Name - ADITYA ADARSH

LinkedIn - https://www.linkedin.com/in/aditya-adarsh-657320188/ (https://www.linkedin.com/in/aditya-adarsh-657320188/)

Workflow

1. Problem Statement

2. Machine learning Formulation

3. Dataset Loading and Description

4. Exploratory Data Analyses and Preprocessing

5. Making Dataset ready and Modeling

6. Results and Conclusion

1. Problem Statement

Business Requirement - Overview

To build an AI/ML model to extract data from the rental agreements. The rental agreements will be in different data formats and available in the form of PDFs to perform the extraction.

The model should be able to extract the following fields from all the documents,

1. Agreement Value
2. Agreement Start Date
3. Agreement End Date
4. Renewal Notice (Days)
5. Party One
6. Party Two

2. Machine learning Formulation

Refer - Research Paper https://arxiv.org/abs/2002.01861 (https://arxiv.org/abs/2002.01861)

This problem can be solved using Name Entity Recognition - NLP

The named entity recognition (NER) is one of the most data preprocessing task. It involves the identification of key information in the text and classification into a set of predefined categories. An entity is basically the thing that is consistently talked about or refer to in the text.

NER is the form of NLP.

At its core, NLP is just a two-step process, below are the two steps that are involved:

- Detecting the entities from the text
- Classifying them into different categories

Some of the categories that are the most important architecture in NER such that:

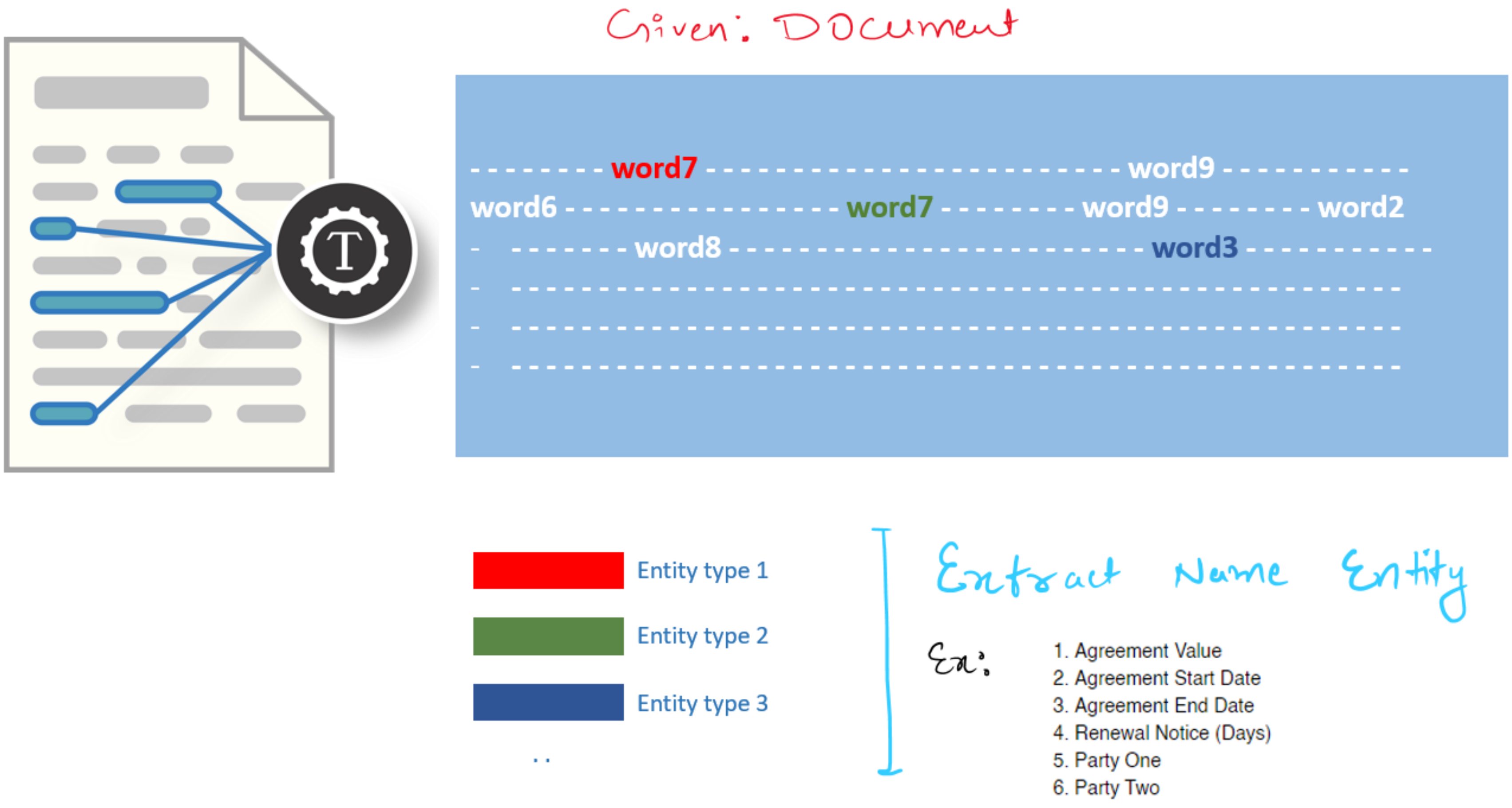
- Person
- Organization
- Place/ location

Source: https://www.geeksforgeeks.org/named-entity-recognition (https://www.geeksforgeeks.org/named-entity-recognition)

In [1113]:

```python
from IPython.display import Image
Image(filename='../images/illustration.png',width=1200, height=900)
```

Out[1113]:



1. Agreement Value
2. Agreement Start Date
3. Agreement End Date
4. Renewal Notice (Days)
5. Party One
6. Party Two

3. Dataset Loading and Description

Description -

The rental agreements are in the docx format. The training and evaluation datasets are available at the following location.

- $workspace/data$

The above directory has 2 sub-directories:

- training/: contains a total of 43 rental agreements
- eval/: contains a total of 8 rental agreements

For each training rental agreements docx file we have 6 entities (Aggrement Values, Aggrement Start Date, Aggrement End Date, Rebewal Notice(Days), Party One, Party Two) in a "data/TrainingTestSet .csv" for training and for validation rental agreements docx file there is "data/ValidationSet.csv".

Dataset Loading

In [911]:

```python
# Ignore all your warnings

%matplotlib inline
import warnings
warnings.filterwarnings("ignore")

import os
import pandas as pd
import docx
from tqdm import tqdm
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import spacy
from spacy.util import minibatch, compounding
from spacy.matcher import PhraseMatcher
from spacy.gold import GoldParse
from spacy.scorer import Scorer
```

In [2]:

```python
# Load TrainingTestSet and ValidationSet
TrainingTestSet = pd.read_csv('../data/TrainingTestSet .csv')
ValidationSet = pd.read_csv('../data/ValidationSet.csv')

# Sample
print('Shape of TrainingTestSet: ', TrainingTestSet.shape)
print('Shape of ValidationSet: ', ValidationSet.shape)
TrainingTestSet.sample(5)
```

```
Shape of TrainingTestSet:  (55, 7)
Shape of ValidationSet:  (8, 7)
```

Out[2]:

	File Name	Aggrement Value	Aggrement Start Date	Aggrement End Date	Renewal Notice (Days)	Party One	Party Two
17	54770958-Rental-Agreement	8000.0	01.04.2011	31.03.2012	90.0	K. Parthasarathy	Veerabrahmam Bathini
39	228094620-Rental-Agreement	15000.0	07.07.2013	06.06.2014	30.0	KAPIL MEHROTRA	.B.Kishore
33	156155545-Rental-Agreement-Kns-Home	12000.0	15.12.2012	14.11.2013	30.0	V.K.NATARAJ	VYSHNAVI DAIRY SPECIALITIES Private Ltd
11	24158401-Rental-Agreement	12000.0	01.04.2008	31.03.2009	60.0	Hanumaiah	Vishal Bhardwaj
16	50070534-RENTAL-AGREEMENT (1)	10000.0	01.04.2010	30.03.2011	90.0	P. JohnsonRavikumar	Saravanan BV

Note: Files of ValidationSet are overlapping in TrainingTestSet csv file.

In [3]:

```python
# Load Training rental agreements docx file
Training_data_docx = os.listdir('../data/Training_data')
Validation_Data_docx = os.listdir('../data/Validation_Data')

print('Number of training docx: ', len(Training_data_docx))
print('Number of validation docx: ', len(Validation_Data_docx))
```

```
Number of training docx:  43
Number of validation docx:  8
```

In [4]:

```python
# Rent Agreement Sample
idx = 0
file_name = Training_data_docx[idx]
doc = docx.Document(f"../data/Training_data/{file_name}")
full_text_list = [paragraph.text for paragraph in doc.paragraphs]
full_text = " ".join(full_text_list)

print('Sampple\n', '-'*120, '\n', full_text)

# Corresponding Entities
print('\n\n', 'Corresponding Entities\n',  '-'*120,)
TrainingTestSet[TrainingTestSet['File Name'] == file_name.rstrip('.pdf.docx')]
```

```
Sampple
 -----------------------------------------------------------------------------------------------------------------------
----
 HOUSE RENTAL AGREEMENT Rental Agreement made on Jan 10, 2011, between Namashivayam, Plat No-182, Door No 16 New/10 O
ld, 24th East Street, Kamaraj Nagar, Thiruvanmiyur, Chennai (hereinafter referred to as landlord) of the house and Mr
s.Thenmalar, W/O Xavier, Kottaikadu (PO), Vadakadu (Via), Pudukkottai (Dt) (hereinafter referred to as tenant) of the
First floor portion of the building. WHERE IT IS AGREED AND DECLARED AS FOLLOWS: The Landlord agrees to let out and t
he tenant agrees to take on rent the First floor portion of the building Plat No-182, Door No 16 New/10 Old, 24th Eas
t Street, Kamaraj Nagar, Thiruvanmiyur along with electrical and sanitary fittings and other accessories fittings and
structures (hereinafter called the premises) from Jan 10 2011 at the monthly rent of Rs. 14500 (Fourteen thousand and
five hundred rupees) being payable on or before 5th of every month to the Landlord. The period of this agreement shal
l be twelve months w.e.f Jan 10, 2011. The tenant has paid Rs. 100000(0ne lack Rupees) as advance amount for the abov
e building and the landlord shall pay this said advance without interest to the tenant at the time of vacating the pr
emises or within 12 months of commencement of this agreement whichever is earlier. At the termination of the period o
f tenancy the tenant agrees to surrender to the Landlord the vacant possession of the premises without raising any ob
jection. This rental agreement can be terminated at any time by three months notice on either side and on such termin
ation the tenant shall surrender the vacant possession of the premises to the Landlord. If for by any reason the tena
nt occupies the building for a period that includes part of a month, it is agreed that the rent will be charged on a 
pro-rated basis for that month. The landlord shall pay all existing and future taxes, rates and assessments in respec
t of the lease hold including the municipal or other tax assessed by a local authority on the value of the building o
r annual letting value of the building and all other rates, taxes and assessments levied by any authority whatsoever.
The tenant shall pay the electricity and water supply charges for the period of time he occupies the premises. The te
nant agrees to leave at the end of tenancy the premises in good condition as they are now, subject to reasonable wear
and tear. The tenant also agreed not to let out the building or a portion of it to anybody else. The tenant shall not
commit any act of waste in the premises. 11 The tenant also agrees to make any maintenance on the building as mutuall
y agreed upon by the tenant and the landlord and the said expenses shall be adjusted against the rent amount due to t
he landlord. 12.The landlord shall retain the original of this agreement and the tenant shall retain its duplicate. W
ITNESS WHERE OF Namashivayam , the landlord and Mrs.Thenmalar, the tenant have affixed their signatures on Jan 10, 20
```

```
11. (Land Lord) (Tenant) Witness:


 Corresponding Entities
 ---------------------------------------------------------------------------------------------------
----
```

Out[4]:

	File Name	Aggrement Value	Aggrement Start Date	Aggrement End Date	Renewal Notice (Days)	Party One	Party Two
29	100999172-House-Rental-Agreement	14500.0	10.01.2011	09.01.2012	90.0	Namashivayam	Thenmalar

Merge Rental-Agreement doc files and entities from csv files

In [70]:

```python
# Merge Rental-Agreement doc files and entities
def extract_doc_text(doc_list):
    """ Given doc file names, return dataframe of extracted text"""
    file_names = []
    texts = []

    for file in tqdm(doc_list):
        file_name = file.rstrip('.pdf.docx').split('/')[-1]

        doc = docx.Document(file)
        full_text_list = [paragraph.text for paragraph in doc.paragraphs]
        full_text = " ".join(full_text_list)

        file_names.append(file_name)
        texts.append(full_text)

    return pd.DataFrame({'File Name': file_names, 'text': texts})
```

In [1103]:

```python
training_data_docx_list = ["../data/Training_data/" + docx for docx in Training_data_docx]
Validation_Data_docx_list = ["../data/Validation_Data/" + docx for docx in Validation_Data_docx]

# Extracting text from doc file
train_text = extract_doc_text(training_data_docx_list)
val_text = extract_doc_text(Validation_Data_docx_list)

# Sample
print("-"*50)
print("Samples - ")
train_text.sample(5)
```

```
100%|██████████████████████████████████████████████████████████| 43/43 [00:00<00:00, 298.13it/s]
100%|████████████████████████████████████████████████████████████| 8/8 [00:00<00:00, 266.43it/s]

--------------------------------------------------
Samples - 
```

Out[1103]:

	File Name	text
9	216973836-Rental-Agreement-Sample	THIS RENTAL AGREEMENT is made on this, the M...
3	142106117-Rental-Agreement	RENTAL AGREEMENT This agreement of Tenancy is ...
1	116950326-December-2012-Rental-Agreement	ROOM RENTAL AGREEMENT This is a legally bind...
0	100999172-House-Rental-Agreement	HOUSE RENTAL AGREEMENT Rental Agreement made o...
6	18325926-Rental-Agreement-1	RENTAL AGREEMENT This deed of rental...

