CS-522

Project Progress Report I – Data/Implementation

Team 3

Abstract

This project aims to build a model that enables researchers to make quantitative arguments about environmental awareness and cultural change. The dataset has been taken from reddit. We'll be using an open-sourced neural network-based technique for Natural Language Processing called BERT (Bidirectional Encoder Representations from Transformers). BERT has been pre-trained on the whole English Wikipedia and can be fine-tuned on smaller task-specific datasets. The quality of the model will further be tested on synonymy and also the Stanford test.

1. Data

1.1. Data Source

This dataset was collected from reddit, searching for articles that contain the keyword 'humanities' . The data is in the format of compressed tar.gz file which has 118,059 files inside it in the format of .txt and has an actual size of 322MB .To ensure that the program doesn't run out of memory, we will train our dataset in multiple batches.

1.2. Data Format

This data is comprised of .txt files which are in natural text language and processed with the Illinois Wikifier. The datasets are all in English except one which is in Norwegian. The data is

1.3. Programming Language, packages

The language Python will be used for fine-tuning and testing the model on Google Colaboratory which allows us to write and execute arbitrary python code through the browser. Google Colab is a Jupyter Notebook environment that doesn't require any setup and runs entirely on cloud. We will be using many machine learning libraries like PyTorch, Pandas, TensorFlow, Transformers by hugging face etc. to pre-train and run the model.

2. Preliminary Experiments

2.1. Preliminary Results

The data set contains files(.txt) which contains various paragraphs and has already been wikified using Illinois Wikifier. The data has been pre-processed by removing punctuation, stop words and performing tokenization and stemming. All the text files have been combined together into a csv file and then converted into a format(tsv) with different columns recognized by BERT. The data has been then loaded into BERT and a language model based on the vocabulary of the corpus has been built. It will then then be tested for wikientries and also on the SQUAD (Stanford Question Answering Dataset*)*.