In [19]:
```
from hdfs import InsecureClient
import pandas as pd
client = InsecureClient('http://datalake:50070')
```

In [20]:
```
with client.read('/shared/EmailSamples/email1.txt') as reader:
    df = pd.read_csv(reader, sep="\n");
print(df)
```

```
                                          Dear all,
0   New week, new prizes: yesterday we had the joy...
1   Tomorrow, Hélène Carrère d'Encausse will recei...
2        Other good news that we are delighted about:
3   Next week marks the start of Aurélie Valognes ...
4   La Ritournelle 4551395 will be on RTL from nex...
5   Il nous restera ça 4702273 will accompany holi...
6   Cardinal Robert Sarah (Catechism of the Spirit...
7   Challenges published this week a portrait of N...
8   Sarah Briand's novel, Les pépins de Grenade 12...
9   Janine Boissard's new novel, Quand la belle se...
10  And all the press review of the week in attach...
11                              Have a good weekend,
12                                  Sincerely yours,
13                                           Pauline
```

In [21]:
```
import re
import spacy
from spacy import displacy
```

In [22]:
```
nlp = spacy.load("en_core_web_sm")
```

In [23]:
```
import pickle
pickle.dump(nlp, open("en_core_web_sm.pickle.dat", "wb"))
```

In [24]:
```
nlp = pickle.load(open("en_core_web_sm.pickle.dat", "rb"))
```

In [25]:
```
from datetime import datetime
import parsedatetime as pdt # $ pip install parsedatetime

def get_date(date_list, current_date):
  cal = pdt.Calendar()
  now = current_date
  ans = []
  #print("now: %s" % now)
  for time_string in date_list:
    if(time_string.find('to') != -1):
        a,b = time_string.split('to')
        a = cal.parseDT(a,now)[0]
        b = cal.parseDT(b,now)[0]
        return (a.strftime("%d/%m/%Y"),b.strftime("%d/%m/%Y"))

    x = cal.parseDT(time_string, now)[0]
    #print("%s:\t%s" % (time_string, x))
    ans.append(x.strftime("%d/%m/%Y"))

  return get_start_end_date(ans, current_date)
```

In [26]:
```
def get_start_end_date(date_list, current_date):

  if(len(date_list) == 0):
    return (current_date.strftime("%d/%m/%Y"), current_date.strftime("%d/%m/%Y"))
  elif(len(date_list) == 1):
    return (date_list[0], date_list[0])
  else:
    listt = []
    for l in date_list:
      listt.append(datetime.strptime(l, '%d/%m/%Y').date())
    x = min(listt)
    y = max(listt)
    return (x.strftime("%d/%m/%Y"), y.strftime("%d/%m/%Y"))
```

In [27]:
```
pattern = pattern = r"(\d{%d})"%7
email = []
for s in df.values.tolist():
    if(re.findall(pattern, s[0])):
        email.append(s[0])
        print(s[0],"***")
print(len(email))
```

```
New week, new prizes: yesterday we had the joy of learning that Henriette Michaud ha
d received the Prix de l'Essai 2022 from the Académie Française for Freud in Bloomsb
ury 3123559: https://www.academie-francaise.fr/sites/academie-francaise.fr/files/pal
mares_2022_vf.pdf ***
Tomorrow, Hélène Carrère d'Encausse will receive an honorary prize for her work at t
he Hossegor book fair (Alexandra Kollontaï large format: 7757456 and plural to be pu
blished in November: 6912274). ***
La Ritournelle 4551395 will be on RTL from next Monday until 10 July and then on Fra
nce Bleu for the biggest summer run from Thursday 28 July to Wednesday 3 August. ***
Il nous restera ça 4702273 will accompany holidaymakers during RTL's most popular ho
urs from Thursday 28 July to Sunday 31 July and then from Friday 12 August to Monday
15 August. ***
Cardinal Robert Sarah (Catechism of the Spiritual Life 2504824) is in the spotlight
this weekend in Le Figaro Magazine with a major four-page interview (enclosed). Next
week, he will receive exceptional media visibility. ***
Challenges published this week a portrait of Nicolas Forissier for L'ennemi intérieu
r 6180412 (en pj). ***
Sarah Briand's novel, Les pépins de Grenade 1282225, is receiving good media coverag
e. Femme Actuelle gave it a special mention this week (en pj), while Anne-Marie Revo
l praised the novel in Patricia Loison's 23H last night: https://www.francetvinfo.f
r/replay-jt/franceinfo/21h-minuit/23-heures/jt-le-23h-jeudi-30-juin-2022_5230810.htm
l ***
Janine Boissard's new novel, Quand la belle se réveillera 4082000, continues to be h
onoured, as shown by the fine reviews in the newspaper Centre presse (en pj) and Bru
xelle culture (en pj). ***
8
```

