Quiz 8

1. As I have some side projects relevant to Knowledge Graph and take Human-Language-Technologies class; hence, I found the dataset named [**Open Research Corpus**](https://allenai.org/data/open-research-corpus) interesting. The dataset is available at <https://allenai.org/data/open-research-corpus>. This dataset is collected and distributed by Allen Institute for AI in Washington, U.S.A. This dataset is a collection of 39 million published research papers over various research topics: Neuroscience, Biomedical, Computer Science, etc. Hence, this dataset is guaranteed to be diverse and well-attributed (in scientific research papers, there are many distinct words to describe a document/paper).
2. This dataset is a good dataset for Open Information Extraction problem that I could build a model to extract keywords to describe a document; and based on extracted keywords, we could build a model to cluster documents that the model could search for documents given user-typed keywords.

Hence, given the dataset, my project is to provide a Google-like search engine for researchers and general users to search scientific papers based on their typed keywords. Google works well in-general to look for research papers. However, intensive amount of ads distracts users from looking for papers. Steps to build the project:

* Natural-Language-Processing Engine
  + Build a IE-based engine to automatically extract keywords from each document to describe each paper/document. Then, build a look-up program to display documents which contain most keywords.
    - For IE-based approaches,
      * Distinct words are collected; and the word frequency are computed
      * Implement IE methods like TF-IDF (term frequency – inverse document frequency) to find less frequent of a document to describe the document.
  + Instead of keywords, build a NLP-based Deep Learning model to extract features that represents the document and cluster papers based on extract features.
* Then, setup a simple API to receive REQUESTS with input keywords and respond with most similar documents

1. In IE chapter, Important words (e.g. TF-IDF) and NER could be used. After inspecting dataset, instead of using the entire research paper, I would use “Abstract section” and “Title” as the corpus which contains concise information and keywords of a research paper. Hence, implementation of TF-IDF would help find the scientific keywords (e.g. Deep Learning, Knowledge Graph) . Then, NER could be used to extract author and organization names which contributed to the research paper (e.g. Ian Goodfellow in Generative Adversarial Network). Based on the extracted scientific keywords and proper name entities, users could search for documents more precisely by either author’s names or documents’ content.