به نام خداوند

توصیف پروژه:

در این پروژه از شما خواسته می‌شود تا مجموعه‌ای از Taskهای پردازش زبان طبیعی را در زبان برنامه‌نویسی دلخواه خود پیاده‌سازی کنید و در قالب یک مجموعه نرم‌افزاری یکپارچه (Suite) پیاده‌سازی کنید. به طور کلی، این پروژه از سه بخش اصلی تشکیل شده است:

1. بخش پیش پردازش داده‌های متنی (انجام پردازش‌های پایه‌ای روی داده‌های متنی).
2. بخش تشخیص و تصحیح غلط املایی در یک داده متنی ورودی.
3. بخش دسته‌بندی داده‌های متنی.

**بخش پیش پردازش داده‌های متنی (انجام پردازش‌های پایه‌ای روی داده‌های متنی)**

در این بخش، از شما خواسته خواهد شد تا مجموعه‌ای از فرایندهای پیش پردازشی را روی داده متنی ورودی (فایل متنی موجود در پوشه “Dataset\TextProcessing”) خاصی پیاده‌سازی کند. کاربر باید قادر باشد تا با انتخاب فرایند پردازشی مد نظرش، خروجی مرتبط با انجام فرایند پیش پردازشی را روی داده متنی مشاهده کند. فرایندهای پیش پردازشی مورد نظر در این بخش عبارتند از:

* توکن بندی رشته‌های متنی موجود در فایل (بخش بندی داده متنی به واحدهای زبانی تشکیل دهنده).
* تبدیل حروف داده‌های متنی به حروف کوچک (Lowercase Folding)
* مشخص کردن تعداد دفعات تکرار هر کلمه یا توکن در داده متنی ورودی.
* انجام فرایند پیش پردازشی Stemming با استفاده از الگوریتم پرتر (Porter) روی داده‌های متنی.

نکته: برای پیاده‌سازی الگوریتم Porter می‌توانید از کتابخانه‌ها یا بسته‌های Third Party ارائه شده برای زبان‌های برنامه‌نویسی مختلف استفاده کنید.

**بخش تشخیص و تصحیح غلط املایی در یک داده متنی ورودی**

در این بخش، هدف این است تا از طریق بارگذاری یک فایل حاوی غلط املایی، سیستم پیاده‌سازی شده قادر باشد تا کلمات حاوی غلط املایی را تشخیص و در خروجی به همراه شکل صحیح املایی آن نمایش دهد. داده مورد نیاز برای پیاده‌سازی این بخش در پوشه (“Dataset\Spelling Dataset\test”) موجود است. برای پیاده‌سازی این بخش، از مطالب ارائه شده در این درس (مدل احتمالی Noisy Channel) استفاده شود.

نکات:

* برای پیاده‌سازی مدل احتمالی Noisy channel، نیازی به محاسبه Confusion Matrix نیست. مقادیر عددی Confusion Matrix برای اعمال Edit چهارگانه Insert و Delete و Substitution و Transposition از پیش محاسبه شده‌اند و در پوشه (“Dataset\Spelling Dataset\test\Confusion Matrix”) در اختیار شما قرار داده شده‌اند.
* دقت داشته باشید که برای پیاده‌سازی سیستم تشخیص و تصحیح غلط املایی از چهار مؤلفه اساسی باید استفاده شود:
  + Language Models
  + Noisy Channel Model
  + Error Confusion Matrix
  + Damerau-Levenshtein Edit Distance
  + دیکشنری و مجموعه داده لازم برای محاسبه احتمال کلمات یا P(W) در پوشه (“Dataset\Spelling Dataset\test\Dictionary”) قرار دارد.
  + کدهای لازم برای محاسبه Damerau-Levenshtein Edit Distance را می‌توانید از منابع موجود در سطح اینترنت نیز تهیه کنید.
  + به غیر از کد لازم برای محاسبه Edit Distance، از کدهای موجود در سطح اینترنت برای پیاده‌سازی این بخش به هیج وجه نباید استفاده شود.

**بخش دسته‌بندی داده‌های متنی**

در این بخش هدف این است تا یک سیستم دسته‌بندی متن با استفاده از روش بیز ساده (Naïve Bayes) پیاده‌سازی شود. در این بخش، مجموعه‌ای متشکل از اسناد متنی در اختیار شما قرار داده شده‌اند که در پنج کلاس مختلف دسته‌بندی شده‌اند. همچنین، داده‌های لازم برای آموزش مدل بیز ساده (یادگیری پارامترهای احتمالی) و تست این مدل در اختیار شما قرار داده شده‌اند. این داده‌ها، در پوشه (“Dataset\Classification-Train And Test”) قرار دارند. برای پیاده‌سازی این بخش لازم است ابتدا:

1. دیکشنری کلمات موجود در اسناد متنی این مجموعه داده را بسازید (برای ساختن دیکشنری از داده‌های آموزش استفاده کنید).
2. احتمالات کلاس‌ها یا P(C) را محاسبه کنید (فقط از داده‌های آموزش برای محاسبه این احتمالات استفاده کنید).
3. احتمالات شرطی هر کدام از کلمات به شرط کلاس‌ها (P(W|C)) را محاسبه کنید.
4. از رابطه قانون بیز برای دسته‌بندی اسناد موجود در مجموعه تست (test) در یکی از کلاس‌های پنج‌گانه استفاده کنید.