In [28]:
```
import re
final_struct_list = []
for s in email:
  struct_list = []
  doc2 = nlp(s)
  displacy.render(doc2, style='ent', jupyter=True)
  pattern = r"(\d{%d})"%7
  print("Product ID: ", end=" ")
  prod_id = re.findall(pattern, str(doc2))
  struct_list.append(prod_id)
  print(prod_id)
  date_list = []
  for ent in filter(lambda e : e.label_ == 'DATE', doc2.ents):
    date_list.append(ent.text)
```

```
    start_date, end_date = get_date(date_list,datetime(2022,7,28))
    struct_list.append(start_date)
    print("Start Date: ", start_date)
    struct_list.append(end_date)
    print("End Date: ", end_date)
    print("Description: ", end=" ")
    temp_str = ""
    for e in doc2.ents:
      temp_str = temp_str + str(e)+" "
    struct_list.append(temp_str)
    print(temp_str)
    final_struct_list.append(struct_list)
    print("\n**********************************************************************************
#print(final_struct_list)
```

New week **DATE** , new prizes: yesterday **DATE** we had the joy of learning that Henriette Michaud **PERSON** had received the Prix de l'Essai 2022 **ORG** from the Académie **ORG** Française for Freud in Bloomsbury 3123559 **DATE** : https://www.academie-francaise.fr/sites/academie-francaise.fr/files/palmares_2022_vf.pdf

```
Product ID:  ['3123559']
Start Date:  27/07/2022
End Date:  28/07/2022
Description:  New week yesterday Henriette Michaud the Prix de l'Essai 2022 Académie
Bloomsbury 3123559

**********************************************************************************
**********************************************************************************
**********************************************************************************
```

Tomorrow **DATE** , Hélène Carrère d'Encausse **ORG** will receive an honorary prize for her work at the Hossegor book fair **EVENT** (Alexandra Kollontaï **PERSON** large format: 7757456 **DATE** and plural to be published in November **DATE** : 6912274 **CARDINAL**).

```
Product ID:  ['7757456', '6912274']
Start Date:  28/07/2022
End Date:  01/11/2022
Description:  Tomorrow Hélène Carrère d'Encausse the Hossegor book fair Alexandra Ko
llontaï 7757456 November 6912274

**********************************************************************************
**********************************************************************************
**********************************************************************************
```

La Ritournelle **PERSON** 4551395 **DATE** will be on RTL **ORG** from next Monday **DATE** until 10 July **DATE** and then on France Bleu **ORG** for the biggest summer run from Thursday 28 July to Wednesday 3 August **DATE** .

```
Product ID:  ['4551395']
Start Date:  28/07/2022
End Date:  03/08/2022
Description:  La Ritournelle 4551395 RTL next Monday 10 July France Bleu Thursday 28
July to Wednesday 3 August

**********************************************************************************
**********************************************************************************
**********************************************************************************
```

Il nous restera ça 4702273 **DATE** will accompany holidaymakers during RTL **ORG** 's most popular hours **TIME** from Thursday 28 July to Sunday 31 July **DATE** and then from Friday 12 August to Monday 15 August **DATE** .

```
Product ID:  ['4702273']
Start Date:  28/07/2022
End Date:  31/07/2022
Description:  4702273 RTL most popular hours Thursday 28 July to Sunday 31 July Frid
ay 12 August to Monday 15 August

**********************************************************************************
**********************************************************************************
**********************************************************************
```

Cardinal Robert Sarah **PERSON** (Catechism of the Spiritual Life 2504824 **CARDINAL**) is in the spotlight this weekend **DATE** in Le Figaro Magazine **ORG** with a major four **CARDINAL** -page interview (enclosed). Next week **DATE** , he will receive exceptional media visibility.

```
Product ID:  ['2504824']
Start Date:  28/07/2022
End Date:  04/08/2022
Description:  Robert Sarah 2504824 this weekend Le Figaro Magazine four Next week

**********************************************************************************
**********************************************************************************
**********************************************************************
```

Challenges published this week **DATE** a portrait of Nicolas Forissier **PERSON** for L'ennemi **CARDINAL** intérieur 6180412 **DATE** (en pj).

```
Product ID:  ['6180412']
Start Date:  28/07/2022
End Date:  29/07/2022
Description:  this week Nicolas Forissier L'ennemi 6180412

**********************************************************************************
**********************************************************************************
**********************************************************************
```

Sarah Briand's **PERSON** novel, Les pépins de Grenade 1282225 **DATE** , is receiving good media coverage. Femme Actuelle **PERSON** gave it a special mention this week **DATE** (en pj), while Anne-Marie Revol **PERSON** praised the novel in Patricia Loison's **PERSON** 23H last night **TIME** : https://www.francetvinfo.fr/replay-jt/franceinfo/21h-minuit/23-heures/jt-le-23h-jeudi-30-juin-2022_5230810.html

```
Product ID:  ['1282225', '5230810']
Start Date:  28/07/2022
End Date:  29/07/2022
Description:  Sarah Briand's 1282225 Actuelle this week Anne-Marie Revol Patricia Lo
ison's 23H last night

**********************************************************************************
**********************************************************************************
**********************************************************************
```

