In [17]:

```python
import nltk
from nltk.tokenize import sent_tokenize
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
import re
import spacy
from spacy import displacy
import pandas as pd
from nltk.corpus import wordnet
import sys
```

In [38]:

```python
nlp=spacy.blank("en")
nlp =spacy.load("en_core_web_sm")
raw_file = open('sample.txt')
data = raw_file.read()
list=[]
list=sys.argv
stats=[[]]
```

Names

In [39]:

```python
def names(data):
# nlp = spacy.load("en_core_web_sm")
# doc = nlp(data)
# for ent in doc.ents:
#     if ent.label_ == "PROPN":
#displacy.render(doc, style='ent') #visualization for entity names using displacy
# name = []
# ent for ent in doc.ents if ent.label == spacy.symbols.PERSON:
    text = data
    doc=nlp(text)
    text = re.sub('\n',' ',text)
    text = re.sub('\t',' ',text)
    for ent in doc.ents:
#        print(ent.text, ent.start_char, ent.end_char, ent.label_)
        if ent.label_=='PERSON':
            text=text.replace(ent.text,'\u2588'*len(ent.text))
            stats.append([ent.text,len(ent.text),'Name'])

    return text
```

In [40]:

```python
names(data)
```

Out[40]:

```
"Message-ID: <14794734.1075840323704.JavaMail.evans@thyme> Date: Mon, 19 Nov 2001 12:30:
44 -0800 (PST) From: chance.rabon@enron.com To: eric.bass@enron.com Subject: RE: Mime-Ve
rsion: 1.0 Content-Type: text/plain; charset=us-ascii Content-Transfer-Encoding: 7bit X-
From: Rabon, Chance </O=ENRON/OU=NA/CN=RECIPIENTS/CN=CRABON> X-To: [redacted], [redacted] </O=ENRON/
OU=NA/CN=RECIPIENTS/CN=Ebass> X-cc:  X-bcc:  X-Folder: \\ExMerge - [redacted], [redacted]\\Inbox\\Pa
intball X-Origin: BASS-E X-FileName: [redacted] 6-25-02.PST  im in 123-456-7892 he, she,
him, her 224 Belmont Street APT 220  225 N Belmont St 220  4 Saffron Hill Road  -----Ori
ginal Message----- From:  [redacted], [redacted]   Sent: Friday, November 16, 2001 1:11 PM To: Love,
[redacted]; Blanchard, Timothy; Ryder, Patrick; Farmer, [redacted]; Smith, [redacted]; Olsen, [redacted]
[redacted]; [redacted], [redacted]; Baumbach, David; Hull, [redacted]; 'val.generes@accenture.com'; Lenhart, Ma
```

```
tthew; 'kevin.a.boone@accenture.com'; Winfree, O'Neal D.; Rabon, Chance; Mills, Bruce Su
bject:   I'm trying to get a feel for everyone's desire to play PAINTBALL in the next fe
w weeks.  This would obviously not be sponsored by Enron b/c Enron doesn't have enough c
ash to buy anything right now.  So let me know if you would be interested and, if so, wh
en you would be available to go.   The cost should be around $30-40 a person.  Please fo
rward to anyone I have forgotten or might be interested.   -[redacted]Message-ID: <14794734.10
75840323704.JavaMail.evans@thyme> Date: Mon, 19 Nov 2001 12:30:44 -0800 (PST) From: chan
ce.rabon@enron.com To: eric.bass@enron.com Subject: RE: Mime-Version: 1.0 Content-Type:
text/plain; charset=us-ascii Content-Transfer-Encoding: 7bit X-From: Rabon, Chance </O=E
NRON/OU=NA/CN=RECIPIENTS/CN=CRABON> X-To: [redacted], [redacted] </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Eb
ass> X-cc:  X-bcc:  X-Folder: \\ExMerge - [redacted], [redacted]\\Inbox\\Paintball X-Origin: BASS-E
X-FileName: [redacted] 6-25-02.PST  im in 123-456-7892 he, she, him, her 224 Belmont Stre
et APT 220 225 N Belmont St 220 4 Saffron Hill Road   -----Original Message----- From:
[redacted], [redacted]   Sent: Friday, November 16, 2001 1:11 PM To: Love, [redacted]; Blanchard, Ti
mothy; Ryder, Patrick; Farmer, [redacted]; Smith, [redacted]; Olsen, [redacted]; [redacted], [redacted]; Baumbac
h, David; Hull, [redacted]; 'val.generes@accenture.com'; Lenhart, Matthew; 'kevin.a.boone@acc
enture.com'; Winfree, O'Neal D.; Rabon, Chance; Mills, Bruce Subject:   I'm trying to ge
t a feel for everyone's desire to play PAINTBALL in the next few weeks.  This would obvi
ously not be sponsored by Enron b/c Enron doesn't have enough cash to buy anything right
now.  So let me know if you would be interested and, if so, when you would be available
to go.   The cost should be around $30-40 a person.  Please forward to anyone I have for
gotten or might be interested.   -[redacted]"
```

