L5: Information extraction

Information extraction (IE) is the task of identifying named entities and semantic relations between these entities in text data. In this lab we will focus on two sub-tasks in IE, **named entity recognition** (identifying mentions of entities) and **entity linking** (matching these mentions to entities in a knowledge base).

Students:

Weng Hang Wong (wenwo535)

Zuxiang Li(zuxli371)

We start by loading spaCy:

In [74]:
```
import spacy
from spacy.gold import GoldParse

nlp = spacy.load('en_core_web_sm')
```

The data that we will be using has been tokenized following the conventions of the Penn Treebank, and we need to prevent spaCy from using its own tokenizer on top of this. We therefore override spaCy's tokenizer with one that simply splits on space.

In [75]:
```
from spacy.tokens import Doc

class WhitespaceTokenizer(object):
    def __init__(self, vocab):
        self.vocab = vocab

    def __call__(self, text):
        return Doc(self.vocab, words=text.split(' '))

nlp.tokenizer = WhitespaceTokenizer(nlp.vocab)
```

Data set

The main data set for this lab is a collection of news wire articles in which mentions of named entities have been annotated with page names from the English Wikipedia. The next code cell loads the training and the development parts of the data into Pandas data frames.

In [76]:
```
import bz2
import csv
import pandas as pd

with bz2.open('ner-train.tsv.bz2', 'rt',encoding="utf-8") as source:
    df_train = pd.read_csv(source, sep='\t', quoting=csv.QUOTE_NONE,encoding

with bz2.open('ner-dev.tsv.bz2', 'rt',encoding="utf-8") as source:
    df_dev = pd.read_csv(source, sep='\t', quoting=csv.QUOTE_NONE,encoding=
```

Each row in these two data frames corresponds to one mention of a named entity and has five columns:

1. a unique identifier for the sentence containing the entity mention
2. the pre-tokenized sentence, with tokens separated by spaces
3. the start position of the token span containing the entity mention
4. the end position of the token span (exclusive, as in Python list indexing)
5. the entity label; either a Wikipedia page name or the generic label `--NME--`

The following cell prints the first five samples from the training data:

In [77]:
```
df_train.head()
```

Out[77]:

	sentence_id	sentence	beg	end	label
0	0000-000	EU rejects German call to boycott British lamb .	0	1	--NME--
1	0000-000	EU rejects German call to boycott British lamb .	2	3	Germany
2	0000-000	EU rejects German call to boycott British lamb .	6	7	United_Kingdom
3	0000-001	Peter Blackburn	0	2	--NME--
4	0000-002	BRUSSELS 1996-08-22	0	1	Brussels

In this sample, we see that the first sentence is annotated with three entity mentions:

- the span 0–1 ‘EU’ is annotated as a mention but only labelled with the generic `--NME--`
- the span 2–3 ‘German’ is annotated with the page Germany
- the span 6–7 ‘British’ is annotated with the page United_Kingdom

Problem 1: Evaluation measures

To warm up, we ask you to write code to print the three measures that you will be using for evaluation:

In [78]:
```
def evaluation_report(gold, pred):
    """Print precision, recall, and F1 score.

    Args:
        gold: The set with the gold-standard values.
        pred: The set with the predicted values.

    Returns:
        Nothing, but prints the precision, recall, and F1 values computed
        based on the specified sets.
    """
    # TODO: Replace the next line with your own code
    precision=len(gold.intersection(pred))/len(pred)*100
    recall=len(gold.intersection(pred))/len(gold)*100
    F1=2*(recall*precision)/(precision+recall)
    print(precision,recall,F1)
```

To test your code, you can run the following cell:

In [79]:
```
evaluation_report(set(range(3)), set(range(5)))
```

```
60.0 100.0 75.0
```

This should give you a precision of 60%, a recall of 100%, and an F1-value of 75%.

Problem 2: Span recognition

One of the first tasks that an information extraction system has to solve is to locate and classify (mentions of) named entities, such as persons and organizations. Here we will tackle the simpler task of recognizing **spans** of tokens that contain an entity mention, without the actual entity label.

The English language model in spaCy features a full-fledged named entity recognizer that identifies a variety of entities, and can be updated with new entity types by the user. Your task in this problem is to evaluate the performance of this component when predicting entity spans in the development data.

Start by implementing a generator function that yields the gold-standard spans in a given data frame.

Hint: The Pandas method `itertuples()` is useful when iterating over the rows in a DataFrame.

In [80]:
```
def gold_spans(df):
    """Yield the gold-standard mention spans in a data frame.

    Args:
        df: A data frame.

    Yields:
        The gold-standard mention spans in the specified data frame as
        triples consisting of the sentence id, start position, and end
        position of each span.
    """
    # TODO: Replace the next line with your own code
    #yield
    for x in df.itertuples():
        yield x[1],x[3],x[4]
```

To test your code, you can count the spans yielded by your function. When called on the development data, you should get a total of 5,917 unique triples. The first triple and the last triple should be

```
('0946-000', 2, 3)
('1161-010', 1, 3)
```

In [81]:
```
spans_dev_gold = set(gold_spans(df_dev))
print(len(spans_dev_gold))
```

```
5917
```

Your next task is to write code that calls spaCy to predict the named entities in the development data, and to evaluate the accuracy of these predictions in terms of precision, recall, and F1. Print these scores using the function that you wrote for Problem 1.

In [82]:
```
# TODO: Write code here to run and evalp;uate the spaCy NER on the developme
ent_pred=[]
for x in df_dev.itertuples(index=False,name=None):
    doc = nlp(x[1])
    for ent in doc.ents:
        #print(ent)
        ent_pred.append((x[0],ent.start,ent.end))
        #print(ent.text, ent.start_char, ent.end_char, ent.label_)
print(ent_pred)
```

```
 18), ('1152-016', 22, 24), ('1152-016', 25, 26), ('1152-016', 26, 27), ('11
52-016', 1, 2), ('1152-016', 2, 3), ('1152-016', 8, 9), ('1152-016', 11, 1
2), ('1152-016', 14, 15), ('1152-016', 16, 17), ('1152-016', 17, 18), ('1152
-016', 22, 24), ('1152-016', 25, 26), ('1152-016', 26, 27), ('1152-016', 1,
2), ('1152-016', 2, 3), ('1152-016', 8, 9), ('1152-016', 11, 12), ('1152-016
', 14, 15), ('1152-016', 16, 17), ('1152-016', 17, 18), ('1152-016', 22, 2
4), ('1152-016', 25, 26), ('1152-016', 26, 27), ('1152-016', 1, 2), ('1152-0
16', 2, 3), ('1152-016', 8, 9), ('1152-016', 11, 12), ('1152-016', 14, 15),
('1152-016', 16, 17), ('1152-016', 17, 18), ('1152-016', 22, 24), ('1152-016
', 25, 26), ('1152-016', 26, 27), ('1152-016', 1, 2), ('1152-016', 2, 3), ('
1152-016', 8, 9), ('1152-016', 11, 12), ('1152-016', 14, 15), ('1152-016', 1
6, 17), ('1152-016', 17, 18), ('1152-016', 22, 24), ('1152-016', 25, 26), ('
```

```
1152-016', 26, 27), ('1152-017', 3, 4), ('1152-017', 6, 7), ('1152-017', 9,
14), ('1152-017', 3, 4), ('1152-017', 6, 7), ('1152-017', 9, 14), ('1152-018
', 1, 3), ('1152-018', 9, 10), ('1152-018', 16, 17), ('1152-018', 18, 19),
('1152-018', 22, 23), ('1152-018', 27, 28), ('1152-018', 1, 3), ('1152-018',
9, 10), ('1152-018', 16, 17), ('1152-018', 18, 19), ('1152-018', 22, 23), ('
1152-018', 27, 28), ('1152-018', 1, 3), ('1152-018', 9, 10), ('1152-018', 1
6, 17), ('1152-018', 18, 19), ('1152-018', 22, 23), ('1152-018', 27, 28), ('
1153-000', 0, 1), ('1153-000', 2, 3), ('1153-001', 0, 1), ('1153-002', 0,
1), ('1153-002', 1, 2), ('1153-002', 6, 7), ('1153-002', 15, 16), ('1153-003
', 4, 5), ('1153-003', 6, 7), ('1153-003', 4, 5), ('1153-003', 6, 7), ('1153
-004', 9, 10), ('1153-006', 2, 3), ('1153-006', 5, 6), ('1153-006', 8, 10),
('1153-006', 11, 12), ('1153-006', 14, 16), ('1153-006', 17, 18), ('1153-006
', 36, 37), ('1153-006', 38, 39), ('1153-006', 2, 3), ('1153-006', 5, 6), ('
1153-006', 8, 10), ('1153-006', 11, 12), ('1153-006', 14, 16), ('1153-006',
17, 18), ('1153-006', 36, 37), ('1153-006', 38, 39), ('1153-006', 2, 3), ('1
153-006', 5, 6), ('1153-006', 8, 10), ('1153-006', 11, 12), ('1153-006', 14,
16), ('1153-006', 17, 18), ('1153-006', 36, 37), ('1153-006', 38, 39), ('115
3-007', 30, 31), ('1153-007', 32, 36), ('1153-007', 30, 31), ('1153-007', 3
2, 36), ('1154-000', 0, 1), ('1154-000', 4, 5), ('1154-001', 0, 1), ('1154-0
02', 8, 9), ('1154-002', 10, 11), ('1154-002', 12, 13), ('1154-002', 16, 1
7), ('1154-002', 20, 21), ('1154-002', 8, 9), ('1154-002', 10, 11), ('1154-0
02', 12, 13), ('1154-002', 16, 17), ('1154-002', 20, 21), ('1154-003', 0,
1), ('1154-003', 8, 9), ('1154-003', 13, 14), ('1154-003', 29, 30), ('1154-0
03', 0, 1), ('1154-003', 8, 9), ('1154-003', 13, 14), ('1154-003', 29, 30),
('1154-004', 1, 3), ('1154-004', 12, 13), ('1154-004', 16, 17), ('1154-005',
0, 1), ('1154-006', 0, 3), ('1154-006', 16, 18), ('1154-006', 19, 21), ('115
4-006', 22, 24), ('1154-006', 26, 27), ('1154-006', 0, 3), ('1154-006', 16,
18), ('1154-006', 19, 21), ('1154-006', 22, 24), ('1154-006', 26, 27), ('115
4-008', 0, 1), ('1154-008', 3, 8), ('1154-008', 9, 10), ('1154-008', 20, 2
1), ('1154-008', 22, 24), ('1154-008', 33, 34), ('1154-008', 36, 40), ('1154
-008', 0, 1), ('1154-008', 3, 8), ('1154-008', 9, 10), ('1154-008', 20, 21),
('1154-008', 22, 24), ('1154-008', 33, 34), ('1154-008', 36, 40), ('1154-008
', 0, 1), ('1154-008', 3, 8), ('1154-008', 9, 10), ('1154-008', 20, 21), ('1
154-008', 22, 24), ('1154-008', 33, 34), ('1154-008', 36, 40), ('1154-008',
0, 1), ('1154-008', 3, 8), ('1154-008', 9, 10), ('1154-008', 20, 21), ('1154
-008', 22, 24), ('1154-008', 33, 34), ('1154-008', 36, 40), ('1154-009', 0,
3), ('1154-009', 7, 8), ('1154-009', 11, 12), ('1154-009', 17, 19), ('1154-0
09', 29, 30), ('1154-009', 0, 3), ('1154-009', 7, 8), ('1154-009', 11, 12),
('1154-009', 17, 19), ('1154-009', 29, 30), ('1154-009', 0, 3), ('1154-009',
7, 8), ('1154-009', 11, 12), ('1154-009', 17, 19), ('1154-009', 29, 30), ('1
155-000', 0, 1), ('1155-000', 3, 4), ('1155-000', 8, 9), ('1155-000', 0, 1),
('1155-000', 3, 4), ('1155-000', 8, 9), ('1155-000', 0, 1), ('1155-000', 3,
4), ('1155-000', 8, 9), ('1155-001', 0, 1), ('1155-002', 0, 1), ('1155-002',
9, 10), ('1155-002', 13, 14), ('1155-002', 14, 17), ('1155-002', 24, 25), ('
1155-002', 30, 31), ('1155-002', 0, 1), ('1155-002', 9, 10), ('1155-002', 1
3, 14), ('1155-002', 14, 17), ('1155-002', 24, 25), ('1155-002', 30, 31), ('
1155-002', 0, 1), ('1155-002', 9, 10), ('1155-002', 13, 14), ('1155-002', 1
4, 17), ('1155-002', 24, 25), ('1155-002', 30, 31), ('1155-002', 0, 1), ('11
55-002', 9, 10), ('1155-002', 13, 14), ('1155-002', 14, 17), ('1155-002', 2
4, 25), ('1155-002', 30, 31), ('1155-003', 12, 13), ('1155-003', 14, 15), ('
1155-003', 18, 19), ('1155-003', 25, 27), ('1155-003', 29, 30), ('1155-003',
30, 31), ('1155-003', 12, 13), ('1155-003', 14, 15), ('1155-003', 18, 19),
('1155-003', 25, 27), ('1155-003', 29, 30), ('1155-003', 30, 31), ('1155-003
', 12, 13), ('1155-003', 14, 15), ('1155-003', 18, 19), ('1155-003', 25, 2
7), ('1155-003', 29, 30), ('1155-003', 30, 31), ('1155-003', 12, 13), ('1155
-003', 14, 15), ('1155-003', 18, 19), ('1155-003', 25, 27), ('1155-003', 29,
30), ('1155-003', 30, 31), ('1155-004', 0, 1), ('1155-005', 5, 6), ('1155-00
5', 14, 15), ('1155-005', 5, 6), ('1155-005', 14, 15), ('1155-006', 1, 2),
('1155-006', 10, 11), ('1155-006', 1, 2), ('1155-006', 10, 11), ('1155-007',
0, 1), ('1155-009', 3, 4), ('1155-009', 15, 16), ('1155-009', 19, 20), ('115
5-009', 21, 23), ('1155-009', 3, 4), ('1155-009', 15, 16), ('1155-009', 19,
20), ('1155-009', 21, 23), ('1155-009', 3, 4), ('1155-009', 15, 16), ('1155-
```

```
009', 19, 20), ('1155-009', 21, 23), ('1155-010', 7, 8), ('1155-010', 17, 1
8), ('1155-010', 25, 26), ('1155-010', 27, 28), ('1155-010', 31, 32), ('1155
-010', 34, 35), ('1155-010', 7, 8), ('1155-010', 17, 18), ('1155-010', 25, 2
6), ('1155-010', 27, 28), ('1155-010', 31, 32), ('1155-010', 34, 35), ('1155
-010', 7, 8), ('1155-010', 17, 18), ('1155-010', 25, 26), ('1155-010', 27, 2
8), ('1155-010', 31, 32), ('1155-010', 34, 35), ('1155-010', 7, 8), ('1155-0
10', 17, 18), ('1155-010', 25, 26), ('1155-010', 27, 28), ('1155-010', 31, 3
2), ('1155-010', 34, 35), ('1155-010', 7, 8), ('1155-010', 17, 18), ('1155-0
10', 25, 26), ('1155-010', 27, 28), ('1155-010', 31, 32), ('1155-010', 34, 3
5), ('1155-011', 0, 1), ('1155-011', 3, 4), ('1155-011', 6, 7), ('1155-011',
15, 16), ('1155-011', 16, 18), ('1155-011', 0, 1), ('1155-011', 3, 4), ('115
5-011', 6, 7), ('1155-011', 15, 16), ('1155-011', 16, 18), ('1155-011', 0,
1), ('1155-011', 3, 4), ('1155-011', 6, 7), ('1155-011', 15, 16), ('1155-011
', 16, 18), ('1155-011', 0, 1), ('1155-011', 3, 4), ('1155-011', 6, 7), ('11
55-011', 15, 16), ('1155-011', 16, 18), ('1155-012', 7, 8), ('1155-012', 12,
13), ('1155-012', 16, 17), ('1155-012', 7, 8), ('1155-012', 12, 13), ('1155-
012', 16, 17), ('1155-012', 7, 8), ('1155-012', 12, 13), ('1155-012', 16, 1
7), ('1155-013', 1, 2), ('1155-014', 0, 1), ('1155-014', 5, 6), ('1155-014',
9, 11), ('1155-014', 17, 18), ('1155-014', 0, 1), ('1155-014', 5, 6), ('1155
-014', 9, 11), ('1155-014', 17, 18), ('1155-014', 0, 1), ('1155-014', 5, 6),
('1155-014', 9, 11), ('1155-014', 17, 18), ('1155-014', 0, 1), ('1155-014',
5, 6), ('1155-014', 9, 11), ('1155-014', 17, 18), ('1155-015', 0, 1), ('1155
-015', 8, 9), ('1155-015', 12, 13), ('1155-015', 0, 1), ('1155-015', 8, 9),
('1155-015', 12, 13), ('1155-015', 0, 1), ('1155-015', 8, 9), ('1155-015', 1
2, 13), ('1156-000', 0, 1), ('1156-000', 2, 3), ('1156-000', 4, 5), ('1156-0
00', 0, 1), ('1156-000', 2, 3), ('1156-000', 4, 5), ('1156-000', 0, 1), ('11
56-000', 2, 3), ('1156-000', 4, 5), ('1156-000', 0, 1), ('1156-000', 2, 3),
('1156-000', 4, 5), ('1156-002', 0, 1), ('1156-002', 3, 5), ('1156-002', 6,
7), ('1156-002', 9, 10), ('1156-002', 11, 13), ('1156-002', 18, 19), ('1156-
002', 0, 1), ('1156-002', 3, 5), ('1156-002', 6, 7), ('1156-002', 9, 10), ('
1156-002', 11, 13), ('1156-002', 18, 19), ('1156-002', 0, 1), ('1156-002',
3, 5), ('1156-002', 6, 7), ('1156-002', 9, 10), ('1156-002', 11, 13), ('1156
-002', 18, 19), ('1156-002', 0, 1), ('1156-002', 3, 5), ('1156-002', 6, 7),
('1156-002', 9, 10), ('1156-002', 11, 13), ('1156-002', 18, 19), ('1156-002
', 0, 1), ('1156-002', 3, 5), ('1156-002', 6, 7), ('1156-002', 9, 10), ('115
6-002', 11, 13), ('1156-002', 18, 19), ('1156-003', 3, 4), ('1156-003', 4,
6), ('1156-003', 15, 16), ('1156-003', 19, 20), ('1156-003', 21, 23), ('1156
-003', 25, 30), ('1156-003', 32, 33), ('1156-003', 33, 35), ('1156-003', 40,
41), ('1156-003', 3, 4), ('1156-003', 4, 6), ('1156-003', 15, 16), ('1156-00
3', 19, 20), ('1156-003', 21, 23), ('1156-003', 25, 30), ('1156-003', 32, 3
3), ('1156-003', 33, 35), ('1156-003', 40, 41), ('1156-003', 3, 4), ('1156-0
03', 4, 6), ('1156-003', 15, 16), ('1156-003', 19, 20), ('1156-003', 21, 2
3), ('1156-003', 25, 30), ('1156-003', 32, 33), ('1156-003', 33, 35), ('1156
-003', 40, 41), ('1156-003', 3, 4), ('1156-003', 4, 6), ('1156-003', 15, 1
6), ('1156-003', 19, 20), ('1156-003', 21, 23), ('1156-003', 25, 30), ('1156
-003', 32, 33), ('1156-003', 33, 35), ('1156-003', 40, 41), ('1156-003', 3,
4), ('1156-003', 4, 6), ('1156-003', 15, 16), ('1156-003', 19, 20), ('1156-0
03', 21, 23), ('1156-003', 25, 30), ('1156-003', 32, 33), ('1156-003', 33, 3
5), ('1156-003', 40, 41), ('1156-003', 3, 4), ('1156-003', 4, 6), ('1156-003
', 15, 16), ('1156-003', 19, 20), ('1156-003', 21, 23), ('1156-003', 25, 3
0), ('1156-003', 32, 33), ('1156-003', 33, 35), ('1156-003', 40, 41), ('1156
-003', 3, 4), ('1156-003', 4, 6), ('1156-003', 15, 16), ('1156-003', 19, 2
0), ('1156-003', 21, 23), ('1156-003', 25, 30), ('1156-003', 32, 33), ('1156
-003', 33, 35), ('1156-003', 40, 41), ('1156-003', 3, 4), ('1156-003', 4,
6), ('1156-003', 15, 16), ('1156-003', 19, 20), ('1156-003', 21, 23), ('1156
-003', 25, 30), ('1156-003', 32, 33), ('1156-003', 33, 35), ('1156-003', 40,
41), ('1156-003', 3, 4), ('1156-003', 4, 6), ('1156-003', 15, 16), ('1156-00
3', 19, 20), ('1156-003', 21, 23), ('1156-003', 25, 30), ('1156-003', 32, 3
3), ('1156-003', 33, 35), ('1156-003', 40, 41), ('1156-004', 6, 7), ('1156-0
04', 14, 16), ('1156-004', 23, 24), ('1156-004', 30, 32), ('1156-004', 6,
7), ('1156-004', 14, 16), ('1156-004', 23, 24), ('1156-004', 30, 32), ('1156
-004', 6, 7), ('1156-004', 14, 16), ('1156-004', 23, 24), ('1156-004', 30, 3
```

```
2), ('1156-004', 6, 7), ('1156-004', 14, 16), ('1156-004', 23, 24), ('1156-0
04', 30, 32), ('1156-006', 9, 12), ('1156-006', 15, 21), ('1156-006', 22, 2
3), ('1156-006', 9, 12), ('1156-006', 15, 21), ('1156-006', 22, 23), ('1156-
006', 9, 12), ('1156-006', 15, 21), ('1156-006', 22, 23), ('1156-007', 2,
3), ('1156-007', 4, 5), ('1156-008', 0, 4), ('1156-008', 5, 6), ('1156-008',
11, 12), ('1156-008', 13, 15), ('1156-008', 18, 19), ('1156-008', 28, 29),
('1156-008', 0, 4), ('1156-008', 5, 6), ('1156-008', 11, 12), ('1156-008', 1
3, 15), ('1156-008', 18, 19), ('1156-008', 28, 29), ('1156-008', 0, 4), ('11
56-008', 5, 6), ('1156-008', 11, 12), ('1156-008', 13, 15), ('1156-008', 18,
19), ('1156-008', 28, 29), ('1156-008', 0, 4), ('1156-008', 5, 6), ('1156-00
8', 11, 12), ('1156-008', 13, 15), ('1156-008', 18, 19), ('1156-008', 28, 2
9), ('1156-009', 2, 3), ('1156-009', 11, 12), ('1156-009', 18, 19), ('1156-0
09', 2, 3), ('1156-009', 11, 12), ('1156-009', 18, 19), ('1156-009', 2, 3),
('1156-009', 11, 12), ('1156-009', 18, 19), ('1156-010', 0, 2), ('1156-010',
10, 11), ('1156-010', 12, 13), ('1156-010', 17, 18), ('1156-010', 20, 21),
('1156-010', 0, 2), ('1156-010', 10, 11), ('1156-010', 12, 13), ('1156-010',
17, 18), ('1156-010', 20, 21), ('1157-000', 0, 1), ('1157-000', 6, 7), ('115
7-001', 0, 1), ('1157-002', 0, 6), ('1157-002', 8, 9), ('1157-002', 19, 20),
('1157-002', 0, 6), ('1157-002', 8, 9), ('1157-002', 19, 20), ('1157-003',
2, 3), ('1157-003', 5, 6), ('1157-003', 2, 3), ('1157-003', 5, 6), ('1157-00
4', 2, 3), ('1157-004', 5, 7), ('1157-004', 10, 11), ('1157-004', 13, 16),
('1157-004', 17, 18), ('1157-004', 25, 26), ('1157-004', 28, 29), ('1157-004
', 2, 3), ('1157-004', 5, 7), ('1157-004', 10, 11), ('1157-004', 13, 16), ('
1157-004', 17, 18), ('1157-004', 25, 26), ('1157-004', 28, 29), ('1157-004',
2, 3), ('1157-004', 5, 7), ('1157-004', 10, 11), ('1157-004', 13, 16), ('115
7-004', 17, 18), ('1157-004', 25, 26), ('1157-004', 28, 29), ('1157-004', 2,
3), ('1157-004', 5, 7), ('1157-004', 10, 11), ('1157-004', 13, 16), ('1157-0
04', 17, 18), ('1157-004', 25, 26), ('1157-004', 28, 29), ('1157-004', 2,
3), ('1157-004', 5, 7), ('1157-004', 10, 11), ('1157-004', 13, 16), ('1157-0
04', 17, 18), ('1157-004', 25, 26), ('1157-004', 28, 29), ('1157-004', 2,
3), ('1157-004', 5, 7), ('1157-004', 10, 11), ('1157-004', 13, 16), ('1157-0
04', 17, 18), ('1157-004', 25, 26), ('1157-004', 28, 29), ('1158-000', 5,
6), ('1158-001', 0, 1), ('1158-002', 0, 3), ('1158-002', 7, 8), ('1158-002',
17, 18), ('1158-002', 22, 23), ('1158-002', 0, 3), ('1158-002', 7, 8), ('115
8-002', 17, 18), ('1158-002', 22, 23), ('1158-003', 5, 6), ('1158-003', 8,
9), ('1158-003', 10, 11), ('1158-003', 26, 27), ('1158-003', 27, 29), ('1158
-003', 5, 6), ('1158-003', 8, 9), ('1158-003', 10, 11), ('1158-003', 26, 2
7), ('1158-003', 27, 29), ('1158-003', 5, 6), ('1158-003', 8, 9), ('1158-003
', 10, 11), ('1158-003', 26, 27), ('1158-003', 27, 29), ('1158-003', 5, 6),
('1158-003', 8, 9), ('1158-003', 10, 11), ('1158-003', 26, 27), ('1158-003',
27, 29), ('1158-005', 1, 2), ('1158-005', 5, 7), ('1158-005', 14, 15), ('115
8-005', 17, 22), ('1158-005', 1, 2), ('1158-005', 5, 7), ('1158-005', 14, 1
5), ('1158-005', 17, 22), ('1158-005', 1, 2), ('1158-005', 5, 7), ('1158-005
', 14, 15), ('1158-005', 17, 22), ('1158-005', 1, 2), ('1158-005', 5, 7), ('
1158-005', 14, 15), ('1158-005', 17, 22), ('1159-000', 3, 4), ('1159-000',
7, 8), ('1159-001', 0, 1), ('1159-002', 0, 1), ('1159-002', 4, 5), ('1159-00
2', 11, 12), ('1159-002', 20, 23), ('1159-003', 15, 16), ('1159-003', 19, 2
0), ('1159-003', 15, 16), ('1159-003', 19, 20), ('1159-004', 2, 3), ('1159-0
04', 4, 5), ('1159-004', 12, 13), ('1159-004', 14, 16), ('1159-004', 17, 1
8), ('1159-004', 20, 22), ('1159-004', 23, 24), ('1159-004', 33, 34), ('1159
-004', 35, 36), ('1159-004', 38, 39), ('1159-004', 2, 3), ('1159-004', 4,
5), ('1159-004', 12, 13), ('1159-004', 14, 16), ('1159-004', 17, 18), ('1159
-004', 20, 22), ('1159-004', 23, 24), ('1159-004', 33, 34), ('1159-004', 35,
36), ('1159-004', 38, 39), ('1159-004', 2, 3), ('1159-004', 4, 5), ('1159-00
4', 12, 13), ('1159-004', 14, 16), ('1159-004', 17, 18), ('1159-004', 20, 2
2), ('1159-004', 23, 24), ('1159-004', 33, 34), ('1159-004', 35, 36), ('1159
-004', 38, 39), ('1159-004', 2, 3), ('1159-004', 4, 5), ('1159-004', 12, 1
3), ('1159-004', 14, 16), ('1159-004', 17, 18), ('1159-004', 20, 22), ('1159
-004', 23, 24), ('1159-004', 33, 34), ('1159-004', 35, 36), ('1159-004', 38,
39), ('1159-004', 2, 3), ('1159-004', 4, 5), ('1159-004', 12, 13), ('1159-00
4', 14, 16), ('1159-004', 17, 18), ('1159-004', 20, 22), ('1159-004', 23, 2
4), ('1159-004', 33, 34), ('1159-004', 35, 36), ('1159-004', 38, 39), ('1159
```

