Assignment 5: Putting it all together with Emojis (due 9 May 2020, 11.59 pm EST)

In this last assignment, the goal is to do a real-world NLP task (from SemEval: Semantic Evaluation workshop[1]), using any tools or knowledge you have learned throughout the semester.

The goals of this assignment are:

(1) To predict, given a tweet in English, its most likely associated emoji (**20** points)

(2) To predict, given a tweet in Spanish, its most likely associated emoji (**20** points)

(3) To use *any* type of multilingual transfer learning to see if you can use English data to improve Spanish emoji prediction or vice versa (**20** points)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
USA	❤️	😍	😂	💕	🔥	😊	😎	✨	💙	😘	📷	🇺🇸	☀️	💜	😉	💯	😁	🎄	📸	😜
ESP	❤️	😍	😂	💕	😊	😘	💪	😉	👌	🇪🇸	😎	💙	💜	😜	💞	✨	🎶	💘	😁	🔝

Figure 1: Emojis per language

NOTE: the class labels for English and Spanish are different (see Figure 1). For example, class label *5* in English is class label *4* in Spanish, class label *6* in English is class label *10* in Spanish, etc. Furthermore, some emojis in Spanish don't have corresponding labels in English e.g., class label *6* in Spanish. To do the multilingual transfer learning, you have to align the labels beforehand, and be aware that the English data (or Spanish data) will be useful for only some (and not all) of the Spanish emoji class prediction (or English emoji class prediction respectively).

[1] http://alt.qcri.org/semeval2018/index.php?id=tasks

The materials provided in this zip file are:

```
code/
|-- data
|   |-- mapping
|   |   |-- english_mapping.txt
|   |   `-- spanish_mapping.txt
|   |-- test
|   |   |-- english_test.labels
|   |   |-- english_test.text
|   |   |-- spanish_test.labels
|   |   `-- spanish_test.text
|   `-- train
|       |-- english_train.labels
|       |-- english_train.text
|       |-- spanish_train.labels
|       `-- spanish_train.text
`-- scorer_semeval18.py
```

(1) Train and test set for English (90k and 10k) and Spanish (19k and 1k) adapted from the multilingual emoji prediction task[2] from SemEval 2018.

(2) The mapping from class label to emojis (i.e., the information in Figure 1) in the folder `mapping/` for English and Spanish.

(3) The evaluation script, `scorer_semeval18.py` that will give the Macro- and Micro-accuracy. Given the gold labels and your predicted labels, you can run the script with:

```
python scorer_semeval18.py gold_labels_file predicted_labels_file
```

Deliverables

Here are the deliverables that you will need to submit.

```
|-- code/
|   |-- ...
|   `-- README.txt
`-- Writeup.pdf
```

Multilingual Resources

- OpenNMT[3]: Open source neural machine translation.
- Facebook MUSE[4]: A library for multilingual embeddings and word-to-word translations.
- MultiBERT[5]: BERT, trained on a 104 languages.
- Europarl[6]: Parallel corpus for machine translation.

[2] https://competitions.codalab.org/competitions/17344#learn_the_details-overview
[3] http://opennmt.net/
[4] https://github.com/facebookresearch/MUSE
[5] https://github.com/google-research/bert/blob/master/multilingual.md
[6] http://www.statmt.org/europarl/