Name: Joseph Jinn

Date: 4-18-19

Course: CS-344 Artificial Intelligence

Instructor: Professor VanderLinden

Assignment: Final Project

Project Draft #1 – Vision Statement

Project Domain:

- Use Keras (TensorFlow) deep neural networks to do SLO topic classification over the standard TBL topics using Tweets relating to mining companies:
  - Environmental
  - Economic
  - Social
- Social License to Operate (SLO) – refers to the ongoing acceptance of a company or industry's standard business practices and operating procedures by its employees, stakeholders, and the general public.
- Triple Bottom Line (TBL) – a framework or theory that recommends that companies commit to focus on social and environmental concerns just as they do on profits.
  - Gauges a corporation's level of commitment to corporate social responsibility and its impact on the environment over time.
- Resources:
  - https://www.investopedia.com/terms/s/social-license-slo.asp
  - https://www.investopedia.com/terms/t/triple-bottom-line.asp
  - Classifying Stance Using Profile Texts – Anonymous ACL submission
- If necessary, manually hand-tag a new dataset to create training, validation, and test subsets for use in SLO topic classification.
- I guess what I'll actually be doing will depend on what the next step is in taking this idea further.

Framing the Problem (from Google Machine Learning Problem Framing):

The machine learning model should be able to correctly identify whether the Tweet is positively oriented or negatively oriented as an environmental topic, economic topic, or social topic.

The ideal outcome is such that the model predicts with 90%+ confidence the orientation (positive or negative) and type of topic (environmental, economic, or social) of the Tweet.

The success metric is that the probability value for the correct label is the highest among all values given by a softmax layer.

The model is deemed a failure if the probability value for the correct label is not the highest among all values given by a softmax layer.

The output of the machine learning model will be an array of probabilities, providing confidence values in how sure the model is that the Tweet is one of the triple-bottom-line topics. The values should sum to a total of 1.

The results will be used to help determine the Social License to Operate (SLO) of mining companies, where SLO is a measure of the company's level of support from their constituencies.

Formulating the Problem (from Google Machine Learning Problem Framing):

The problem is best suited to the use of a multi-class single label classification training model.

Simplifying, the training model should output for each class the probability that the Tweet belongs to that class. This process should be performed for all classes that define all possible SLO classifications.

Data (from Classifying Stance Using Profile Texts – Anonymous ACL submission .pdf document):

Twitter data – raw Tweets collected from the Twitter API.

- Pre-processing:
  - Tokenize texts using the CMU Tweet Tagger. Stop words are retained.
  - Remove "RT" tags marking re-tweets.
  - Shrink character elongations except in usernames.
  - Replace URL's, mentions, year/time/cash/hashtag items with placeholders.
  - Down-case all text.
  - Remove tweets that are not labelled as some variant of English either by the Twitter or by Polyglot.
  - Remove the tweets found to not be associated with any company.
- Inputs:
  - Input 1: N-gram counts.
  - Input 2: target company name presence.
  - Input 3: word embeddings.
- Outputs:
  - SLO classification:
    - Environmental
    - Economic
    - Social

We should initially start with:

- Input 1: N-gram counts.