

## بسمه تعالی

### پروژه پایانی درس بازیابی اطلاعات

#### • عنوان : پیاده سازی و ارزیابی سیستم بازیابی اطلاعات

در این پروژه قصد داریم یک سیستم بازیابی اطلاعات پیاده سازی کرده و با استفاده از مجموعه داده برچسب گذاری شده ( در این مجموعه اسناد مرتبط با هر پرس و جو مشخص شده اند)، دقت سیستم پیاده سازی شده را بررسی نمائیم. پارامترهای مورد استفاده شامل دقت، فراخوانی و معیار F می باشند. در این پروژه مراحل کار شامل موارد زیر است:

- ۱) دریافت متون اسناد (CISI.ALL)، نرمال سازی، توکن سازی و ساخت جدول posting list و آمار فراوانی کلمات در اسناد.
- ۲) اعمال پرس و جوهای موجود در مجموعه داده (CISI.QRY) و محاسبه معیارهای ارزیابی برای هر کدام و میانگین گیری برای محاسبه میزان دقت کلی مدل های پیاده سازی شده. اسناد مرتبط با هر پرس و جو در فایل (CISI.REL) نگهداری می شوند. با توجه به اینکه در روش های برداری لیستی از اسناد به همراه امتیاز برگردانده می شود، برای ارزیابی می توانید با در نظر گرفتن محدودیت بر روی تعداد اسناد بازیابی شده ( مثلا ۱۰ سند اول) ارزیابی را انجام داد.
- ۳) بررسی تاثیر هر کدام از بخش های پیش پردازش در نتیجه نهایی.  
- به عنوان مثال بخش آیا حذف کلمات توقف تاثیر مثبت دارد یا خیر. برای پیدا کردن جواب می تواند ابتدا با حذف کلمات توقف میانگین امتیاز F را محاسبه کرد و در ادامه این کار را بدون حذف کلمات انجام داد و نتیجه را با هم مقایسه کرد.
- ۴) بررسی تاثیر روش های مختلف وزن دهی بر روی نتیجه نهایی.  
- به عنوان مثال اگر تنها از معیار فرکانس کلمات استفاده کنیم با زمانی که فرکانس معکوس سند را هم استفاده کنیم چه تغییری در نتایج خواهیم داشت.

در نهایت جدول پیوست را تکمیل نمائید.

#### • مجموعه داده CISI :

لینک جهت دانلود: [http://ir.dcs.gla.ac.uk/resources/test\\_collections/cisi](http://ir.dcs.gla.ac.uk/resources/test_collections/cisi)

لینک جهت توضیحات : <https://www.pragmalingu.de/docs/guides/data-comparison>

#### • تاریخ تحویل:

بارگزاری در سامانه VU: یکشنبه ۹ بهمن ساعت ۲۳:۵۹

تحویل حضوری : دوشنبه ۱۰ بهمن ساعت ۱۰:۰۰

پیوست:

F-measure	recall	Precision	پیش پردازش	روش استفاده شده
			حذف S.W با ریشه یابی	فقط TF
			عدم حذف S.W با ریشه یابی	
			حذف S.W بدون ریشه یابی	
			عدم حذف S.W بدون ریشه یابی	
			حذف S.W با ریشه یابی	TF*IDF بدون نرمال سازی
			عدم حذف S.W با ریشه یابی	
			حذف S.W بدون ریشه یابی	
			عدم حذف S.W بدون ریشه یابی	
			حذف S.W با ریشه یابی	TF*IDF نرمال سازی شده
			عدم حذف S.W با ریشه یابی	
			حذف S.W بدون ریشه یابی	
			عدم حذف S.W بدون ریشه یابی	

با ارزیابی موفقیت

سلطانی