**NLP application (April 27, 2022) Second section**

**Tutorial:** Codes for the VISION project (violence, health and society)

**Preparation:**

* Codes location: <https://github.com/KHP-Informatics/vision>
* Before using these notebook, you should download your python, jupyter notebook and install them:

Following links are for your kind reference:

1. <https://realpython.com/installing-python/>
2. <https://medium.com/@GalarnykMichael/install-python-anaconda-on-windows-2020-f8e188f9a63d>
3. <https://problemsolvingwithpython.com/01-Orientation/01.04-Installing-Anaconda-on-MacOS/>

* If not, you may use the google service to run the codes: <https://colab.research.google.com/>

To use this online service, please create a google account first.

**Explanations of the tutorials:**

1. Try to focus on "NLP\_supervised learning.ipynb" first rather than other notebooks. This notebook aims to help you learn how to use supervised learning methods to tackle a classification problem. The data applied in this notebook are: 'test2.csv' and 'train2.csv'. They were generated from 'potentially\_abuse\_V2.csv' which aims to conduct a sentence level classification. If label=1, means certain sentence include words that suggest the existence of violence behavior. label = 0, indicates no violence or not sure whether there exists violence. (Decision Tree Classifier (>85%) performs better than other algorithms.)
2. If you want to try a post level classifier, replace the datasets test2.csv' and 'train2.csv into 'test.csv' and 'train.csv' separately. These two datasets were generated from 'emotional\_physical.csv'
3. We prepared two sampled datasets for you to practice the labelling work which may help us understand the basic idea of NLP. The first one "sampled\_100\_emotional\_physical.xls" is the posts level coding example, add one column for yourself to annotate them based on your own judgement. The second one "sampled\_100\_potentially\_abuse.xls" is the sentence level coding example, add one column for yourself to annotate them based on your own judgement.

**Additional practices:**

1. The unsupervised learning methods applied the emotional and physical violence dataset (one post as the analyze unit, emotional\_physical.csv, see notebook, NLP-unsupervised learing.ipynb).
2. The supervised and CNN RNN applied the violence dataset (one sentence as the analyze unit, 'potentially\_abuse\_V2.csv', see notebooks RNN\_text\_processing\_WorkShop.ipynb and CNN\_text\_processing\_WorkShop.ipynb).