In [93]:

```
# Merge Rental-Agreement doc files and entities from csv
train_data = pd.merge(TrainingTestSet, train_text, on = 'File Name', how = 'inner')
val_data = pd.merge(ValidationSet, val_text, on = 'File Name', how = 'inner')

print('shape of train_data: ',  train_data.shape)
print('shape of val_data: ',  val_data.shape)

# Samples
train_data.sample(5)
```

```
shape of train_data:  (43, 8)
shape of val_data:  (8, 8)
```

Out[93]:

	File Name	Aggrement Value	Aggrement Start Date	Aggrement End Date	Renewal Notice (Days)	Party One	Party Two	text
14	63793679-Rental-Agreement	9000.0	01.09.2011	31.08.2012	NaN	S Parthasarathy	Hari Kiran Tholeti	RENTAL AGREEMENT THIS RENTAL A...
37	294331674-Rental-Agreement	3500.0	17.07.2014	17.06.2015	30.0	MICHAEL DELA CRUZ	CATHERINE CABOCHA	RENTAL AGREEMENT THE AGREEMENT The landlord ag...
27	203615996-Rental-Agreement-Format	3500.0	01.02.2008	31.01.2009	30.0	T.RADHA KRISHNAN	ABHIJIT BHARADWAJ	RENTAL AGREEMENT This Agre...
28	216973836-Rental-Agreement-Sample	15000.0	23.03.2013	23.03.2014	60.0	Kamal	V.Arun Kumar	THIS RENTAL AGREEMENT is made on this, the M...
11	62126501-Rental-Agreement	4200.0	23.05.2011	22.04.2012	90.0	M.V Srinivas & M.V Madhumathi	M.V Thirumalesh	RENTAL AGREEMENT This Rental Agreement is made...

4. Exploratory Data Analyses and Preprocessing

In [147]:

```
# Utility function to plot lineplot and distplot using seaborn
def plot_sns(data,feature,color='lightblue',title=None,subtitle=None):

    """
    Utility function to plot lineplot and distplot using seaborn

    plot_sns(data,feature,color='lightblue',title=None,subtitle=None):

    data = data
    feature = coulum name
    color = color of plot
    title = Either 'length' or 'number' based on which to plot. Otherwise by default='None'
    subtitle = Either 'train_df' or 'val_df'. Otherwise by default='None'

    """
    f, (ax1, ax2) = plt.subplots(1, 2, figsize=(18, 6))

    # line plot
    sns.lineplot(np.arange(len(data)),data,ax=ax1,color=color)
    if title=='number':
        ax1.set(xlabel=f"Idx of {feature}", ylabel=f"Number of words in {feature}", title=f'Number of words in {feat
    elif title=='length':
        ax1.set(xlabel=f"Idx of {feature}", ylabel=f"Length of {feature}", title=f'Length of {feature} in {subtitle}
    ax1.grid()

    # distribution plot
    sns.distplot(data,ax=ax2,color=color)
    if title=='number':
        ax2.set(xlabel=f"Number of words in {feature}", ylabel="pdf", title=f'Number of words in {feature} in {subti
    elif title=='length':
        ax2.set(xlabel=f"Length of {feature}", ylabel="pdf", title=f'Length of {feature} in {subtitle}\n')
    ax2.grid()
    plt.show()
    return None

# Utility function to plot frequency of most popular words
def word_frequency_plot(dataframe, title=None):
    list_of_all_words = []
    for sent in dataframe:
        list_of_all_words.extend(sent.split())

    top_50_words = pd.Series(list_of_all_words).value_counts()[:50]
    top_50_words_prob_dist = top_50_words.values/sum(top_50_words.values)

    #  plot of frequency of polpular words in train
    plt.figure(figsize=(16,7))
    sns.barplot(top_50_words.index, top_50_words_prob_dist)
    plt.xlabel("words")
    plt.ylabel("frequency")
    plt.title(f"Frequency of most popular words {title}\n")
    plt.xticks(rotation=70)
    plt.grid()
    plt.show()
    return None

# Utility function to plot frequency of number of words
def plot_distibution_diff(df1, df2):
    """ Given 2 text dataframe, plot word frequency dist."""
    # Calculating the length of text before and after preprocessing
    len_after_cleaning = df1.apply(lambda x: len(x.split()))
    len_before_cleaning = df2.apply(lambda x: len(x.split()))

    # ploting
    plt.figure(figsize=(9, 6))
    sns.distplot(len_before_cleaning, label='len_before_cleaning')
    sns.distplot(len_after_cleaning, label='len_after_cleaning')
    plt.title(f" Distribution of number of words in text before v/s after preprocessing\n", fontsize=15)
    plt.ylabel("distribtion")
    plt.xlabel(f"number of words in text")
    plt.legend()
    plt.grid()
    plt.show()
    return None

```

4.1 How many entities are missing?

In [107]:

```
print('Number of entities missing from training data -')
np.sum(train_data.drop(labels=['File Name', 'text'], axis=1).isnull())
```

```
Number of entities missing from training data -
```

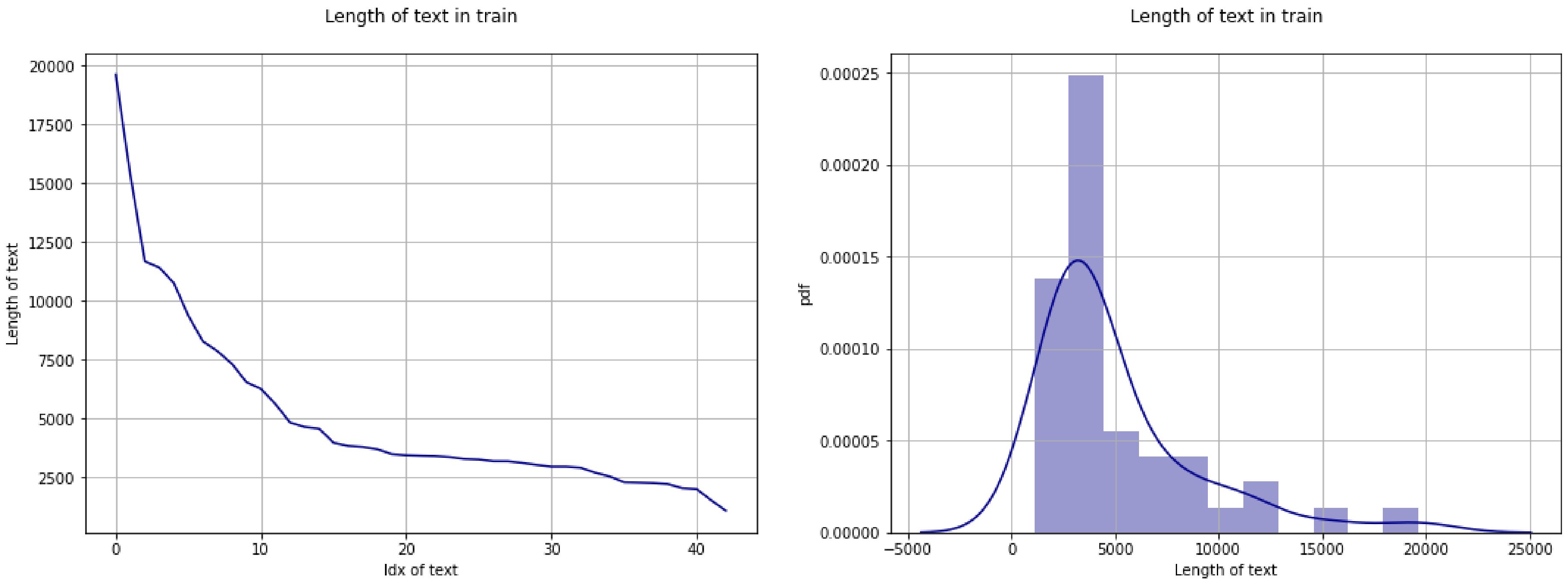
Out[107]:

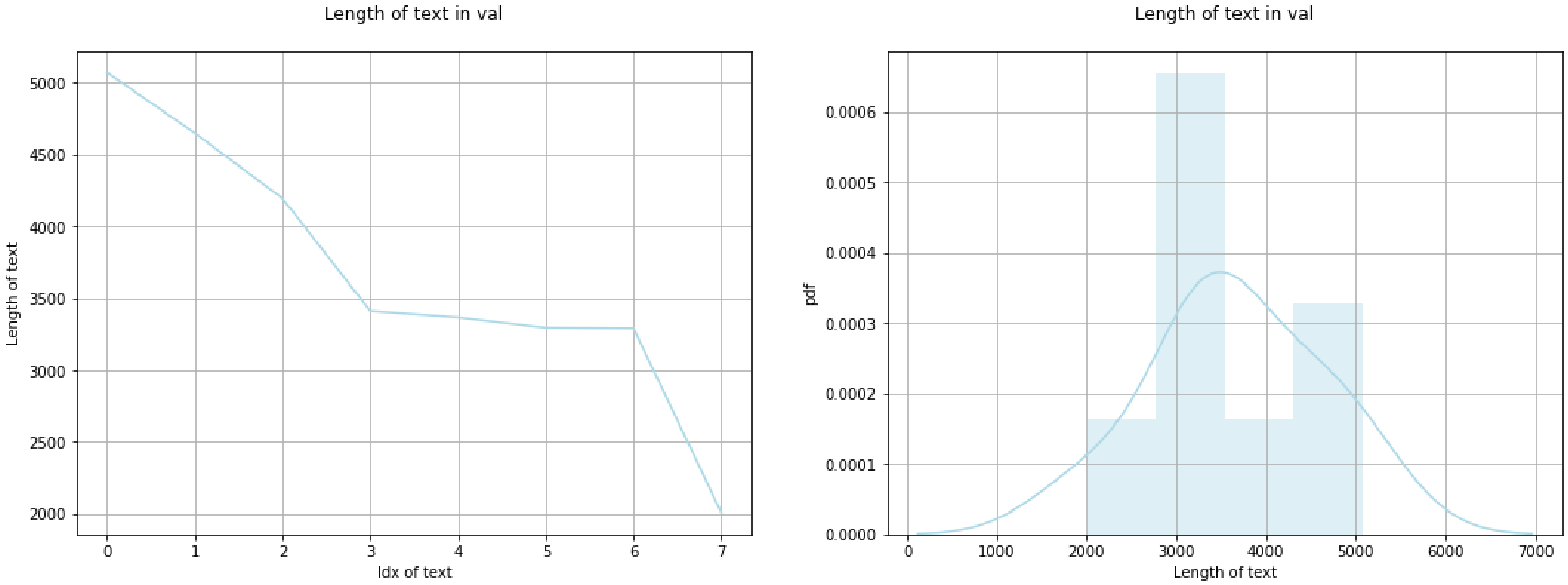
```
Aggrement Value          1
Aggrement Start Date     3
Aggrement End Date       6
Renewal Notice (Days)   11
Party One                0
Party Two                1
dtype: int64
```

4.2 Distribution of length of text in train and val data

In [141]:

```
# Length of question_title in train
len_of_text_train = sorted(train_data['text'].apply(lambda x: len(x)),reverse=True)

# Length of question_title in test
len_of_text_val = sorted(val_data['text'].apply(lambda x: len(x)),reverse=True)

# plot for train_df
plot_sns(len_of_text_train,"text",color='darkblue',title='length',subtitle='train')

# plot for test_df
plot_sns(len_of_text_val,"text",color='lightblue',title='length',subtitle='val')
```