-007', 28, 29), ('1159-007', 30, 34), ('1159-007', 28, 29), ('1159-007', 30, 34), ('1160-000', 5, 6), ('1160-000', 5, 6), ('1160-001', 0, 2), ('1160-002', 0, 1), ('1160-002', 2, 3), ('1160-003', 2, 4), ('1160-003', 16, 18), ('1160-003', 25, 26), ('1160-003', 31, 32), ('1160-003', 33, 37), ('1160-003', 2, 4), ('1160-003', 16, 18), ('1160-003', 25, 26), ('1160-003', 31, 32), ('1160-003', 33, 37), ('1160-003', 2, 4), ('1160-003', 16, 18), ('1160-003', 25, 26), ('1160-003', 31, 32), ('1160-003', 33, 37), ('1160-003', 2, 4), ('1160-003', 16, 18), ('1160-003', 25, 26), ('1160-003', 31, 32), ('1160-003', 33, 37), ('1160-004', 4, 6), ('1160-004', 7, 9), ('1160-004', 14, 16), ('1160-004', 17, 22), ('1160-004', 28, 29), ('1160-004', 31, 34), ('1160-004', 4, 6), ('1160-004', 7, 9), ('1160-004', 14, 16), ('1160-004', 17, 22), ('1160-004', 28, 29), ('1160-004', 31, 34), ('1160-004', 4, 6), ('1160-004', 7, 9), ('1160-004', 14, 16), ('1160-004', 17, 22), ('1160-004', 28, 29), ('1160-004', 31, 34), ('1160-004', 4, 6), ('1160-004', 7, 9), ('1160-004', 14, 16), ('1160-004', 17, 22), ('1160-004', 28, 29), ('1160-004', 31, 34), ('1160-004', 4, 6), ('1160-004', 7, 9), ('1160-004', 14, 16), ('1160-004', 17, 22), ('1160-004', 28, 29), ('1160-004', 31, 34), ('1160-005', 6, 7), ('1160-005', 8, 11), ('1160-005', 26, 28), ('1160-005', 6, 7), ('1160-005', 8, 11), ('1160-005', 26, 28), ('1160-005', 6, 7), ('1160-005', 8, 11), ('1160-005', 26, 28), ('1160-006', 0, 1), ('1160-006', 15, 16), ('1160-007', 6, 7), ('1160-007', 9, 10), ('1160-007', 21, 22), ('1160-007', 28, 30), ('1160-007', 6, 7), ('1160-007', 9, 10), ('1160-007', 21, 22), ('1160-007', 28, 30), ('1160-007', 6, 7), ('1160-007', 9, 10), ('1160-007', 21, 22), ('1160-007', 28, 30), ('1160-007', 6, 7), ('1160-007', 9, 10), ('1160-007', 21, 22), ('1160-007', 28, 30), ('1160-010', 0, 1), ('1160-010', 3, 4), ('1160-010', 5, 10), ('1160-010', 27, 28), ('1160-010', 0, 1), ('1160-010', 3, 4), ('1160-010', 5, 10), ('1160-010', 27, 28), ('1160-010', 0, 1), ('1160-010', 3, 4), ('1160-010', 5, 10), ('1160-010', 27, 28), ('1160-010', 0, 1), ('1160-010', 3, 4), ('1160-010', 5, 10), ('1160-010', 27, 28), ('1160-011', 17, 19), ('1160-011', 20, 21), ('1160-011', 22, 24), ('1160-011', 25, 26), ('1160-011', 28, 29), ('1160-011', 17, 19), ('1160-011', 20, 21), ('1160-011', 22, 24), ('1160-011', 25, 26), ('1160-011', 28, 29), ('1160-011', 17, 19), ('1160-011', 20, 21), ('1160-011', 22, 24), ('1160-011', 25, 26), ('1160-011', 28, 29), ('1160-011', 17, 19), ('1160-011', 20, 21), ('1160-011', 22, 24), ('1160-011', 25, 26), ('1160-011', 28, 29), ('1160-011', 17, 19), ('1160-011', 20, 21), ('1160-011', 22, 24), ('1160-011', 25, 26), ('1160-011', 28, 29), ('1160-012', 1, 2), ('1160-012', 12, 13), ('1160-014', 0, 1), ('1160-014', 3, 4), ('1160-014', 11, 12), ('1160-014', 20, 24), ('1160-014', 30, 31), ('1160-014', 0, 1), ('1160-014', 3, 4), ('1160-014', 11, 12), ('1160-014', 20, 24), ('1160-014', 30, 31), ('1160-014', 0, 1), ('1160-014', 3, 4), ('1160-014', 11, 12), ('1160-014', 20, 24), ('1160-014', 30, 31), ('1160-014', 0, 1), ('1160-014', 3, 4), ('1160-014', 11, 12), ('1160-014', 20, 24), ('1160-014', 30, 31), ('1160-014', 0, 1), ('1160-014', 3, 4), ('1160-014', 11, 12), ('1160-014', 20, 24), ('1160-014', 30, 31), ('1160-015', 0, 2), ('1160-015', 3, 5), ('1160-015', 6, 7), ('1160-015', 17, 19), ('1160-015', 22, 23), ('1160-015', 0, 2), ('1160-015', 3, 5), ('1160-015', 6, 7), ('1160-015', 17, 19), ('1160-015', 22, 23), ('1160-015', 0, 2), ('1160-015', 3, 5), ('1160-015', 6, 7), ('1160-015', 17, 19), ('1160-015', 22, 23), ('1160-015', 0, 2), ('1160-015', 3, 5), ('1160-015', 6, 7), ('1160-015', 17, 19), ('1160-015', 22, 23), ('1160-015', 0, 2), ('1160-015', 3, 5), ('1160-015', 6, 7), ('1160-015', 17, 19), ('1160-015', 22, 23), ('1160-016', 9, 10), ('1160-016', 11, 12), ('1160-016', 13, 14), ('1160-016', 9, 10), ('1160-016', 11, 12), ('1160-016', 13, 14), ('1160-018', 4, 5), ('1160-018', 16, 17), ('1160-018', 4, 5), ('1160-018', 16, 17), ('1160-022', 0, 1), ('1160-022', 16, 17), ('1160-022', 31, 32), ('1160-022', 37, 38), ('1160-022', 0, 1), ('1160-022', 16, 17), ('1160-022', 31, 32), ('1160-022', 37, 38), ('1160-023', 18, 20), ('1160-023', 33, 34), ('1160-023', 18, 20), ('1160-023', 33, 34), ('1160-024', 5, 6), ('1160-024', 19, 20), ('1160-024', 23, 24), ('1161-000', 0, 1), ('1161-002', 2, 6), ('1161-002', 37, 38), ('1161-002', 38, 40), ('1161-002', 2, 6), ('1161-002', 37, 38), ('1161-002', 38, 40), ('1161-002', 2, 6), ('1161-002', 37, 38), ('1161-002', 38, 40), ('1161-004', 12, 14), ('1161-005', 0, 3), ('1161-007', 7, 9), ('1161-007', 10, 12), ('1161-007', 13, 15), ('1161-007', 7, 9), ('1161-007', 10, 12), ('1161-

```
007'  13  15)  ('1161-007'  7  9)  ('1161-007'  10  12)  ('1161-007'  13  1
```

In [83]:

```
spans_dev_pred=set(ent_pred)
print(len(spans_dev_pred))
evaluation_report(spans_dev_gold,spans_dev_pred)
```

```
8034
51.53099327856609 69.9678891330066 59.35058418751344
```

Problem 3: Error analysis

As you were able to see in Problem 2, the span accuracy of the named entity recognizer is far from perfect. In particular, only slightly more than half of the predicted spans are correct according to the gold standard. Your next task is to analyse this result in more detail.

