شیوه‌نامه‌ی پروژه تشخیص زبان فارسی از غیرفارسی

هدف از این پروژه، ارائه یک مدل کلاسه‌بندی برای تشخیص زبان فارسی از غیرفارسی در بستر شبکه‌های اجتماعی است.

فایل‌های برنامه:

language_classification.ipynb : این فایل شامل کتابخانه‌ها، کلاس و متدهای مورد نیاز است که مدل اصلی تشخیص زبان در آن پیاده‌سازی شده است.

run.py : این فایل داده‌ها را می‌خواند، آن‌ها را پیش‌پردازش می‌کند، بردارهای جاسازی کلمات را به دست می‌آورد و سپس زبان متن را با استفاده از یک مدل از پیش آموزش‌دیده KNN پیش‌بینی می‌کند.

new_final_dataset.xlsx : مجموعه داده مورد استفاده در این پروژه دارای ۲۲۵۹۳ نمونه از زبان‌های فارسی، پشتو، دری، عربی، اردو، انگلیسی، روسی و اوکراینی است. برچسب‌گذاری این مجموعه داده به این صورت است که برچسب ۱ نشان‌دهنده متن فارسی و برچسب ۰ نشان‌دهنده متن غیرفارسی است.

insta_wchr.vec : این فایل حاوی بردارهای جاسازی کلمات است که برای تبدیل متن به بردار استفاده می‌شود.

KNN_model.pkl : این فایل مدل K-Nearest Neighbors از پیش آموزش‌دیده، حاصل از اجرای فایل language_classification.ipynb برای کلاسه‌بندی است.

requirements.txt : این فایل دارای پکیج‌های مورد نیاز پروژه، برای نصب است.

README.md : این فایل دارای نسخه پایتون مورد استفاده و راهنمایی برای اجرای فایل run.py است.

معرفی کلاس‌ها و توابع farsivsnfarsiimplementation.py:

۱. **تابع preprocess :** این تابع برای پیش‌پردازش متنی که در یک دیتافریم ورودی قرار دارد، استفاده می‌شود.

۲. **کلاس Word2VecVectorizer :** یک کلاس تعریف شده است که برای تبدیل متن به بردارهای واژه با استفاده از فایل insta_wchr.vec می‌باشد. این کلاس سه متد اصلی دارد:

- **تابع __init__ :** در این تابع، مدل Word2Vec و اندازه جاسازی (embedding size) به‌عنوان ورودی گرفته می‌شود. این متد وقتی اجرا می‌شود، ابتدا یک پیام چاپ می‌کند که نشان می‌دهد داده‌ها در حال بارگذاری هستند، سپس مدل Word2Vecو اندازه جاسازی را ذخیره می‌کند و درنهایت یک پیام دیگر چاپ می‌شود که نشان می‌دهد بارگذاری داده‌ها تمام شده است.
- **تابع fit :** این تابع نیازی به انجام عملیات خاصی ندارد.
- **تابع transform :** این تابع برای تبدیل داده‌های ورودی به بردارهای عددی کلمات استفاده می‌شود.
- **تابع fit_transform :** این تابع ابتدا fit را صدا می‌زند تا داده‌ها آماده شوند و سپس transform را صدا می‌زند تا داده‌ها تبدیل شوند. در نتیجه، داده‌های ورودی از دو مرحله fit و transform عبور می‌کنند و بردارهای معادل کلمات تولید می‌شود.

۳. **تابع language_detect :** این تابع برای تشخیص زبان متن‌ها در یک مجموعه داده استفاده می‌شود. این تابع شامل مراحل پیش‌پردازش داده، آموزش و ارزیابی مدل برای تشخیص زبان متن‌ها است. این تابع یک فایل اکسل به نام final_dataset.xlsx را می‌خواند و دو ستون به نام `textcontent` و `target` را از آن استخراج می‌کند. سپس از تابع preprocess() برای پیش‌پردازش متن‌ها استفاده می‌کند و موارد تکراری و مقادیر تهی را هم حذف می‌کند. داده‌ها به دو بخش آموزش و آزمون تقسیم می‌شوند. متن‌ها با استفاده از مدل به بردارهای عددی تبدیل می‌شوند. مدل پیشنهادی آموزش و ارزیابی می‌شود. این مدل به‌صورت فایل Pickle ذخیره می‌شود. همچنین برای استفاده از این تابع، باید فایل اکسل مربوطه به همراه فایل Word2Vec موجود باشد و این تابع یک سری وابستگی به کتابخانه‌های مختلف دارد که باید قبل از اجرا نصب شده باشند.