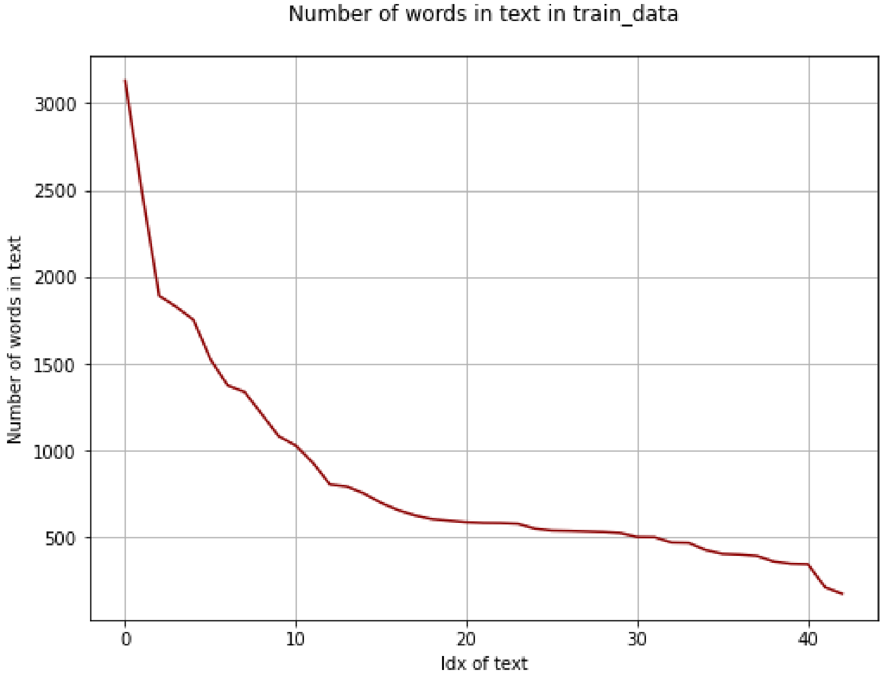


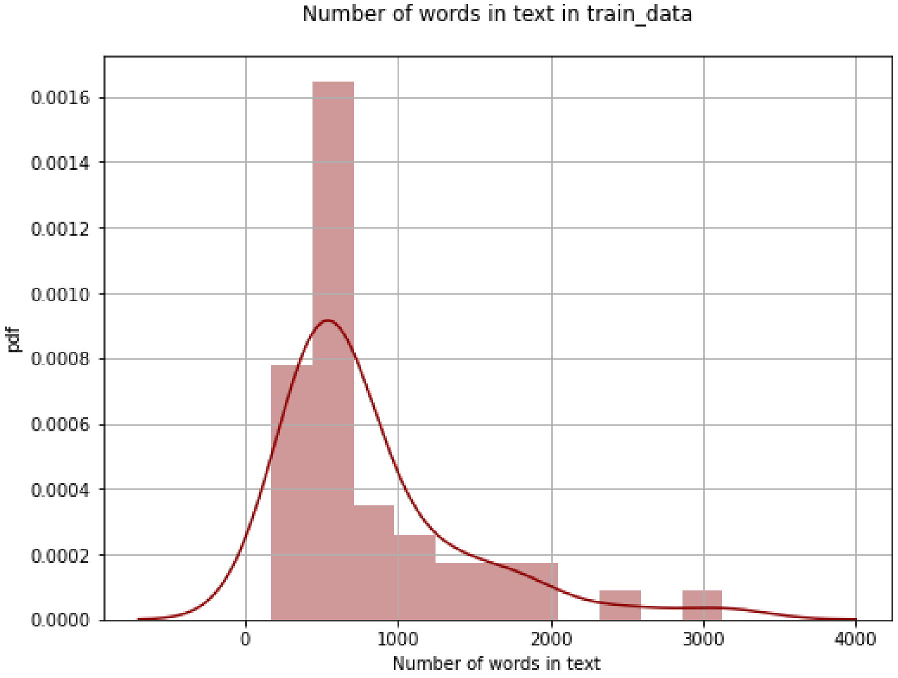


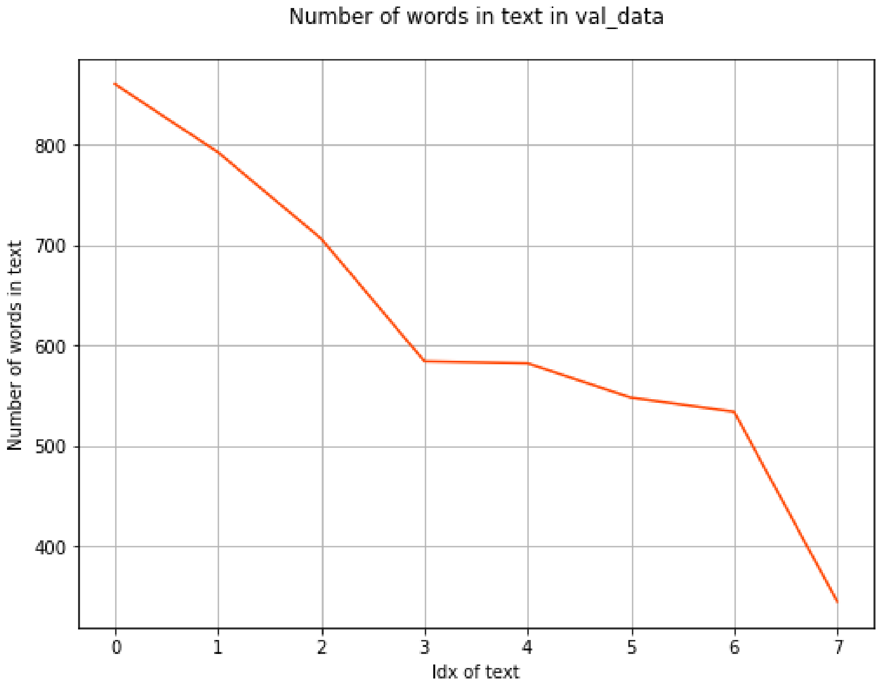
4.3 Distribution of number of words in train and val text data

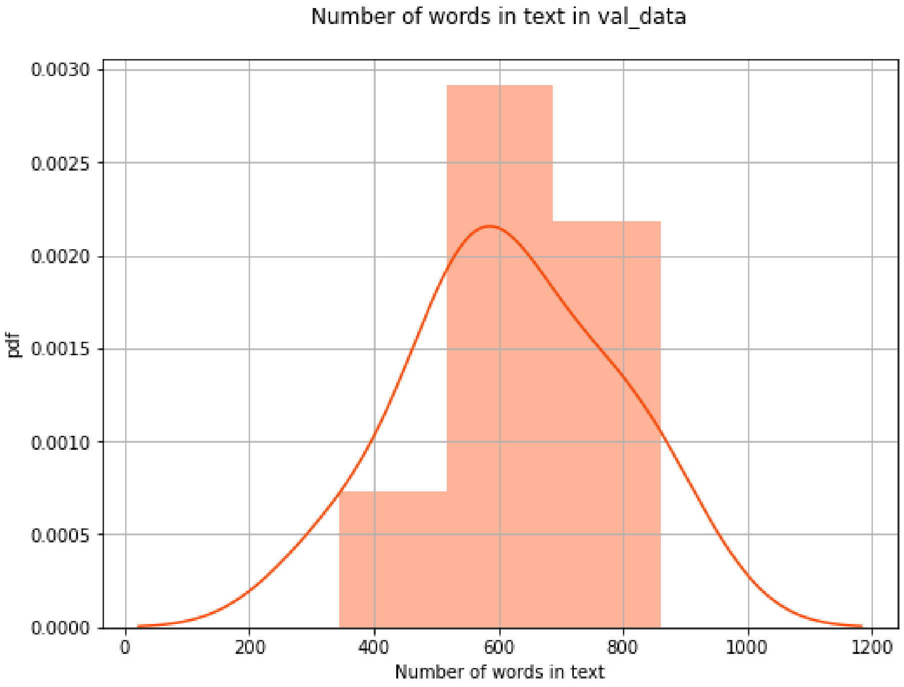
In [145]:

```
# number of words in question_title in train
n_words_train = sorted(train_data['text'].apply(lambda x: len(x.split(" "))),reverse=True)

# number of words in question_title in test
n_words_val = sorted(val_data['text'].apply(lambda x: len(x.split(" "))),reverse=True)

# plot for train_df
plot_sns(n_words_train,"text",color='darkred',title='number',subtitle='train_data')

# plot for test_df
plot_sns(n_words_val,"text",color='orangered',title='number',subtitle='val_data')
```


Number of words in text in train_data


Number of words in text in train_data

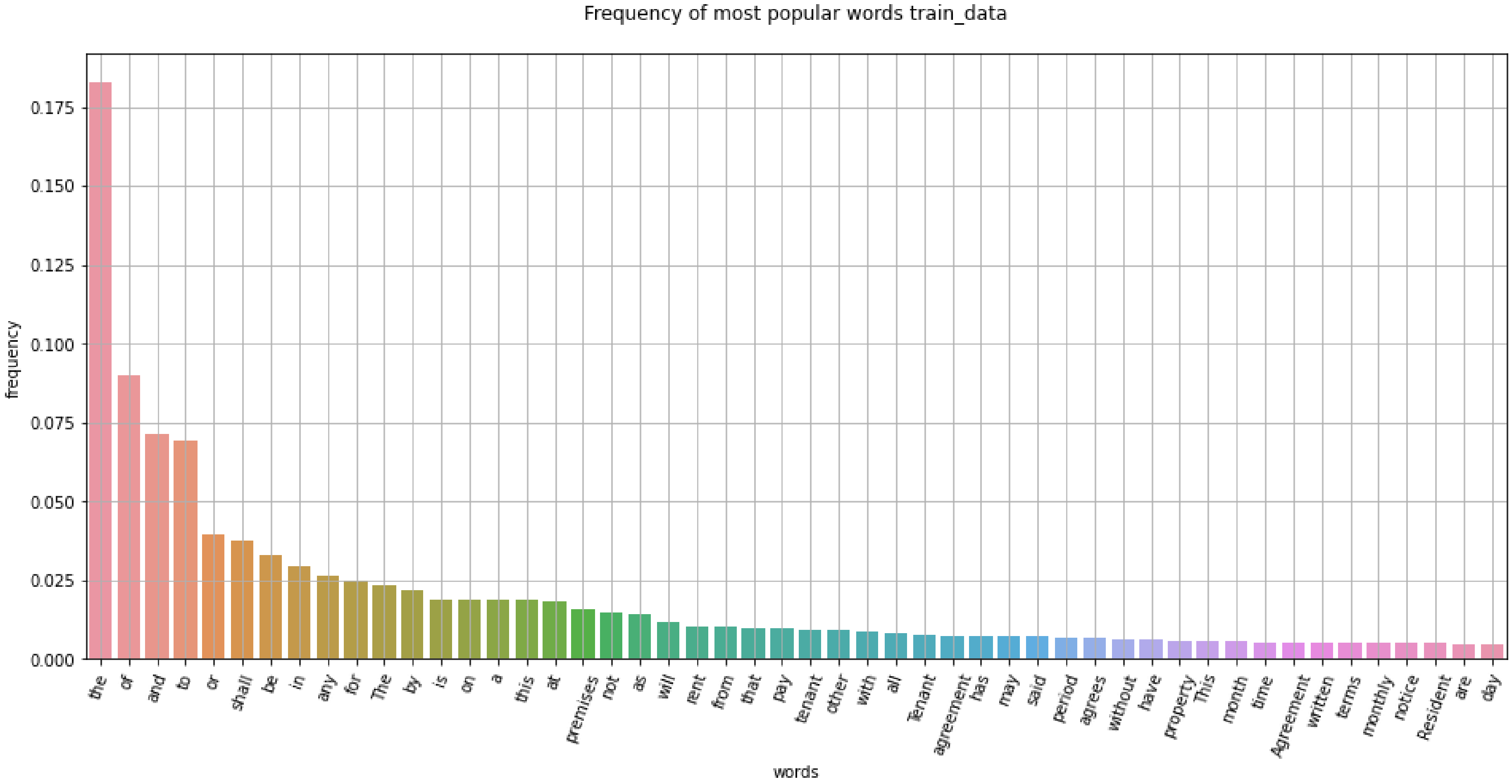

Number of words in text in val_data
Number of words in text
800
700
600
500
400
0
1
2
3
4
5
6
7
Idx of text