Janine Boissard's **PERSON** new novel, Quand la belle **PERSON** se réveillera 4082000 **DATE** , continues to be honoured, as shown by the fine reviews in the newspaper Centre **PRODUCT** presse (en pj) and Bruxelle **GPE** culture (en pj).

```
Product ID:  ['4082000']
Start Date:  28/07/2022
End Date:  28/07/2022
Description:  Janine Boissard's Quand la belle 4082000 Centre Bruxelle

********************************************************************************
********************************************************************************
********************************************************************************
```

In [29]:

```
import pandas as pd

df = pd.DataFrame(final_struct_list, columns = ['Product ID', 'Start Date', 'End Dat
df
```

Out[29]:

	Product ID	Start Date	End Date	Description
0	[3123559]	27/07/2022	28/07/2022	New week yesterday Henriette Michaud the Prix ...
1	[7757456, 6912274]	28/07/2022	01/11/2022	Tomorrow Hélène Carrère d'Encausse the Hossego...
2	[4551395]	28/07/2022	03/08/2022	La Ritournelle 4551395 RTL next Monday 10 July...
3	[4702273]	28/07/2022	31/07/2022	4702273 RTL most popular hours Thursday 28 Jul...
4	[2504824]	28/07/2022	04/08/2022	Robert Sarah 2504824 this weekend Le Figaro Ma...
5	[6180412]	28/07/2022	29/07/2022	this week Nicolas Forissier L'ennemi 6180412
6	[1282225, 5230810]	28/07/2022	29/07/2022	Sarah Briand's 1282225 Actuelle this week Anne...
7	[4082000]	28/07/2022	28/07/2022	Janine Boissard's Quand la belle 4082000 Centr...

In [30]:

```
Requirement already satisfied: spacy in /opt/conda/lib/python3.7/site-packages (2.3.
4)
Requirement already satisfied: catalogue<1.1.0,>=0.0.7 in /opt/conda/lib/python3.7/s
ite-packages (from spacy) (1.0.0)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /opt/conda/lib/python3.7/site-
packages (from spacy) (4.64.0)
Requirement already satisfied: numpy>=1.15.0 in /opt/conda/lib/python3.7/site-packag
es (from spacy) (1.21.6)
Requirement already satisfied: blis<0.8.0,>=0.4.0; python_version >= "3.6" in /opt/c
onda/lib/python3.7/site-packages (from spacy) (0.7.3)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /opt/conda/lib/python3.7/site-
packages (from spacy) (2.0.4)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /opt/conda/lib/python3.
7/site-packages (from spacy) (1.0.4)
Requirement already satisfied: srsly<1.1.0,>=1.0.2 in /opt/conda/lib/python3.7/site-
packages (from spacy) (1.0.4)
Requirement already satisfied: setuptools in /opt/conda/lib/python3.7/site-packages
(from spacy) (65.6.0)
Requirement already satisfied: plac<1.2.0,>=0.9.6 in /opt/conda/lib/python3.7/site-p
ackages (from spacy) (0.9.6)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /opt/conda/lib/python3.7/s
ite-packages (from spacy) (2.23.0)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /opt/conda/lib/python3.7/sit
e-packages (from spacy) (3.0.4)
Requirement already satisfied: thinc<7.5.0,>=7.4.1 in /opt/conda/lib/python3.7/site-
```

```
packages (from spacy) (7.4.3)
Requirement already satisfied: wasabi<1.1.0,>=0.4.0 in /opt/conda/lib/python3.7/site
-packages (from spacy) (0.10.1)
Requirement already satisfied: importlib-metadata>=0.20; python_version < "3.8" in /
opt/conda/lib/python3.7/site-packages (from catalogue<1.1.0,>=0.0.7->spacy) (4.11.4)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.7/site-p
ackages (from requests<3.0.0,>=2.13.0->spacy) (2022.9.24)
Requirement already satisfied: idna<3,>=2.5 in /opt/conda/lib/python3.7/site-package
s (from requests<3.0.0,>=2.13.0->spacy) (2.8)
Requirement already satisfied: chardet<4,>=3.0.2 in /opt/conda/lib/python3.7/site-pa
ckages (from requests<3.0.0,>=2.13.0->spacy) (3.0.4)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /opt/cond
a/lib/python3.7/site-packages (from requests<3.0.0,>=2.13.0->spacy) (1.24.3)
Requirement already satisfied: typing-extensions>=3.6.4; python_version < "3.8" in /
opt/conda/lib/python3.7/site-packages (from importlib-metadata>=0.20; python_version
< "3.8"->catalogue<1.1.0,>=0.0.7->spacy) (4.2.0)
Requirement already satisfied: zipp>=0.5 in /opt/conda/lib/python3.7/site-packages
(from importlib-metadata>=0.20; python_version < "3.8"->catalogue<1.1.0,>=0.0.7->spa
cy) (3.8.0)
```

