**CS 410 Project Proposal**

**Yiao Ding**

**yiaod2@illinois.edu**

1. **What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.**

   My team has only one member, my name is Yiao Ding and my NetID is yiaod2. I will also be the captain in my team.

2. **What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?**

   I select my free topic as creating model to judge the similarity between two sentences. The task is given two sentences, syntactic structure of both sentences (subject, predicate, and object) should be extracted and compared to determine if the two sentences are similar or not. Investigation in the similarity between sentences is interesting because it can be applied to many areas like determining the plagiarism, or making some predictions based on existing descriptions.
   To achieve the goals, the planned approach is to first find some training data (sentences we already know the similarity), creating an algorithm to parse the sentence and extract the syntactic structure, then transferring the extracted data into some input vectors, and finally input those vectors into some machine learning model to train and make the predictions of similarity. I planned to use some dataset from Microsoft as the training data, which can be found here: https://www.microsoft.com/en-us/download/details.aspx?id=52398
   I plan to use the accuracy to evaluate my model to determine its capacity of judging the similarity between sentences. Expected outcome is set to 70% at this proposal stage.

3. **Which programming language do you plan to use?**

   Python will be used for this project.

4. **Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.**

   Since only one team member, workload should be greater or equal than 20 hours. Here is a initial estimation of main tasks to be completed:

   a). Working on an algorithm to extract the syntactic structure of a sentence (~10 hours)

   b). Working on transforming the extracted structure into vector representation (~10 hours)

c). Working on creating a machine learning model to train the data and make predictions (~10 hours)

Each step costs about 10 hours and the total estimation is about 30 hours.