Here is a function that prints the false positives as well as the false negatives spans for a data frame, given a reference set of gold-standard spans and a candidate set of predicted spans.

In [84]:

```
from collections import defaultdict

def error_report(df, spans_gold, spans_pred):
    false_pos = defaultdict(list)
    for s, b, e in spans_pred - spans_gold:
        false_pos[s].append((b, e))
    false_neg = defaultdict(list)
    for s, b, e in spans_gold - spans_pred:
        false_neg[s].append((b, e))
    for row in df.drop_duplicates('sentence_id').itertuples():
        if row.sentence_id in false_pos or row.sentence_id in false_neg:
            print('Sentence:', row.sentence)
            for b, e in false_pos[row.sentence_id]:
                print('  FP:', ' '.join(row.sentence.split()[b:e]))
            for b, e in false_neg[row.sentence_id]:
                print('  FN:', ' '.join(row.sentence.split()[b:e]))
```

Use this function to inspect and analyse the errors that the automated prediction makes. Can you see any patterns? Base your analysis on the first 500 rows of the training data. Summarize your observations in a short text.

In [85]:

```
# TODO: Write code here to do your analysis
ent_pred=[]
print(1111)
for x in df_train.iloc[:500].itertuples():
    doc = nlp(x[2])
    for ent in doc.ents:
        ent_pred.append((x[1],ent.start,ent.end))
spans_train_pred=set(ent_pred)
error_report(df_train.iloc[:500],set(gold_spans(df_train)),spans_train_pred)
```

```
tified alarm through " dangerous generalisation . "
  FP: Spanish Farm
  FN: Spanish
Sentence: The EU 's scientific veterinary and multidisciplinary committees a
re due to re-examine the issue early next month and make recommendations to
```

```
the senior veterinary officials .
  FP: early next month
Sentence: British farmers denied on Thursday there was any danger to human h
ealth from their sheep , but expressed concern that German government advice
to consumers to avoid British lamb might influence consumers across Europe .
  FP: Thursday
Sentence: " What we have to be extremely careful of is how other countries a
re going to take Germany 's lead , " Welsh National Farmers ' Union ( NFU )
chairman John Lloyd Jones said on BBC radio .
  FP: BBC
  FN: BBC radio
Sentence: Bonn has led efforts to protect public health after consumer confi
dence collapsed in March after a British report suggested humans could contr
act an illness similar to mad cow disease by eating contaminated beef .
  FP: March
Sentence: Germany imported 47,600 sheep from Britain last year , nearly half
of total imports .
  FP: 47,600
  FP: last year
  FP: nearly half
Sentence: It brought in 4,275 tonnes of British mutton , some 10 percent of
overall imports .
  FP: some 10 percent
  FP: 4,275
Sentence: Rare Hendrix song draft sells for almost $ 17,000 .
  FP: Rare Hendrix
  FP: almost $ 17,000
  FN: Hendrix
Sentence: A rare early handwritten draft of a song by U.S. guitar legend Jim
i Hendrix was sold for almost $ 17,000 on Thursday at an auction of some of
the late musician 's favourite possessions .
  FP: Thursday
  FP: almost $ 17,000
Sentence: A Florida restaurant paid 10,925 pounds ( $ 16,935 ) for the draft
of " Ai n't no telling " , which Hendrix penned on a piece of London hotel s
tationery in late 1966 .
  FP: 16,935
  FP: late 1966
  FP: 10,925 pounds
  FN: Ai n't no telling
Sentence: At the end of a January 1967 concert in the English city of Nottin
gham he threw the sheet of paper into the audience , where it was retrieved
by a fan .
  FP: the end of a January 1967
Sentence: Buyers also snapped up 16 other items that were put up for auction
by Hendrix 's former girlfriend Kathy Etchingham , who lived with him from 1
966 to 1969 .
  FP: 16
  FP: 1969
  FP: 1966
Sentence: They included a black lacquer and mother of pearl inlaid box used
by Hendrix to store his drugs , which an anonymous Australian purchaser boug
ht for 5,060 pounds ( $ 7,845 ) .
  FP: 7,845
  FP: 5,060 pounds
Sentence: China on Thursday accused Taipei of spoiling the atmosphere for a
resumption of talks across the Taiwan Strait with a visit to Ukraine by Taiw
anese Vice President Lien Chan this week that infuriated Beijing .
  FP: Thursday
  FP: the Taiwan Strait
  FP: this week
  FN: Taiwan Strait
```

```
Sentence: Speaking only hours after Chinese state media said the time was ri
ght to engage in political talks with Taiwan , Foreign Ministry spokesman Sh
en Guofang told Reuters : " The necessary atmosphere for the opening of the
talks has been disrupted by the Taiwan authorities . "
  FP: only hours
Sentence: State media quoted China 's top negotiator with Taipei , Tang Shub
ei , as telling a visiting group from Taiwan on Wednesday that it was time f
or the rivals to hold political talks .
  FP: Wednesday
Sentence: that is to end the state of hostility , " Thursday 's overseas edi
tion of the People 's Daily quoted Tang as saying .
  FP: Thursday
  FP: the People 's Daily
  FN: People 's Daily
Sentence: The foreign ministry 's Shen told Reuters Television in an intervi
ew he had read reports of Tang 's comments but gave no details of why the ne
gotiator had considered the time right for talks with Taiwan , which Beijing
considers a renegade province .
  FN: Shen
Sentence: China , which has long opposed all Taipei efforts to gain greater
international recognition , was infuriated by a visit to Ukraine this week b
y Taiwanese Vice President Lien .
  FP: this week
Sentence: Consultations should be held to set the time and format of the tal
ks , the official Xinhua news agency quoted Tang Shubei , executive vice cha
irman of the Association for Relations Across the Taiwan Straits , as saying
late on Wednesday .
  FP: Xinhua news agency
  FP: the Association for Relations Across
  FP: Wednesday
  FN: Xinhua
  FN: Association for Relations Across the Taiwan Straits
Sentence: German July car registrations up 14.2 pct yr / yr .
  FP: 14.2
  FP: German July
  FN: German
Sentence: German first-time registrations of motor vehicles jumped 14.2 perc
ent in July this year from the year-earlier period , the Federal office for
motor vehicles said on Thursday .
  FP: July this year
  FP: 14.2 percent
  FP: Thursday
  FP: Federal
  FN: Federal office for motor vehicles
Sentence: Volkswagen AG won 77,719 registrations , slightly more than a quar
ter of the total .
  FP: 77,719
  FP: more than a quarter
Sentence: Opel AG together with General Motors came in second place with 49,
269 registrations , 16.4 percent of the overall figure .
  FP: second
  FP: 16.4 percent
  FP: 49,269
Sentence: Third was Ford with 35,563 registrations , or 11.7 percent .
  FP: Third
  FP: 11.7 percent
  FP: 35,563
Sentence: Only Seat and Porsche had fewer registrations in July 1996 compare
d to last year 's July .
  FP: last year 's July
  FP: July 1996
  FN: Seat
```

```
Sentence: Seat posted 3,420 registrations compared with 5522 registrations i
n July a year earlier .
  FP: 3,420
  FP: 5522
  FP: July
  FP: a year earlier
  FN: Seat
Sentence: Porsche 's registrations fell to 554 from 643 .
  FP: 643
  FP: 554
Sentence: GREEK SOCIALISTS GIVE GREEN LIGHT TO PM FOR ELECTIONS .
  FP: GREEK SOCIALISTS
  FN: GREEK
Sentence: ATHENS 1996-08-22
  FN: ATHENS
Sentence: Prime Minister Costas Simitis is going to make an official announc
ement after a cabinet meeting later on Thursday , said Skandalidis .
  FP: Thursday
Sentence: -- Dimitris Kontogiannis , Athens Newsroom +301 3311812-4
  FP: Athens Newsroom +301 3311812-4
  FP: -- Dimitris Kontogiannis
  FN: Dimitris Kontogiannis
  FN: Athens Newsroom
Sentence: BayerVB sets C$ 100 million six-year bond .
  FP: C$ 100 million
  FN: BayerVB
  FN: C$
Sentence: BORROWER BAYERISCHE VEREINSBANK
  FN: BAYERISCHE VEREINSBANK
Sentence: AMT C$ 100 MLN COUPON 6.625 MATURITY 24.SEP.02
  FP: 6.625
  FN: C$
Sentence: S&P = DENOMS ( K ) 1-10-100 SALE LIMITS US / UK / CA
  FP: UK / CA
  FP: 1-10-100
  FN: CA
  FN: UK
  FN: S&P
Sentence: GOV LAW GERMAN HOME CTRY = TAX PROVS STANDARD
  FN: GERMAN
Sentence: NOTES BAYERISCHE VEREINSBANK IS JOINT LEAD MANAGER
  FN: BAYERISCHE VEREINSBANK
Sentence: -- London Newsroom +44 171 542 7658
  FP: -- London Newsroom +44
  FN: London Newsroom
Sentence: Venantius sets $ 300 million January 1999 FRN .
  FP: $ 300 million
  FP: FRN
  FP: January 1999
  FN: Venantius
Sentence: BORROWER VENANTIUS AB ( SWEDISH NATIONAL MORTGAGE AGENCY )
  FP: SWEDISH NATIONAL MORTGAGE AGENCY
  FP: BORROWER VENANTIUS AB
  FN: SWEDISH
  FN: VENANTIUS AB
Sentence: TYPE FRN BASE 3M LIBOR PAY DATE S23.SEP.96
  FN: 3M
Sentence: LAST S&P AA+ REOFFER =
  FN: S&P
Sentence: LISTING LONDON DENOMS ( K ) 1-10-100 SALE LIMITS US / UK / JP / FR
  FP: 1-10-100
  FN: FR
```

```
  FN: JP
Sentence: GOV LAW ENGLISH HOME CTRY SWEDEN TAX PROVS STANDARD
  FN: SWEDEN
Sentence: -- London Newsroom +44 171 542 8863
  FP: 542
  FP: -- London Newsroom +44
  FP: 171
  FN: London Newsroom
Sentence: Port conditions update - Syria - Lloyds Shipping .
  FN: Lloyds Shipping
Sentence: LATTAKIA , Aug 10 - waiting time at Lattakia and Tartous presently
24 hours .
  FP: 24 hours
  FP: 10
Sentence: Israel 's outgoing peace negotiator with Syria said on Thursday cu
rrent tensions between the two countries appeared to be a storm in a teacup
.
  FP: Thursday
  FP: two
Sentence: Rabinovich is winding up his term as ambassador .
  FN: Rabinovich
Sentence: Israel on Wednesday sent Syria a message , via Washington , saying
it was committed to peace and wanted to open negotiations without preconditi
ons .
  FP: Wednesday
Sentence: Syria accused Israel on Wednesday of launching a hysterical campai
gn against it after Israeli television reported that Damascus had recently t
est fired a missile .
  FP: Wednesday
Sentence: Tension has mounted since Israeli Prime Minister Benjamin Netanyah
u took office in June vowing to retain the Golan Heights Israel captured fro
m Syria in the 1967 Middle East war .
  FP: June
  FP: 1967
Sentence: Israeli-Syrian peace talks have been deadlocked over the Golan sin
ce 1991 despite the previous government 's willingness to make Golan concess
ions .
  FP: 1991
Sentence: We do not want a war , God forbid .
  FN: God
Sentence: The television also said that Netanyahu had sent messages to reass
ure Syria via Cairo , the United States and Moscow .
  FP: the United States
  FN: United States
Sentence: TUNIS 1996-08-22
  FN: TUNIS
Sentence: A Polish diplomat on Thursday denied a Polish tabloid report this
week that Libya was refusing exit visas to 100 Polish nurses trying to retur
n home after working in the North African country .
  FP: 100
  FP: Thursday
  FP: this week
Sentence: Up to today , we have no knowledge of any nurse stranded or kept i
n Libya without her will , and we have not received any complaint , " the Po
lish embassy 's charge d'affaires in Tripoli , Tadeusz Awdankiewicz , told R
euters by telephone .
  FP: today
Sentence: Poland 's labour ministry said this week it would send a team to L
ibya to investigate , but Awdankiewicz said the probe was prompted by some n
urses complaining about their work conditions such as non-payment of their s
alaries .
  FP: this week
```

```
Sentence: He said that there are an estimated 800 Polish nurses working in L
ibya .
  FP: an estimated 800
Sentence: Two Iranian opposition leaders meet in Baghdad .
  FP: Two
Sentence: An Iranian exile group based in Iraq vowed on Thursday to extend s
upport to Iran 's Kurdish rebels after they were attacked by Iranian troops
deep inside Iraq last month .
  FP: Thursday
  FP: last month
Sentence: A Mujahideen Khalq statement said its leader Massoud Rajavi met in
Baghdad the Secretary-General of the Kurdistan Democratic Party of Iran ( KD
PI ) Hassan Rastegar on Wednesday and voiced his support to Iran 's rebel Ku
rds .
  FP: the Kurdistan Democratic Party
  FP: Wednesday
  FN: Kurdistan Democratic Party of Iran
  FN: KDPI
Sentence: " Rajavi emphasised that the Iranian Resistance would continue to
stand side by side with their Kurdish compatriots and the resistance movemen
t in Iranian Kurdistan , " it said .
  FP: Kurdistan
  FP: Iranian
  FP: the Iranian Resistance
  FN: Resistance
  FN: Iranian
  FN: Iranian Kurdistan
Sentence: A spokesman for the group said the meeting " signals a new level o
f cooperation between Mujahideen Khalq and the Iranian Kurdish oppositions "
.
  FP: Iranian
  FP: Kurdish
  FN: Iranian Kurdish
Sentence: Iran heavily bombarded targets in northern Iraq in July in pursuit
of KDPI guerrillas based in Iraqi Kurdish areas outside the control of the g
overnment in Baghdad .
  FP: July
  FP: Kurdish
  FP: Iraqi
  FN: Iraqi Kurdish
Sentence: Iraqi Kurdish areas bordering Iran are under the control of guerri
llas of the Iraqi Kurdish Patriotic Union of Kurdistan ( PUK ) group .
  FP: Kurdish
  FP: the Iraqi Kurdish Patriotic Union of Kurdistan
  FP: Iraqi
  FN: Iraqi Kurdish
  FN: Iraqi Kurdish Patriotic Union of Kurdistan
Sentence: PUK and Iraq 's Kurdistan Democratic Party ( KDP ) the two main Ir
aqi Kurdish factions , have had northern Iraq under their control since Iraq
i forces were ousted from Kuwait in the 1991 Gulf War .
  FP: two
  FP: 1991
Sentence: Clashes between the two parties broke out at the weekend in the mo
st serious fighting since a U.S.-sponsored ceasefire last year .
  FP: last year
  FP: two
  FP: the weekend
  FN: U.S.-sponsored
Sentence: Mujahideen Khalq said Iranian troops had also been shelling KDP po
sitions in Qasri region in Suleimaniya province near the Iranian border over
the last two days .
  FP: the last two days
```

```
  FP: Suleimaniya province
  FN: Suleimaniya
Sentence: It said about 100 Iraqi Kurds were killed or wounded in the attack
.
  FP: about 100
Sentence: A U.S.-led air force in southern Turkey protects Iraqi Kurds from
possible attacks by Baghdad troops .
  FP: Kurds
  FP: Iraqi
  FN: U.S.-led
  FN: Iraqi Kurds
Sentence: Saudi riyal rates steady in quiet summer trade .
  FP: summer
Sentence: The spot Saudi riyal against the dollar and riyal interbank deposi
t rates were mainly steady this week in quiet summer trade , dealers in the
kingdom said .
  FP: summer
  FP: this week
Sentence: One-month interbank deposits were at 5-1/2 , 3/8 percent , three m
onths were 5-5/8 , 1/2 percent and six months were 5-3/4 , 5/8 percent .
  FP: 5-5/8
  FP: six months
  FP: 1/2 percent
  FP: 5-1/2
  FP: 3/8 percent
  FP: 5/8 percent
  FP: 5-3/4
  FP: three months
  FN: One-month
Sentence: One-year funds were at six , 5-7/8 percent .
  FP: 5-7/8 percent
  FP: six
  FN: One-year
Sentence: Israel gave Palestinian President Yasser Arafat permission on Thur
sday to fly over its territory to the West Bank , ending a brief Israeli-PLO
crisis , an Arafat adviser said .
  FP: Thursday
  FP: the West Bank
  FN: Israeli-PLO
  FN: West Bank
Sentence: The president 's aircraft has received permission to pass through
Israeli airspace but the president is not expected to travel to the West Ban
k before Monday , " Nabil Abu Rdainah told Reuters .
  FP: the West Bank
  FP: Monday
  FN: West Bank
Sentence: Arafat had been scheduled to meet former Israeli prime minister Sh
imon Peres in the West Bank town of Ramallah on Thursday but the venue was c
hanged to Gaza after Israel denied flight clearance to the Palestinian leade
r 's helicopters .
  FP: Thursday
Sentence: Arafat subsequently cancelled a meeting between Israeli and PLO of
ficials , on civilian affairs , at the Allenby Bridge crossing between Jorda
n and the West Bank .
  FP: the West Bank
  FP: the Allenby Bridge
  FN: West Bank
  FN: Allenby Bridge
Sentence: Abu Rdainah said Arafat had decided against flying to the West Ban
k on Thursday , after Israel lifted the ban , because he had a busy schedule
in Gaza and would not be free until Monday .
  FP: Thursday
```

```
  FP: Monday
  FP: the West Bank
  FN: West Bank
Sentence: Yasser Arafat will meet Shimon Peres in Gaza on Thursday after Pal
estinians said the right-wing Israeli government had barred the Palestinian
leader from flying to the West Bank for talks with the former prime minister
.
  FP: the West Bank
  FP: Thursday
  FN: West Bank
Sentence: Palestinian officials said the Israeli government had barred Arafa
t from overflying Israel in a Palestinian helicopter to the West Bank in an
attempt to bar the meeting with Peres .
  FP: the West Bank
  FN: West Bank
Sentence: Israeli Prime Minister Benjamin Netanyahu has accused opposition l
eader Peres , who he defeated in May elections , of trying to undermine his
Likud government 's authority to conduct peace talks .
  FP: May
Sentence: Afghan UAE embassy says Taleban guards going home .
  FN: Taleban
  FN: UAE
Sentence: DUBAI 1996-08-22
  FN: DUBAI
Sentence: Three Afghan guards brought to the United Arab Emirates last week
by Russian hostages who escaped from the Taleban militia will return to Afgh
anistan in a few days , the Afghan embassy in Abu Dhabi said on Thursday .
  FP: a few days
  FP: the United Arab Emirates
  FP: Three
  FP: last week
  FP: Thursday
  FN: United Arab Emirates
Sentence: Their return to Afghanistan will take place in two or three days ,
" an embassy official said .
  FP: two or three days
Sentence: The three Islamic Taleban guards were overpowered by seven Russian
aircrew who escaped to UAE state Sharjah last Friday on board their own airc
raft after a year in the captivity of Taleban militia in Kandahar in souther
n Afghanistan .
  FP: seven
  FP: three
  FP: a year
  FP: Taleban
  FP: Islamic
  FP: last Friday
  FN: Islamic Taleban
Sentence: The UAE said on Monday it would hand over the three to the Interna
tional Red Crescent , possibly last Tuesday .
  FP: Monday
  FP: the International Red Crescent
  FP: last Tuesday
  FP: three
  FN: International Red Crescent
Sentence: When asked whether the three guards would travel back to Kandahar
or the Afghan capital Kabul , the embassy official said : " That has not bee
n decided , but possibly Kandahar . "
  FP: three
Sentence: The embassy official said the three men , believed to be in their
20s , were currently in Abu Dhabi .
  FP: three
  FP: their 20s
```

```
Sentence: The Russians , working for the Aerostan firm in the Russian republ
ic of Tatarstan , were taken hostage after a Taleban MiG-19 fighter forced t
heir cargo plane to land in August 1995 .
  FP: August 1995
  FN: MiG-19
  FN: Taleban
Sentence: The Russians , who said they overpowered the guards -- two armed w
ith Kalashnikov automatic rifles -- while doing regular maintenance work on
their Ilyushin 76 cargo plane last Friday , left the UAE capital Abu Dhabi f
or home on Sunday .
  FP: Sunday
  FP: 76
  FP: two
  FP: last Friday
  FN: Ilyushin 76
Sentence: Iraqi President Saddam Hussein has told visiting Russian ultra-nat
ionalist Vladimir Zhirinovsky that Baghdad wanted to maintain " friendship a
nd cooperation " with Moscow , official Iraqi newspapers said on Thursday .
  FP: Thursday
Sentence: They said Zhirinovsky told Saddam before he left Baghdad on Wednes
day that his Liberal Democratic party and the Russian Duma ( parliament ) "
are calling for an immediate lifting of the embargo " imposed on Iraq after
its 1990 invasion of Kuwait .
  FP: Wednesday
  FP: 1990
  FP: the Russian Duma
  FN: Russian
  FN: Duma
Sentence: Zhirinovsky said on Tuesday he would press the Russian government
to help end U.N. trade sanctions on Iraq and blamed Moscow for delaying esta
blishment of good ties with Baghdad .
  FP: Tuesday
Sentence: Zhirinovsky visited Iraq twice in 1995 .
  FP: 1995
Sentence: Last October he was invited to attend the referendum held on Iraq
's presidency , which extended Saddam 's term for seven more years .
  FP: Last October
  FP: seven more years
Sentence: PRESS DIGEST - Iraq - Aug 22 .
  FP: PRESS DIGEST - Iraq
  FN: Iraq
Sentence: These are some of the leading stories in the official Iraqi press
on Thursday .
  FP: Thursday
Sentence: THAWRA
  FN: THAWRA
Sentence: - Turkish foreign minister says Turkey will take part in the Baghd
ad trade fair that will be held in November .
  FP: November
Sentence: IRAQ
  FN: IRAQ
Sentence: - A shipload of 12 tonnes of rice arrives in Umm Qasr port in the
Gulf .
  FP: Umm
  FP: 12
  FN: Umm Qasr
Sentence: PRESS DIGEST - Lebanon - Aug 22 .
  FP: PRESS DIGEST - Lebanon - Aug 22
  FN: Lebanon
Sentence: BEIRUT 1996-08-22
  FN: BEIRUT
Sentence: These are the leading stories in the Beirut press on Thursday .
```

```
  FP: Thursday
Sentence: AN-NAHAR
  FN: AN-NAHAR
Sentence: AS-SAFIR
  FN: AS-SAFIR
Sentence: - Parliament Speaker Berri : Israel is preparing for war against S
yria and Lebanon .
  FP: Parliament
Sentence: AL-ANWAR
  FN: AL-ANWAR
Sentence: - Continued criticism of law violation incidents -- which occurred
in the Mount Lebanon elections last Sunday .
  FP: last Sunday
Sentence: AD-DIYAR
  FN: AD-DIYAR
Sentence: - Hariri to step into the election battle with an incomplete list
.
  FN: Hariri
Sentence: NIDA'A AL-WATAN
  FN: NIDA'A AL-WATAN
Sentence: - Maronite Patriarch Sfeir expressed sorrow over the violations in
Sunday ' elections .
  FP: Sunday
  FN: Maronite
  FN: Sfeir
Sentence: Early calls on CME live and feeder cattle futures ranged from 0.20
0 cent higher to 0.100 lower , livestock analysts said .
  FP: 0.200 cent
  FP: 0.100
Sentence: KinderCare Learning Centers Inc said on Thursday that a debt buyba
ck would mean an extraordinary loss of $ 1.2 million in its fiscal 1997 firs
t quarter .
  FP: buyback
  FP: Thursday
  FP: fiscal 1997 first quarter
```

