Importing Libraries

In [246]:

```
import pandas as pd
import numpy as np
import nltk
import docx
import os
import re
import nltk
from nltk.stem import WordNetLemmatizer
nltk.download('stopwords')
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
nltk.download('punkt')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package stopwords to C:\Users\ashok
[nltk_data]     kumar\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to C:\Users\ashok
[nltk_data]     kumar\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to C:\Users\ashok
[nltk_data]     kumar\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

Out[246]:

```
True
```

Reading File Names

In [247]:

```
from os import listdir
from os.path import isfile, join
file_names = [f for f in listdir('Training_data') ]
print(file_names)
```

```
['100999172-House-Rental-Agreement.pdf.docx', '116950326-December-2012-Renta
l-Agreement.pdf.docx', '136441742-Rental-Agreement-Format.pdf.docx', '142106
117-Rental-Agreement.pdf.docx', '170499354-Anand-Nagar-Agreement.pdf.docx',
'175488575-House-Rental-Agreement.pdf.docx', '18325926-Rental-Agreement-1.pd
f.docx', '195231682-This-RENTAL-AGREEMENT-is-Made-and-Executed-on-24th-Day-o
f-September.pdf.docx', '203615996-Rental-Agreement-Format.pdf.docx', '216973
836-Rental-Agreement-Sample.pdf.docx', '248636461-Rental-Agreement.pdf.doc
x', '249104436-House-Rental-Agreement.pdf.docx', '251798216-Rental-Agreement
-Format1.pdf.docx', '267005869-Rental-Tenant.pdf.docx', '269135973-Udaya-Ren
tal-Agreement.pdf.docx', '269137702-Rental-Agreement.pdf.docx', '269760124-9
7646-41223-Rental-Agreement.pdf.docx', '288024755-Rental-Agreement-1.pdf.doc
x', '294331674-Rental-Agreement.pdf.docx', '308044452-Rental-Agreement.pdf.d
ocx', '323828497-Rental-Agreement-Micky.pdf.docx', '334060786-House-Rental-A
greement.pdf.docx', '343492954-Rent-Agreement-3E.pdf.docx', '36199312-Rental
-Agreement.pdf.docx', '392810415-RENT-AGREEMENT.pdf.docx', '44737744-Maddire
ddy-Bhargava-Reddy-Rental-Agreement.pdf.docx', '46239065-Standard-Rental-Agr
eement-Rental-With-Performance-Fee.pdf.docx', '47854715-RENTAL-AGREEMENT.pd
f.docx', '50070534-RENTAL-AGREEMENT (1).pdf.docx', '54770958-Rental-Agreemen
t.pdf.docx', '54945838-Rental-Agreement.pdf.docx', '56736420-Rental-Agreemen
t.pdf.docx', '62126501-Rental-Agreement.pdf.docx', '62144960-Rental-Agreemen
t.pdf.docx', '63057680-Rental-Agreement.pdf.docx', '63793679-Rental-Agreemen
t.pdf.docx', '6683127-House-Rental-Contract-GERALDINE-GALINATO-v2-Page-1.pd
f.docx', '6683129-House-Rental-Contract-Geraldine-Galinato-v2.pdf.docx', '77
112358-Jaggu-Rental-Agreemnt.pdf.docx', '81655723-Rental-Agreement.pdf.doc
x', '95421373-Agreement.pdf.docx', '95980236-Rental-Agreement.pdf.docx', '99
699504-Rental-Agreement-English-Model.pdf.docx']
```

Creating a list of Documnets

In [248]:

```
df_data=[]
for x in file_names:
    doc = docx.Document('Training_data'+"/"+str(x))  #Parsing through all address to get th
    data = ""
    fullText = []
    for para in doc.paragraphs:#Parsing Paragraphs.
        fullText.append(para.text)
        data = '\n'.join(fullText)
    df_data.append(data)# Adding Documents to the list.
```

In [249]:

```
df = pd.DataFrame (df_data,columns=['Agreement'])#Converting list to DataFrame
```

In [250]:

```
df.insert(0,'Filename',value=onlyfiles)#Inseting File Names
```

In [251]:

```
df.head()
```

Out[251]:

	Filename	Agreement
0	100999172-House-Rental-Agreement.pdf.docx	HOUSE RENTAL AGREEMENT\nRental Agreement made ...
1	116950326-December-2012-Rental-Agreement.pdf.docx	\n\nROOM RENTAL AGREEMENT\nThis is a legally b...
2	136441742-Rental-Agreement-Format.pdf.docx	\nTENANCY AGREEMENT\n\nThis Tenancy Agreement ...
3	142106117-Rental-Agreement.pdf.docx	RENTAL AGREEMENT\nThis agreement of Tenancy is...
4	170499354-Anand-Nagar-Agreement.pdf.docx	\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\nRENTAL A...

In [252]:

```
df_1=pd.read_csv('TrainingTestSet.csv')
df_2=pd.read_csv('ValidationSet.csv')
```

In [253]:

```
df_2
```

Out[253]:

	File Name	Aggrement Value	Aggrement Start Date	Aggrement End Date	Renewal Notice (Days)	Party One	Party Two
0	24158401-Rental-Agreement	12000	01.04.2008	31.03.2009	60.0	Hanumaiah	Vishal Bhardwaj
1	63793679-Rental-Agreement	9000	01.09.2011	31.08.2012	NaN	S Parthasarathy	Hari Kiran Tholeti
2	95980236-Rental-Agreement	9000	01.04.2010	31.03.2011	30.0	S.Sakunthala	V.V.Ravi Kian
3	156155545-Rental-Agreement-Kns-Home	12000	15.12.2012	14.11.2013	30.0	V.K.NATARAJ	VYSHNAVI DAIRY SPECIALITIES Private Ltd
4	195231682-This-RENTAL-AGREEMENT-is-Made-and-Ex...	13000	06.04.2013	05.03.2014	30.0	C.BHAGYAMMA	JP INTERIO
5	228094620-Rental-Agreement	15000	07.07.2013	06.06.2014	30.0	KAPIL MEHROTRA	.B.Kishore
6	239419594-Rental-Agreement	9000	07.07.2014	06.06.2015	90.0	Abraham	Annamalai
7	269135973-Udaya-Rental-Agreement	8300	01.04.2014	31.02.2013	30.0	Giddappa	Pottumurthi Udayalaxmi

In [254]:

```
df_1=df_1.rename(columns={"File Name": "Filename"})
```

In [255]:

```
for x in range(len(df)):
    df['Filename'][x]=df['Filename'][x].rstrip('.pdf.docx')
```

In [256]:

```
DF_final=pd.merge(df, df_1, on="Filename")#Merging two DataFrames
```

In [257]:

```
DF_final['Agreement'][0].find('$Agreement Value$')
```

Out[257]:

```
886
```

In [258]:

```
DF_final.head(10)
```

Out[258]:

	Filename	Agreement	Aggrement Value	Aggrement Start Date	Aggremen End Dat
0	100999172-House-Rental-Agreement	HOUSE RENTAL AGREEMENT\nRental Agreement made ...	14500.0	10.01.2011	09.01.201
1	116950326-December-2012-Rental-Agreement	\n\nROOM RENTAL AGREEMENT\nThis is a legally b...	600.0	NaN	NaN
2	136441742-Rental-Agreement-Format	\nTENANCY AGREEMENT\n\nThis Tenancy Agreement ...	15000.0	11.04.2012	10.03.201
3	142106117-Rental-Agreement	RENTAL AGREEMENT\nThis agreement of Tenancy is...	7200.0	01.03.2011	31.02.201
4	170499354-Anand-Nagar-Agreement	\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\nRENTAL A...	8000.0	05.04.2011	04.04.201
5	175488575-House-Rental-Agreement	LEASE AGREEMENT\nThis DEED OF RENTAL AGREEMENT...	600.0	01.09.2012	31.08.201
6	18325926-Rental-Agreement-1	\n\n\n\n\n\n\n\n\nRENTAL AGREEMENT\nThis dee...	4000.0	05.12.2008	31.11.200
7	195231682-This-RENTAL-AGREEMENT-is-Made-and-Ex...	\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n...	13000.0	06.04.2013	05.03.201
8	203615996-Rental-Agreement-Format	\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\nRENTAL...	3500.0	01.02.2008	31.01.200
9	216973836-Rental-Agreement-Sample	\n\nTHIS RENTAL AGREEMENT is made on this, the...	15000.0	23.03.2013	23.03.201

In [245]:

```
for x in range(len(DF_final)):
    DF_final['Agreement'][x]=re.sub(r'[\n|$|.|!|\t|\|/]',' ',DF_final['Agreement'][x])#Remo
```

```
C:\Users\ashok kumar\Anaconda3\lib\site-packages\ipykernel_launcher.py:2: Se
ttingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/s
table/user_guide/indexing.html#returning-a-view-versus-a-copy (http://panda
s.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ve
rsus-a-copy)
```

In [260]:

```
[DF_final.columns]
```

Out[260]:

```
[Index(['Filename', 'Agreement', 'Aggrement Value', 'Aggrement Start Date',
        'Aggrement End Date', 'Renewal Notice (Days)', 'Party One',
        'Party Two'],
       dtype='object')]
```

Preparing Data For Spacy

In [276]:

```
TRAIN_DATA =[(DF_final['Agreement'][0],{'entities': [(904, 909, 'Agreement Value'),(543, 55
             (DF_final['Agreement'][2],{'entities': [(1427, 1433, 'Agreement Value'),(404, 4
             (DF_final['Agreement'][3],{'entities': [(808, 814, 'Agreement Value'),(2345, 23
             (DF_final['Agreement'][4],{'entities': [(771, 777, 'Agreement Value'),(241, 254
             (DF_final['Agreement'][5],{'entities': [(992, 998, 'Agreement Value'),(634, 645
             (DF_final['Agreement'][6],{'entities': [(1229, 1235, 'Agreement Value'),(2656,
             (DF_final['Agreement'][7],{'entities': [(2086, 2092, 'Agreement Value'),(675, 6
             (DF_final['Agreement'][8],{'entities': [(815, 821, 'Agreement Value'),(1243, 12
             (DF_final['Agreement'][9],{'entities': [(2008, 2014, 'Agreement Value'),(1908,

             ]
```

Training Spacy Model

In [277]:

```
import spacy
import random
def train_spacy(data,iterations):
    TRAIN_DATA = data
    nlp = spacy.blank('en')  # create blank Language class
    # create the built-in pipeline components and add them to the pipeline
    # nlp.create_pipe works for built-ins that are registered with spaCy
    if 'ner' not in nlp.pipe_names:
        ner = nlp.create_pipe('ner')
        nlp.add_pipe(ner, last=True)


    # add labels
    for _, annotations in TRAIN_DATA:
         for ent in annotations.get('entities'):
            ner.add_label(ent[2])

    # get names of other pipes to disable them during training
    other_pipes = [pipe for pipe in nlp.pipe_names if pipe != 'ner']
    with nlp.disable_pipes(*other_pipes):  # only train NER
        optimizer = nlp.begin_training()
        for itn in range(iterations):
            print("Statring iteration " + str(itn))
            random.shuffle(TRAIN_DATA)
            losses = {}
            for text, annotations in TRAIN_DATA:
                nlp.update(
                    [text],  # batch of texts
                    [annotations],  # batch of annotations
                    drop=0.2,  # dropout - make it harder to memorise data
                    sgd=optimizer,  # callable to update weights
                    losses=losses)
            print(losses)
    return nlp