Phone

In [41]:

```python
def phone(data):
    text = data
    doc=nlp(text)
    text = re.sub('\n',' ',text)
    text = re.sub('\t',' ',text)
    phone = []
    regex_phone=r'\d{3}[-\.\s]??\d{3}[-\.\s]??\d{4}|\(\d{3}\)\s*\d{3}[-\.\s]?\d{4}|\d{3
    phone=re.findall(regex_phone, text)
    for j in phone:
        text=text.replace(j,'\u2588'*len(j))
        stats.append([j,len(j),'Phone'])
    return text
```

In [43]:

```python
phone(data)
```

Out[43]:

```
"Message-ID: <[redacted]4.[redacted]704.JavaMail.evans@thyme> Date: Mon, 19 Nov 2001 12:30:
44 -0800 (PST) From: chance.rabon@enron.com To: eric.bass@enron.com Subject: RE: Mime-Ve
rsion: 1.0 Content-Type: text/plain; charset=us-ascii Content-Transfer-Encoding: 7bit X-
From: Rabon, Chance </O=ENRON/OU=NA/CN=RECIPIENTS/CN=CRABON> X-To: Bass, Eric </O=ENRON/
OU=NA/CN=RECIPIENTS/CN=Ebass> X-cc:  X-bcc:  X-Folder: \\ExMerge - Bass, Eric\\Inbox\\Pa
intball X-Origin: BASS-E X-FileName: eric bass 6-25-02.PST  im in [redacted] he, she,
him, her 224 Belmont Street APT 220  225 N Belmont St 220  4 Saffron Hill Road  -----Ori
ginal Message----- From:  Bass, Eric   Sent: Friday, November 16, 2001 1:11 PM To: Love,
Phillip M.; Blanchard, Timothy; Ryder, Patrick; Farmer, Daren J.; Smith, Jay; Olsen, Mic
hael; Parks, Joe; Baumbach, David; Hull, Bryan; 'val.generes@accenture.com'; Lenhart, Ma
tthew; 'kevin.a.boone@accenture.com'; Winfree, O'Neal D.; Rabon, Chance; Mills, Bruce Su
bject:   I'm trying to get a feel for everyone's desire to play PAINTBALL in the next fe
w weeks.  This would obviously not be sponsored by Enron b/c Enron doesn't have enough c
ash to buy anything right now.  So let me know if you would be interested and, if so, wh
en you would be available to go.   The cost should be around $30-40 a person.  Please fo
rward to anyone I have forgotten or might be interested.   -EricMessage-ID: <[redacted]4.[redacted]
[redacted]704.JavaMail.evans@thyme> Date: Mon, 19 Nov 2001 12:30:44 -0800 (PST) From: chan
```

```
ce.rabon@enron.com To: eric.bass@enron.com Subject: RE: Mime-Version: 1.0 Content-Type:
text/plain; charset=us-ascii Content-Transfer-Encoding: 7bit X-From: Rabon, Chance </O=E
NRON/OU=NA/CN=RECIPIENTS/CN=CRABON> X-To: Bass, Eric </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Eb
ass> X-cc:  X-bcc:  X-Folder: \\ExMerge - Bass, Eric\\Inbox\\Paintball X-Origin: BASS-E
X-FileName: eric bass 6-25-02.PST  im in [redacted] he, she, him, her 224 Belmont Stre
et APT 220 225 N Belmont St 220 4 Saffron Hill Road   -----Original Message----- From:
Bass, Eric   Sent: Friday, November 16, 2001 1:11 PM To: Love, Phillip M.; Blanchard, Ti
mothy; Ryder, Patrick; Farmer, Daren J.; Smith, Jay; Olsen, Michael; Parks, Joe; Baumbac
h, David; Hull, Bryan; 'val.generes@accenture.com'; Lenhart, Matthew; 'kevin.a.boone@acc
enture.com'; Winfree, O'Neal D.; Rabon, Chance; Mills, Bruce Subject:   I'm trying to ge
t a feel for everyone's desire to play PAINTBALL in the next few weeks.  This would obvi
ously not be sponsored by Enron b/c Enron doesn't have enough cash to buy anything right
now.  So let me know if you would be interested and, if so, when you would be available
to go.   The cost should be around $30-40 a person.  Please forward to anyone I have for
gotten or might be interested.   -Eric"
```