TODO: Write a short text that summarises the errors that you observed

False Positive is the wrongly predicted entities and we can see from previous result that most of the FPs are numbers, locations, time and date.

Now, use the insights from your error analysis to improve the automated prediction that you implemented in Problem 2. While the best way to do this would be to update spaCy's NER model using domain-specific training data, for this lab it suffices to write code to post-process the output produced by spaCy. You should be able to improve the F1 score from Problem 2 by at last 15 percentage points.

In [86]:

```
# TODO: Write code here to improve the span prediction from Problem 2

ent_pred=[]
for x in df_dev.itertuples():
    doc = nlp(x[2])
    for ent in doc.ents:
        #print(ent)
        if ent.label_ not in ['DATE', 'TIME', 'PERCENT', 'MONEY', 'QUANTITY
            ent_pred.append((x[1],ent.start,ent.end))
print(ent_pred)
```

```
50-006', 2, 3), ('1150-006', 12, 16), ('1150-006', 2, 3), ('1150-006', 12, 1
```

```
6), ('1151-000', 6, 7), ('1151-001', 0, 1), ('1151-002', 6, 7), ('1151-006',
20, 21), ('1152-000', 0, 1), ('1152-000', 5, 6), ('1152-000', 7, 8), ('1152-
000', 0, 1), ('1152-000', 5, 6), ('1152-000', 7, 8), ('1152-000', 0, 1), ('1
152-000', 5, 6), ('1152-000', 7, 8), ('1152-001', 0, 2), ('1152-002', 0, 1),
('1152-003', 3, 4), ('1152-003', 15, 17), ('1152-003', 21, 22), ('1152-003',
24, 25), ('1152-003', 29, 30), ('1152-003', 31, 32), ('1152-003', 3, 4), ('1
152-003', 15, 17), ('1152-003', 21, 22), ('1152-003', 24, 25), ('1152-003',
29, 30), ('1152-003', 31, 32), ('1152-003', 3, 4), ('1152-003', 15, 17), ('1
152-003', 21, 22), ('1152-003', 24, 25), ('1152-003', 29, 30), ('1152-003',
31, 32), ('1152-003', 3, 4), ('1152-003', 15, 17), ('1152-003', 21, 22), ('1
152-003', 24, 25), ('1152-003', 29, 30), ('1152-003', 31, 32), ('1152-003',
3, 4), ('1152-003', 15, 17), ('1152-003', 21, 22), ('1152-003', 24, 25), ('1
152-003', 29, 30), ('1152-003', 31, 32), ('1152-003', 3, 4), ('1152-003', 1
5, 17), ('1152-003', 21, 22), ('1152-003', 24, 25), ('1152-003', 29, 30), ('
1152-003', 31, 32), ('1152-004', 12, 14), ('1152-004', 19, 20), ('1152-004',
29, 30), ('1152-004', 49, 52), ('1152-004', 12, 14), ('1152-004', 19, 20),
('1152-004', 29, 30), ('1152-004', 49, 52), ('1152-004', 12, 14), ('1152-004
', 19, 20), ('1152-004', 29, 30), ('1152-004', 49, 52), ('1152-004', 12, 1
4), ('1152-004', 19, 20), ('1152-004', 29, 30), ('1152-004', 49, 52), ('1152
-004', 12, 14), ('1152-004', 19, 20), ('1152-004', 29, 30), ('1152-004', 49,
52), ('1152-005', 5, 9), ('1152-005', 32, 33), ('1152-005', 34, 36), ('1152-
005', 5, 9), ('1152-005', 32, 33), ('1152-005', 34, 36), ('1152-005', 5, 9),
('1152-005', 32, 33), ('1152-005', 34, 36), ('1152-007', 0, 1), ('1152-007',
7, 8), ('1152-007', 11, 12), ('1152-007', 24, 25), ('1152-007', 27, 28), ('1
152-007', 31, 32), ('1152-007', 0, 1), ('1152-007', 7, 8), ('1152-007', 11,
12), ('1152-007', 24, 25), ('1152-007', 27, 28), ('1152-007', 31, 32), ('115
2-007', 0, 1), ('1152-007', 7, 8), ('1152-007', 11, 12), ('1152-007', 24, 2
5), ('1152-007', 27, 28), ('1152-007', 31, 32), ('1152-007', 0, 1), ('1152-0
07', 7, 8), ('1152-007', 11, 12), ('1152-007', 24, 25), ('1152-007', 27, 2
8), ('1152-007', 31, 32), ('1152-007', 0, 1), ('1152-007', 7, 8), ('1152-007
', 11, 12), ('1152-007', 24, 25), ('1152-007', 27, 28), ('1152-007', 31, 3
2), ('1152-007', 0, 1), ('1152-007', 7, 8), ('1152-007', 11, 12), ('1152-007
', 24, 25), ('1152-007', 27, 28), ('1152-007', 31, 32), ('1152-008', 10, 1
1), ('1152-008', 17, 18), ('1152-008', 19, 20), ('1152-008', 10, 11), ('1152
-008', 17, 18), ('1152-008', 19, 20), ('1152-008', 10, 11), ('1152-008', 17,
18), ('1152-008', 19, 20), ('1152-010', 1, 2), ('1152-010', 16, 17), ('1152-
010', 1, 2), ('1152-010', 16, 17), ('1152-011', 0, 1), ('1152-011', 2, 3),
('1152-011', 9, 10), ('1152-011', 18, 19), ('1152-011', 21, 22), ('1152-011
', 26, 27), ('1152-011', 0, 1), ('1152-011', 2, 3), ('1152-011', 9, 10), ('1
152-011', 18, 19), ('1152-011', 21, 22), ('1152-011', 26, 27), ('1152-011',
0, 1), ('1152-011', 2, 3), ('1152-011', 9, 10), ('1152-011', 18, 19), ('1152
-011', 21, 22), ('1152-011', 26, 27), ('1152-011', 0, 1), ('1152-011', 2,
3), ('1152-011', 9, 10), ('1152-011', 18, 19), ('1152-011', 21, 22), ('1152-
011', 26, 27), ('1152-011', 0, 1), ('1152-011', 2, 3), ('1152-011', 9, 10),
('1152-011', 18, 19), ('1152-011', 21, 22), ('1152-011', 26, 27), ('1152-011
', 0, 1), ('1152-011', 2, 3), ('1152-011', 9, 10), ('1152-011', 18, 19), ('1
152-011', 21, 22), ('1152-011', 26, 27), ('1152-012', 5, 6), ('1152-012', 1
1, 12), ('1152-012', 15, 16), ('1152-012', 5, 6), ('1152-012', 11, 12), ('11
52-012', 15, 16), ('1152-012', 5, 6), ('1152-012', 11, 12), ('1152-012', 15,
16), ('1152-013', 0, 1), ('1152-013', 5, 6), ('1152-013', 9, 10), ('1152-013
', 17, 18), ('1152-013', 0, 1), ('1152-013', 5, 6), ('1152-013', 9, 10), ('1
152-013', 17, 18), ('1152-013', 0, 1), ('1152-013', 5, 6), ('1152-013', 9, 1
0), ('1152-013', 17, 18), ('1152-013', 0, 1), ('1152-013', 5, 6), ('1152-013
', 9, 10), ('1152-013', 17, 18), ('1152-014', 5, 8), ('1152-014', 18, 19),
('1152-014', 23, 25), ('1152-014', 46, 47), ('1152-014', 49, 52), ('1152-014
', 5, 8), ('1152-014', 18, 19), ('1152-014', 23, 25), ('1152-014', 46, 47),
('1152-014', 49, 52), ('1152-014', 5, 8), ('1152-014', 18, 19), ('1152-014',
23, 25), ('1152-014', 46, 47), ('1152-014', 49, 52), ('1152-014', 5, 8), ('1
152-014', 18, 19), ('1152-014', 23, 25), ('1152-014', 46, 47), ('1152-014',
49, 52), ('1152-014', 5, 8), ('1152-014', 18, 19), ('1152-014', 23, 25), ('1
152-014', 46, 47), ('1152-014', 49, 52), ('1152-014', 5, 8), ('1152-014', 1
8, 19), ('1152-014', 23, 25), ('1152-014', 46, 47), ('1152-014', 49, 52), ('
```

