MAIS 202 – PROJECT DELIVERABLE 1

Anthony Dumais & Alberto Lopez

Part 1 – Choose your dataset

Getting Started competitions are semi-permanent competitions on Kaggle that are meant to be used by new users just starting in the field of machine learning. They are designed to be good first or second ML projects. The one we chose is called "Contradictory, My Dear Watson" and is about detecting contradiction and entailment in multilingual text using TPUs. We selected this competition because it seemed more challenging than the other ones from the same category and because the subject was somewhat related to our program (Cognitive Science). For more details about the competition: https://www.kaggle.com/c/contradictory-my-dear-watson.

Part 2 – Methodology

Data processing: Since the dataset comes from a Kaggle competition, we are confident that it is feasible. The file used for training contains the ID, premise, hypothesis, and label, as well as the language of the text and its two-letter abbreviation. We think that all this information will be useful. Non-English languages will need to be translated to English and the textual data will be tokenized.

Machine Learning Model: The goal is to predict whether a given hypothesis is related to its premise by contradiction, entailment, or whether neither of those is true (neutral). We are considering different ideas like classification models and NLP models that do not require Tensor Processing Unit (such as LSTMs, simple RNNs), or pre-trained language models such as BERT. We may try TPU computing if everything goes well. We'll have to learn more about these models before making more precise decisions.

Evaluation Metric: The percentage of relationships correctly predicted will be used as a score to measure the accuracy of our model.

Final Conceptualization: We will integrate our model in a sample landing page webapp.

Part 3 – Application

The webapp will take a premise and a hypothesis as input from the user and will output the relationship between the two (contradiction, entailment or neutral/unrelated). To make that possible, we plan on attending the Intro to Flask Workshop.