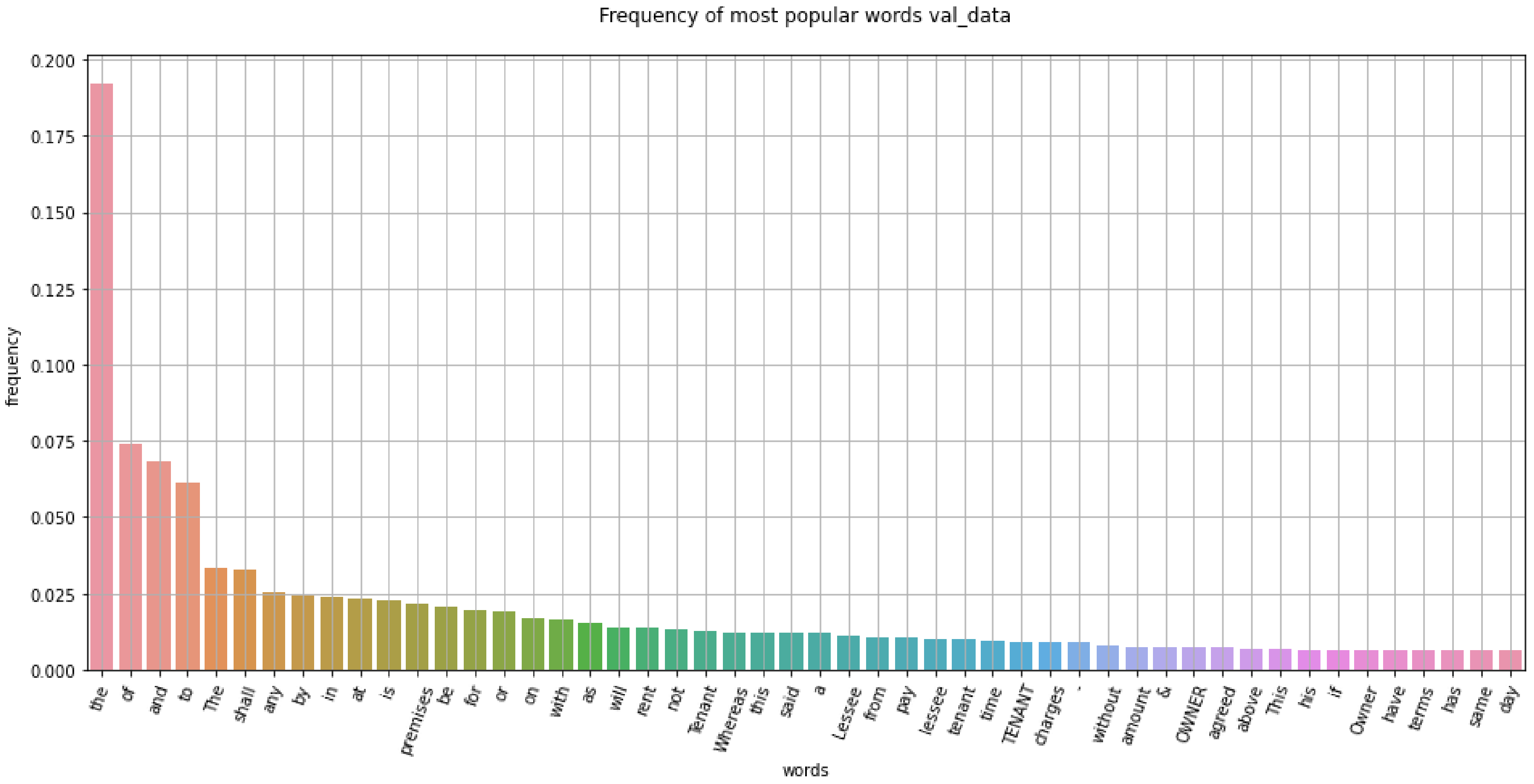

Number of words in text in val_data
pdf
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
200
400
600
800
1000
1200
Number of words in text

4.4 Frequency of most popular 50 words

In [148]:

```python
# Frequency of most popular 50 words in train_df
word_frequency_plot(train_data['text'], title='train_data')

# Frequency of most popular words in val_df
word_frequency_plot(val_data['text'], title='val_data')
```





Preprocessing and Cleaning data

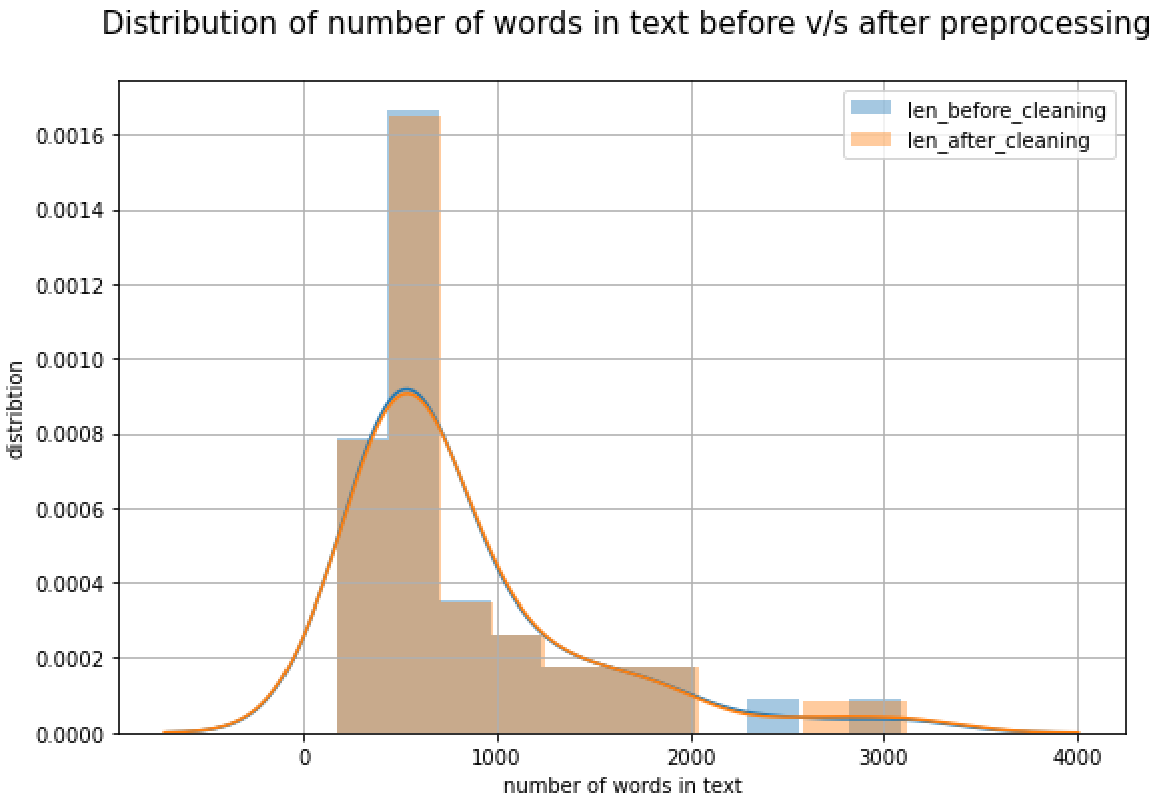
In [244]:

```
# Basic Cleaning in text data
# Train
train_data['cleaned_text'] = train_data['text'].apply(lambda x: re.sub(r'[^A-Za-z0-9-,/.]',' ', x))
train_data['cleaned_text'] = train_data['cleaned_text'].apply(lambda x: " ".join(x.split()))

# Val
val_data['cleaned_text'] = val_data['text'].apply(lambda x: re.sub(r'[^A-Za-z0-9-,/.]',' ', x))
val_data['cleaned_text'] = val_data['cleaned_text'].apply(lambda x: " ".join(x.split()))
```

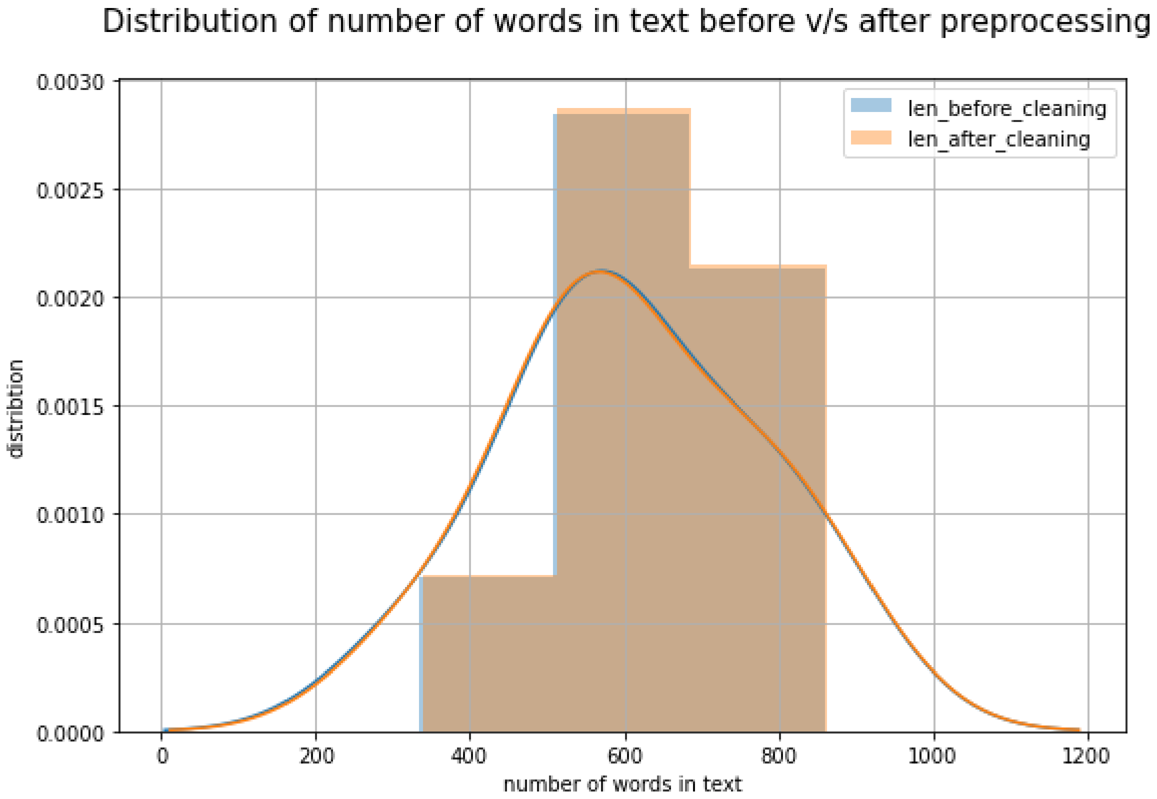
In [247]:

```
# Distribution before and after performing basic cleaning
print('Train -')
plot_distibution_diff(train_data['cleaned_text'], train_data['text'])
print('Val -')
plot_distibution_diff(val_data['cleaned_text'], val_data['text'])
```

```
Train -
```



```
Val -
```



- Plots are self explanatory

5. Making Dataset ready and Modeling

Approach 1 -

1. Use preprocessed text and find the span of each entity in the rental-agreements docx text file using spacy pattern matcher.
2. Convert the data into required format and finetune Spacy NER.

5.1.1. Converting the data into required format for spacy

In [255]:

```
def change_format(df):
    file_names = []
    texts = []
    entities = []

    for idx in tqdm(range(df.shape[0])):
        file_name = df.iloc[idx]['File Name']
        full_text = df.iloc[idx]['cleaned_text']

        entity = {
            'agreement_value': str(df.iloc[idx]['Aggrement Value']),
            'agreement_start_data': str(df.iloc[idx]['Aggrement Start Date']),
            'agreement_end_data': str(df.iloc[idx]['Aggrement End Date']),
            'renewal_notice': str(df.iloc[idx]['Renewal Notice (Days)']),
            'party_one': str(df.iloc[idx]['Party One']),
            'party_two': str(df.iloc[idx]['Party Two'])
        }

        file_names.append(file_name)
        texts.append(full_text)
        entities.append(entity)

    return pd.DataFrame({'filename': file_names, 'text': texts, 'entities': entities})
```

In [258]:

```
# Change Format
train_data_changes_format_df = change_format(train_data)
val_data_changes_format_df = change_format(val_data)

# Sample
train_data_changes_format_df.sample(5)
```

```
100%|████████████████████████████████████████| 43/43 [00:00<00:00, 1074.03it/s]
100%|█████████████████████████████████████████| 8/8 [00:00<00:00, 1141.85it/s]
```

Out[258]:

	filename	text	entities
1	6683129-House-Rental-Contract-Geraldine-Galina...	House Rental Contract KNOWN ALL MEN BY THESE P...	{'agreement_value': '6500.0', 'agreement_start...
37	294331674-Rental-Agreement	RENTAL AGREEMENT THE AGREEMENT The landlord ag...	{'agreement_value': '3500.0', 'agreement_start...
13	63057680-Rental-Agreement	Rental Agreement This agreement made this 19 d...	{'agreement_value': '450.0', 'agreement_start_...
27	203615996-Rental-Agreement-Format	RENTAL AGREEMENT This Agreement of Tenancy is ...	{'agreement_value': '3500.0', 'agreement_start...
39	323828497-Rental-Agreement-Micky	RENTAL AGREEMENT THIS AGREEMENT made this 27TH...	{'agreement_value': '800.0', 'agreement_start_...