1152-014', 5, 8), ('1152-014', 18, 19), ('1152-014', 23, 25), ('1152-014', 4
6, 47), ('1152-014', 49, 52), ('1152-014', 5, 8), ('1152-014', 18, 19), ('11
52-014', 23, 25), ('1152-014', 46, 47), ('1152-014', 49, 52), ('1152-014',
5, 8), ('1152-014', 18, 19), ('1152-014', 23, 25), ('1152-014', 46, 47), ('1
152-014', 49, 52), ('1152-015', 7, 9), ('1152-016', 2, 3), ('1152-016', 8,
9), ('1152-016', 14, 15), ('1152-016', 17, 18), ('1152-016', 22, 24), ('1152
-016', 26, 27), ('1152-016', 2, 3), ('1152-016', 8, 9), ('1152-016', 14, 1
5), ('1152-016', 17, 18), ('1152-016', 22, 24), ('1152-016', 26, 27), ('1152
-016', 2, 3), ('1152-016', 8, 9), ('1152-016', 14, 15), ('1152-016', 17, 1
8), ('1152-016', 22, 24), ('1152-016', 26, 27), ('1152-016', 2, 3), ('1152-0
16', 8, 9), ('1152-016', 14, 15), ('1152-016', 17, 18), ('1152-016', 22, 2
4), ('1152-016', 26, 27), ('1152-016', 2, 3), ('1152-016', 8, 9), ('1152-016
', 14, 15), ('1152-016', 17, 18), ('1152-016', 22, 24), ('1152-016', 26, 2
7), ('1152-016', 2, 3), ('1152-016', 8, 9), ('1152-016', 14, 15), ('1152-016
', 17, 18), ('1152-016', 22, 24), ('1152-016', 26, 27), ('1152-017', 3, 4),
('1152-017', 9, 14), ('1152-017', 3, 4), ('1152-017', 9, 14), ('1152-018',
1, 3), ('1152-018', 18, 19), ('1152-018', 22, 23), ('1152-018', 1, 3), ('115
2-018', 18, 19), ('1152-018', 22, 23), ('1152-018', 1, 3), ('1152-018', 18,
19), ('1152-018', 22, 23), ('1153-000', 2, 3), ('1153-001', 0, 1), ('1153-00
2', 1, 2), ('1153-003', 4, 5), ('1153-003', 6, 7), ('1153-003', 4, 5), ('115
3-003', 6, 7), ('1153-004', 9, 10), ('1153-006', 8, 10), ('1153-006', 14, 1
6), ('1153-006', 36, 37), ('1153-006', 8, 10), ('1153-006', 14, 16), ('1153-
006', 36, 37), ('1153-006', 8, 10), ('1153-006', 14, 16), ('1153-006', 36, 3
7), ('1153-007', 30, 31), ('1153-007', 30, 31), ('1154-000', 0, 1), ('1154-0
01', 0, 1), ('1154-002', 8, 9), ('1154-002', 16, 17), ('1154-002', 8, 9), ('
1154-002', 16, 17), ('1154-003', 0, 1), ('1154-003', 29, 30), ('1154-003',
0, 1), ('1154-003', 29, 30), ('1154-004', 1, 3), ('1154-005', 0, 1), ('1154-
006', 16, 18), ('1154-006', 26, 27), ('1154-006', 16, 18), ('1154-006', 26,
27), ('1154-008', 0, 1), ('1154-008', 3, 8), ('1154-008', 9, 10), ('1154-008
', 33, 34), ('1154-008', 0, 1), ('1154-008', 3, 8), ('1154-008', 9, 10), ('1
154-008', 33, 34), ('1154-008', 0, 1), ('1154-008', 3, 8), ('1154-008', 9, 1
0), ('1154-008', 33, 34), ('1154-008', 0, 1), ('1154-008', 3, 8), ('1154-008
', 9, 10), ('1154-008', 33, 34), ('1154-009', 7, 8), ('1154-009', 11, 12),
('1154-009', 29, 30), ('1154-009', 7, 8), ('1154-009', 11, 12), ('1154-009',
29, 30), ('1154-009', 7, 8), ('1154-009', 11, 12), ('1154-009', 29, 30), ('1
155-000', 0, 1), ('1155-000', 3, 4), ('1155-000', 8, 9), ('1155-000', 0, 1),
('1155-000', 3, 4), ('1155-000', 8, 9), ('1155-000', 0, 1), ('1155-000', 3,
4), ('1155-000', 8, 9), ('1155-001', 0, 1), ('1155-002', 0, 1), ('1155-002',
9, 10), ('1155-002', 13, 14), ('1155-002', 24, 25), ('1155-002', 0, 1), ('11
55-002', 9, 10), ('1155-002', 13, 14), ('1155-002', 24, 25), ('1155-002', 0,
1), ('1155-002', 9, 10), ('1155-002', 13, 14), ('1155-002', 24, 25), ('1155-
002', 0, 1), ('1155-002', 9, 10), ('1155-002', 13, 14), ('1155-002', 24, 2
5), ('1155-003', 14, 15), ('1155-003', 18, 19), ('1155-003', 25, 27), ('1155
-003', 30, 31), ('1155-003', 14, 15), ('1155-003', 18, 19), ('1155-003', 25,
27), ('1155-003', 30, 31), ('1155-003', 14, 15), ('1155-003', 18, 19), ('115
5-003', 25, 27), ('1155-003', 30, 31), ('1155-003', 14, 15), ('1155-003', 1
8, 19), ('1155-003', 25, 27), ('1155-003', 30, 31), ('1155-004', 0, 1), ('11
55-005', 5, 6), ('1155-005', 14, 15), ('1155-005', 5, 6), ('1155-005', 14, 1
5), ('1155-006', 1, 2), ('1155-006', 10, 11), ('1155-006', 1, 2), ('1155-006
', 10, 11), ('1155-007', 0, 1), ('1155-009', 15, 16), ('1155-009', 19, 20),
('1155-009', 21, 23), ('1155-009', 15, 16), ('1155-009', 19, 20), ('1155-009
', 21, 23), ('1155-009', 15, 16), ('1155-009', 19, 20), ('1155-009', 21, 2
3), ('1155-010', 7, 8), ('1155-010', 17, 18), ('1155-010', 27, 28), ('1155-0
10', 31, 32), ('1155-010', 7, 8), ('1155-010', 17, 18), ('1155-010', 27, 2
8), ('1155-010', 31, 32), ('1155-010', 7, 8), ('1155-010', 17, 18), ('1155-0
10', 27, 28), ('1155-010', 31, 32), ('1155-010', 7, 8), ('1155-010', 17, 1
8), ('1155-010', 27, 28), ('1155-010', 31, 32), ('1155-010', 7, 8), ('1155-0
10', 17, 18), ('1155-010', 27, 28), ('1155-010', 31, 32), ('1155-011', 0,
1), ('1155-011', 3, 4), ('1155-011', 6, 7), ('1155-011', 15, 16), ('1155-011
', 0, 1), ('1155-011', 3, 4), ('1155-011', 6, 7), ('1155-011', 15, 16), ('11
55-011', 0, 1), ('1155-011', 3, 4), ('1155-011', 6, 7), ('1155-011', 15, 1
6), ('1155-011', 0, 1), ('1155-011', 3, 4), ('1155-011', 6, 7), ('1155-011',

```
15, 16), ('1155-012', 7, 8), ('1155-012', 16, 17), ('1155-012', 7, 8), ('115
5-012', 16, 17), ('1155-012', 7, 8), ('1155-012', 16, 17), ('1155-013', 1,
2), ('1155-014', 0, 1), ('1155-014', 5, 6), ('1155-014', 9, 11), ('1155-014
', 17, 18), ('1155-014', 0, 1), ('1155-014', 5, 6), ('1155-014', 9, 11), ('1
155-014', 17, 18), ('1155-014', 0, 1), ('1155-014', 5, 6), ('1155-014', 9, 1
1), ('1155-014', 17, 18), ('1155-014', 0, 1), ('1155-014', 5, 6), ('1155-014
', 9, 11), ('1155-014', 17, 18), ('1155-015', 0, 1), ('1155-015', 8, 9), ('1
155-015', 12, 13), ('1155-015', 0, 1), ('1155-015', 8, 9), ('1155-015', 12,
13), ('1155-015', 0, 1), ('1155-015', 8, 9), ('1155-015', 12, 13), ('1156-00
0', 0, 1), ('1156-000', 2, 3), ('1156-000', 4, 5), ('1156-000', 0, 1), ('115
6-000', 2, 3), ('1156-000', 4, 5), ('1156-000', 0, 1), ('1156-000', 2, 3),
('1156-000', 4, 5), ('1156-000', 0, 1), ('1156-000', 2, 3), ('1156-000', 4,
5), ('1156-002', 0, 1), ('1156-002', 3, 5), ('1156-002', 9, 10), ('1156-002
', 11, 13), ('1156-002', 18, 19), ('1156-002', 0, 1), ('1156-002', 3, 5), ('
1156-002', 9, 10), ('1156-002', 11, 13), ('1156-002', 18, 19), ('1156-002',
0, 1), ('1156-002', 3, 5), ('1156-002', 9, 10), ('1156-002', 11, 13), ('1156
-002', 18, 19), ('1156-002', 0, 1), ('1156-002', 3, 5), ('1156-002', 9, 10),
('1156-002', 11, 13), ('1156-002', 18, 19), ('1156-002', 0, 1), ('1156-002',
3, 5), ('1156-002', 9, 10), ('1156-002', 11, 13), ('1156-002', 18, 19), ('11
56-003', 3, 4), ('1156-003', 15, 16), ('1156-003', 19, 20), ('1156-003', 21,
23), ('1156-003', 25, 30), ('1156-003', 32, 33), ('1156-003', 33, 35), ('115
6-003', 40, 41), ('1156-003', 3, 4), ('1156-003', 15, 16), ('1156-003', 19,
20), ('1156-003', 21, 23), ('1156-003', 25, 30), ('1156-003', 32, 33), ('115
6-003', 33, 35), ('1156-003', 40, 41), ('1156-003', 3, 4), ('1156-003', 15,
16), ('1156-003', 19, 20), ('1156-003', 21, 23), ('1156-003', 25, 30), ('115
6-003', 32, 33), ('1156-003', 33, 35), ('1156-003', 40, 41), ('1156-003', 3,
4), ('1156-003', 15, 16), ('1156-003', 19, 20), ('1156-003', 21, 23), ('1156
-003', 25, 30), ('1156-003', 32, 33), ('1156-003', 33, 35), ('1156-003', 40,
41), ('1156-003', 3, 4), ('1156-003', 15, 16), ('1156-003', 19, 20), ('1156-
003', 21, 23), ('1156-003', 25, 30), ('1156-003', 32, 33), ('1156-003', 33,
35), ('1156-003', 40, 41), ('1156-003', 3, 4), ('1156-003', 15, 16), ('1156-
003', 19, 20), ('1156-003', 21, 23), ('1156-003', 25, 30), ('1156-003', 32,
33), ('1156-003', 33, 35), ('1156-003', 40, 41), ('1156-003', 3, 4), ('1156-
003', 15, 16), ('1156-003', 19, 20), ('1156-003', 21, 23), ('1156-003', 25,
30), ('1156-003', 32, 33), ('1156-003', 33, 35), ('1156-003', 40, 41), ('115
6-003', 3, 4), ('1156-003', 15, 16), ('1156-003', 19, 20), ('1156-003', 21,
23), ('1156-003', 25, 30), ('1156-003', 32, 33), ('1156-003', 33, 35), ('115
6-003', 40, 41), ('1156-003', 3, 4), ('1156-003', 15, 16), ('1156-003', 19,
20), ('1156-003', 21, 23), ('1156-003', 25, 30), ('1156-003', 32, 33), ('115
6-003', 33, 35), ('1156-003', 40, 41), ('1156-004', 6, 7), ('1156-004', 14,
16), ('1156-004', 23, 24), ('1156-004', 6, 7), ('1156-004', 14, 16), ('1156-
004', 23, 24), ('1156-004', 6, 7), ('1156-004', 14, 16), ('1156-004', 23, 2
4), ('1156-004', 6, 7), ('1156-004', 14, 16), ('1156-004', 23, 24), ('1156-0
06', 9, 12), ('1156-006', 15, 21), ('1156-006', 9, 12), ('1156-006', 15, 2
1), ('1156-006', 9, 12), ('1156-006', 15, 21), ('1156-007', 2, 3), ('1156-00
8', 0, 4), ('1156-008', 11, 12), ('1156-008', 13, 15), ('1156-008', 28, 29),
('1156-008', 0, 4), ('1156-008', 11, 12), ('1156-008', 13, 15), ('1156-008',
28, 29), ('1156-008', 0, 4), ('1156-008', 11, 12), ('1156-008', 13, 15), ('1
156-008', 28, 29), ('1156-008', 0, 4), ('1156-008', 11, 12), ('1156-008', 1
3, 15), ('1156-008', 28, 29), ('1156-009', 11, 12), ('1156-009', 18, 19), ('
1156-009', 11, 12), ('1156-009', 18, 19), ('1156-009', 11, 12), ('1156-009',
18, 19), ('1156-010', 10, 11), ('1156-010', 20, 21), ('1156-010', 10, 11),
('1156-010', 20, 21), ('1157-000', 0, 1), ('1157-000', 6, 7), ('1157-001',
0, 1), ('1157-002', 0, 6), ('1157-002', 19, 20), ('1157-002', 0, 6), ('1157-
002', 19, 20), ('1157-003', 2, 3), ('1157-003', 5, 6), ('1157-003', 2, 3),
('1157-003', 5, 6), ('1157-004', 5, 7), ('1157-004', 10, 11), ('1157-004', 1
3, 16), ('1157-004', 17, 18), ('1157-004', 25, 26), ('1157-004', 28, 29), ('
1157-004', 5, 7), ('1157-004', 10, 11), ('1157-004', 13, 16), ('1157-004', 1
7, 18), ('1157-004', 25, 26), ('1157-004', 28, 29), ('1157-004', 5, 7), ('11
57-004', 10, 11), ('1157-004', 13, 16), ('1157-004', 17, 18), ('1157-004', 2
5, 26), ('1157-004', 28, 29), ('1157-004', 5, 7), ('1157-004', 10, 11), ('11
57-004', 13, 16), ('1157-004', 17, 18), ('1157-004', 25, 26), ('1157-004', 2
```

```
8, 29), ('1157-004', 5, 7), ('1157-004', 10, 11), ('1157-004', 13, 16), ('11
57-004', 17, 18), ('1157-004', 25, 26), ('1157-004', 28, 29), ('1157-004',
5, 7), ('1157-004', 10, 11), ('1157-004', 13, 16), ('1157-004', 17, 18), ('1
157-004', 25, 26), ('1157-004', 28, 29), ('1158-000', 5, 6), ('1158-001', 0,
1), ('1158-002', 7, 8), ('1158-002', 17, 18), ('1158-002', 7, 8), ('1158-002
', 17, 18), ('1158-003', 5, 6), ('1158-003', 8, 9), ('1158-003', 10, 11), ('
1158-003', 26, 27), ('1158-003', 5, 6), ('1158-003', 8, 9), ('1158-003', 10,
11), ('1158-003', 26, 27), ('1158-003', 5, 6), ('1158-003', 8, 9), ('1158-00
3', 10, 11), ('1158-003', 26, 27), ('1158-003', 5, 6), ('1158-003', 8, 9),
('1158-003', 10, 11), ('1158-003', 26, 27), ('1158-005', 14, 15), ('1158-005
', 17, 22), ('1158-005', 14, 15), ('1158-005', 17, 22), ('1158-005', 14, 1
5), ('1158-005', 17, 22), ('1158-005', 14, 15), ('1158-005', 17, 22), ('1159
-000', 7, 8), ('1159-001', 0, 1), ('1159-002', 0, 1), ('1159-003', 15, 16),
('1159-003', 19, 20), ('1159-003', 15, 16), ('1159-003', 19, 20), ('1159-004
', 4, 5), ('1159-004', 14, 16), ('1159-004', 20, 22), ('1159-004', 33, 34),
('1159-004', 38, 39), ('1159-004', 4, 5), ('1159-004', 14, 16), ('1159-004',
20, 22), ('1159-004', 33, 34), ('1159-004', 38, 39), ('1159-004', 4, 5), ('1
159-004', 14, 16), ('1159-004', 20, 22), ('1159-004', 33, 34), ('1159-004',
38, 39), ('1159-004', 4, 5), ('1159-004', 14, 16), ('1159-004', 20, 22), ('1
159-004', 33, 34), ('1159-004', 38, 39), ('1159-004', 4, 5), ('1159-004', 1
4, 16), ('1159-004', 20, 22), ('1159-004', 33, 34), ('1159-004', 38, 39), ('
1159-007', 28, 29), ('1159-007', 28, 29), ('1160-000', 5, 6), ('1160-000',
5, 6), ('1160-001', 0, 2), ('1160-002', 0, 1), ('1160-002', 2, 3), ('1160-00
3', 2, 4), ('1160-003', 16, 18), ('1160-003', 25, 26), ('1160-003', 33, 37),
('1160-003', 2, 4), ('1160-003', 16, 18), ('1160-003', 25, 26), ('1160-003',
33, 37), ('1160-003', 2, 4), ('1160-003', 16, 18), ('1160-003', 25, 26), ('1
160-003', 33, 37), ('1160-003', 2, 4), ('1160-003', 16, 18), ('1160-003', 2
5, 26), ('1160-003', 33, 37), ('1160-004', 4, 6), ('1160-004', 7, 9), ('1160
-004', 14, 16), ('1160-004', 17, 22), ('1160-004', 28, 29), ('1160-004', 4,
6), ('1160-004', 7, 9), ('1160-004', 14, 16), ('1160-004', 17, 22), ('1160-0
04', 28, 29), ('1160-004', 4, 6), ('1160-004', 7, 9), ('1160-004', 14, 16),
('1160-004', 17, 22), ('1160-004', 28, 29), ('1160-004', 4, 6), ('1160-004',
7, 9), ('1160-004', 14, 16), ('1160-004', 17, 22), ('1160-004', 28, 29), ('1
160-004', 4, 6), ('1160-004', 7, 9), ('1160-004', 14, 16), ('1160-004', 17,
22), ('1160-004', 28, 29), ('1160-005', 6, 7), ('1160-005', 26, 28), ('1160-
005', 6, 7), ('1160-005', 26, 28), ('1160-005', 6, 7), ('1160-005', 26, 28),
('1160-006', 0, 1), ('1160-007', 6, 7), ('1160-007', 9, 10), ('1160-007', 2
1, 22), ('1160-007', 28, 30), ('1160-007', 6, 7), ('1160-007', 9, 10), ('116
0-007', 21, 22), ('1160-007', 28, 30), ('1160-007', 6, 7), ('1160-007', 9, 1
0), ('1160-007', 21, 22), ('1160-007', 28, 30), ('1160-007', 6, 7), ('1160-0
07', 9, 10), ('1160-007', 21, 22), ('1160-007', 28, 30), ('1160-010', 0, 1),
('1160-010', 5, 10), ('1160-010', 27, 28), ('1160-010', 0, 1), ('1160-010',
5, 10), ('1160-010', 27, 28), ('1160-010', 0, 1), ('1160-010', 5, 10), ('116
0-010', 27, 28), ('1160-010', 0, 1), ('1160-010', 5, 10), ('1160-010', 27, 2
8), ('1160-011', 17, 19), ('1160-011', 20, 21), ('1160-011', 22, 24), ('1160
-011', 25, 26), ('1160-011', 28, 29), ('1160-011', 17, 19), ('1160-011', 20,
21), ('1160-011', 22, 24), ('1160-011', 25, 26), ('1160-011', 28, 29), ('116
0-011', 17, 19), ('1160-011', 20, 21), ('1160-011', 22, 24), ('1160-011', 2
5, 26), ('1160-011', 28, 29), ('1160-011', 17, 19), ('1160-011', 20, 21), ('
1160-011', 22, 24), ('1160-011', 25, 26), ('1160-011', 28, 29), ('1160-011',
17, 19), ('1160-011', 20, 21), ('1160-011', 22, 24), ('1160-011', 25, 26),
('1160-011', 28, 29), ('1160-012', 1, 2), ('1160-014', 0, 1), ('1160-014',
3, 4), ('1160-014', 11, 12), ('1160-014', 20, 24), ('1160-014', 30, 31), ('1
160-014', 0, 1), ('1160-014', 3, 4), ('1160-014', 11, 12), ('1160-014', 20,
24), ('1160-014', 30, 31), ('1160-014', 0, 1), ('1160-014', 3, 4), ('1160-01
4', 11, 12), ('1160-014', 20, 24), ('1160-014', 30, 31), ('1160-014', 0, 1),
('1160-014', 3, 4), ('1160-014', 11, 12), ('1160-014', 20, 24), ('1160-014',
30, 31), ('1160-014', 0, 1), ('1160-014', 3, 4), ('1160-014', 11, 12), ('116
0-014', 20, 24), ('1160-014', 30, 31), ('1160-015', 0, 2), ('1160-015', 3,
5), ('1160-015', 6, 7), ('1160-015', 17, 19), ('1160-015', 22, 23), ('1160-0
15', 0, 2), ('1160-015', 3, 5), ('1160-015', 6, 7), ('1160-015', 17, 19), ('
1160-015', 22, 23), ('1160-015', 0, 2), ('1160-015', 3, 5), ('1160-015', 6,
```

```
7), ('1160-015', 17, 19), ('1160-015', 22, 23), ('1160-015', 0, 2), ('1160-0
15', 3, 5), ('1160-015', 6, 7), ('1160-015', 17, 19), ('1160-015', 22, 23),
('1160-015', 0, 2), ('1160-015', 3, 5), ('1160-015', 6, 7), ('1160-015', 17,
19), ('1160-015', 22, 23), ('1160-016', 9, 10), ('1160-016', 11, 12), ('1160
-016', 9, 10), ('1160-016', 11, 12), ('1160-018', 4, 5), ('1160-018', 16, 1
7), ('1160-018', 4, 5), ('1160-018', 16, 17), ('1160-022', 16, 17), ('1160-0
22', 31, 32), ('1160-022', 16, 17), ('1160-022', 31, 32), ('1160-023', 18, 2
0), ('1160-023', 33, 34), ('1160-023', 18, 20), ('1160-023', 33, 34), ('1160
-024', 19, 20), ('1161-000', 0, 1), ('1161-002', 2, 6), ('1161-002', 37, 3
8), ('1161-002', 2, 6), ('1161-002', 37, 38), ('1161-002', 2, 6), ('1161-002
', 37, 38), ('1161-004', 12, 14), ('1161-005', 0, 3), ('1161-007', 7, 9), ('
1161-007', 10, 12), ('1161-007', 13, 15), ('1161-007', 7, 9), ('1161-007', 1
```