prdnlp = train_spacy(TRAIN_DATA, 20)
```

```
Statring iteration 0

C:\Users\ashok kumar\Anaconda3\lib\site-packages\spacy\language.py:479: Us
erWarning: [W030] Some entities could not be aligned in the text "RENTAL A
GREEMENT
This agreement of Tenancy is made..." with entities "[(808, 814, 'Agreemen
t Value')]". Use `spacy.gold.biluo_tags_from_offsets(nlp.make_doc(text), e
ntities)` to check the alignment. Misaligned entities ('-') will be ignore
d during training.
  gold = GoldParse(doc, **gold)
C:\Users\ashok kumar\Anaconda3\lib\site-packages\spacy\language.py:479: Us
erWarning: [W030] Some entities could not be aligned in the text "LEASE AG
REEMENT
This DEED OF RENTAL AGREEMENT EXEC..." with entities "[(992, 998, 'Agreeme
nt Value')]". Use `spacy.gold.biluo_tags_from_offsets(nlp.make_doc(text),
entities)` to check the alignment. Misaligned entities ('-') will be ignor
ed during training.
  gold = GoldParse(doc, **gold)
C:\Users\ashok kumar\Anaconda3\lib\site-packages\spacy\language.py:479: Us
```

```
erWarning: [W030] Some entities could not be aligned in the text "

THIS RENTAL AGREEMENT is made on this, the March..." with entities "[(200
8, 2014, 'Agreement Value')]". Use `spacy.gold.biluo_tags_from_offsets(nl
p.make_doc(text), entities)` to check the alignment. Misaligned entities
('-') will be ignored during training.
  gold = GoldParse(doc, **gold)
C:\Users\ashok kumar\Anaconda3\lib\site-packages\spacy\language.py:479: Us
erWarning: [W030] Some entities could not be aligned in the text "
```

```
RENTAL AGREEMENT
This deed of rental agr..." with entities "[(1229, 1235, 'Agreement Valu
e')]". Use `spacy.gold.biluo_tags_from_offsets(nlp.make_doc(text), entitie
s)` to check the alignment. Misaligned entities ('-') will be ignored duri
ng training.
  gold = GoldParse(doc, **gold)
C:\Users\ashok kumar\Anaconda3\lib\site-packages\spacy\language.py:479: Us
erWarning: [W030] Some entities could not be aligned in the text "
```

```
RENTAL AGREEMENT
This Agreemen..." with entities "[(815, 821, 'Agreement Value')]". Use `sp
acy.gold.biluo_tags_from_offsets(nlp.make_doc(text), entities)` to check t
he alignment. Misaligned entities ('-') will be ignored during training.
  gold = GoldParse(doc, **gold)
{'ner': 2931.6645906865597}
Statring iteration 1
{'ner': 11.838193838705793}
Statring iteration 2
{'ner': 38.93045370350218}
Statring iteration 3
{'ner': 12.719574446111666}
Statring iteration 4
{'ner': 6.470753840427717}
Statring iteration 5
{'ner': 6.417913334071088}
```

```
Statring iteration 6
{'ner': 3.0149815268557862}
Statring iteration 7
{'ner': 1.8679226398865334}
Statring iteration 8
{'ner': 1.9341167676909772}
Statring iteration 9
{'ner': 1.926356342546345}
Statring iteration 10
{'ner': 1.3506800504453114}
Statring iteration 11
{'ner': 1.5910249880830458}
Statring iteration 12
{'ner': 2.1941321071747635}
Statring iteration 13
{'ner': 2.1976133722026487}
Statring iteration 14
{'ner': 1.8581282199743663}
Statring iteration 15
{'ner': 2.8158914832588624}
Statring iteration 16
{'ner': 0.8012024785849241}
Statring iteration 17
{'ner': 1.0530225103691617}
Statring iteration 18
{'ner': 0.4640535881170406}
Statring iteration 19
{'ner': 0.024327205278681464}
```

Testing Spacy Model

In [321]:

```
doc = prdnlp(DF_final['Agreement'][7])
for ent in doc.ents:
    print(ent.text, ent.label_)
```

```
13,000  Agreement Value
06.04.2013  Aggrement Start Date
05.03.2014 Aggrement End Date
1 month  Renewal Notice (Days)
C.BHAGYAMMA  Party One
JP INTERIO Party Two
```