**Deal Sentence(Amount and Organization) classification pipeline:-**

* **Purpose:-**

The main purpose of the deal sentence(amt and org) classification is to take news article as the input and output the deals sentence (which have preferrably less than 256 characters and do contain 2 or more organizations/persons along with a deal keyword, deal amount) and henceforth also output the list of organizations, deal amount present in the output deal sentence.

* **Package installation and requirements:-**

Here are the list of python packages that need to be installed:-

* 1. common nltk packages basically sent\_tokenize, word\_tokenize, stopwords, WordNetLemmatizer.
  2. Specific usage nltk packages like StanfordNERTagger, CoreNLPParser for ‘POS’ and ‘NER’ tagging.
  3. Packages for machine learning like keras, sklearn, torch, transformers.
  4. Other packages like pymongo, warnings, re, csv that already come installed while installation of python editor.

Since I used BertForSequenceClassification for training my deal sentence classifier, there is usage of bert tokens for the purpose of training. And I used base-uncased version of bert tokens, which should be downloaded prior to its usage. Here is the syntax for it:-

f*rom transformers import BertTokenizer*

*# Load the BERT tokenizer.*

*print('Loading BERT tokenizer...')*

*tokenizer = BertTokenizer.from\_pretrained('bert-base-uncased', do\_lower\_case=True)*

* **Usage:-**
  1. Firstly data in form of strings needs to be input from a source. For my case, I had my input from mongo collection . In that, I considered the ‘title’ and ‘description’ fields of each records.
  2. In case the input contains multiple sentences, they need to be pre-processed using ‘sent\_tokenizers’, and other preprocessing methods, so that we end up having a list where each element of the list is a single element list containg one sentence.
  3. Now stanford NER tagger is run over the obtained list so that we get labels (either ORGANIZATION, PERSON, or MONEY) for each word of each sentence(that are contained in a singular fashion in each sublist of main list).
  4. Then the indices of the sentences that have money labels are filtered. For those sentences stanford POS tagger is ran. From those sentences, stopwords and words having positive NER labels are removed. And for remaining part of the sentences, we check POS labels of each words whether they are verb, adjectives or nouns and filter them out.
  5. Now we change those filtered words into their base words using ‘WordNetLemmatizer’ and check if the base words are present in BOW containing deals keywords. If they are present in BOW , the corrsponding sentences are appended into a sentence list. And finally create Dataframe for this list.
  6. Next the trained bert model is called. Then the token ids and masking ids for each sentence of the dataframe are generated. Now these variables are feeded into the model to get the predicted results.
  7. Thereafter the sentences whose labels are identified ‘1’ are considered and stanford NER tagger is ran over those to get organizations list and deal amount from those.
  8. Finally a dataframe having Sentence, their predicted labels(status) and deal amout, organization given their status is ‘1’ is created.
  9. Once the dataframe is created, a csv can be created from it and stored to a desired location.