In [87]:
```
evaluation_report(spans_dev_gold,set(ent_pred))
```

```
86.05619146722164 69.88338685144498 77.1311322514456
```

Show that you achieve the performance goal by reporting the evaluation measures that you implemented in Problem 1.

Before going on, we ask you to store the outputs of the improved named entity recognizer on the development data in a new data frame. This new frame should have the same layout as the original data frame for the development data that you loaded above, but should contain the *predicted* start and end positions for each token span, rather than the gold positions. As the `label` of each span, you can use the special value `--NME--`.

In [88]:
```
# TODO: Write code here to store the predicted spans in a new data frame
#df_dev[df_dev.sentence_id=="0946-002"].iloc[0].sentence
# sentence_id=[x[0] for x in ent_pred]
for i,e in enumerate(ent_pred):
    ent_pred[i]=(e[0],df_dev[df_dev.sentence_id==e[0]].iloc[0].sentence,e[1]
pred_df=pd.DataFrame(set(ent_pred))
```

Problem 4: Entity linking

Now that we have a method for predicting mention spans, we turn to the task of **entity linking**, which amounts to predicting the knowledge base entity that is referenced by a given mention. In our case, for each span we want to predict the Wikipedia page that this mention references.

Start by extending the generator function that you implemented in Problem 2 to labelled spans.

In [89]:
```
def gold_mentions(df):
    """Yield the gold-standard mentions in a data frame.

    Args:
        df: A data frame.

    Yields:
        The gold-standard mention spans in the specified data frame as
        quadruples consisting of the sentence id, start position, end
        position and entity label of each span.
    """
    # TODO: Replace the next line with your own code
    for x in df.itertuples():
        yield x[1],x[3],x[4],x[5]
```

A naive baseline for entity linking on our data set is to link each mention span to the Wikipedia page name that we get when we join the tokens in the span by underscores, as is standard in Wikipedia page names. Suppose, for example, that a span contains the two tokens

```
Jimi Hendrix
```

The baseline Wikipedia page name for this span would be

```
Jimi_Hendrix
```

Implement this naive baseline and evaluate its performance. Print the evaluation measures that you implemented in Problem 1.

Here and in the remainder of this lab, you should base your entity predictions on the predicted spans that you computed in Problem 3.

In [90]:

```
# TODO: Write code here to implement the baseline
naive_pred=[]

# for x in pred_df.iloc[:2].itertuples():
#     doc = nlp(x[2])
#     for ent in doc.ents:
#         tmp_str=[]
#         for i in range(ent.start,ent.end):
#             tmp_str.append(doc[i].text)

#         tmp_str="_".join(tmp_str)
#         naive_pred.append((x[1],ent.start,ent.end,tmp_str))
# print(naive_pred)


for x in pred_df.itertuples():
    doc = nlp(x[2])
    ents = [(x[1],e.start,e.end,e.text.replace(" ","_")) for e in doc.ents]
    for e in ents:
        naive_pred.append(e)
print(naive_pred)
```

```
, 16, 18, 'Prince_Eugene'), ('1116-006', 20, 21, 'French'), ('1116-006', 31,
32, 'Italy'), ('1046-006', 16, 17, 'ten'), ('1046-006', 19, 21, 'Armando_Alv
arez'), ('1046-006', 24, 26, '15_minutes'), ('1116-006', 0, 1, '1706'), ('11
16-006', 2, 3, 'French'), ('1116-006', 5, 6, 'Duke'), ('1116-006', 7, 8, 'Or
leans'), ('1116-006', 9, 10, 'Turin'), ('1116-006', 14, 15, 'Austrians'), ('
1116-006', 16, 18, 'Prince_Eugene'), ('1116-006', 20, 21, 'French'), ('1116-
006', 31, 32, 'Italy'), ('1072-012', 5, 6, 'Joubert'), ('1072-012', 11, 14,
'the_25th_minute'), ('1072-012', 18, 20, 'Pieter_Hendriks'), ('1072-012', 2
4, 25, 'Joubert'), ('1072-012', 31, 33, 'Derek_Bevan'), ('1072-012', 34, 35,
'Van'), ('1072-012', 36, 37, 'Westhuizen'), ('0957-026', 0, 2, 'Andrei_Medve
dev'), ('0957-026', 3, 4, 'Ukraine'), ('0957-026', 6, 8, 'Jan_Kroslak'), ('0
957-026', 9, 10, 'Slovakia'), ('1072-017', 0, 3, 'Centre_Walter_Little'), ('
1072-017', 5, 6, 'Mehrtens'), ('1072-017', 15, 17, 'Justin_Marshall'), ('105
8-026', 1, 2, 'Orioles'), ('1058-026', 5, 8, 'the_White_Sox'), ('1058-026',
10, 12, 'American_League'), ('1058-026', 16, 17, 'Mariners'), ('1094-034',
0, 1, 'MONTREAL'), ('1094-034', 1, 4, '71_62_.534'), ('1094-034', 4, 6, '12_
1/2'), ('1128-011', 1, 2, 'Poland'), ('1128-011', 3, 7, 'up_to_75_percent'),
('1128-011', 11, 12, 'March'), ('1128-011', 12, 13, 'Kaczmarek'), ('1121-003
', 4, 8, 'the_Iraqi_National_Congress'), ('1121-003', 9, 10, 'Iraqi'), ('112
1-003', 16, 17, 'Iraqi'), ('1121-003', 22, 24, '10_km'), ('1121-003', 25, 2
7, 'six_miles'), ('1121-003', 29, 30, 'Arbil'), ('1121-003', 36, 37, 'Kurdis
h'), ('1121-003', 41, 42, 'Iraq'), ('0957-009', 0, 2, 'Sjeng_Schalken'), ('0
957-009', 3, 4, 'Netherlands'), ('0957-009', 6, 8, 'David_Rikl'), ('0957-009
', 9, 10, 'Czech'), ('0957-009', 12, 13, '6'), ('0991-011', 2, 4, 'Viktor_Ch
ernomyrdin'), ('0991-011', 6, 7, 'Friday'), ('0991-011', 8, 9, 'Yeltsin'),
('0991-011', 14, 15, 'Lebed'), ('0955-013', 0, 1, 'Ulsan'), ('0955-013', 3,
4, '2'), ('0955-013', 4, 5, '8'), ('0955-013', 5, 7, '9_3'), ('0982-002', 0,
1, 'UK'), ('0982-002', 11, 13, 'around_1.50'), ('0982-002', 17, 18, 'Friday
'), ('0982-002', 21, 22, 'Chicago'), ('0982-002', 23, 24, 'Thursday'), ('116
0-011', 17, 19, 'Mao_Tse-Tung'), ('1160-011', 20, 21, 'China'), ('1160-011',
22, 24, 'Yitzak_Shamir'), ('1160-011', 25, 26, 'Israel'), ('1160-011', 28, 2
9, 'Jordan'), ('1003-005', 3, 4, 'second'), ('1003-005', 5, 6, 'Brunswijk'),
('1003-005', 13, 17, 'less_than_two_years'), ('1112-004', 0, 1, 'Klinsmann
'), ('1112-004', 4, 5, 'Germany'), ('1112-004', 8, 9, 'France'), ('1112-004
', 18, 19, 'European'), ('1112-004', 21, 22, 'England'), ('1112-004', 22, 2
4, 'this_summer'), ('0963-017', 0, 1, 'Henke'), ('0963-017', 5, 6, 'one'),
('0963-017', 13, 15, 'Andrew_Magee'), ('0963-017', 16, 17, 'Wednesday'), ('0
```

963-017', 21, 22, 'three'), ('0963-017', 23, 25, 'PGA_Tour'), ('0963-017', 3
1, 32, '1993'), ('0963-017', 32, 34, 'BellSouth_Classic'), ('1059-002', 0,
1, 'Essex'), ('1059-002', 2, 3, 'Kent'), ('1059-002', 8, 9, 'Monday'), ('105
9-002', 18, 19, 'Derbyshire'), ('1059-002', 20, 21, 'Surrey'), ('1059-002',
23, 24, 'three-day'), ('1059-002', 26, 27, 'Saturday'), ('1059-002', 30, 32,
'English_county'), ('1103-006', 10, 12, 'Sergen_Yalcin'), ('1103-006', 14, 1
5, 'Turkish'), ('1103-006', 17, 19, 'the_61st'), ('1103-006', 25, 26, 'Belgi
an'), ('1103-006', 27, 30, 'Filip_De_Wilde'), ('1121-003', 4, 8, 'the_Iraqi_
National_Congress'), ('1121-003', 9, 10, 'Iraqi'), ('1121-003', 16, 17, 'Ira
qi'), ('1121-003', 22, 24, '10_km'), ('1121-003', 25, 27, 'six_miles'), ('11
21-003', 29, 30, 'Arbil'), ('1121-003', 36, 37, 'Kurdish'), ('1121-003', 41,
42, 'Iraq'), ('1062-002', 0, 1, 'Wales'), ('1062-002', 2, 4, 'San_Marino'),
('1062-002', 10, 13, 'a_World_Cup'), ('1062-002', 14, 15, 'European'), ('106
2-002', 16, 17, '7'), ('1062-002', 19, 20, 'Saturday'), ('0960-003', 0, 2, '
Andre_Agassi'), ('0960-003', 5, 6, 'Thursday'), ('0960-003', 7, 8, 'Wimbledo
n'), ('0960-003', 9, 11, 'MaliVai_Washington'), ('0960-003', 12, 14, 'Marcel
o_Rios'), ('0960-003', 19, 21, 'a_night'), ('0960-003', 24, 27, 'the_U.S._Op
en'), ('0982-002', 0, 1, 'UK'), ('0982-002', 11, 13, 'around_1.50'), ('0982-
002', 17, 18, 'Friday'), ('0982-002', 21, 22, 'Chicago'), ('0982-002', 23, 2
4, 'Thursday'), ('1133-006', 15, 16, 'Xiao'), ('1133-006', 18, 19, 'Manila
'), ('1133-006', 22, 24, 'Amnesty_International'), ('1133-006', 29, 30, 'Chi
na'), ('1157-002', 0, 6, "The_North_Atlantic_Treaty_Organisation_'s"), ('115
7-002', 8, 9, 'Saturday'), ('1157-002', 19, 20, 'Iraq'), ('1027-006', 4, 5,
'today'), ('1027-006', 8, 9, 'Israeli'), ('1027-006', 12, 13, 'Jamil'), ('10
27-006', 14, 15, 'Hebrew'), ('1057-007', 0, 1, 'Pakistan'), ('1057-007', 2,
4, 'Aamir_Sohail'), ('1057-007', 5, 7, 'Saeed_Anwar'), ('1057-007', 8, 10, '
Ijaz_Ahmed'), ('1057-007', 11, 13, 'Salim_Malik'), ('1057-007', 16, 18, 'Was
im_Akram'), ('1057-007', 22, 24, 'Moin_Khan'), ('1057-007', 25, 27, 'Saqlain
_Mushtaq'), ('1057-007', 28, 30, 'Mushtaq_Ahmed'), ('1057-007', 31, 33, 'Waq
ar_Younis'), ('1075-001', 0, 1, 'NOUAKCHOTT'), ('1001-000', 0, 1, 'Romania
'), ('1001-000', 7, 8, '50.19'), ('0958-002', 0, 3, 'Major_League_Baseball
'), ('0953-004', 3, 4, 'Duran'), ('0953-004', 5, 6, 'first'), ('0953-004', 1
0, 12, '10_years'), ('0953-004', 26, 27, 'Duran'), ('1073-007', 0, 4, 'New_Z
ealand_-_Tries'), ('1073-007', 5, 7, 'Sean_Fitzpatrick'), ('1073-007', 8, 1
0, 'Walter_Little'), ('1073-007', 11, 13, 'Justin_Marshall'), ('1037-002',
0, 1, 'Swedish'), ('1037-002', 3, 6, 'LM_Ericsson_AB'), ('1037-002', 8, 9, '
Friday'), ('1037-002', 14, 16, '1.2_billion'), ('1037-002', 25, 26, 'Guangdo
ng'), ('1037-002', 28, 29, 'China'), ('1153-003', 4, 5, 'Rachel'), ('1153-00
3', 6, 7, 'Severine'), ('0997-006', 0, 2, 'Naina_Yeltsin'), ('0997-006', 5,
6, 'Moscow'), ('0997-006', 7, 10, 'Central_Clinical_Hospital'), ('0997-006',
18, 20, 'last_year'), ('1022-002', 0, 1, 'Canadian'), ('1022-002', 5, 6, 'Fr
iday'), ('1022-002', 12, 13, 'U.S.'), ('1022-002', 17, 18, 'U.S.'), ('1022-0
02', 21, 22, 'Canadian'), ('1160-002', 0, 1, 'VENICE'), ('1160-002', 2, 3, '
Italy'), ('1011-012', 10, 11, 'Missouri'), ('1011-012', 13, 17, 'First_Bank_
System_Inc.'), ('1011-012', 18, 20, 'Norwest_Corp.'), ('1011-012', 22, 23, '
KeyCorp'), ('1011-012', 24, 27, 'Banc_One_Corp.'), ('1011-012', 29, 33, 'Fir
st_Chicago_NBD_Corp.'), ('1011-012', 34, 35, 'Illinois'), ('1126-004', 1, 4,
'up_to_75'), ('1126-004', 4, 5, 'Moslem'), ('1126-004', 11, 12, 'Mahala'),
('1126-004', 14, 15, 'NATO'), ('1126-004', 16, 19, 'Lieutenant-Colonel_Max_M
arriner'), ('1126-004', 21, 22, 'Sarajevo'), ('1073-007', 0, 4, 'New_Zealand
_-_Tries'), ('1073-007', 5, 7, 'Sean_Fitzpatrick'), ('1073-007', 8, 10, 'Wal
ter_Little'), ('1073-007', 11, 13, 'Justin_Marshall'), ('1097-000', 0, 7, 'T
ENNIS_-_EDBERG_REFUSES_TO_QO_QUIETLY'), ('1119-006', 10, 15, 'the_end_of_the
_1991'), ('1119-006', 15, 17, 'Gulf_War'), ('1119-006', 21, 22, 'British'),
('1119-006', 24, 26, 'John_Major'), ('1119-006', 28, 29, 'Iraqi'), ('1119-00
6', 29, 30, 'Kurds'), ('1119-006', 34, 35, 'Iraqi'), ('1000-018', 6, 7, 'Kec
skemet'), ('1000-018', 10, 11, 'Petofi'), ('1000-018', 26, 27, 'Petofi'), ('
1000-018', 29, 30, 'Frater'), ('1022-002', 0, 1, 'Canadian'), ('1022-002',
5, 6, 'Friday'), ('1022-002', 12, 13, 'U.S.'), ('1022-002', 17, 18, 'U.S.'),
('1022-002', 21, 22, 'Canadian'), ('1004-003', 0, 2, 'Hen_Vipheak'), ('1004-
003', 6, 9, 'the_Sereipheap_Thmei'), ('1004-003', 10, 12, 'New_Liberty'), ('
1004-003', 22, 23, 'French'), ('1004-003', 24, 25, 'T3'), ('1004-003', 26, 2