Date

In [44]:

```python
def date(data):
            text = data
            doc=nlp(text)
            text = re.sub('\n',' ',text)
            text = re.sub('\t',' ',text)
            for ent in doc.ents:
                        if ent.label_=='DATE':
                                    stats.append([ent.text,len(ent.text),'Date'])
                                    text=text.replace(ent.text,'\u2588'*len(ent.text))
            return text
```

In [45]:

```python
date(data)
```

Out[45]:

```
"Message-ID: <14794734.1075840323704.JavaMail.evans@thyme> Date: Mon, [redacted] 12:30:
44 -0800 (PST) From: chance.rabon@enron.com To: eric.bass@enron.com Subject: RE: Mime-Ve
rsion: 1.0 Content-Type: text/plain; charset=us-ascii Content-Transfer-Encoding: 7bit X-
From: Rabon, Chance </O=ENRON/OU=NA/CN=RECIPIENTS/CN=CRABON> X-To: Bass, Eric </O=ENRON/
OU=NA/CN=RECIPIENTS/CN=Ebass> X-cc:  X-bcc:  X-Folder: \\ExMerge - Bass, Eric\\Inbox\\Pa
intball X-Origin: BASS-E X-FileName: eric bass 6-25-02.PST  im in 123-456-7892 he, she,
him, her 224 Belmont Street APT 220  225 N Belmont St 220  4 Saffron Hill Road  -----Ori
ginal Message----- From:  Bass, Eric   Sent: [redacted] 1:11 PM To: Love,
Phillip M.; Blanchard, Timothy; Ryder, Patrick; Farmer, Daren J.; Smith, Jay; Olsen, Mic
hael; Parks, Joe; Baumbach, David; Hull, Bryan; 'val.generes@accenture.com'; Lenhart, Ma
tthew; 'kevin.a.boone@accenture.com'; Winfree, O'Neal D.; Rabon, Chance; Mills, Bruce Su
bject:   I'm trying to get a feel for everyone's desire to play PAINTBALL in [redacted]
.  This would obviously not be sponsored by Enron b/c Enron doesn't have enough c
ash to buy anything right now.  So let me know if you would be interested and, if so, wh
en you would be available to go.   The cost should be around $30-40 a person.  Please fo
rward to anyone I have forgotten or might be interested.   -EricMessage-ID: <14794734.10
75840323704.JavaMail.evans@thyme> Date: Mon, [redacted] 12:30:44 -0800 (PST) From: chan
ce.rabon@enron.com To: eric.bass@enron.com Subject: RE: Mime-Version: 1.0 Content-Type:
text/plain; charset=us-ascii Content-Transfer-Encoding: 7bit X-From: Rabon, Chance </O=E
NRON/OU=NA/CN=RECIPIENTS/CN=CRABON> X-To: Bass, Eric </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Eb
ass> X-cc:  X-bcc:  X-Folder: \\ExMerge - Bass, Eric\\Inbox\\Paintball X-Origin: BASS-E
X-FileName: eric bass 6-25-02.PST  im in 123-456-7892 he, she, him, her 224 Belmont Stre
et APT 220 225 N Belmont St 220 4 Saffron Hill Road   -----Original Message----- From:
Bass, Eric   Sent: [redacted] 1:11 PM To: Love, Phillip M.; Blanchard, Ti
mothy; Ryder, Patrick; Farmer, Daren J.; Smith, Jay; Olsen, Michael; Parks, Joe; Baumbac
h, David; Hull, Bryan; 'val.generes@accenture.com'; Lenhart, Matthew; 'kevin.a.boone@acc
```

```
enture.com'; Winfree, O'Neal D.; Rabon, Chance; Mills, Bruce Subject:   I'm trying to ge
t a feel for everyone's desire to play PAINTBALL in [redacted].  This would obvi
ously not be sponsored by Enron b/c Enron doesn't have enough cash to buy anything right
now.  So let me know if you would be interested and, if so, when you would be available
to go.   The cost should be around $30-40 a person.  Please forward to anyone I have for
gotten or might be interested.   -Eric"
```

