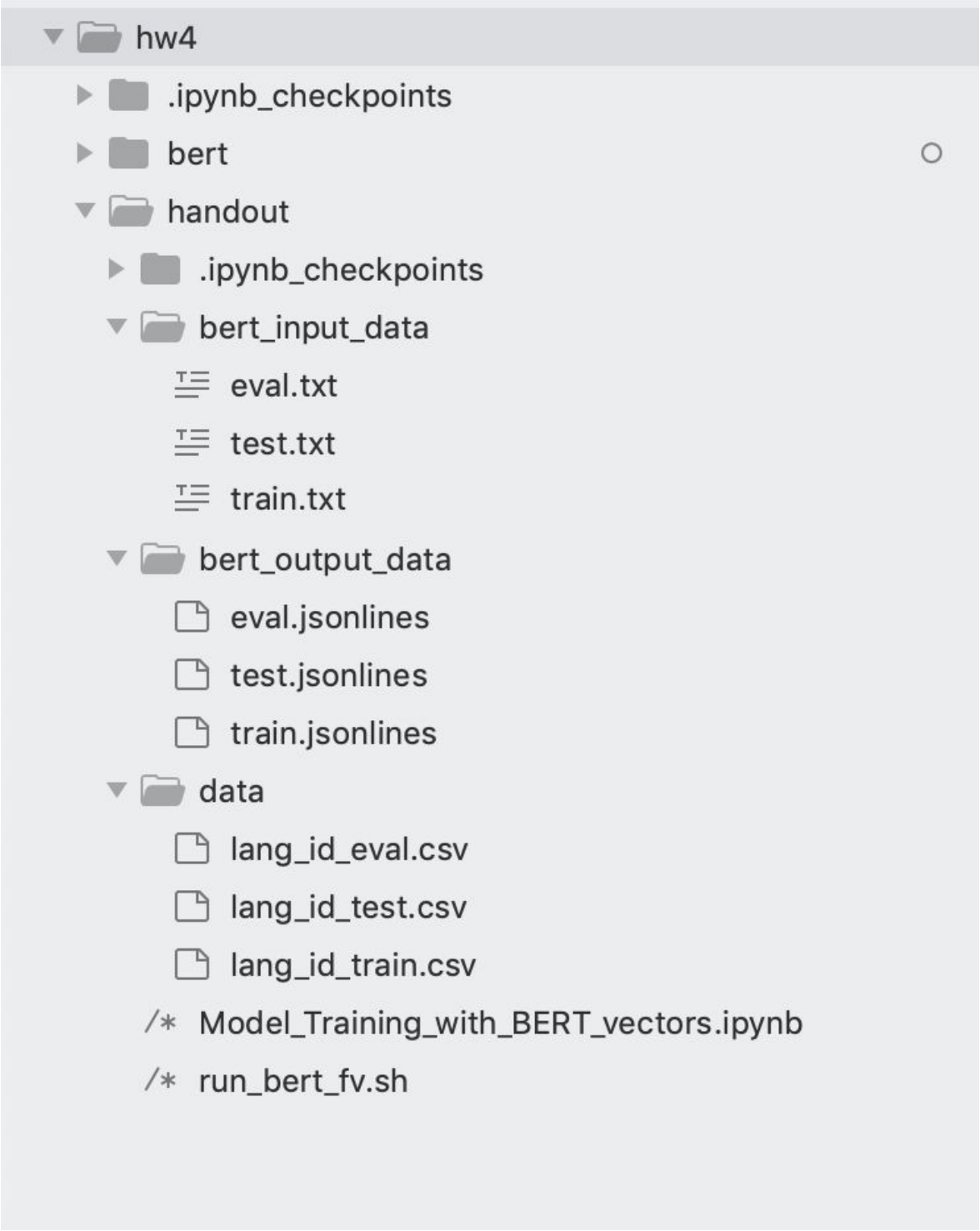
Homework 4: Using pre-trained BERT vectors for text classification

Kalpana Pratapaneni

A20448916

Step1&2:

Cloned and downloaded the BERT directory and handout materials from the provided links.



Folder Structure:

hw4/bert

hw4/handout

- bert_output_data/
- bert_input_data/
- data/

Step 3 & 4:

Remove header of the second column text, and then transfer each line to one text per input file using run_bert_fv.sh file and process these input files by the BERT extract_features.py script.

$ sh run_bert_fv.sh

After training, run Model_Training_with_BERT_vectors.ipynb using **"jupyter"**.

Summary:

Overall performance is 46.6 %.

