**STEPS INCLUDED IN THE PYTHON CODE**

**Step 1: Preprocessing Abstracts**

* Cleaning the abstracts and extracting only the sentences with action(keywords) from the entire abstract
* Sentence Tokenization
* Word Tokenization
* Stemming
* Cleaning sentences which include removing special characters, extra spaces and converting everything to lower case
* Generating a file with containing the sentences as said above.

**Step 2: Training POS Tagger**

* Training the tagger on med post tagger (contains sentences which are already tagged and we are using these sentences as train data for our tagger).
* Using combined tagger.
* No intermediate file is generated in this step as we are training the model

**Step 3: Named Entity Recognition (NER)**

* Running tagger on our data and collecting only NOUN forms (NER).
* Generating a file having the words (Noun forms) and associated Parts of Speech tag

**Step 4: Accessing UMLS data base to find the Semantic Type (Gene/Protein)**

* Accessing the UMLS database created in PhpMyAdmin.
* Generating a file containing only Proteins and Genes with their Semantic type and Parts of Speech tag.

**Step 5: Final data set generation**

* Generating final dataset containing sentences with more than one Gene or Protein and the action between them.
* The above generated dataset is used for Deep learning to train the predictive model.

**Step 6 : Dataset generation for visualization**

* Run the code in the step : 6 to generate a dataset which is useful to create gene-protein-action interaction network