Address

In [46]:

```
def address(data):

    text = data
    doc=nlp(text)
    text = re.sub('\n',' ',text)
    text = re.sub('\t',' ',text)
#     regex_gender = r'^(\d+) ?([A-Za-z](?= ))? (.*?) ([^ ]+?) ?((?<= )APT)? ?((?<= )\d
#     gender = re.findall(regex_gender,text)
#     for i in gender:
#         text = text.replace(i,'\u2588'*len(i))
#     return text

    for ent in doc.ents:
            #print(ent.text, ent.start_char, ent.end_char, ent.label_)
            if ent.label_=='FAC':
                text=text.replace(ent.text,'\u2588'*len(ent.text))
                stats.append([ent.text,len(ent.text),'Date'])
    return text
```

In [47]:

```
address(data)
```

Out[47]:

```
"Message-ID: <14794734.1075840323704.JavaMail.evans@thyme> Date: Mon, 19 Nov 2001 12:30:
44 -0800 (PST) From: chance.rabon@enron.com To: eric.bass@enron.com Subject: RE: Mime-Ve
rsion: 1.0 Content-Type: text/plain; charset=us-ascii Content-Transfer-Encoding: 7bit X-
From: Rabon, Chance </O=ENRON/OU=NA/CN=RECIPIENTS/CN=CRABON> X-To: Bass, Eric </O=ENRON/
OU=NA/CN=RECIPIENTS/CN=Ebass> X-cc:  X-bcc:  X-Folder: \\ExMerge - Bass, Eric\\Inbox\\Pa
intball X-Origin: BASS-E X-FileName: eric bass 6-25-02.PST  im in 123-456-7892 he, she,
him, her 224 [redacted] APT 220  225 N Belmont St 220  4 Saffron Hill Road  -----Ori
ginal Message----- From:  Bass, Eric   Sent: Friday, November 16, 2001 1:11 PM To: Love,
Phillip M.; Blanchard, Timothy; Ryder, Patrick; Farmer, Daren J.; Smith, Jay; Olsen, Mic
hael; Parks, Joe; Baumbach, David; Hull, Bryan; 'val.generes@accenture.com'; Lenhart, Ma
tthew; 'kevin.a.boone@accenture.com'; Winfree, O'Neal D.; Rabon, Chance; Mills, Bruce Su
bject:   I'm trying to get a feel for everyone's desire to play PAINTBALL in the next fe
w weeks.  This would obviously not be sponsored by Enron b/c Enron doesn't have enough c
ash to buy anything right now.  So let me know if you would be interested and, if so, wh
en you would be available to go.   The cost should be around $30-40 a person.  Please fo
rward to anyone I have forgotten or might be interested.   -EricMessage-ID: <14794734.10
75840323704.JavaMail.evans@thyme> Date: Mon, 19 Nov 2001 12:30:44 -0800 (PST) From: chan
ce.rabon@enron.com To: eric.bass@enron.com Subject: RE: Mime-Version: 1.0 Content-Type:
text/plain; charset=us-ascii Content-Transfer-Encoding: 7bit X-From: Rabon, Chance </O=E
NRON/OU=NA/CN=RECIPIENTS/CN=CRABON> X-To: Bass, Eric </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Eb
ass> X-cc:  X-bcc:  X-Folder: \\ExMerge - Bass, Eric\\Inbox\\Paintball X-Origin: BASS-E
X-FileName: eric bass 6-25-02.PST  im in 123-456-7892 he, she, him, her 224 [redacted]
 APT 220 225 N Belmont St 220 4 Saffron Hill Road   -----Original Message----- From:
Bass, Eric   Sent: Friday, November 16, 2001 1:11 PM To: Love, Phillip M.; Blanchard, Ti
mothy; Ryder, Patrick; Farmer, Daren J.; Smith, Jay; Olsen, Michael; Parks, Joe; Baumbac
h, David; Hull, Bryan; 'val.generes@accenture.com'; Lenhart, Matthew; 'kevin.a.boone@acc
enture.com'; Winfree, O'Neal D.; Rabon, Chance; Mills, Bruce Subject:   I'm trying to ge
```

```
t a feel for everyone's desire to play PAINTBALL in the next few weeks.  This would obvi
ously not be sponsored by Enron b/c Enron doesn't have enough cash to buy anything right
now.  So let me know if you would be interested and, if so, when you would be available
to go.   The cost should be around $30-40 a person.  Please forward to anyone I have for
gotten or might be interested.   -Eric"
```

