بسم الله الرحمن الرحیم

پروژه پیشنهادی اینجانب احسان اعظمی برای دروس یادگیری ماشین و داده کاوی پیشرفته

موضوع: دسته‌بندی خودکار کلمات زبان فارسی با استفاده از درخت تصمیم‌گیری (براساس پیکره بی‌جن‌خان)

۱- **صورت مسئله:** دسته‌بندی کلمات فارسی بر اساس پیشوندها و پسوندها. هر کلمه به یک دسته از ادوات سخن در زبان متعلق می‌باشد. با دانش کافی از پیشوندها و پسوندهای زبان فارسی، هر یک از کلمات را دسته بندی می‌کنیم. برچسب مربوط به این دسته‌بندی براساس برچسب‌های پیکره ی دکتر بی‌جن‌خان می‌باشد.

۲- **ویژگی‌های پیشنهادی:** ، تمامی موارد زیر ویژگی‌های(Features) پیشنهادی هستند که ممکن است درطی انجام پروژه به جهت بهبود کیفیت درخت تصمیم‌گیری تغییر کنند.

۱- طول کلمه

۲- طول ساقه: طول قسمتی از کلمه که تمامی پسوندهای آن زدوده شده باشد. تمامی پیشوندها را جزو ساقه محسوب می‌کنیم.

۳- به ازای نقطه‌ی شروع هر یک از پیشوندها و پسوندها (چه صرفی باشند چه اشتقاقی) یک ویژگی تعریف می‌کنیم. به عنوان مثال: اگر کلمه شامل پسوند »ها« می‌باشد، نقطه‌ی شروع پسوند »ها« یک ویژگی می‌باشد. در صورتی که اگر شامل »ها« نباشد، به ازای ویژگی »ها« عدد صفر را می‌گذاریم.

۴- ادوات سخن سه کلمه قبل و سه کلمه بعد از کلمه ی حاضر، هریک به تنهایی یک ویژگی محسوب می‌شوند.

۳- پیش پردازش داده ها:

۱- در صورتی که پیشوند یا پسوند جدا نوشته شده باشد، مانند »کتاب ها« به جای »کتابها« یا »می آیم« به جای »میآیم«، می‌توانیم پیشوند یا پسوند را به کلمه بچسپانیم و محاسبات ویژگی را دوباره انجام دهیم.

۲- محدوده پیشوندها یا پسوندها نباید یکدیگر را قطع کنند.

۳- نقطه شروع هر پسوند باید بلافاصله بعد از پسوند قبلی باشد. در غیر این صورت پسوند قبلی قابل قبول نیست و جزوی از ساقه محسوب می‌شود.

۴- کلمات قصار لاتین و مخلوط کلمات لاتین به همراه پسوندهای فارسی می‌بایست از فهرست ورودی‌های داده حذف شوند.

۵- نیم‌فاصله‌ها در محاسبه نقطه شروع پیشوند و پسوند دخیل نیستند.

۴- **مجموعه داده‌ها:** از پیکره بی‌جن‌خان به عنوان مجموع داده هدف استفاده می شود، پیکره بی‌جن‌خان شامل دسته‌بندی ادوات سخن می‌باشد بنابراین پروژه رویکرد یادگیری بانظارت دارد و بخشی از پیکره جهت یادگیری به کار گرفته می‌شود.

۵- **محاسبه متریک ارزیابی درخت:** مدل یادگیری ما استفاده از درخت تصمیم‌گیری می‌باشد که پس از اتمام فرآیند، یادگیری ماشین توسط متریک‌های ارزیابی درخت تولید شده محاسبه می شود. (مقادیر مثبت و منفی درست و نادرست، دقت، فراخوانی، ماتریس هزینه، ...)

باتشکر