8, 'late_Friday'), ('1004-003', 28, 29, 'afternoon'), ('1004-003', 37, 38, '
Sihanouk'), ('1051-009', 0, 1, '290'), ('1051-009', 1, 3, 'Colin_Montgomerie
'), ('1051-009', 3, 5, '68_76'), ('1051-009', 5, 7, '77_69'), ('1051-009',
8, 10, 'Robert_Coles'), ('1051-009', 10, 12, '74_76'), ('1051-009', 12, 14,
'71_69'), ('1051-009', 15, 17, 'Philip_Walton'), ('1051-009', 18, 19, 'Irela
nd'), ('1051-009', 22, 24, '74_71'), ('1051-009', 25, 27, 'Peter_Mitchell'),
('1051-009', 27, 28, '74'), ('1051-009', 29, 31, '74_71'), ('1051-009', 32,
34, 'Klas_Eriksson'), ('1051-009', 35, 36, 'Sweden'), ('1051-009', 37, 39, '
71_75'), ('1051-009', 39, 41, '72_72'), ('1051-009', 42, 44, 'Pedro_Linhart
'), ('1051-009', 45, 46, 'Spain'), ('1051-009', 47, 49, '72_73'), ('1014-003
', 1, 2, 'Dole'), ('1014-003', 6, 7, 'Friday'), ('1014-003', 9, 10, 'Clinton
'), ('1014-003', 14, 15, 'Dole'), ('1014-003', 17, 19, 'Scott_Reed'), ('0986
-007', 0, 4, '--_Dublin_Newsroom_+353'), ('0983-004', 1, 2, 'Iraqis'), ('098
3-004', 13, 14, 'Saddam'), ('0983-004', 15, 16, 'Hussein'), ('0983-004', 29,
30, 'seven'), ('0983-004', 39, 40, 'asylum'), ('1058-040', 1, 2, 'Milwaukee
'), ('1058-040', 3, 5, 'Marc_Newfield'), ('1058-040', 7, 9, 'Jose_Parra'),
('1058-040', 17, 19, 'the_12th'), ('1058-040', 21, 22, 'Brewers'), ('1069-01
7', 1, 3, 'Tunbridge_Wells'), ('1069-017', 5, 6, '214'), ('1069-017', 7, 8,
'167-6'), ('1069-017', 9, 12, 'C._Tolley_64'), ('1069-017', 16, 18, 'Kent_24
4'), ('1069-017', 19, 22, 'C._Hooper_58'), ('1069-017', 23, 26, 'C._Tolley_4
-68'), ('1069-017', 27, 29, 'K._Evans'), ('1087-004', 4, 5, 'Jansher'), ('10
87-004', 17, 18, 'Hill'), ('1087-004', 21, 22, 'Pakistani'), ('1145-001', 0,
1, 'MANAMA'), ('1049-003', 1, 2, 'Irish'), ('1049-003', 6, 8, 'Mick_McCarthy
'), ('1049-003', 14, 16, '20_minutes'), ('1049-003', 18, 20, 'Andy_Townsend
'), ('1049-003', 24, 26, "Keith_O'Neill"), ('1049-003', 27, 28, 'Sunderland
'), ('1049-003', 29, 31, 'Niall_Quinn'), ('1049-003', 33, 35, 'Ian_Harte'),
('0986-006', 2, 3, 'Breen'), ('0986-006', 4, 5, 'Staunton'), ('0986-006', 6,
7, 'Irwin'), ('0986-006', 8, 9, 'McAteer'), ('0986-006', 10, 11, 'Harte'),
('0986-006', 12, 13, 'McLoughlin'), ('0986-006', 14, 15, 'Houghton'), ('0986
-006', 16, 17, 'Townsend'), ('0986-006', 18, 19, 'Quinn'), ('0986-006', 20,
21, "O'Neill"), ('1026-002', 3, 4, '26'), ('1026-002', 7, 8, 'Egypt'), ('102
6-002', 12, 17, "al-Gama'a_al-Islamiya_(_Islamic_Group"), ('1026-002', 23, 2
4, 'Friday'), ('1097-003', 5, 7, 'the_night'), ('1097-003', 8, 10, 'Stefan_E
dberg'), ('1097-003', 15, 16, '14th'), ('1097-003', 18, 20, 'U.S._Open'), ('
1097-003', 21, 23, 'Bernd_Karbacher'), ('1097-003', 31, 32, 'fourth'), ('109
7-003', 37, 38, 'Friday'), ('1022-000', 0, 1, 'Canadian'), ('1022-000', 7,
8, 'U.S.'), ('1039-007', 0, 1, 'Indonesian'), ('1039-007', 2, 3, 'Suharto'),
('1132-000', 0, 1, 'Mexican'), ('1132-000', 4, 5, 'Michoacan'), ('1124-002',
0, 2, 'South_African'), ('1124-002', 3, 4, 'Afrikaners'), ('1124-002', 5, 6,
'Saturday'), ('1124-002', 20, 22, 'Northern_Cape'), ('1060-016', 1, 3, 'Fran
k_Nobilo'), ('1060-016', 4, 6, 'New_Zealand'), ('1060-016', 7, 8, '209,412
'), ('1071-005', 0, 1, 'Panama'), ('1071-005', 7, 8, '50th'), ('1090-007',
4, 6, 'this_year'), ('1090-007', 10, 11, 'Americans'), ('1090-007', 23, 25,
'last_year'), ('1090-007', 28, 29, 'Spain'), ('1090-007', 30, 32, 'September
_28-29'), ('1090-007', 33, 35, 'Atlantic_City'), ('1125-007', 0, 3, 'Only_ab
out_100'), ('1125-007', 10, 11, 'Rwanda'), ('1125-007', 15, 16, '600'), ('11
25-007', 21, 22, 'weekly'), ('0966-079', 0, 1, '8.'), ('0966-079', 1, 3, 'Sa
lly_Barsosio'), ('0966-079', 4, 5, 'Kenya'), ('1011-012', 10, 11, 'Missouri
'), ('1011-012', 13, 17, 'First_Bank_System_Inc.'), ('1011-012', 18, 20, 'No
rwest_Corp.'), ('1011-012', 22, 23, 'KeyCorp'), ('1011-012', 24, 27, 'Banc_O
ne_Corp.'), ('1011-012', 29, 33, 'First_Chicago_NBD_Corp.'), ('1011-012', 3
4, 35, 'Illinois'), ('0946-004', 2, 3, 'Somerset'), ('0946-004', 5, 6, '83
'), ('0946-004', 7, 10, 'the_opening_morning'), ('0946-004', 11, 13, 'Grace_
Road'), ('0946-004', 14, 15, 'Leicestershire'), ('0946-004', 17, 18, 'first
'), ('0946-004', 20, 21, '94'), ('0946-004', 27, 28, '296'), ('0946-004', 2
9, 30, 'England'), ('0946-004', 31, 33, 'Andy_Caddick'), ('0946-004', 34, 3
5, 'three'), ('0946-004', 36, 37, '83'), ('1126-005', 0, 1, 'Marriner'), ('1
126-005', 3, 4, 'NATO'), ('1126-005', 14, 15, 'Tuzla'), ('1126-005', 16, 17,
'Mahala'), ('1143-003', 3, 4, 'five'), ('1143-003', 10, 11, 'Israel'), ('114
3-003', 12, 13, 'Lebanon'), ('1143-003', 14, 15, 'Syria'), ('1143-003', 16,
17, 'France'), ('1143-003', 18, 21, 'the_United_States'), ('1143-003', 25, 2
7, '11_a.m.'), ('1143-003', 32, 33, 'Naqoura'), ('1143-003', 38, 42, 'the_U.

N._Interim_Force'), ('1143-003', 43, 44, 'Lebanon'), ('1143-003', 45, 46, 'U
NIFIL'), ('1103-002', 0, 1, 'BRUSSELS'), ('1076-040', 0, 1, 'Spartak'), ('10
76-040', 1, 2, '3'), ('1076-040', 2, 3, '1'), ('1076-040', 5, 6, '3'), ('107
6-040', 6, 7, '3'), ('1076-040', 7, 8, '4'), ('0970-004', 0, 1, 'Lentini'),
('0970-004', 2, 3, '27'), ('0970-004', 5, 6, 'Milan'), ('0970-004', 7, 8, 'T
orino'), ('0970-004', 10, 13, '$_12_million'), ('1094-045', 0, 2, 'SAN_DIEGO
'), ('1094-045', 3, 4, '60'), ('1134-005', 0, 1, 'Jakarta'), ('1134-005', 1
8, 20, 'last_Friday'), ('0966-047', 0, 1, '3.'), ('0966-047', 1, 3, 'William
_Tanui'), ('0966-047', 4, 5, 'Kenya'), ('1014-010', 8, 9, 'Texas'), ('1014-0
10', 10, 12, 'Ross_Perot'), ('1014-010', 13, 16, 'the_Reform_Party'), ('1014
-010', 31, 36, 'the_Commission_on_Presidential_Debates'), ('1014-010', 48, 4
9, '1988'), ('1014-010', 50, 55, 'the_League_of_Women_Voters'), ('0990-005',
4, 6, 'Boris_Nikolayevich'), ('0990-005', 7, 8, 'Yeltsin'), ('0990-005', 9,
10, 'yesterday'), ('0966-154', 0, 1, '3.'), ('0966-154', 1, 3, 'Isel_Lopez
'), ('0966-154', 4, 5, 'Cuba'), ('0966-154', 6, 7, '65.10'), ('0961-008', 1,
4, 'his_11_years'), ('0961-008', 5, 6, 'Philadelphia'), ('0961-008', 7, 8, '
Randall'), ('0961-008', 13, 14, 'Eagles'), ('0961-008', 24, 25, 'Philadelphi
a'), ('0961-008', 30, 31, 'NFL'), ('1003-003', 0, 1, 'Brunswijk'), ('1003-00
3', 7, 9, '10_days'), ('1003-003', 10, 12, 'Freddy_Pinas'), ('1003-003', 18,
19, 'Netherlands'), ('1003-003', 21, 22, 'Brunswijk'), ('1003-003', 37, 39,
'56_miles'), ('1003-003', 40, 42, '90_km'), ('1003-003', 45, 46, 'Paramaribo
'), ('1066-042', 0, 2, 'Bristol_Rovers'), ('1066-042', 2, 4, '3_1'), ('1066-
042', 6, 7, '2'), ('1066-042', 8, 9, '5'), ('0953-012', 0, 1, 'Duran'), ('09
53-012', 6, 8, 'three_decades'), ('0953-012', 22, 25, "Puerto_Rico_'s"), ('1
066-070', 0, 1, 'Darlington'), ('1066-070', 1, 2, '4'), ('1066-070', 4, 5, '
2'), ('1066-070', 5, 6, '9'), ('1066-070', 6, 7, '8'), ('1066-070', 7, 8, '4
'), ('1103-009', 1, 3, 'Luis_Oliveira'), ('1103-009', 10, 12, 'Rustu_Recber
'), ('1103-009', 22, 24, 'seven_minutes'), ('1103-009', 28, 29, 'Turkey'),
('1103-009', 41, 43, 'Ogun_Temizkanoglu'), ('1103-009', 44, 46, 'De_Wilde'),
('1117-008', 5, 6, 'Belgium'), ('1117-008', 8, 9, 'two'), ('1117-008', 9, 1
0, 'eight-year-old'), ('1117-008', 14, 15, 'two'), ('1023-000', 0, 5, 'Aw_Co
mputer_Systems_Inc_Q2'), ('1012-006', 0, 4, 'The_U.S._Treasury_Department'),
('1012-006', 5, 6, 'Wednesday'), ('1012-006', 7, 8, 'Farrakhan'), ('1012-006
', 14, 15, '250,000'), ('1012-006', 18, 21, '$_1_billion'), ('1012-006', 22,
23, 'Libyan'), ('1012-006', 24, 26, 'Muammar_Gaddafi'), ('1012-006', 29, 30,
'Farrakhan'), ('1012-006', 33, 34, 'Islam'), ('1012-006', 38, 40, 'last_Janu
ary'), ('0972-029', 0, 2, 'H._Tillekeratne'), ('0972-029', 5, 6, '1'), ('102
5-007', 3, 4, 'Arabs'), ('1025-007', 15, 16, 'Israeli'), ('0984-021', 0, 2,
'Next_week'), ('0984-021', 2, 5, 'Kansai_Electric_Power'), ('0984-021', 6,
9, 'Kansai_International_Airport'), ('1139-002', 0, 1, 'Chinese'), ('1139-00
2', 5, 7, 'Wang_Donghai'), ('1139-002', 9, 11, 'New_York-based'), ('1139-002
', 13, 15, 'Human_Rights'), ('1139-002', 16, 17, 'China'), ('1139-002', 19,
20, 'Saturday'), ('1150-001', 0, 1, 'ROCHESTER'), ('1150-001', 2, 3, 'N.
H.'), ('1066-057', 0, 3, 'Fulham_4_3'), ('1066-057', 3, 5, '0_1'), ('1066-05
7', 5, 6, '5'), ('1066-057', 6, 7, '3'), ('1066-057', 7, 8, '9'), ('1152-012
', 5, 6, 'Iran'), ('1152-012', 11, 12, 'Iraq'), ('1152-012', 15, 16, 'Iraq
'), ('1012-007', 0, 1, 'Farrakhan'), ('1012-007', 2, 5, "last_October_'s"),
('1012-007', 10, 11, 'thousands'), ('1012-007', 15, 16, 'Washington'), ('096
6-015', 0, 1, '4.'), ('0966-015', 1, 3, 'Yekaterina_Podkopayeva'), ('0966-01
5', 4, 5, 'Russia'), ('0953-002', 0, 1, 'Panamanian'), ('0953-002', 3, 4, 'R
oberto'), ('0953-002', 5, 8, 'Hands_of_Stone'), ('0953-002', 9, 10, 'Duran
'), ('0953-002', 15, 16, 'Saturday'), ('1015-006', 2, 3, 'Titanic'), ('1015-
006', 13, 17, 'April_14_,_1912'), ('1015-006', 22, 23, '1,523'), ('1015-006
', 25, 26, '2,200'), ('1034-004', 1, 2, 'Dutch'), ('1034-004', 13, 14, 'Dutc
h'), ('1034-004', 18, 21, 'May_this_year'), ('1037-002', 0, 1, 'Swedish'),
('1037-002', 3, 6, 'LM_Ericsson_AB'), ('1037-002', 8, 9, 'Friday'), ('1037-0
02', 14, 16, '1.2_billion'), ('1037-002', 25, 26, 'Guangdong'), ('1037-002',
28, 29, 'China'), ('1019-003', 4, 5, 'Nicole'), ('1019-003', 9, 10, 'Groning
en'), ('1019-003', 15, 16, 'Dutch'), ('1130-001', 0, 1, 'HAVANA'), ('1119-00
0', 0, 1, 'Britain'), ('1119-000', 2, 3, 'Iraq'), ('1119-000', 5, 6, 'Arbil
'), ('1096-019', 1, 2, 'Chicago'), ('1096-019', 4, 5, 'Braves'), ('1096-019
', 6, 7, 'Cubs'), ('1160-011', 17, 19, 'Mao_Tse-Tung'), ('1160-011', 20, 21,

```
'China'), ('1160-011', 22, 24, 'Yitzak_Shamir'), ('1160-011', 25, 26, 'Israe
l'), ('1160-011', 28, 29, 'Jordan'), ('1116-038', 0, 1, '1994'), ('1116-038
', 2, 6, 'The_Stars_and_Stripes'), ('1116-038', 14, 15, 'U.S.'), ('1116-038
', 18, 19, 'Berlin'), ('1116-038', 23, 24, 'American'), ('1116-038', 30, 34,
'nearly_half_a_century'), ('0983-004', 1, 2, 'Iraqis'), ('0983-004', 13, 14,
'Saddam'), ('0983-004', 15, 16, 'Hussein'), ('0983-004', 29, 30, 'seven'),
('0983-004', 39, 40, 'asylum'), ('1055-037', 5, 7, 'Mullally_6-0-30-1'), ('1
055-037', 8, 9, 'Headley'), ('1069-018', 1, 2, 'Bristol'), ('1069-018', 4,
5, '183'), ('1069-018', 6, 7, '249'), ('1069-018', 8, 11, 'J._Russell_75'),
('1069-018', 13, 15, 'Northamptonshire_190'), ('1069-018', 16, 17, '218-9'),
('1118-002', 0, 3, 'British_Labour_Party'), ('1118-002', 4, 6, 'Tony_Blair
'), ('1118-002', 11, 12, 'Saturday'), ('1118-002', 16, 17, 'Scottish'), ('11
18-002', 19, 20, '21-18'), ('1118-002', 33, 34, 'Scotland'), ('1025-009', 6,
7, 'Israel'), ('1025-009', 18, 19, 'Israeli'), ('1037-004', 3, 4, 'AXE'), ('
1037-004', 7, 8, 'ISDN'), ('1037-004', 10, 12, 'Intelligent_Network'), ('103
7-004', 27, 28, 'Ericsson'), ('1037-004', 29, 31, 'Per_Zetterquist'), ('1037
-004', 32, 33, 'Reuters'), ('1103-004', 0, 1, 'Turkish'), ('1103-004', 14, 1
5, 'dozens'), ('1070-016', 0, 4, '10._Gil_de_Ferran'), ('1070-016', 5, 6, 'B
razil'), ('1070-016', 8, 10, 'Reynard_Honda'), ('1070-016', 11, 12, '54.774
'), ('1100-002', 0, 1, 'American'), ('1100-002', 1, 3, 'Tom_Johnson'), ('110
0-002', 6, 7, 'IBF'), ('1100-002', 17, 18, 'Venezuela'), ('1100-002', 19, 2
1, 'Ramon_Guzman'), ('1100-002', 22, 23, 'Saturday'), ('1101-005', 5, 8, 'th
e_second_half'), ('1101-005', 9, 10, 'France'), ('1101-005', 13, 14, 'Mexica
ns'), ('1101-005', 15, 16, 'three'), ('1012-006', 0, 4, 'The_U.S._Treasury_D
epartment'), ('1012-006', 5, 6, 'Wednesday'), ('1012-006', 7, 8, 'Farrakhan
'), ('1012-006', 14, 15, '250,000'), ('1012-006', 18, 21, '$_1_billion'), ('
1012-006', 22, 23, 'Libyan'), ('1012-006', 24, 26, 'Muammar_Gaddafi'), ('101
2-006', 29, 30, 'Farrakhan'), ('1012-006', 33, 34, 'Islam'), ('1012-006', 3
8, 40, 'last_January'), ('1116-037', 4, 5, 'Soviet'), ('1116-037', 7, 8, 'Ru
ssia'), ('1116-037', 9, 10, 'Belarus'), ('1116-037', 11, 12, 'Kazakhstan'),
('1116-037', 13, 14, 'Uzbekistan'), ('1116-037', 15, 16, 'Armenia'), ('1116-
037', 17, 18, 'Tajikistan'), ('1116-037', 25, 26, 'Russian'), ('1027-002',
2, 4, 'West_Bank'), ('1045-003', 7, 8, 'Palestinian'), ('1045-003', 8, 9, '
U.N.'), ('1045-003', 12, 14, 'Security_Council'), ('1045-003', 15, 17, 'Tono
_Eitel'), ('1045-003', 18, 19, 'Germany'), ('1045-003', 29, 30, 'Israel'),
('1045-003', 34, 36, 'David_Peleg'), ('1155-005', 5, 6, 'German'), ('1155-00
5', 14, 15, 'Tehran'), ('1071-000', 0, 4, 'SOCCER_-_CANADA_BEAT'), ('1087-02
0', 5, 6, 'Jansher'), ('1087-020', 19, 20, 'Eyles'), ('1089-026', 1, 2, 'Wal
```