Concept

In [48]:

```python
def concept(data, concept):
    concept_list = []
    concept_count = 0
    synonyms = wordnet.synsets(concept)
    for i in synonyms:
        temp = i.lemma_names()
        for f in temp:
            if f not in concept_list:
                concept_list.append(f)

    for k in nltk.sent_tokenize(data):
        for l in concept_list:
            if l.lower() in k.lower():
                data = data.replace(k, len(k)*'\u2588')
                stats.append([k,len(k),'Concept'])
#                 concept_count += 1
    return data, concept_list
```

In [49]:

```python
concept(data,'interest')
```

Out[49]:

```
("Message-ID: <14794734.1075840323704.JavaMail.evans@thyme>\nDate: Mon, 19 Nov 2001 12:3
0:44 -0800 (PST)\nFrom: chance.rabon@enron.com\nTo: eric.bass@enron.com\nSubject: RE:\nM
ime-Version: 1.0\nContent-Type: text/plain; charset=us-ascii\nContent-Transfer-Encoding:
7bit\nX-From: Rabon, Chance </O=ENRON/OU=NA/CN=RECIPIENTS/CN=CRABON>\nX-To: Bass, Eric
</O=ENRON/OU=NA/CN=RECIPIENTS/CN=Ebass>\nX-cc: \nX-bcc: \nX-Folder: \\ExMerge - Bass, Er
ic\\Inbox\\Paintball\nX-Origin: BASS-E\nX-FileName: eric bass 6-25-02.PST\n\nim in\n123-
456-7892\nhe, she, him, her\n224 Belmont Street APT 220\n\n225 N Belmont St 220\n\n4 Saf
fron Hill Road\n -----Original Message-----\nFrom: \tBass, Eric  \nSent:\tFriday, Novemb
er 16, 2001 1:11 PM\nTo:\tLove, Phillip M.; Blanchard, Timothy; Ryder, Patrick; Farmer,
Daren J.; Smith, Jay; Olsen, Michael; Parks, Joe; Baumbach, David; Hull, Bryan; 'val.gen
eres@accenture.com'; Lenhart, Matthew; 'kevin.a.boone@accenture.com'; Winfree, O'Neal
D.; Rabon, Chance; Mills, Bruce\nSubject:\t\n\nI'm trying to get a feel for everyone's d
esire to play PAINTBALL in the next few weeks.  This would obviously not be sponsored by
Enron b/c Enron doesn't have enough cash to buy anything right now.\n\n[redacted]
  The cost shoul
d be around $30-40 a person.\n\n[redacted]
\n\n\n-EricMessage-ID: <14794734.1075840323704.JavaMail.evans@thyme>\nDate: Mo
n, 19 Nov 2001 12:30:44 -0800 (PST)\nFrom: chance.rabon@enron.com\nTo: eric.bass@enron.c
om\nSubject: RE:\nMime-Version: 1.0\nContent-Type: text/plain; charset=us-ascii\nContent
-Transfer-Encoding: 7bit\nX-From: Rabon, Chance </O=ENRON/OU=NA/CN=RECIPIENTS/CN=CRABON>
\nX-To: Bass, Eric </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Ebass>\nX-cc: \nX-bcc: \nX-Folder:
