Technology Review

NLPmaps

Meenal, Andrew, Prithviraj, Shijie

Background

- We are developing an NLP selection framework that can analyze text data from various sources in chemical engineering plants to provide real-time insights and predictions.
- Our goal is to build a flexible and scalable framework that can effectively select the best algorithm.
- The framework will also include various techniques for pre-processing and analyzing the text data, such as data cleaning, feature extraction, and model training.
- Aim to improve operational efficiency, reduce maintenance costs, and enhance product quality and safety in the chemical engineering plant.

Technologies Considered

- NLP Algorithms
  - Computers don’t understand words
  - NLP embeds text for downstream models
  - Dozens of NLP models exist
    - Select diverse but not redundant set of models
- Data Management
  - Data presented in a variety of forms
  - Algorithms needed for pre-processing
  - Numpy, Pandas
- Visualization
  - Convey results in succinct figures
  - Pyplot, Seaborn

Example Models

- Word2Vec
  - Converts words to vectors that are understandable by the computer
  - Similar to one-hot encoding
  - Uses probability functions to:
    - Predict a word based on surrounding words (bag of words model)
    - Predict surrounding words based on target word (skip-gram model)
- GLOVE
  - Unsupervised ML algorithm
  - Uses KNN to discern similarity of words
  - Computes distance between word clusters for nuanced understanding
- ELMO
  - Highly Advanced NN
  - Billions of neurons and millions of parameters
  - Entire context of a word is used

Advantages

- Word2Vec
  - Can handle large amounts of text data
  - Capture the semantic relationships between words, such as synonyms and antonyms
  - Has many pre-trained word embeddings are available, making it easy to incorporate into NLP pipelines
- GLOVE
  - Can handle rare words well
  - Can capture not only semantic relationships between words but also the syntactic relationships between them
- ELMO
  - Captures the nuances of word meanings in different contexts.
  - High-quality embeddings for rare words
  - Can outperform other embedding models on certain NLP tasks, such as question answering and named entity recognition.

Disadvantages/Drawbacks

- Word2Vec
  - Inability to handle unknown or OOV (out-of-vocabulary) words
  - No shared representations at sub-word levels
  - Scaling to new languages requires new embedding metrics
- GLOVE
  - Unsupervised ML algorithm is not effective in identifying homographs
  - Inability to handle unknown or OOV (out-of-vocabulary) words
  - Both Word2Vec and GLOVE are not contextual
- ELMO
  - Although ELMO takes into account of the context for a particular word but they are two separate vectors (i.e. it simply concatenate the left-to-right and right-to-left information), and therefore it can't take advantage of both left and right contexts simultaneously

Thank You