Chris Huber: 902339417 02/16/2022
CSC820-1, Project Update #1

---

1) Project topic/goal
My goal for this project is to devise an algorithm that will correctly classify a corpus of 433,000 sentences from Multi-Genre Natural Language Inference dataset as either an entailment, a contradiction, or neutral. The input will come from the dataset described at https://cims.nyu.edu/~sbowman/multinli/ (and possibly other datasets I am able to find) and the output will be a .csv file with a sentence id and a corresponding classifier for now or potentially a web front end to submit corpuses for analysis.

2) Why this project topic?
I chose this topic foremost because I have never done this type of classification before and quite frankly had never heard the term entailment before researching this as a potential project. I feel that it is a good way to deepen my understanding of NLP by broaching the topic of inference. It is also quanifiable using an accuracy score and was previously a Kaggle competition, so there is a testing framework already set up that will be useful in evaluating my models.

3) What resources exist for this topic?
dataset: 433K sentences with matching classifiers at https://cims.nyu.edu/~sbowman/multinli/
codebase: https://github.com/sebastianruder/NLP-progress/blob/master/english/natural_language_inference.md
scientific papers: EdgeBERT: Sentence-Level Energy Optimizations for Latency-Aware Multi-Task NLP Inference, Semantic Lifiing of Unstructured Data Based on NLP Inference of Annotations, Briding the Gap Between Training and Inference for Neural Machine Translation

4) How will I know/evaluate/assess that I have succeeded at meeting my project goal?
The score from submitting my results against the Kaggle competition's submissions should not only tell me my acccuracy score, but also how it rates against other people's submissions. Since the submission is simply a .csv file, it is a model-independent approach that should allow me to compare the effectiveness of different algorithms. (I should also be able to derive accuracy on my own, but the Kaggle submissions will streamline the process.)

5) Timeline of project development:
2/16: Meet with Prof. Kulkari to discuss project
February: begin researching Inference as it relates to NLP, identify what technologies I will need to learn to complete the project
3/1: begin writing code
3/16: Meet with Prof. Kulkari with prototype of project to demo
March: refine prototype based on feedback
April: write models and refine them to improve accuracy

May: finish refining models and write Term Paper describing my methodologies and results