا با آشنایی بیش‌تری بتوانید با موتورهای جستجو کار کنید.

۴. خروجی سامانه‌ی شما باید به یک واسط کاربری متصل شود. به عبارت دیگر همه‌ی تمرین‌های سری سوم، چهارم و پنجم شما و همچنین همه‌ی بخش‌های پروژه‌ی پایانی شما باید توسط یک واسط کاربری نمایش داده شوند. شما می‌توانید واسط کاربری خود را مبتنی بر وب بنویسید و یا مبتنی بر هر پتلفرمی که راحت‌تر هستید. (برای نمونه موبایل، دسکتاپ و ...) واسط کاربری شما باید قابلیت دریافت یک پرس‌وجو داشته باشد و سپس بتواند با روش‌های مختلف مانند بولین، ترنسفورمر، tfidf و fasttext خروجی را نمایش دهد. در صورتی که در هر کدام از این روش‌ها، سامانه‌ی شما ضعف‌هایی از پیش دارد باید برطرف شده باشد. همچنین خروجی موتور جستجوی elastic و یا milvus نیز باید در این واسط کاربری نمایش داده شود. همچنین در این واسط کاربری باید گزینه‌هایی برای خوشه‌بندی و یا دسته‌بندی نیز داشته باشید. منظور از «داشتن گزینه‌هایی برای دسته‌بندی یا خوشه‌بندی» آن است که با دادن یک ورودی سامانه‌ی شما بتواند دسته‌ی مربوط به آن ورودی و همچنین خوشه‌ی مربوط به ورودی را مشخص کند. افزون‌بر این موارد، باید بتوان پرس‌وجوی گسترش‌یافته را نیز به عنوان یک گزینه‌ی دیگر در این واسط کاربری مشاهده کرد

همچنین توجه شود که با توجه به متفاوت بودن موضوع‌های پروپوزال‌ها، در صورتی که گروهی دارای پرسش و یا ابهامی پیرامون موارد خواسته شده در پروژه‌ی پایانی بود می‌توانند در پست مربوط به پروپوزال‌شان در کوئرا در قالب کامنت موارد خود را مطرح کنند.

توجه: در پروژه‌ی پایانی نیز می‌توانید از حاصل کار عزیزان ترم گذشته که با زحمات تدریس‌یاران درس در قالب کتابخانه parsi.io ایجاد شده بهره ببرید. به امید خدا در ترم‌های آینده حاصل تجمیع زحمات شما عزیزان در قالب محصولات متن‌باز (البته با ذکر نام خودتان) در اختیار دیگر دانشجویان و بلکه جامعه ایرانی قرار می‌گیرد تا در اثر این تلاش‌ها محصولاتی ارزشمند برای پردازش متن‌های فارسی و بلکه زبان‌های ایرانی و فراتر از آن داشته باشیم. می‌توانید به این کتابخانه از طریق این لینک دسترسی داشته باشید