دانشگاه صنعتی شریف

گزارش تمرین دوم درس پردازش زبان‌های طبیعی

آرمان مظلوم‌زاده

فرانک کریمی

استاد درس

دکتر احسان‌الدین عسگری

آذر ماه ۱۴۰۱

فهرست مطالب

تغییرات lemmetizer

قواعد صرف کهن افعال برای حالات ماضی، مضارع و امر با توجه به قواعد زیر به تابع ***conjugations*** اضافه شد.

قواعد دستوری اعمال شده(صرف افعال کهن):

تمامی موارد زیر در صرف افعال به کتابخانه‌ی هضم اضافه شده و تابع lemmatizer این کتابخانه پیش از این برای هیچ یک از فرم های صرف فعل زیر به درستی کار نمی‌کرد

۱. ماضی استمراری-فرم۱: همی+ بن ماضی + شناسه‌های ماضی
   a. مثال: همی‌گفتند
۲. ماضی استمراری-منفی-فرم۱: همی+ ن +بن ماضی + شناسه‌های ماضی
   a. مثال: همی‌نگفتند
۳. ماضی استمراری فرم ۲: بن ماضی + ی + شناسه های ماضی
   a. مثال: گفتندی
۴. ماضی استمراری فرم ۲: می+ ب + بن ماضی + شناسه های ماضی
   a. مثال: می‌بگفتی
۵. ماضی استمراری فرم -۲-منفی: می+ ن + بن ماضی + شناسه های ماضی
   a. مثال: می‌نگفتی
۶. مضارع استمراری فرم۲- منفی: ن + بن ماضی + شناسه‌های ماضی
   a. مثال: نگفتندی
۷. مضارع استمراری- فرم ۱: همی + بن مضارع + شناسه‌ی فعل مضارع
   a. مثال: همی‌گویند
۸. مضارع استمراری- منفی-فرم۱: بن مضارع + شناسه فعل مضارع
   a. مثال: همی‌نآیی(همی‌نیایی)
۹. مضارع استمراری-فرم۲
   a. می‌بروم
۱۰. مضارع استمراری-فرم۲-منفی
   a. می‌نروم
۱۱. ماضی التزامی-فرم۱: ب+ بن ماضی+ شناسه ماضی + ی
   a. برفتندی

۱۲. ماضی التزامی-فرم۱: ب+ بن ماضی+ شناسه ماضی
   a. برفتند
۱۳. مضارع ساده منفی: م + بن مضارع + شناسه
   a. مگویند
۱۴. نهی: م + بن مضارع
   a. مگو
۱۵. ماضی ساده منفی + بن ماضی + شناسه
   a. مگفتم
۱۶. وجه دعایی
   a. مکناد

رفع ایراد تابع lemmetizer

اصلاح ماضی التزامی سوم شخص مفرد:

در کتابخانه‌ی هضم شناسه‌ی ه برای این حالت در نظر گرفته شده بود که به است تغییر کرد

قبل از اصلاح کتابخانه		پس از اصلاح کتابخانه	
رفته‌ام	رفته‌ای	رفته‌ام	رفته‌ای
رفته	رفته‌ایم	رفته است	رفته‌ایم
رفته‌اید	رفته‌اند	رفته‌اید	رفته‌اند

تغییرات Normalizer

اسامی

اسقاط ه آخر کلمه

در فارسی کهن گاه رخ می‌داده که ه پایان کلمه حذف شود به خصوص در نثر و برای حفظ وزن و قافیه

برای مثال کلمه‌ی پادشاه به صورت پادشا و کلمه‌ی گیاه به صورت گیا میامده است. چنین مواردی پس از نرمال‌سازی به فرم کامل خود درمیایند.

ب به جای و در آغاز کلمه

برای مثال در فارسی کهن استفاده از برزیدن به جای ورزیدن و باژگون به جای واژگون رواج داشته

این واژگان پس از نرمال‌سازی به شکل آشنای امروزی خود درمی‌آیند.

کوتاه‌سازی

در هجای آخر کلماتی که به ه (ملفوظ) ختم شده است، غالباً الف ممدود به فتحه تبدیل می‌شود.

مثال:

کوتاه- کوته

سپاه- سپه

سیاه-سیه

راه-ره

این کلمات کوتاه‌شده، پس از نرمال‌سازی با شکل الف‌دار خود ذخیره می‌شوند.

حذف الف ابتدای افعال

برخی از افعال و صفات در فارسی کهن به دو شکل کاربرد دارند. با الف ابتدا و با حذف آن. مثال:

افراشتن- فراشتن

افکندن- فکندن

افروختن- فروختن

در نرمال‌سازی تمام شکل با الف ترجیح داده شده چرا که در فارسی امروز رایج‌تر است.

حروف اضافه

ترمیم حروف اضافه‌ای که بر اساس قوانین ادغام به کلماتی که با مصوت آغاز می‌شوند می‌چسبند.

مثال:

پیش از نرمال‌سازی	پس از نرمال‌سازی
کان	که آن
کاخر/ کآخر	که آخر
کاندر	که در

پیدا کردن و جایگزینی فرم استاندارد حروف اضافه‌ای که در متون کهن و به خصوص در نظم برای رعایت وزن و قافیه به صورت خلاصه به کار می روند.

مثال:

پیش از نرمال‌سازی	پس از نرمال‌سازی
ز	از
ار	اگر
گر	اگر
اندر	در

بهبود تابع affix_spacing_pattern

در کتابخانه‌ی Hazm تمام فاصله‌های بین می و کلمات بعد از آن به نیم‌فاصله تبدیل می‌شود.

این امر به خصوص در نظم کهن که کلمه‌ی مِی بسیار پرتکرار است، مشکل‌زاست. برای مثال تابع اولیه‌ی نرمال‌ساز این کتابخانه بیت زیر را به این صورت تغییر می‌دهد.

پیش از نرمال‌سازی	پس از نرمال‌سازی با Hazm	پس از بهبود نرمال‌ساز Hazm
به می سجاده رنگین کن	به می‌سجاده رنگین کن	به می سجاده رنگین کن

این تغییر از اشتباه آینده در تابع tokenizer جلوگیری می‌کند.

تغییرات Stop word removal

لیستی از کلمات زائد از منظر پردازشی، که در متون نظم کهن پرتکرار است به فایل stopwords.dat اضافه شده.

این کلمات شامل موارد زیر می‌شود.

اندر

مر

همی

کاین

و ...