Percentage of missing entities in training data

In [858]:

```python
def count_percentage_of_missing_entities(df, n_entity=6):
    df_len = df.shape[0]
    count = 0
    for idx in range(df_len):
        for value in df['entities'][idx].values():
            if value == 'nan':
                count += 1

    return count, np.round((count / (df_len * n_entity))*100, 2)
```

In [888]:

```python
# Count the percentage of nan/missing entities from training data
n_entity = 6
n_missing_enitity, percentage = count_percentage_of_missing_entities(train_data_changes_format_df, n_entity)
print(f"Percentage of nan/missing entities from training data: {n_missing_enitity}/{train_data_changes_format_df.sha
```

```
Percentage of nan/missing entities from training data: 22/258 8.53 %
```

In [287]:

```python
# Convert the data into required spacy format
def change_spacy_format(df):
    """
    Given df, convert into spacy format
    code refer: https://github.com/chawla201/Custom-Named-Entity-Recognition
    """
    training_data = []
    id_ent = []

    nlp_match = spacy.load('en_core_web_sm')
    matcher = PhraseMatcher(nlp_match.vocab)
    for index in tqdm(range(df.shape[0])):

        ent_dic = df.iloc[index]["entities"]
        ent = []
        phrases = list(ent_dic.values())
        patterns = [nlp_match.make_doc(phrase) for phrase in phrases]
        matcher.add("EntityList", None, *patterns)

        doc = nlp_match(df.iloc[index]["text"])
        matches = matcher(doc)
        for match_id, start, end in matches:
            try:
                span = doc[start:end]
                if start > 0:
                    sb = doc[0:start]
                    start_index = len(sb.text) + 1
                else:
                    start_index = 0
                end_index = start_index + len(span.text)
            except:
                pass

            for key, value in ent_dic.items():
                if value == span.text:
                    ent_tup = (start_index, end_index, key)
                    ent.append(ent_tup)

        id_ent.append(len(ent))
        entity_dictionary = {"entities": ent}
        train_tup = (df.iloc[index]["text"], entity_dictionary)
        training_data.append(train_tup)

    return training_data
```

In [856]:

```
# Converting the data into required format for spacy
training_data_spacy_format = change_spacy_format(train_data_changes_format_df)
val_data_spacy_format = change_spacy_format(val_data_changes_format_df)
```

```
100%|██████████████████████████████████████████████████| 43/43 [00:03<00:00, 10.85it/s]
100%|████████████████████████████████████████████████████| 8/8 [00:00<00:00, 13.85it/s]
```

Percentage of missing entities in from spacy_format training data

In [884]:

```
def count_percentage_of_missing_entities_from_spacy_format(spacy_format_data):
    count = 0
    for train_data_file in spacy_format_data:
        count += len(train_data_file[1]['entities'])

    return count
```

In [896]:

```
# Calculate Percentage of missing entities in spacy_format training data
n_entity_in_spacy_format_training_data = count_percentage_of_missing_entities_from_spacy_format(training_data_spacy_

n_entity_in_spacy_format_training_data_per = np.round((n_entity_in_spacy_format_training_data /
                                            (train_data_changes_format_df.shape[0]*n_entity - n_missing_enitity)*100

print(f'Percentage containing entity in training data: {n_entity_in_spacy_format_training_data_per} %')

```

```
Percentage containing entity in training data: 21.186 %
```

Observation -

- Only 21.186 % entity is able to preserve in the training data when converted the data into spacy required format using spacy patter matcher.
- Clearly spacy patter matcher is not able to annotate the entity properly.
- Let's see how training validation is performing on this half baked annotated data.

5.1.2. Finetune Spacy-NER Model

In [961]:

```
import random

def train_model(train_data, n_iter, drop_rate):
    """
    Finetune Spacy-NER model.
    Code Refer: https://www.machinelearningplus.com/nlp/training-custom-ner-model-in-spacy/
    """
    if 'ner' not in nlp.pipe_names:
        ner = nlp.create_pipe('ner')
        nlp.add_pipe(ner, last = True)

    for _, annotation in train_data:
        for ent in annotation['entities']:
            ner.add_label(ent[2])


    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 tqdm(range(n_iter)):
            random.shuffle(train_data)
            losses = {}
            index = 0
            for text, annotations in train_data:
                try:
                    nlp.update( [text],
                        [annotations],
                        drop=drop_rate,
                        sgd=optimizer,
                        losses=losses)
                except Exception as e:
                    pass
            print("Iteration " + str(itn+1) + f" -- {str(losses)}")
    return None
```

In [963]:

```
# Training spacy model

n_iter = 100 # Number of iteration to train the model
drop_rate = 0.4 # Drop Rate
nlp = spacy.blank('en')
train_model(training_data_spacy_format, n_iter, drop_rate)
```

```
Iteration 82 -- {'ner': 11.266264691817957}

 83%|███████████████████████████████████████████████████████████████████▏              | 83/100 [06:08<01:17,  4.54s/
it]

Iteration 83 -- {'ner': 5.051446836846079}

 84%|████████████████████████████████████████████████████████████████████             | 84/100 [06:12<01:12,  4.50s/i
t]

Iteration 84 -- {'ner': 5.027146308191458}

 85%|█████████████████████████████████████████████████████████████████████▌            | 85/100 [06:17<01:07,  4.48s/
it]

Iteration 85 -- {'ner': 1.9417035257235007}

 86%|██████████████████████████████████████████████████████████████████████▍           | 86/100 [06:21<01:02,  4.48s/
it]

Iteration 86 -- {'ner': 25.754415197955485}
```

In [ ]:

```
# Saving the model
nlp.to_disk('model1')
```

In [1107]:

```
# Loading the model
nlp_model = spacy.load('model1')
```

Recall on training data and val data (Given only 21.186 % entity is able to preserve in the training data when converted the data into spacy required format using spacy patter matcher)

In [965]:

```python
def score(spacy_format_data, model):
    """ Function to clacluate recall metric of a model"""
    scorer = Scorer()
    try:
        for input_, annot in spacy_format_data:
            doc_gold_text = model.make_doc(input_)
            gold = GoldParse(doc_gold_text, entities=annot['entities'])
            pred_value = model(input_)
            scorer.score(pred_value, gold)
    except Exception as e: print(e)
    return scorer.scores['ents_r']
```

In [966]:

```python
# Recall on training data
training_recall_model1 = score(training_data_spacy_format, nlp_model)
print(f"Recall on validation data: {training_recall_model1}")

# Recall on val data
val_recall_model1 = score(val_data_spacy_format, nlp_model)
print(f"Recall on validation data: {val_recall_model1}")
```

```
Recall on validation data: 100.0
Recall on validation data: 90.9090909090909
```

In [967]:

```python
# Prediction on val data
for idx, data_point in enumerate(val_data_spacy_format):
    print(f'{idx+1}', 'filename- ', val_data['File Name'].iloc[idx])
    doc = nlp_model(data_point[0])
    for ent in doc.ents:
        print(f'{ent.label_.upper():{30}}- {ent.text}')
    print('--'*50)
```

```
1 filename-  24158401-Rental-Agreement
----------------------------------------------------------------------------------------------------
2 filename-  63793679-Rental-Agreement
PARTY_ONE                     - S Parthasarathy
PARTY_TWO                     - Hari Kiran Tholeti
PARTY_ONE                     - S Parthasarathy
PARTY_ONE                     - S Parthasarathy
PARTY_TWO                     - Hari Kiran Tholeti
----------------------------------------------------------------------------------------------------
3 filename-  95980236-Rental-Agreement
PARTY_TWO                     - V.V.Ravi Kian
----------------------------------------------------------------------------------------------------
4 filename-  156155545-Rental-Agreement-Kns-Home
----------------------------------------------------------------------------------------------------
5 filename-  195231682-This-RENTAL-AGREEMENT-is-Made-and-Executed-on-24th-Day-of-September
PARTY_ONE                     - C.BHAGYAMMA
PARTY_TWO                     - JP INTERIO
PARTY_TWO                     - JP INTERIO
----------------------------------------------------------------------------------------------------
6 filename-  228094620-Rental-Agreement
----------------------------------------------------------------------------------------------------
7 filename-  239419594-Rental-Agreement
----------------------------------------------------------------------------------------------------
8 filename-  269135973-Udaya-Rental-Agreement
PARTY_TWO                     - Pottumurthi Udayalaxmi
----------------------------------------------------------------------------------------------------
```

Observation -

1. As expected approach 1 has badly failed to extract metainfor from rental agreement.

2. Although training recall and validation recall of model is high but while prediction six fields on validation data, model is hardly able to predict PARTY_ONE and PARTY_TWO and in some scenario even those fileds are not getting predicted.

3. Reason in simple. We have poorly annotated training data which we have created using spacy pattern matcher.

4. We have ambigious entities in TrainingSet csv and given corresponding docx file. E.g, data 10.5.2011 != 10-05-2011 != 10/05/2011 != 10th may of 2011.

5. Potential Solution - Prepare better annotated data for training.

Approach 2 -

1. Use pretrained model which is trained on similar dataset.
2. Perform pseudo annotation of training text and val text.
3. Use Pseudo annotated data for training and observe the prediction.

Pretrained model which is trained on similar dataset.

Source: https://github.com/sanghavi-vemulapati/Rental-Agreement-Metadata-extraction-using-spacy (https://github.com/sanghavi-vemulapati/Rental-Agreement-Metadata-extraction-using-spacy)

- Author of this person has gone through mannual annotation on training data using Label Studio and trained Spacy-NER model on the top of that.
- But only problem is that author has used 8 entity fileds rather that 6 entity fileds which is our case. Let's see how can we tackle that problem.

In [944]:

```python
def pseudo_annotation(df, pseudo_model):
    """
    Give a pretrained model and dataframe containing cleaned_text,
    this function pseudo_annotate the data.
    """
    pseudo_annotated_data = []

    df_len = df.shape[0]
    for idx in range(df_len):
        doc = pseudo_model(df['cleaned_text'].iloc[idx])

        entity = []
        for ent in doc.ents:

            if ent.label_.upper() == 'PARTYONE':
                entity.append((ent.start_char, ent.end_char, 'party_one'))

            if ent.label_.upper() == 'PARTYTWO':
                entity.append((ent.start_char, ent.end_char, 'party_two'))

            if ent.label_.upper() == 'STARTDATE':
                entity.append((ent.start_char, ent.end_char, 'agreement_start_data'))

            if ent.label_.upper() == 'ENDDATE':
                entity.append((ent.start_char, ent.end_char, 'agreement_end_data'))

            if ent.label_.upper() == 'NOTICE':
                entity.append((ent.start_char, ent.end_char, 'renewal_notice'))

            if ent.label_.upper() == 'AGREEMENTVALUE':
                entity.append((ent.start_char, ent.end_char, 'agreement_value'))

        pseudo_annotated_data.append((df['cleaned_text'].iloc[idx], {'entities': entity}))
    return pseudo_annotated_data
```

In [945]:

```python
# Loading the pretrained model foe pseudo annotation of data
nlp_pretrained_model_pseudo = spacy.load('pretrained_model_pseudo_annotation')
```

5.2.1. Pseudo Annotatiion

Pseudo Annotation

In [951]:

```
# Pseudo Annotation
pseudo_annotated_data_training_data = pseudo_annotation(train_data, nlp_pretrained_model_pseudo)
pseudo_annotated_data_val_data = pseudo_annotation(val_data, nlp_pretrained_model_pseudo)
```

In [955]:

```python
# Sample
print('Pseudo annotated data Sample - ')
pseudo_annotated_data_training_data[0]
```

```
Pseudo annotated data Sample - 
```

Out[955]:

```
('House Rental Contract KNOWN ALL MEN BY THESE PRESENTS This House Rental Contract, made and entered into this 20th day
of May 2007 at Manila by and between Antonio Levy S. Ingles. Jr. and/or Mary Rose C. Ingles, of legal age, with residen
ce and postal address at Unit 2006 EGI Taft Tower 2339 Taft Avenue, Malate, Manila, And herein referred to as the Owner
s , And GERALDINE O. GALINATO. of legal age, with residence and postal address at 6 Manganese Road, Pilar Village, Las
Pinas, Metro Manila, And herein referred to as the Resident s , WITNESSETH In consideration of the agreements of the Re
sident s , known as GERALDINE O. GALINATO. the Owner s , known as Antonio Levy S. Ingles. Jr. and/or Mary Rose C. Ingle
s, hereby rent their the dwelling/house located at Lot 6, Block 20, Royal South Townhomes, Marcos Alvarez Avenue, Talon
5, Las Pinas City, Metro Manila for the period commencing on the 20th day of May, 2007, and monthly thereafter until th
e 20th day of May, 2008, at which time this Agreement is terminated. Resident s , in consideration of Owner s permittin
g them to occupy the above property, hereby agrees to the following terms RENT To pay as rental the sum of SIX THOUSAND
FIVE HUNDRED PESOS IP 6.500.001 per month, due and payable in advance from the 20th day of every month. FAILURE TO PAY
ON TIME Failure to pay the rent will result in being served a Notice to End Residential Tenancy. This Notice may be ser
ved if you have an outstanding balance from failure to pay your rent. This Notice may also be served from being habitua
lly late in paying your rent regardless of the balance owed. Once the Notice to End Residential Tenancy is received, yo
u will have a prescribed time to pay all of the amount overdue on your rent. A three-dav grace period will be allowed f
or late payment. Failure to pay the monthly rental within the grace period is subject to FIVE 5 PERCENT interest per mo
nth of delay as penalty. Habitual failure of the Resident s to pay within the prescribed time shall result in the Owner
s taking immediate legal action to evict the Resident s from the premises and seize the security deposit. SECURITY DEPO
SIT Resident s agrees to pay a deposit in the amount of SIX THOUSAND FIVE HUNDRED PESOS P 6.500.001 to secure Resident
s s pledge of full compliance with the terms of this agreement. Note THE DEPOSIT MAY NOT BE USED BY TENANT TO PAY THE R
ENT DURING THE TENANCY. The security deposit will be used at the end of the tenancy to compensate the Owner s for any d
amages or unpaid rent or charges, and will be repaired or replaced at Resident s s expense with funds other than the de
posit. METHOD OF PAYMENT The initial advance payment of rent and deposit under this contract be PAID IN CASH at least 7
days before the date of moving-in. Thereafter, monthly rent payments must be paid by POST DATED CHECKS payable to ANTON
IO LEVY S. INGLES. JR. until a first check is dishonored and returned unpaid. Regardless of cause, no other additional
payments may afterwards be made by check. Checks returned will not be redeposited. The Resident s will be notified by a
3 day notice, and will be required to pay the amount due in cash.',
 {'entities': [(155, 209, 'party_one'),
   (365, 386, 'party_two'),
   (899, 920, 'agreement_start_data'),
   (955, 976, 'agreement_end_data'),
   (1271, 1294, 'agreement_start_data')]})
```

5.2.2. Training the model on pseudo_annotated_data_training_data

In [956]:

```python
import random

# Number of iteration to train the model
n_iter = 100
drop_rate = 0.4 # Drop Rate
nlp = spacy.blank('en')

# Training spacy model on pseudo annotated data
train_model(pseudo_annotated_data_training_data, n_iter, drop_rate)

# Saving the model
nlp.to_disk('model2')

# Loading the mode
nlp_model2 = spacy.load('model2')
```

```
Iteration 84 -- {'ner': 394.34162059116574}

 85%|██████████████████████████████████████████████▌        | 85/100 [06:15<01:07,  4.53s/
it]

Iteration 85 -- {'ner': 345.8932800400572}

 86%|███████████████████████████████████████████████▍       | 86/100 [06:20<01:03,  4.55s/
it]

Iteration 86 -- {'ner': 428.4442598853578}

 87%|███████████████████████████████████████████████▊       | 87/100 [06:24<00:59,  4.55s/
it]


Iteration 87 -- {'ner': 386.31626133986146}

 88%|████████████████████████████████████████████████▍      | 88/100 [06:29<00:54,  4.57s/
it]

Iteration 88 -- {'ner': 318.0317877506144}
```

In [959]:

```
# Recall on training data - Pseudo annotated data
training_recall_model2 = score(pseudo_annotated_data_training_data, nlp_model2)
print(f"Recall on validation data: {training_recall_model2}")

# Recall on val data - Pseudo annotated data
val_recall_model2 = score(pseudo_annotated_data_val_data, nlp_model2)
print(f"Recall on validation data: {val_recall_model2}")
```

```
Recall on validation data: 96.21621621621622
Recall on validation data: 75.75757575757575
```

Prediction on val data

In [960]:

```python
# Prediction on val data - Pseudo annotated data
for idx, data_point in enumerate(val_data_spacy_format):
    print(f'{idx+1}', 'filename- ', val_data['File Name'].iloc[idx])
    doc = nlp_model2(data_point[0])
    for ent in doc.ents:
        print(f'{ent.label_.upper():{30}}- {ent.text}')
    print('--'*50)
```

```
1 filename-  24158401-Rental-Agreement
PARTY_TWO                     - Sri Vishal Bhardwaj
AGREEMENT_VALUE               - Rs 12000 Twelve thousand
AGREEMENT_START_DATA          - 1st April 2008
RENEWAL_NOTICE                - two months
----------------------------------------------------------------------------------------------------
2 filename-  63793679-Rental-Agreement
PARTY_ONE                     - Mr. S Parthasarathy
PARTY_TWO                     - Mr. Hari Kiran Tholeti
AGREEMENT_VALUE               - Rs.9,000/- Rupees Nine Thousand only
AGREEMENT_START_DATA          - 1st September 2011
----------------------------------------------------------------------------------------------------
3 filename-  95980236-Rental-Agreement
PARTY_ONE                     - Mrs. S.Sakunthala
PARTY_TWO                     - V.V.Ravi Kian
AGREEMENT_START_DATA          - 1st April 2010
AGREEMENT_VALUE               - Rs. 9,000/- Nine thousand and two hundred rupees only
RENEWAL_NOTICE                - one month
----------------------------------------------------------------------------------------------------
4 filename-  156155545-Rental-Agreement-Kns-Home
PARTY_TWO                     - SRI VYSHNAVI DAIRY SPECIALITIES Private Ltd.
AGREEMENT_START_DATA          - only
RENEWAL_NOTICE                - one month
----------------------------------------------------------------------------------------------------
5 filename-  195231682-This-RENTAL-AGREEMENT-is-Made-and-Executed-on-24th-Day-of-September
PARTY_TWO                     - 06th day of March 2013
PARTY_ONE                     - Smt C.BHAGYAMMA
PARTY_TWO                     - M/S. JP INTERIO
AGREEMENT_START_DATA          - 06th day of April 2013
AGREEMENT_VALUE               - RS. 13,000/- Rupees Thirteen Thousand Only
RENEWAL_NOTICE                - ONE month
----------------------------------------------------------------------------------------------------
6 filename-  228094620-Rental-Agreement
```

```
PARTY_ONE                      - Mr. KAPIL MEHROTRA
PARTY_TWO                      - Mr.B.Kishore ,
AGREEMENT_VALUE                - Rs. 15,000.00 Rupees Fifteen Thousand Only
AGREEMENT_VALUE                - thousand Rupees Only
AGREEMENT_END_DATA             - 6th June, 2014
RENEWAL_NOTICE                 - one months
----------------------------------------------------------------------------------------------------
7 filename-  239419594-Rental-Agreement
AGREEMENT_VALUE                - Rs. 9000/- Rupees Nine Thousand Only
AGREEMENT_VALUE                - Rs. 90,000/- Rupees Ninety Thousand Only
RENEWAL_NOTICE                 - 3 Three months
----------------------------------------------------------------------------------------------------
8 filename-  269135973-Udaya-Rental-Agreement
PARTY_ONE                      - Mr .Giddappa
PARTY_TWO                      - Ms Pottumurthi Udayalaxmi
AGREEMENT_VALUE                - thousand three hundred only
AGREEMENT_START_DATA           - date of Agreement
RENEWAL_NOTICE                 - Two Month
----------------------------------------------------------------------------------------------------
```

Observation -

- Recall on pseudo annoted labeled data is dropped in approach 2 as compare to approach 1 but prediction on validation data is definitely improved.
- Now, It is able to to predict more than 'PARTY_ONE' and 'PARTY_TWO' entity field.

Let's try retraining the same model (approach - 2) with (drop_rate = 0.2)

5.2.3. Training the model on pseudo_annotated_data_training_data (drop_rate = 0.2)

In [969]:

```python
import random

# Number of iteration to train the model
n_iter = 100
drop_rate = 0.2 # Drop Rate
nlp = spacy.blank('en')

# Training spacy model on pseudo annotated data
train_model(pseudo_annotated_data_training_data, n_iter, drop_rate)

# Saving the model
nlp.to_disk('model2')

# Loading the model
nlp_model2 = spacy.load('model2')
```

```
 91%|█████████████████████████████████████████████████████▋     | 91/100 [06:50<00:41,  4.60s/
it]

Iteration 91 -- {'ner': 86.68351462906045}

 92%|██████████████████████████████████████████████████████▏    | 92/100 [06:55<00:36,  4.61s/i
t]

Iteration 92 -- {'ner': 35.97828843279325}

 93%|██████████████████████████████████████████████████████▊    | 93/100 [07:00<00:32,  4.59s/
it]

Iteration 93 -- {'ner': 85.52451682575928}

 94%|███████████████████████████████████████████████████████▍   | 94/100 [07:04<00:27,  4.61s/
it]

Iteration 94 -- {'ner': 32.867785095587095}

 95%|████████████████████████████████████████████████████████   | 95/100 [07:09<00:23,  4.66s/
it]
```

In [970]:

```python
# Recall on training data - Pseudo annotated data
training_recall_model2 = score(pseudo_annotated_data_training_data, nlp_model2)
print(f"Recall on validation data: {training_recall_model2}")

# Recall on val data - Pseudo annotated data
val_recall_model2 = score(pseudo_annotated_data_val_data, nlp_model2)
print(f"Recall on validation data: {val_recall_model2}")
```

```
Recall on validation data: 97.83783783783784
Recall on validation data: 81.81818181818183
```

Prediction on val data

In [971]:

```python
# Prediction on val data - Pseudo annotated data
for idx, data_point in enumerate(val_data_spacy_format):
    print(f'{idx+1}', 'filename- ', val_data['File Name'].iloc[idx])
    doc = nlp_model2(data_point[0])
    for ent in doc.ents:
        print(f'{ent.label_.upper():{30}}- {ent.text}')
    print('--'*50)
```

```
1 filename-  24158401-Rental-Agreement
PARTY_TWO                     - Sri Vishal Bhardwaj
AGREEMENT_VALUE               - Rs 12000 Twelve thousand
AGREEMENT_START_DATA          - 1st April 2008
RENEWAL_NOTICE                - two months
----------------------------------------------------------------------------------------------------
2 filename-  63793679-Rental-Agreement
PARTY_ONE                     - Mr. S Parthasarathy
PARTY_TWO                     - Mr. Hari Kiran Tholeti
AGREEMENT_VALUE               - Rs.9,000/- Rupees Nine Thousand only
AGREEMENT_START_DATA          - 1st September 2011
----------------------------------------------------------------------------------------------------
3 filename-  95980236-Rental-Agreement
PARTY_ONE                     - Mrs. S.Sakunthala
PARTY_TWO                     - V.V.Ravi Kian
AGREEMENT_START_DATA          - 1st April 2010
AGREEMENT_VALUE               - Rs. 9,000/- Nine thousand and two hundred rupees only
RENEWAL_NOTICE                - one month
----------------------------------------------------------------------------------------------------
4 filename-  156155545-Rental-Agreement-Kns-Home
AGREEMENT_START_DATA          - date of this agreement
RENEWAL_NOTICE                - one month
----------------------------------------------------------------------------------------------------
5 filename-  195231682-This-RENTAL-AGREEMENT-is-Made-and-Executed-on-24th-Day-of-September
PARTY_TWO                     - 06th day of March 2013
PARTY_ONE                     - Smt C.BHAGYAMMA
PARTY_TWO                     - M/S. JP INTERIO
AGREEMENT_START_DATA          - 06th day of April 2013
AGREEMENT_VALUE               - RS. 13,000/- Rupees Thirteen Thousand Only per month
RENEWAL_NOTICE                - ONE month
----------------------------------------------------------------------------------------------------
6 filename-  228094620-Rental-Agreement
PARTY_ONE                     - Mr. KAPIL MEHROTRA
```

```
PARTY_TWO                     - Mr.B.Kishore
AGREEMENT_VALUE               - Rs. 15,000.00 Rupees Fifteen Thousand Only
AGREEMENT_VALUE               - thousand Rupees Only
RENEWAL_NOTICE                - one months
----------------------------------------------------------------------------------------------------
7 filename-  239419594-Rental-Agreement
AGREEMENT_VALUE               - Rs. 9000/- Rupees Nine Thousand Only
AGREEMENT_VALUE               - Rs. 90,000/- Rupees Ninety Thousand Only
RENEWAL_NOTICE                - 3 Three months
----------------------------------------------------------------------------------------------------
8 filename-  269135973-Udaya-Rental-Agreement
PARTY_ONE                     - Mr .Giddappa
PARTY_TWO                     - Ms Pottumurthi Udayalaxmi
AGREEMENT_VALUE               - thousand three hundred only
AGREEMENT_START_DATA          - date of Agreement
RENEWAL_NOTICE                - Two Month
----------------------------------------------------------------------------------------------------
```

Observation -

- Looks like with reducing drop rate, recall on pseudo validation data has been improved, so does the prediction on val data.

What if annotate whole data (training + val) mannualy, will it improve the score and prediction?

Let's find out -

Approach 3 -

1. Mannualy annotate training and validation document using spacy annotator.
2. Finetune Spacy-NER model and observe the performance.

Note: Mannual annotating data could take some extra time but it would be worth exploring this experiment.

In [ ]:

```
import pandas as pd
import spacy_annotator as spa
```

5.3.1. Mannual annotation of training data using spacy ner annotator

In [608]:

```
df_labels = annotator.annotate(df=train_data, col_text="cleaned_text")
```

43 examples annotated, 0 examples left

Aggrement…	ent one, ent two, ent three
Aggrement…	ent one, ent two, ent three
Aggrement…	ent one, ent two, ent three
Renewal N…	ent one, ent two, ent three
Party One	ent one, ent two, ent three
Party Two	ent one, ent two, ent three

submit skip finish

That's all folks!

5.3.2. Mannual annotation of validation data using spacy ner annotator

In [804]:

```python
df_val_labels = annotator.annotate(df=val_data, col_text="cleaned_text")
```

8 examples annotated, 0 examples left

Aggrement… ent one, ent two, ent three

Aggrement… ent one, ent two, ent three

Aggrement… ent one, ent two, ent three

Renewal N… ent one, ent two, ent three

Party One ent one, ent two, ent three

Party Two ent one, ent two, ent three

submit skip finish

That's all folks!

In [ ]:

```python
# df_labels['annotations'].to_csv('../data/mannual_annotated/train_annotations.csv')
# df_val_labels['annotations'].to_csv('../data/mannual_annotated/val_annotations.csv')
```

In [1048]:

```
import pickle

# Save training annotated data
with open('../data/mannual_annotated/train_annotations.pkl', "wb") as f:
    pickle.dump(df_labels['annotations'], f)

# Save validation annotated data
with open('../data/mannual_annotated/val_annotations.pkl', "wb") as f:
    pickle.dump(df_val_labels['annotations'], f)

# loading training annotated data
with open('../data/mannual_annotated/train_annotations.pkl', 'rb') as f:
    mannual_annotated_train_df = pickle.load(f)

# loading validation annotated data
with open('../data/mannual_annotated/val_annotations.pkl', 'rb') as f:
    mannual_annotated_val_df = pickle.load(f)
```

5.3.3. Convert into spacy format

In [1074]:

```
train_data['mannal_annotation'] = mannual_annotated_train_df
val_data['mannal_annotation'] = mannual_annotated_val_df

mannual_annotated_train_df_without_null = [x for x in train_data['mannal_annotation'] if len(x)>0]
mannual_annotated_val_df_without_null = [x for x in val_data['mannal_annotation'] if len(x)>0]
```

5.3.4. Modeling on Mannualy Annotated data

In [1077]:

```python
import random

# Number of iteration to train the model
n_iter = 100
drop_rate = 0.4 # Drop Rate
nlp = spacy.blank('en')

# Training spacy model on pseudo annotated data
train_model(mannual_annotated_train_df_without_null, n_iter, drop_rate)

# Saving the model
nlp.to_disk('model3')

# Loading the model
nlp_model3 = spacy.load('model3')
```

```
 71%|█████████████████████████████████████████████████████████                       | 71/100 [04:12<01:43,  3.58s/i
t]

Iteration 71 -- {'ner': 538.4821247363601}

 72%|█████████████████████████████████████████████████████████▋                       | 72/100 [04:16<01:39,  3.57s/
it]

Iteration 72 -- {'ner': 705.6375217683752}

 73%|██████████████████████████████████████████████████████████▏                      | 73/100 [04:19<01:35,  3.52s/
it]

Iteration 73 -- {'ner': 495.53078146957523}

 74%|████████████████████████████████████████████████████████████▌                    | 74/100 [04:23<01:31,  3.51s/
it]

Iteration 74 -- {'ner': 827.4730336696688}

 75%|█████████████████████████████████████████████████████████████▌                   | 75/100 [04:26<01:27,  3.48s/
```

In [1104]:

```
# Recall on training data - Pseudo annotated data
training_recall_model3 = score(mannual_annotated_train_df_without_null, nlp_model3)
print(f"Recall on validation data: {training_recall_model3}")

# Recall on val data - Pseudo annotated data
val_recall_model3 = score(mannual_annotated_val_df_without_null, nlp_model3)
print(f"Recall on validation data: {val_recall_model3}")
```

```
[E103] Trying to set conflicting doc.ents: '(999, 1009, 'Aggrement Start Date')' and '(968, 1009, 'Aggrement End Dat
e')'. A token can only be part of one entity, so make sure the entities you're setting don't overlap.
Recall on validation data: 78.125
[E103] Trying to set conflicting doc.ents: '(1162, 1176, 'Aggrement Start Date')' and '(1129, 1176, 'Aggrement End Dat
e')'. A token can only be part of one entity, so make sure the entities you're setting don't overlap.
Recall on validation data: 46.666666666666664
```

Prediction on val data

In [1091]:

```python
# Prediction on val data - Pseudo annotated data
for idx, data_point in enumerate(mannual_annotated_val_df_without_null):
    print(f'{idx+1}', 'filename- ', val_data['File Name'].iloc[idx])
    doc = nlp_model3(data_point[0])
    for ent in doc.ents:
        print(f'{ent.label_.upper():{30}}- {ent.text}')
    print('--'*50)
```

```
1 filename-  24158401-Rental-Agreement
AGGREMENT START DATE          - 1st day of April 2008
AGGREMENT START DATE          - 1-04-08 by and between Sri Hanumaiah No 12, 1st Floor, 6th Cross, Balajinagar DRC Pos
t, Bangalore 560029 Hereinafter referred to as the owner Lesser of the one part and in favour of Sri Vishal Bhardwaj
S/O Charnel Singh Village Pandol Road PO and Tehsil Baijnath Dist Kangra H.P. Himachal Pradesh 176125 Hereinafter ref
erred to as the Tenant Lessee of the other part Where as the terms both the lesser and the Lessee shall mean and incl
ude their respective heirs executors legal representatives administrators and assigns. Whereas the lesser herein is t
he absolute owner of the schedule premises situated at No 12, Ground Floor, 6th Cross, Balajinagar, DRC Post, and Ban
galore 560029. Whereas the lessee approached with the lesser let out the schedule premises and the lesser has agreed
to let out the schedule premises under the following terms and conditions The lesser agrees to let out the above prem
ises to the lessee on a monthly rent of Rs 12000 Twelve thousand the lessee has agreed to pay the same to the lesser
regularly. This lease is effective from 1st April 2008
AGGREMENT END DATE            - period of 12 months
----------------------------------------------------------------------------------------------------
2 filename-  63793679-Rental-Agreement
AGGREMENT START DATE          - 01-09-2011
PARTY ONE                     - S Parthasarathy
AGGREMENT VALUE               - 9
PARTY TWO                     - Hari Kiran Tholeti
PARTY ONE                     - S Parthasarathy
AGGREMENT END DATE            - 11 eleven months
PARTY ONE                     - S Parthasarathy
PARTY TWO                     - Hari Kiran Tholeti
----------------------------------------------------------------------------------------------------
3 filename-  95980236-Rental-Agreement
PARTY ONE                     - S.Sakunthala
AGGREMENT END DATE            - period of 11 Eleven months commencing from 1st April 2010
----------------------------------------------------------------------------------------------------
4 filename-  156155545-Rental-Agreement-Kns-Home
RENEWAL NOTICE                - 9 2 18 numbers . Front gate key - 1 No., IN WITHNESSES WHEROF the parties affix their
signature hereunder on this. V.K.NATARAJ WITNESSES LESSOR For SRI VYSHNAVI DAIRY SPECIALITIES Pvt., Ltd., Authorized
signatory LESSEE
----------------------------------------------------------------------------------------------------
```

```
5 filename-  195231682-This-RENTAL-AGREEMENT-is-Made-and-Executed-on-24th-Day-of-September
AGGREMENT START DATE          - 06th day of March 2013
PARTY ONE                     - C.BHAGYAMMA
PARTY TWO                     - JP INTERIO
PARTY TWO                     - JP INTERIO
AGGREMENT END DATE            - 11 months commencing from 06th day of April 2013
RENEWAL NOTICE                - ONE month notice on either side. The sad flat will be used only for Business purpose
and it will be in a proper Tenantable Condition. The said Flat shall keep the premises with all fixtures Electrical i
nstallations etc... In condition as it is let subject to reasonable wear tear. That the tenant herein shall not store
any dangerous or highly inflammable material in the demised premises at any time. At all times, during the term of te
nancy to keep and maintain the premises clean, Tidy, Healthy in and watertight in all seasons and further in good and
substantial repair reasonable wear tear expected. The terms and conditions arrived at above by both the parties are w
ith their own free will and consent and without any coercion or dues from anybody. WITNESSES OWNER l. TENANT
----------------------------------------------------------------------------------------------------
6 filename-  228094620-Rental-Agreement
PARTY ONE                     - B.Kishore
PARTY TWO                     - B. Pampaiah
AGGREMENT END DATE            - 11 months
----------------------------------------------------------------------------------------------------
7 filename-  239419594-Rental-Agreement
AGGREMENT END DATE            - period of 11 Eleven months effective from 07-072014
RENEWAL NOTICE                - 3 Three
PARTY TWO                     - SECOND PARTY
----------------------------------------------------------------------------------------------------
8 filename-  269135973-Udaya-Rental-Agreement
PARTY TWO                     - Pottumurthi Udayalaxmi
AGGREMENT END DATE            - 11 Eleven months from this date of Agreement
----------------------------------------------------------------------------------------------------
```

5.3.5. Modeling on Mannualy Annotated data (drop_rate = 0.2)

In [1097]:

```python
import random

# Number of iteration to train the model
n_iter = 100
drop_rate = 0.2 # Drop Rate
nlp = spacy.blank('en')

# Training spacy model on pseudo annotated data
train_model(mannual_annotated_train_df_without_null, n_iter, drop_rate)

# Saving the model
nlp.to_disk('model4')

# Loading the model
nlp_model4 = spacy.load('model4')
```

```
 45%|████████████████████████████████████▌                                             | 45/100 [02:38<03:17,  3.60s/
it]

Iteration 45 -- {'ner': 74.62215992462495}

 46%|█████████████████████████████████████▌                                            | 46/100 [02:42<03:13,  3.59s/
it]

Iteration 46 -- {'ner': 119.10439395573638}

 47%|██████████████████████████████████████                                           | 47/100 [02:45<03:10,  3.60s/i
t]

Iteration 47 -- {'ner': 99.99710005552109}

 48%|██████████████████████████████████████▉                                           | 48/100 [02:49<03:07,  3.60s/
it]

Iteration 48 -- {'ner': 121.8036911508726}

 49%|███████████████████████████████████████▍                                          | 49/100 [02:52<03:04,  3.61s/
it]
```

In [1100]:

```
# Recall on training data - Pseudo annotated data
training_recall_model4 = score(mannual_annotated_train_df_without_null, nlp_model4)
print(f"Recall on validation data: {training_recall_model4}")

# Recall on val data - Pseudo annotated data
val_recall_model4 = score(mannual_annotated_val_df_without_null, nlp_model4)
print(f"Recall on validation data: {val_recall_model4}")
```

```
[E103] Trying to set conflicting doc.ents: '(999, 1009, 'Aggrement Start Date')' and '(968, 1009, 'Aggrement End Dat
e')'. A token can only be part of one entity, so make sure the entities you're setting don't overlap.
Recall on validation data: 100.0
[E103] Trying to set conflicting doc.ents: '(1162, 1176, 'Aggrement Start Date')' and '(1129, 1176, 'Aggrement End Dat
e')'. A token can only be part of one entity, so make sure the entities you're setting don't overlap.
Recall on validation data: 53.333333333333336
```

Prediction on val data

In [1101]:

```python
# Prediction on val data - Pseudo annotated data
for idx, data_point in enumerate(mannual_annotated_val_df_without_null):
    print(f'{idx+1}', 'filename- ', val_data['File Name'].iloc[idx])
    doc = nlp_model3(data_point[0])
    for ent in doc.ents:
        print(f'{ent.label_.upper():{30}}- {ent.text}')
    print('--'*50)
```

```
1 filename-  24158401-Rental-Agreement
AGGREMENT START DATE          - 1st day of April 2008
AGGREMENT START DATE          - 1-04-08 by and between Sri Hanumaiah No 12, 1st Floor, 6th Cross, Balajinagar DRC Pos
t, Bangalore 560029 Hereinafter referred to as the owner Lesser of the one part and in favour of Sri Vishal Bhardwaj
S/O Charnel Singh Village Pandol Road PO and Tehsil Baijnath Dist Kangra H.P. Himachal Pradesh 176125 Hereinafter ref
erred to as the Tenant Lessee of the other part Where as the terms both the lesser and the Lessee shall mean and incl
ude their respective heirs executors legal representatives administrators and assigns. Whereas the lesser herein is t
he absolute owner of the schedule premises situated at No 12, Ground Floor, 6th Cross, Balajinagar, DRC Post, and Ban
galore 560029. Whereas the lessee approached with the lesser let out the schedule premises and the lesser has agreed
to let out the schedule premises under the following terms and conditions The lesser agrees to let out the above prem
ises to the lessee on a monthly rent of Rs 12000 Twelve thousand the lessee has agreed to pay the same to the lesser
regularly. This lease is effective from 1st April 2008
AGGREMENT END DATE            - period of 12 months
----------------------------------------------------------------------------------------------------
2 filename-  63793679-Rental-Agreement
AGGREMENT START DATE          - 01-09-2011
PARTY ONE                     - S Parthasarathy
AGGREMENT VALUE               - 9
PARTY TWO                     - Hari Kiran Tholeti
PARTY ONE                     - S Parthasarathy
AGGREMENT END DATE            - 11 eleven months
PARTY ONE                     - S Parthasarathy
PARTY TWO                     - Hari Kiran Tholeti
----------------------------------------------------------------------------------------------------
3 filename-  95980236-Rental-Agreement
PARTY ONE                     - S.Sakunthala
AGGREMENT END DATE            - period of 11 Eleven months commencing from 1st April 2010
----------------------------------------------------------------------------------------------------
4 filename-  156155545-Rental-Agreement-Kns-Home
RENEWAL NOTICE                - 9 2 18 numbers . Front gate key - 1 No., IN WITHNESSES WHEROF the parties affix their
signature hereunder on this. V.K.NATARAJ WITNESSES LESSOR For SRI VYSHNAVI DAIRY SPECIALITIES Pvt., Ltd., Authorized
signatory LESSEE
----------------------------------------------------------------------------------------------------
```

```
5 filename-  195231682-This-RENTAL-AGREEMENT-is-Made-and-Executed-on-24th-Day-of-September
AGGREMENT START DATE          - 06th day of March 2013
PARTY ONE                     - C.BHAGYAMMA
PARTY TWO                     - JP INTERIO
PARTY TWO                     - JP INTERIO
AGGREMENT END DATE            - 11 months commencing from 06th day of April 2013
RENEWAL NOTICE                - ONE month notice on either side. The sad flat will be used only for Business purpose
and it will be in a proper Tenantable Condition. The said Flat shall keep the premises with all fixtures Electrical i
nstallations etc... In condition as it is let subject to reasonable wear tear. That the tenant herein shall not store
any dangerous or highly inflammable material in the demised premises at any time. At all times, during the term of te
nancy to keep and maintain the premises clean, Tidy, Healthy in and watertight in all seasons and further in good and
substantial repair reasonable wear tear expected. The terms and conditions arrived at above by both the parties are w
ith their own free will and consent and without any coercion or dues from anybody. WITNESSES OWNER l. TENANT
----------------------------------------------------------------------------------------------------
6 filename-  228094620-Rental-Agreement
PARTY ONE                     - B.Kishore
PARTY TWO                     - B. Pampaiah
AGGREMENT END DATE            - 11 months
----------------------------------------------------------------------------------------------------
7 filename-  239419594-Rental-Agreement
AGGREMENT END DATE            - period of 11 Eleven months effective from 07-072014
RENEWAL NOTICE                - 3 Three
PARTY TWO                     - SECOND PARTY
----------------------------------------------------------------------------------------------------
8 filename-  269135973-Udaya-Rental-Agreement
PARTY TWO                     - Pottumurthi Udayalaxmi
AGGREMENT END DATE            - 11 Eleven months from this date of Agreement
----------------------------------------------------------------------------------------------------
```

Observation -

1. Neither the recall of mannually annotated validation data nor the prediction of validation data is upto mark.
2. Reason could be not able to annotate the data correctly.
3. To verify this, redo the approach 3 again.

