# Hands-on 1

Due Date: 15:10, October 13 Thursday, 2017

TA: 鄒澔仁 howardtzou.eecs03@nctu.edu.tw

1. Please encode each movie line into a vector of dimension 300 where each dimension represents a word using the below **TF-IDF** formula.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = frequency of i in j

$df_i$ = number of sentences containing i

$N$ = total number of sentences

**Submit your answer of task 1 to: https://goo.gl/forms/8MWFBoP7TSVOgFBB2**

2. Please encode each movie line into a vector of dimension 300 where each dimension represents a word and its POS tagging using the **TF-IDF** formula given in Task 1.

**Submit your answer of task 2 to: https://goo.gl/forms/1Tqj3yikaWsmemvt2**

3. Find out which movies are different from the others using **cosine similarity** and **another method you learned from class**. For each method, you need to list a set of 3 movies that are similar to one another, and a movie that is different from the set of movies.

**Submit your answer of task 3 to: https://goo.gl/forms/gEB9Zw3QeJoIlrpk1**

4. Simplify each sentence by reducing its dependency parse tree to first-degree dependency relation (i.e. words that are not directly connected to the root of the tree should are excludes); then, build a bigram model and calculate its entropy. In addition, compare its value with the entropy of the bigram model for the original sentences.

**Submit your answer of task 4 to: https://goo.gl/forms/JrUIwvRw3vpZFDAz1**

**Reminder**

- For the rules on formatting your output, refer to **slides.pdf**
- **Check your demo progress here.**
- You many use any packages such as NumPy, pandas to accomplish the tasks.

- Cheating is strictly prohibited. You'll get zero marks if you're caught copying others' work.

- The deadline for Late Submission is 23:59, October 25 Wednesday, 2017. Submissions after hands-on section will be considered Late Submission, and no submissions are accepted after the Later Submission deadline.

- Raise your hand or come to TA if you have any questions.