\\ExMerge - Bass, Eric\\Inbox\\Paintball\nX-Origin: BASS-E\nX-FileName: eric bass 6-25-0
2.PST\n\nim in\n123-456-7892\nhe, she, him, her\n224 Belmont Street APT 220\n225 N Belmo
nt St 220\n4 Saffron Hill Road\n\n -----Original Message-----\nFrom: \tBass, Eric  \nSen
t:\tFriday, November 16, 2001 1:11 PM\nTo:\tLove, Phillip M.; Blanchard, Timothy; Ryder,
Patrick; Farmer, Daren J.; Smith, Jay; Olsen, Michael; Parks, Joe; Baumbach, David; Hul
l, Bryan; 'val.generes@accenture.com'; Lenhart, Matthew; 'kevin.a.boone@accenture.com';
Winfree, O'Neal D.; Rabon, Chance; Mills, Bruce\nSubject:\t\n\nI'm trying to get a feel
for everyone's desire to play PAINTBALL in the next few weeks.  This would obviously not
be sponsored by Enron b/c Enron doesn't have enough cash to buy anything right now.\n\n[redacted]
The cost should be around $30-40 a person.\n\n[redacted]\n\n\n-Eric",
 ['interest',
  'involvement',
  'sake',
  'interestingness',
  'stake',
  'interest_group',
  'pastime',
  'pursuit',
  'concern',
  'occupy',
  'worry',
  'matter_to'])
```

In [50]:

```
df=pd.DataFrame(stats)
df.to_csv(r'stats.txt',header=None, index=True, sep=' ')
```

In [51]:

```
df
```

Out[51]:

	0	1	2
0	None	NaN	None
1	Bass	4.0	Name
2	Eric	4.0	Name
3	Bass	4.0	Name
4	eric bass	9.0	Name
5	Bass	4.0	Name
6	Eric \nSent	11.0	Name
7	Phillip M.	10.0	Name
8	Daren J.	8.0	Name
9	Jay	3.0	Name
10	Michael	7.0	Name
11	Parks	5.0	Name
12	Joe	3.0	Name
13	Bryan	5.0	Name
14	Bass	4.0	Name
15	Eric	4.0	Name
16	Bass	4.0	Name
17	eric bass	9.0	Name
18	Bass	4.0	Name
19	Eric \nSent	11.0	Name

	0	1	2
20	Phillip M.	10.0	Name
21	Daren J.	8.0	Name
22	Jay	3.0	Name
23	Michael	7.0	Name
24	Parks	5.0	Name
25	Joe	3.0	Name
26	Bryan	5.0	Name
27	1479473	7.0	Phone
28	1075840323	10.0	Phone
29	123-456-7892	12.0	Phone
30	1479473	7.0	Phone
31	1075840323	10.0	Phone
32	123-456-7892	12.0	Phone
33	19 Nov 2001	11.0	Date
34	Friday, November 16, 2001	25.0	Date
35	the next few weeks	18.0	Date
36	19 Nov 2001	11.0	Date
37	Friday, November 16, 2001	25.0	Date
38	the next few weeks	18.0	Date
39	Belmont Street	14.0	Date
40	Belmont Street	14.0	Date
41	So let me know if you would be interested and,...	88.0	Concept
42	Please forward to anyone I have forgotten or m...	65.0	Concept
43	So let me know if you would be interested and,...	88.0	Concept
44	Please forward to anyone I have forgotten or m...	65.0	Concept

```
In [ ]:
```