Prediction of all 3 approaches

1. Prediction by approach -1

In [1108]:

```python
# Prediction by approach 1

# loading the model approach -1
nlp_model = spacy.load('model1')

for idx, data_point in enumerate(val_data_spacy_format):
    print(f'{idx+1}', 'filename- ', val_data['File Name'].iloc[idx])
    doc = nlp_model(data_point[0])
    for ent in doc.ents:
        print(f'{ent.label_.upper():{30}}- {ent.text}')
    print('--'*50)
```

```
1 filename-  24158401-Rental-Agreement
----------------------------------------------------------------------------------------------------
2 filename-  63793679-Rental-Agreement
PARTY_ONE                     - S Parthasarathy
PARTY_TWO                     - Hari Kiran Tholeti
PARTY_ONE                     - S Parthasarathy
PARTY_ONE                     - S Parthasarathy
PARTY_TWO                     - Hari Kiran Tholeti
----------------------------------------------------------------------------------------------------
3 filename-  95980236-Rental-Agreement
PARTY_TWO                     - V.V.Ravi Kian
----------------------------------------------------------------------------------------------------
4 filename-  156155545-Rental-Agreement-Kns-Home
----------------------------------------------------------------------------------------------------
5 filename-  195231682-This-RENTAL-AGREEMENT-is-Made-and-Executed-on-24th-Day-of-September
PARTY_ONE                     - C.BHAGYAMMA
PARTY_TWO                     - JP INTERIO
PARTY_TWO                     - JP INTERIO
----------------------------------------------------------------------------------------------------
6 filename-  228094620-Rental-Agreement
----------------------------------------------------------------------------------------------------
7 filename-  239419594-Rental-Agreement
----------------------------------------------------------------------------------------------------
8 filename-  269135973-Udaya-Rental-Agreement
PARTY_TWO                     - Pottumurthi Udayalaxmi
----------------------------------------------------------------------------------------------------
```

2. Prediction by approach -2

In [1109]:

```
# Prediction by approach 2

# loading the model approach -2
nlp_model2 = spacy.load('model2')

# Prediction on val data - Pseudo annotated data
for idx, data_point in enumerate(val_data_spacy_format):
    print(f'{idx+1}', 'filename- ', val_data['File Name'].iloc[idx])
    doc = nlp_model2(data_point[0])
    for ent in doc.ents:
        print(f'{ent.label_.upper():{30}}- {ent.text}')
    print('--'*50)
```

```
1 filename-  24158401-Rental-Agreement
PARTY_TWO                     - Sri Vishal Bhardwaj
AGREEMENT_VALUE               - Rs 12000 Twelve thousand
AGREEMENT_START_DATA          - 1st April 2008
RENEWAL_NOTICE                - two months
----------------------------------------------------------------------------------------------------
2 filename-  63793679-Rental-Agreement
PARTY_ONE                     - Mr. S Parthasarathy
PARTY_TWO                     - Mr. Hari Kiran Tholeti
AGREEMENT_VALUE               - Rs.9,000/- Rupees Nine Thousand only
AGREEMENT_START_DATA          - 1st September 2011
----------------------------------------------------------------------------------------------------
3 filename-  95980236-Rental-Agreement
PARTY_ONE                     - Mrs. S.Sakunthala
PARTY_TWO                     - V.V.Ravi Kian
AGREEMENT_START_DATA          - 1st April 2010
AGREEMENT_VALUE               - Rs. 9,000/- Nine thousand and two hundred rupees only
RENEWAL_NOTICE                - one month
----------------------------------------------------------------------------------------------------
4 filename-  156155545-Rental-Agreement-Kns-Home
AGREEMENT_START_DATA          - date of this agreement
RENEWAL_NOTICE                - one month
----------------------------------------------------------------------------------------------------
5 filename-  195231682-This-RENTAL-AGREEMENT-is-Made-and-Executed-on-24th-Day-of-September
PARTY_TWO                     - 06th day of March 2013
PARTY_ONE                     - Smt C.BHAGYAMMA
PARTY_TWO                     - M/S. JP INTERIO
AGREEMENT_START_DATA          - 06th day of April 2013
```

```
AGREEMENT_VALUE              - RS. 13,000/- Rupees Thirteen Thousand Only per month
RENEWAL_NOTICE               - ONE month
----------------------------------------------------------------------------------------------------
6 filename-  228094620-Rental-Agreement
PARTY_ONE                    - Mr. KAPIL MEHROTRA
PARTY_TWO                    - Mr.B.Kishore
AGREEMENT_VALUE              - Rs. 15,000.00 Rupees Fifteen Thousand Only
AGREEMENT_VALUE              - thousand Rupees Only
RENEWAL_NOTICE               - one months
----------------------------------------------------------------------------------------------------
7 filename-  239419594-Rental-Agreement
AGREEMENT_VALUE              - Rs. 9000/- Rupees Nine Thousand Only
AGREEMENT_VALUE              - Rs. 90,000/- Rupees Ninety Thousand Only
RENEWAL_NOTICE               - 3 Three months
----------------------------------------------------------------------------------------------------
8 filename-  269135973-Udaya-Rental-Agreement
PARTY_ONE                    - Mr .Giddappa
PARTY_TWO                    - Ms Pottumurthi Udayalaxmi
AGREEMENT_VALUE              - thousand three hundred only
AGREEMENT_START_DATA         - date of Agreement
RENEWAL_NOTICE               - Two Month
----------------------------------------------------------------------------------------------------
```

2. Prediction by approach -3

In [1110]:

```python
# Prediction by approach 3

# loading the model approach -3
nlp_model3 = spacy.load('model3')

# Prediction on val data - Pseudo annotated data
for idx, data_point in enumerate(mannual_annotated_val_df_without_null):
    print(f'{idx+1}', 'filename- ', val_data['File Name'].iloc[idx])
    doc = nlp_model3(data_point[0])
    for ent in doc.ents:
        print(f'{ent.label_.upper():{30}}- {ent.text}')
    print('--'*50)
```

```
1 filename-  24158401-Rental-Agreement
AGGREMENT START DATE          - 1st day of April 2008
AGGREMENT START DATE          - 1-04-08 by and between Sri Hanumaiah No 12, 1st Floor, 6th Cross, Balajinagar DRC Post, 
Bangalore 560029 Hereinafter referred to as the owner Lesser of the one part and in favour of Sri Vishal Bhardwaj S/O C
harnel Singh Village Pandol Road PO and Tehsil Baijnath Dist Kangra H.P. Himachal Pradesh 176125 Hereinafter referred t
o as the Tenant Lessee of the other part Where as the terms both the lesser and the Lessee shall mean and include their
respective heirs executors legal representatives administrators and assigns. Whereas the lesser herein is the absolute
owner of the schedule premises situated at No 12, Ground Floor, 6th Cross, Balajinagar, DRC Post, and Bangalore 560029.
Whereas the lessee approached with the lesser let out the schedule premises and the lesser has agreed to let out the sc
hedule premises under the following terms and conditions The lesser agrees to let out the above premises to the lessee
on a monthly rent of Rs 12000 Twelve thousand the lessee has agreed to pay the same to the lesser regularly. This lease
is effective from 1st April 2008
AGGREMENT END DATE            - period of 12 months
----------------------------------------------------------------------------------------------------
2 filename-  63793679-Rental-Agreement
AGGREMENT START DATE          - 01-09-2011
PARTY ONE                     - S Parthasarathy
AGGREMENT VALUE               - 9
PARTY TWO                     - Hari Kiran Tholeti
PARTY ONE                     - S Parthasarathy
AGGREMENT END DATE            - 11 eleven months
PARTY ONE                     - S Parthasarathy
PARTY TWO                     - Hari Kiran Tholeti
----------------------------------------------------------------------------------------------------
3 filename-  95980236-Rental-Agreement
PARTY ONE                     - S.Sakunthala
AGGREMENT END DATE            - period of 11 Eleven months commencing from 1st April 2010
----------------------------------------------------------------------------------------------------
```

```
4 filename-  156155545-Rental-Agreement-Kns-Home
RENEWAL NOTICE                - 9 2 18 numbers . Front gate key - 1 No., IN WITHNESSES WHEROF the parties affix their s
ignature hereunder on this. V.K.NATARAJ WITNESSES LESSOR For SRI VYSHNAVI DAIRY SPECIALITIES Pvt., Ltd., Authorized sig
natory LESSEE
----------------------------------------------------------------------------------------------------
5 filename-  195231682-This-RENTAL-AGREEMENT-is-Made-and-Executed-on-24th-Day-of-September
AGGREMENT START DATE          - 06th day of March 2013
PARTY ONE                     - C.BHAGYAMMA
PARTY TWO                     - JP INTERIO
PARTY TWO                     - JP INTERIO
AGGREMENT END DATE            - 11 months commencing from 06th day of April 2013
RENEWAL NOTICE                - ONE month notice on either side. The sad flat will be used only for Business purpose an
d it will be in a proper Tenantable Condition. The said Flat shall keep the premises with all fixtures Electrical insta
llations etc... In condition as it is let subject to reasonable wear tear. That the tenant herein shall not store any d
angerous or highly inflammable material in the demised premises at any time. At all times, during the term of tenancy t
o keep and maintain the premises clean, Tidy, Healthy in and watertight in all seasons and further in good and substant
ial repair reasonable wear tear expected. The terms and conditions arrived at above by both the parties are with their 
own free will and consent and without any coercion or dues from anybody. WITNESSES OWNER l. TENANT
----------------------------------------------------------------------------------------------------
6 filename-  228094620-Rental-Agreement
PARTY ONE                     - B.Kishore
PARTY TWO                     - B. Pampaiah
AGGREMENT END DATE            - 11 months
----------------------------------------------------------------------------------------------------
7 filename-  239419594-Rental-Agreement
AGGREMENT END DATE            - period of 11 Eleven months effective from 07-072014
RENEWAL NOTICE                - 3 Three
PARTY TWO                     - SECOND PARTY
----------------------------------------------------------------------------------------------------
8 filename-  269135973-Udaya-Rental-Agreement
PARTY TWO                     - Pottumurthi Udayalaxmi
AGGREMENT END DATE            - 11 Eleven months from this date of Agreement
----------------------------------------------------------------------------------------------------
```

Observation

- Approach 2 Seems to work best among all.
- If we annotate date properly then approach 3 might have better chance to perform well.

Saving Prediction of apprach 2 (nlp_model2)

In [1111]:

```
# for idx in range(len(val_data.shape[0])):
#     file_name = val_data[Name].iloc[idx]
#     doc = nlp_model3(data_point[0])
#     result = {'File Name': file_name, 'Aggrement Value': np.nan, 'Aggrement Start Date': np.nan,
#     'Renewal Notice (Days)': np.nan, 'Party Two': np.nan}
#     for ent in doc.ents:
#         if ent.label_.upper() == 'AGGREMENT VALUE':
#             result['Aggrement Value'] = ent.text
#         elif ent.label_.upper() == 'AGGREMENT START DATE':
#             result['Aggrement Start Date'] = ent.text

#         elif ent.label_.upper() == 'AGGREMENT START DATE':
#             result['Aggrement Start Date'] = ent.text
```

In [1112]:

```
# # Prediction on val data - Pseudo annotated data
# for idx, data_point in enumerate(mannual_annotated_val_df_without_null):
#     print(f'{idx+1}', 'filename- ', val_data['File Name'].iloc[idx])
#     doc = nlp_model3(data_point[0])
#     for ent in doc.ents:
#         print(f'{ent.label_.upper():{30}}- {ent.text}')
#     print('--'*50)
```

To be continue ..