DA5401 Assignment 6

Amazon has released a 51-languages parallel dataset called MASSIVE to the public domain. The same dataset is also available in Huggingface at https://huggingface.co/datasets/qanastek/MASSIVE. The dataset typically consists of sentences from 51 languages structured in a JSON format. The JSON structure contains the following headings (features).

['id', 'locale', 'partition', 'scenario', 'intent', 'utt', 'annot_utt', 'tokens', 'ner_tags', 'worker_id', 'slot_method', 'judgments']

From those headings, we are interested only in the following subset {'locale', 'partition', 'utt', 'tokens'}. 'locale' represents the language-country pair, 'partition' represents where the sentence is coming from amidst {'train', 'test', 'validation'}, 'utt' represents the actual sentence, and finally 'tokens' represents the split tokens from the sentence.

We are going to build a language classifier the covers all the languages with roman letters. There is already a classifier built on this dataset for all the 51 languages using transformers, which appears to be SOTA. https://huggingface.co/qanastek/51-languages-classifier. Our goal is not to compete with transformers, rather we are going to use this exercise to learn and overcome the challenges in dealing with multilingual datasets.

Task 1 [15 points]

Let's construct a dataset ourselves for with a subset of languages that are roman-script based. The following are the locales that we want to consider in our dataset [27 languages].

af-ZA, da-DK, de-DE, en-US, es-ES, fr-FR, fi-FI, hu-HU, is-IS, it-IT, jv-ID, lv-LV, ms-MY, nb-NO, nl-NL, pl-PL, pt-PT, ro-RO, ru-RU, sl-SL, sv-SE, sq-AL, sw-KE, tl-PH, tr-TR, vi-VN, cy-GB

Programmatically, extract the utterances "utt" from the dataset for each of the above languages. You can choose between your tokenization vs the preexisting tokens. By the end of this step, you should have 27 files (one for each language) with one sentence per line. Typically, all the 27 files will end up have the same number of lines as the dataset is a parallel-corpus.

Besides simple English like characters, you may encounter other characters and characters with accents. You may choose to deaccent the characters if accents are not useful in your method. Choose wisely!

Task 2 [15 points]

Build a multinomial Naive Bayes classifier on your 27 language dataset using the 'training' partition of the dataset. Finetune the model with the validation partition. Finally, report the performance metrics for all the three partitions.

Task 3 [20 points]

Convert further your 27 language dataset into language groups, where the grouping is via their respective continent names. It appears that the dataset has Asia, Africa, Europe, and North america. So, you will have four classes now. Collapse the dataset into 4 classes by appending the files into large files.

Build a Regularized Discriminant Analysis (RDA) model, which has a hyper-parameter lambda to tradeoff between LDA and QDA. You may use bag-of-words via CountVectorizer or Tfidf Vectorizer to create the feature space of your dataset. It will be a huge feature space, but LDA/QDA can handle large feature spaces, so no worries there. Of course, you may use some clever feature elimination methods such as low frequency pruning, noise removal, etc.