پروژه‌ی درس پردازش زبان طبیعی

« غنی‌سازی متن به کمک شعر، غزل، نقل‌قول و حدیث»

اعضای تیم: مهدی سعیدی، فرانک کریمی، محمدرضا کمالی، آرمان مظلوم‌زاده

تعریف مساله

ورودی: یک پرس‌وجوی به زبان فارسی که قرار است با محتوای چندزبانه غنی شود.

خروجی: تجمیع نتایج روش‌های مختلف تعیین شباهت که پرس‌وجوی ورودی را با محتوای مناسب (شعر، غزل، نقل‌قول و حدیث) که از نظر معنایی مرتبط هستند، غنی می‌کند.

اهمیت و کاربرد: این روش با به‌کارگیری مدل‌های مختلف برای تعیین موارد مرتبط به ورودی به صورت چندزبانه مانند Weighted TF-IDF، LaBSE و FastText Multilingual، سعی در غنی‌کردن پرس‌وجوی ورودی دارد. تجمیع نتایج این روش‌ها باعث می‌شود که هم خروجی پوشاتری بدست آید و هم در نهایت با انجام ارزیابی مشخص شود که کدام روش نقش بیشتری در خروجی نهایی داشته است. بنابراین می‌توان گفت اهمیت اصلی این روش، ایجاد یک مجموعه داده‌ی برچسب‌گذاری‌شده به همراه ارزیابی بر روی روش‌های پایه‌ای گفته‌شده است که امکان مقایسه‌ی آن‌ها را در چارچوبی علمی ممکن کرده است.

روش

داده و تقسیم‌بندی: مجموعه‌داده‌های گنجور به زبان فارسی شامل چندین شعر و غزل، مجموعه‌داده‌ی Goodreads به زبان انگلیسی با ۲۵۰۸ نقل‌قول از نویسندگان سرشناس و مجموعه‌داده‌های نهج‌البلاغه به زبان عربی شامل ۴۸۰ حکمت. با توجه به اینکه ارزیابی supervised است این داده‌ها تماما به عنوان ورودی به این روش داده می‌شوند.

روش و داده ارزیابی: از طریق تعیین relevance score به صورت supervised به این شکل که به ازای هر جمله‌ی پرس‌وجو، از هر روش ۱۰ قطعه‌ی غنی‌کننده دریافت می‌شود و بنابراین برای هر ورودی روی هم چندین قطعه خواهیم داشت. سپس به ازای هر جفت پرس‌وجو و هر قطعه‌ی غنی‌کننده‌ی آن، دو نفر عددی بین ۱- و ۱ به آن اختصاص می‌دهند و نهایتا امتیاز نهایی با توافق بین آن‌ها تعیین می‌شود. سپس از روش‌هایی مانند Fleiss' kappa صحت امتیازدهی مشخص می‌شود. در آخر روش‌های مختلف بر حسب اینکه در مجموع چه امتیازهایی گرفته‌اند رتبه‌بندی می‌شوند. داده‌ی ارزیابی نیز داده‌ای است که به شکل supervised امتیازدهی شده است.

مدل‌های پایه: Weighted TF-IDF، LaBSE و FastText Multilingual.

مدل اصلی: تجمیع مدل‌های پایه.

لینک گیت‌هاب پروژه - **لینک تلگرام پروژه**