In [31]:

```
import os
import subprocess
import logging
import time
import sys
```

In [43]:

```
process = subprocess.Popen("python -m spacy download en_core_web_lg", shell=True,
                           stdout=subprocess.PIPE,
                           stderr=subprocess.PIPE, universal_newlines=True)
```

In [44]:

```
out, err = process.communicate()
```

In [46]:

```
print(out,err)
```

```
Requirement already satisfied: en_core_web_lg==2.3.1 from https://github.com/explosi
on/spacy-models/releases/download/en_core_web_lg-2.3.1/en_core_web_lg-2.3.1.tar.gz#e
gg=en_core_web_lg==2.3.1 in /opt/conda/lib/python3.7/site-packages (2.3.1)
Requirement already satisfied: spacy<2.4.0,>=2.3.0 in /opt/conda/lib/python3.7/site-
packages (from en_core_web_lg==2.3.1) (2.3.4)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /opt/conda/lib/python3.7/sit
e-packages (from spacy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1) (3.0.4)
Requirement already satisfied: thinc<7.5.0,>=7.4.1 in /opt/conda/lib/python3.7/site-
packages (from spacy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1) (7.4.3)
Requirement already satisfied: numpy>=1.15.0 in /opt/conda/lib/python3.7/site-packag
es (from spacy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1) (1.21.6)
Requirement already satisfied: catalogue<1.1.0,>=0.0.7 in /opt/conda/lib/python3.7/s
ite-packages (from spacy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1) (1.0.0)
Requirement already satisfied: wasabi<1.1.0,>=0.4.0 in /opt/conda/lib/python3.7/site
-packages (from spacy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1) (0.10.1)
Requirement already satisfied: plac<1.2.0,>=0.9.6 in /opt/conda/lib/python3.7/site-p
ackages (from spacy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1) (0.9.6)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /opt/conda/lib/python3.
7/site-packages (from spacy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1) (1.0.4)
Requirement already satisfied: srsly<1.1.0,>=1.0.2 in /opt/conda/lib/python3.7/site-
packages (from spacy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1) (1.0.4)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /opt/conda/lib/python3.7/site-
packages (from spacy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1) (4.64.0)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /opt/conda/lib/python3.7/site-
packages (from spacy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1) (2.0.4)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /opt/conda/lib/python3.7/s
ite-packages (from spacy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1) (2.23.0)
Requirement already satisfied: blis<0.8.0,>=0.4.0; python_version >= "3.6" in /opt/c
onda/lib/python3.7/site-packages (from spacy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1)
```

```
(0.7.3)
Requirement already satisfied: setuptools in /opt/conda/lib/python3.7/site-packages
(from spacy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1) (65.6.0)
Requirement already satisfied: importlib-metadata>=0.20; python_version < "3.8" in /
opt/conda/lib/python3.7/site-packages (from catalogue<1.1.0,>=0.0.7->spacy<2.4.0,>=
2.3.0->en_core_web_lg==2.3.1) (4.11.4)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /opt/cond
a/lib/python3.7/site-packages (from requests<3.0.0,>=2.13.0->spacy<2.4.0,>=2.3.0->en
_core_web_lg==2.3.1) (1.24.3)
Requirement already satisfied: idna<3,>=2.5 in /opt/conda/lib/python3.7/site-package
s (from requests<3.0.0,>=2.13.0->spacy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1) (2.8)
Requirement already satisfied: chardet<4,>=3.0.2 in /opt/conda/lib/python3.7/site-pa
ckages (from requests<3.0.0,>=2.13.0->spacy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1)
(3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.7/site-p
ackages (from requests<3.0.0,>=2.13.0->spacy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1)
(2022.9.24)
Requirement already satisfied: zipp>=0.5 in /opt/conda/lib/python3.7/site-packages
(from importlib-metadata>=0.20; python_version < "3.8"->catalogue<1.1.0,>=0.0.7->spa
cy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1) (3.8.0)
Requirement already satisfied: typing-extensions>=3.6.4; python_version < "3.8" in /
opt/conda/lib/python3.7/site-packages (from importlib-metadata>=0.20; python_version
< "3.8"->catalogue<1.1.0,>=0.0.7->spacy<2.4.0,>=2.3.0->en_core_web_lg==2.3.1) (4.2.
0)
✓ Download and installation successful
You can now load the model via spacy.load('en_core_web_lg')
```

In [ ]: