Tally Solutions Pvt. Ltd.

Named Entity Recognition (NER) using Natural Language Tool Kit (NLTK)

NLTK :

NLTK is a toolkit build for working with NLP in Python. It provides us various text processing libraries with a lot of test datasets. A variety of tasks can be performed using NLTK such as tokenizing, parse tree visualization, etc.

It is used to process the text for NLP tasks.

Various techniques used in NLTK :

- Stopwords Removal - [“i”, “me”, “are”, ...]
- Tokenization - Extract unique tokens from a text
- Stemming - All the words like “playing, played, player, . . .” converted to “play”
- Lemmatizing - Returns a base word for each word
- Parse Trees - Used to extract entities like noun, verb, adjective from a text
- Parts Of Speech (POS) Tagging - Tags each token with its part of speech.

```
[('European', 'JJ'),
 ('authorities', 'NNS'),
 ('fined', 'VBD'),
 ('Google', 'NNP'),
 ('a', 'DT'),
 ('record', 'NN'),
 ('$', '$'),
 ('5.1', 'CD'),
 ('billion', 'CD'),
 ('on', 'IN'),
 ('Wednesday', 'NNP'),
 ('for', 'IN'),
 ('abusing', 'VBG'),
 ('its', 'PRP$'),
 ('power', 'NN'),
 ('in', 'IN'),
 ('the', 'DT'),
 ('mobile', 'JJ'),
 ('phone', 'NN'),
 ('market', 'NN'),
 ('and', 'CC'),
 ('ordered', 'VBD'),
 ('the', 'DT'),
 ('company', 'NN'),
 ('to', 'TO'),
 ('alter', 'VB'),
 ('its', 'PRP$'),
 ('practices', 'NNS')]
```

Tokens with coressponding POS tag

- IOB Tagging (Inside, Outside, Beginning) - “B” means the token begins an entity, ”I”means it is inside an entity, “O” means it is outside an entity, and “ “ means no entity tag is set.

  Example = 'European authorities fined Google a record $5.1 billion on Wednesday for abusing its power in the mobile phone market and ordered the company to alter its practices'

  Here,

  $ 5.1 Billion is an entity,

  IOB tagging = [., ., ., ., ., ,
  (“$”, “B”),
  (“5.1”, “I”),
  (“Billion”, “I”),
  (“on”, “O”),
  ., ., ., ., ., .,.]

NLTK with SPACY :

SpaCy recognizes the following built-in entity types:

PERSON - People, including fictional.

NORP - Nationalities or religious or political groups.

FAC - Buildings, airports, highways, bridges, etc.

ORG - Companies, agencies, institutions, etc.

GPE - Countries, cities, states.

LOC - Non-GPE locations, mountain ranges, bodies of water.

PRODUCT - Objects, vehicles, foods, etc. (Not services.)

EVENT - Named hurricanes, battles, wars, sports events, etc.

WORK_OF_ART - Titles of books, songs, etc.

LAW - Named documents made into laws.

LANGUAGE - Any named language.

DATE - Absolute or relative dates or periods.

TIME - Times smaller than a day.

PERCENT - Percentage, including "%".

MONEY - Monetary values, including unit.

QUANTITY - Measurements, as of weight or distance.

ORDINAL - "first", "second", etc.

CARDINAL - Numerals that do not fall under another type.

However, spacy and NLTK models can be optimized by adding different labels to the pre trained models.

Limitations :

- Layout information won’t be taken into account.
- Wont able to differentiate among invoice_date-due_date, sender_name, reciever_name, etc.
- These models uses parts of speech rather than context to predict the output.