Experiment 1

*Start Date: Oct. 12, 2022
*Report Due Date: Oct. 26, 2022

Notes:

- Each experiment must be done individually. You can search material through Internet but remember to mark it.
- Write an experiment report to describe and analyze the experimental observations.
- You should pack your files including **source code, report, readme file, and other related files** in one .zip/.rar/.7z file.
- Please submit on the Web Learning platform. Do NOT Email or Wechat your report to the instructor or TAs.
- No late report is accepted. No exceptions.
- You can write the report using English or Chinese.

Task 1:

Learn about how to get the dense representations of words via off-the-shelf packages.

Goal:

In this task, you are expected to learn how to get dense representations via off-the-shelf packages. Besides, you will also learn how to use the jieba package for Chinese word segmentation.

Data:

Here we provide a subset of the Chinese wiki. You can download it via this link.

Experiment Step:

- Install the package jieba. Using the *jieba.cut* function to segment any sentence you choose from the corpus we provided. Try to change the parameter such as **cut_all** and **HMM**, then compare the result. Please write down your conclusion.
- Choose arbitrary paragraph from the corpus. Use the *jieba.analyse.extract_tags* and *jieba.analyse.textrank* function to individually extract keywords. Try to improve the quality of keywords.
- Use the word2vec function in the GENSIM package to get the dense word vectors from the provided data. Before your training, you need to use jieba to split the sequences into words. After your training, you can use the built-in *most_similar* function to show your results by listing the words similar to the word you choose. Change the parameter of a function, compare the difference in results and try to analyze them.

Task 2:

Learn about how to use the dense representations of words.

Goal:

In this task, you are expected to use the LSTM and CRF for sequence labeling.

Data:

Here we provide the train and test dataset. You can download it via this link. We use the B/I/S to mark the start/indepent/single for the chunk words in sentence. For example:

```
<s> 成功 入侵 民主党 的 电脑系统 </s>
<s> B  I  B  I  B  I  I  S  B  I  I  I</s>
```

B notes the beginning of a chunk while I for other parts. S notes the chunk containing only one word.

Experiment Step:

- Preprocess the train and test set by converting the original sequence label into BIS label format.
- You can download a pretrained word embedding file from here. There is an example to use it.
- Built your sequence labeling model using BiLSTM and CRF. Analyze the performance of your model and try to improve it.