CS286 Spring 2016 Lab 1:
Multi-lingual Translator
MapReduce Program

Exercise 1: Write Translator from English to Multiple Languages

Objective

The objective of this exercise is to write a MapReduce program that merges a set of to-English dictionaries into a single file. The languages include French, German, Italian, Portuguese, and Spanish.

Data

The data you will use for this program is located at the "I Love Languages" Web site: http://www.ilovelanguages.com/IDP/IDPfiles.html

The data files have been provided to you in the DATA directory of your ZIP file.

Output

The output of your program a single text file with lines of the following format:

```
English-word: [part of speech] french:french-translation|
german:german-translation|italian:italian-
translation|portuguese:portuguese-translation|spanish:
spanish-translation
```

For example,

```
hello: [Noun] french:bonjour|german:guten tag|italian:
ciao|portuguese:N/A|spanish:hola
```

Note that not all words are represented in all the dictionaries. For example, the word "hello" does not have a translation in the given Portuguese dictionary. For all such words, put "N/A" to indicate the word is "not available" in the translator.

Conversely, some dictionaries have multiple translations for a given word. For example, the word "abandon" as a noun in the French dictionary has two translations – "abandon" and "laisser-aller". In such cases, include all the translations separated by a comma.

The lines in the output file are to be sorted on the English word, so words beginning with the letter 'a' appear in the beginning of the file, and those words beginning with the letter 'z' appear at the end of the file.

Exercise 2: Enrich translation with Esperanto

Objective

The objective of this exercise is to write a MapReduce program that enriches the dictionary you created in Exercise 1 with words in Esperanto.

Data

The Esperanto data you will use for this program is located in the DATA directory provided to you by the instructor. Note that you are required to use the distributed cache in your mapper to read in the Esperanto data. You will also be using the output from Exercise 1 as input to this program.

Output

The output of your program is a single text file with lines of the following format:

```
English-word: [part of speech] french:french-
translation|german:german-translation|italian:italian-
translation|portuguese:portuguese-translation|spanish:
spanish-translation|esperanto:esperanto-translation
```

For example,

```
hello: [Noun] french:bonjour|german:guten
tag|italian:ciao|portuguese:N/A|spanish:hola|esperanto:ha
lo
```

Deliverable

Provide the following artifacts for your solution in a single ZIP file named "`CS286_spring2016_lab1_fname_lname.zip`". Note that I will provide a ZIP file of this format containing starter code for the driver, mapper, and reducer.

1. `rebuild.sh`
2. `rerun.sh`
3. `DictionaryDriver.java`
4. `DictionaryMapper.java`
5. `DictionaryReducer.java`
6. `EsperantoDriver.java`
7. `EsperantoMapper.java`
8. `EsperantoReducer.java`
9. `DATA` (directory containing 6 .txt files)

Notes

Do not do any "pre-processing" or "post-processing" of the data files. All your code must run in the driver, mapper, and reducer Java programs for your solution.