Class metrics and frequency of errors:

Class	Misclassification Rate	Precision	Recall	Fscore
Arabic	0.505	0.498	0.495	0.496
Cantonese	0.659	0.327	0.34	0.333
Japanese	0.505	0.461	0.495	0.477
Korean	0.545	0.484	0.455	0.469
Mandarin	0.685	0.318	0.315	0.316
Polish	0.49	0.470	0.51	0.489
Russian	0.430	0.513	0.57	0.540
Spanish	0.48	0.517	0.52	0.519
Thai	0.4	0.659	0.6	0.628
Vietnamese	0.64	0.423	0.36	0.389

Arabic Language class metrics:

Actual Native Language - Predicted Language	Misclassification Count	Misclassification Rate
Arabic - Cantonese	11	0.055

Arabic - Japanese	9	0.045
Arabic - Korean	7	0.035
Arabic - Mandarin	12	0.06
Arabic - Polish	14	0.07
Arabic - Russian	10	0.05
Arabic - Spanish	16	0.08
Arabic - Thai	9	0.045
Arabic - Vietnamese	13	0.065

Cantonese Language class metrics:

Actual Native Language - Predicted Language	**Misclassification Count**	**Misclassification Rate**
Cantonese-Arabic	11	0.055
Cantonese-Japanese	14	0.07
Cantonese-Korean	12	0.06
Cantonese-Mandarin	46	0.23
Cantonese-Polish	12	0.06
Cantonese-Russian	10	0.05
Cantonese-Spanish	4	0.02
Cantonese-Thai	9	0.045
Cantonese-Vietnamese	14	0.07

Japanese Language class metrics:

Actual Native Language - Predicted Language	**Misclassification Count**	**Misclassification Rate**
Japanese-Arabic	8	0.04
Japanese-Cantonese	13	0.065
Japanese-Korean	23	0.115

Japanese-Mandarin	11	0.055
Japanese-Polish	17	0.085
Japanese-Russian	10	0.05
Japanese-Spanish	5	0.025
Japanese-Thai	5	0.025
Japanese-Vietnamese	9	0.045

Korean Language class metrics:

Actual Native Language - Predicted Language	**Misclassification Count**	**Misclassification Rate**
Korean-Arabic	7	0.035
Korean-Cantonese	17	0.085
Korean-Japanese	23	0.115
Korean-Mandarin	13	0.065
Korean-Polish	8	0.04
Korean-Russian	12	0.06
Korean-Spanish	9	0.045
Korean-Thai	12	0.06
Korean-Vietnamese	8	0.04

Mandarin Language class metrics:

Actual Native Language - Predicted Language	**Misclassification Count**	**Misclassification Rate**
Mandarin-Arabic	12	0.06
Mandarin-Cantonese	36	0.18
Mandarin-Japanese	19	0.095
Mandarin-Korean	14	0.07
Mandarin-Polish	9	0.045

Mandarin-Russian	10	0.05
Mandarin-Spanish	14	0.07
Mandarin-Thai	6	0.03
Mandarin-Vietnamese	17	0.085

Polish Language class metrics:

Actual Native Language - Predicted Language	**Misclassification Count**	**Misclassification Rate**
Polish-Arabic	11	0.055
Polish-Cantonese	14	0.07
Polish-Japanese	8	0.04
Polish-Korean	3	0.015
Polish-Mandarin	5	0.025
Polish-Russian	29	0.145
Polish-Spanish	11	0.055
Polish-Thai	8	0.04
Polish-Vietnamese	9	0.045

Russian Language class metrics:

Actual Native Language - Predicted Language	**Misclassification Count**	**Misclassification Rate**
Russian-Arabic	10	0.05
Russian-Cantonese	8	0.04
Russian-Japanese	14	0.07
Russian-Korean	5	0.025
Russian-Mandarin	6	0.03
Russian-Polish	23	0.115
Russian-Spanish	14	0.07

Russian-Thai	0	0
Russian-Vietnamese	6	0.03

Spanish Language class metrics:

Actual Native Language - Predicted Language	Misclassification Count	Misclassification Rate
Spanish-Arabic	17	0.085
Spanish-Cantonese	2	0.01
Spanish-Japanese	11	0.055
Spanish-Korean	5	0.025
Spanish-Mandarin	14	0.07
Spanish-Polish	17	0.085
Spanish-Russian	14	0.07
Spanish-Thai	5	0.025
Spanish-Vietnamese	11	0.055

Thai Language class metrics:

Actual Native Language - Predicted Language	Misclassification Count	Misclassification Rate
Thai-Arabic	12	0.06
Thai-Cantonese	10	0.05
Thai-Japanese	7	0.035
Thai-Korean	15	0.075
Thai-Mandarin	9	0.045
Thai-Polish	2	0.01
Thai-Russian	1	0.005
Thai-Spanish	13	0.065
Thai-Vietnamese	11	0.055

Vietnamese Language class metrics:

Actual Native Language - Predicted Language	Misclassification Count	Misclassification Rate
Vietnamese-Arabic	12	0.06
Vietnamese-Cantonese	29	0.145
Vietnamese-Japanese	11	0.055
Vietnamese-Korean	13	0.065
Vietnamese-Mandarin	19	0.095
Vietnamese-Polish	13	0.065
Vietnamese-Russian	12	0.06
Vietnamese-Spanish	11	0.055
Vietnamese-Thsi	8	0.04