In [91]:
```
evaluation_report(set(gold_mentions(df_dev)),set(naive_pred))
```

```
20.426788350556972 25.7224945073517 22.77079593058049
```

Problem 5: Extending the training data using the knowledge base

State-of-the-art approaches to entity linking exploit information in knowledge bases. In our case, where Wikipedia is the knowledge base, one particularly useful type of information are links to other Wikipedia pages. In particular, we can interpret the anchor texts (the highlighted texts that you click on) as mentions of the entities (pages) that they link to. This allows us to harvest long lists of mention–entity pairings.

The following cell loads a data frame summarizing anchor texts and page references harvested from the first paragraphs of the English Wikipedia. The data frame also contains all entity mentions in the training data (but not the development or the test data).

In [92]:
```
with bz2.open('kb.tsv.bz2', 'rt',encoding="utf-8") as source:
    df_kb = pd.read_csv(source, sep='\t', quoting=csv.QUOTE_NONE)
```

To understand what information is available in this data, the following cell shows the entry for the anchor text `Sweden`.

In [93]:
```
df_kb.loc[df_kb.mention == 'Sweden']
```

Out[93]:

	mention	entity	prob
17436	Sweden	Sweden	0.985768
17437	Sweden	Sweden_national_football_team	0.014173
17438	Sweden	Sweden_men's_national_ice_hockey_team	0.000059

As you can see, each row of the data frame contains a pair (m, e) of a mention m and an entity e, as well as the conditional probability $P(e|m)$ for mention m referring to entity e. These probabilities were estimated based on the frequencies of mention–entity pairs in the knowledge base. The example shows that the anchor text 'Sweden' is most often used to refer to the entity Sweden, but in a few cases also to refer to Sweden's national football and ice hockey teams. Note that references are sorted in decreasing order of probability, so that the most probable pairing come first.

Implement an entity linking method that resolves each mention to the most probable entity in the data frame. If the mention is not included in the data frame, you can predict the generic label `--NME--`. Print the precision, recall, and F1 of your method using the function that you implemented for Problem 1.

In [94]:

```
# TODO: Write code here to implement the "most probable entity" method.
wiki_pred=[]
# for x in pred_df.itertuples():
#     doc = nlp(x[2])
#     for ent in doc.ents:
#         tmp_str=[]
#         for i in range(ent.start,ent.end):
#             tmp_str.append(doc[i].text)
#         tmp_str=" ".join(tmp_str)
#         tmp_df=df_kb.loc[df_kb.mention == tmp_str]
#         if len(tmp_df)!=0:
#             wiki_pred.append((x[1],ent.start,ent.end,tmp_df.iloc[0].entity
#         else:
#             wiki_pred.append((x[1],ent.start,ent.end,"--NME--"))

for x in pred_df.itertuples():
    doc=nlp(x[2])
    tmp_str=str(doc[x[3]:x[4]])
    tmp_df=df_kb.loc[df_kb.mention==tmp_str]
    if len(tmp_df)!=0:
        new_ent=(x[1],x[3],x[4],tmp_df.iloc[0].entity)
    else:
        new_ent=(x[1],x[3],x[4],"--NME--")
    wiki_pred.append(new_ent)


print(wiki_pred)
```

```
'), ('1146-006', 7, 8, 'Saddam_Hussein'), ('1152-004', 49, 52, 'United_State
s'), ('0946-014', 9, 11, 'Paul_Johnson_(squash_player)'), ('0974-011', 3, 5,
'Muttiah_Muralitharan'), ('1039-011', 29, 30, 'Moro_National_Liberation_Fron
t'), ('1097-001', 0, 2, '--NME--'), ('1132-004', 0, 2, '--NME--'), ('1141-00
8', 0, 1, 'Guangzhou'), ('1151-000', 6, 7, 'Texas'), ('0984-008', 15, 16,
'--NME--'), ('1111-005', 1, 4, '--NME--'), ('1103-020', 17, 18, 'Italy'), ('
1144-002', 35, 36, '--NME--'), ('1152-011', 21, 22, 'Baghdad'), ('1160-010',
27, 28, 'HarperCollins'), ('0969-001', 0, 1, 'OGC_Nice'), ('0965-009', 16, 1
9, '--NME--'), ('0966-057', 1, 3, 'Vladimir_Dubrovshchik'), ('1056-012', 0,
3, '--NME--'), ('0953-003', 7, 8, 'Mexico'), ('0972-036', 9, 10, 'Glenn_McGr
ath'), ('1043-004', 0, 1, 'Singapore'), ('0957-007', 6, 8, 'Doug_Flach'), ('
0965-003', 8, 10, 'Donovan_Bailey'), ('0966-083', 4, 5, '--NME--'), ('1016-0
07', 22, 25, '--NME--'), ('1031-016', 9, 10, 'China'), ('1053-003', 19, 22,
'--NME--'), ('1063-002', 14, 15, 'Europe'), ('1112-008', 1, 3, 'Berti_Vogts
'), ('1137-003', 10, 12, 'North_Korea'), ('1146-005', 50, 51, 'Jalal_Talaban
i'), ('1063-001', 0, 1, '--NME--'), ('1094-028', 2, 3, 'Texas_Rangers_(baseb
all)'), ('0946-006', 0, 1, 'Essex'), ('1155-011', 3, 4, 'Germany'), ('0957-0
06', 11, 12, 'Australia'), ('1089-001', 2, 3, 'Wisconsin'), ('0966-087', 1,
3, 'Rohan_Robinson'), ('0948-003', 4, 5, 'England'), ('1056-009', 4, 5, 'Uni
ted_Kingdom'), ('0955-009', 0, 1, '--NME--'), ('1030-040', 1, 2, 'Tokyo'),
('1058-003', 6, 8, '--NME--'), ('1142-005', 17, 18, 'Kurdistan_Democratic_Pa
rty'), ('1152-017', 3, 4, 'United_States_Marine_Corps'), ('0984-006', 29, 3
0, 'Missouri'), ('1056-043', 3, 4, 'France'), ('1058-009', 15, 16, 'Texas'),
('0981-000', 0, 3, '--NME--'), ('1017-002', 19, 21, 'Air_France'), ('0957-01
2', 0, 2, 'Alexander_Volkov_(tennis)'), ('0957-020', 35, 36, 'Australia'),
('1075-002', 0, 1, 'Mauritania'), ('0955-014', 0, 1, 'FC_Seoul'), ('1018-004
', 14, 15, 'Berlin'), ('1086-002', 0, 1, 'England'), ('1099-008', 17, 18, 'D
uncan_Ferguson'), ('1116-035', 8, 9, '--NME--'), ('1152-003', 29, 30, 'Iraq
'), ('0977-008', 0, 1, 'Canada'), ('1139-002', 9, 11, '--NME--'), ('1142-002
', 16, 17, 'Baghdad'), ('0968-004', 24, 25, 'Bosnia_and_Herzegovina'), ('099
2-002', 3, 4, 'Islam'), ('1056-040', 0, 2, 'Marty_Nothstein'), ('1057-003',
```

13, 15, 'Graham_Lloyd'), ('1090-010', 21, 22, 'Nagoya'), ('1016-014', 7, 9, 'Yitzhak_Mordechai'), ('1051-009', 32, 34, 'Klas_Eriksson'), ('0997-004', 14, 15, 'Russia'), ('1031-007', 35, 37, '--NME--'), ('0972-005', 0, 3, '--NME--'), ('1117-003', 0, 2, '--NME--'), ('1139-003', 2, 3, '--NME--'), ('0992-000', 1, 2, 'Islam'), ('0991-002', 10, 12, 'Aslan_Maskhadov'), ('1009-014', 0, 1, 'Mexico'), ('1056-017', 16, 17, 'Germany'), ('1109-007', 7, 8, '--NME--'), ('1125-005', 19, 20, 'Rwanda'), ('1047-000', 8, 9, '--NME--'), ('1099-009', 9, 11, 'Craig_Brown'), ('1146-007', 17, 18, 'Iraq'), ('0972-015', 0, 2, '--NME--'), ('0999-002', 1, 2, 'Moscow'), ('1099-018', 14, 16, 'Colin_Hendry'), ('0994-007', 0, 1, 'Alexander_Lukashenko'), ('0966-068', 4, 5, 'The_Bahamas'), ('1160-012', 1, 2, 'HarperCollins'), ('0966-146', 4, 5, '--NME--'), ('1145-002', 10, 11, 'Bahrain'), ('1156-002', 0, 1, 'Italy'), ('0956-002', 2, 4, 'South_Korea'), ('1066-047', 0, 2, 'Notts_County_F.C.'), ('0963-003', 12, 14, 'Billy_Andrade'), ('1012-009', 10, 11, 'Libya'), ('1154-006', 16, 18, '--NME--'), ('1027-009', 9, 12, '--NME--'), ('1092-004', 7, 8, 'Germany'), ('1084-002', 0, 1, 'Sweden'), ('1101-008', 33, 35, '--NME--'), ('1033-007', 10, 11, 'Switzerland'), ('1009-001', 0, 2, '--NME--'), ('1009-011', 0, 1, 'Mexico'), ('0947-011', 19, 21, '--NME--'), ('0949-005', 24, 25, 'Southampton'), ('1058-007', 10, 13, '--NME--'), ('0966-081', 1, 3, 'Torrance_Zellner'), ('1033-010', 6, 7, 'Chechnya'), ('0972-014', 0, 3, '--NME--'), ('0976-002', 0, 5, '--NME--'), ('1033-009', 0, 1, '--NME--'), ('1044-004', 30, 31, 'Arabs'), ('1068-026', 4, 5, 'Wrexham_A.F.C.'), ('1089-005', 0, 2, '--NME--'), ('1058-019', 5, 6, 'New_York_Yankees'), ('1072-019', 67, 69, 'Robin_Brooke'), ('1044-004', 11, 12, 'State_of_Palestine'), ('1117-004', 7, 8, 'Netherlands'), ('0957-026', 3, 4, 'Ukraine'), ('1108-000', 0, 3, '--NME--'), ('1077-003', 16, 17, 'England'), ('1155-010', 31, 32, 'Berlin'), ('0959-012', 0, 1, 'Cincinnati'), ('0960-021', 11, 12, 'Jeff_Tarango'), ('0960-039', 7, 8, 'Jeff_Tarango'), ('1008-018', 1, 3, '--NME--'), ('1056-034', 21, 23, 'Philippe_Ermenault'), ('1076-010', 0, 4, '--NME--'), ('1099-017', 35, 38, '--NME--'), ('1010-021', 20, 22, '--NME--'), ('1056-030', 1, 3, '--NME--'), ('1103-017', 0, 1, 'Belgium'), ('1126-002', 33, 34, 'Serbs'), ('1095-005', 0, 1, 'Detroit_Tigers'), ('0955-012', 0, 1, '--NME--'), ('1056-043', 7, 9, '--NME--'), ('1136-002', 19, 20, 'Tianjin'), ('0961-004', 6, 7, '--NME--'), ('1099-018', 23, 25, 'Stuart_McCall'), ('1006-007', 5, 6, 'Niger'), ('0963-017', 23, 25, 'PGA_Tour'), ('1036-014', 3, 4, 'Norway'), ('0966-084', 1, 3, 'Fabrizio_Mori'), ('1056-031', 1, 3, '--NME--'), ('1121-001', 0, 1, 'London'), ('1142-006', 20, 21, 'Kurdistan_Democratic_Party'), ('1068-013', 6, 8, 'Manchester_City_F.C.'), ('0993-006', 14, 19, '--NME--'), ('1064-005', 28, 29, 'Gloucester_Rugby'), ('1056-021', 1, 3, 'Annett_Neumann'), ('1099-004', 0, 1, 'Scotland'), ('1049-004', 5, 6, 'Ireland'), ('0946-001', 0, 1, 'London'), ('0962-010', 10, 12, 'Ken_Green_(golfer)'), ('0966-034', 1, 3, 'Patrick_Stevens'), ('0966-143', 4, 5, 'United_Kingdom'), ('1051-008', 1, 3, 'Joakim_Haeggman'), ('1094-036', 0, 2, 'New_York_City'), ('1072-018', 49, 51, 'Gary_Teichmann'), ('0960-007', 9, 10, 'Andre_Agassi'), ('1051-011', 15, 17, 'Peter_Hedblom'), ('0949-010', 2, 4, 'UEFA_Euro_1996'), ('1133-004', 33, 35, 'Deng_Xiaoping'), ('1133-011', 0, 1, '--NME--'), ('1116-006', 2, 3, 'France'), ('1028-002', 8, 9, 'Michigan'), ('0962-009', 25, 27, '--NME--'), ('1115-004', 6, 8, '--NME--'), ('1035-008', 12, 13, 'Armenia'), ('0966-048', 4, 5, 'Kenya'), ('0980-013', 19, 22, '--NME--'), ('1054-023', 5, 6, 'Netherlands'), ('1068-021', 0, 4, '--NME--'), ('0962-010', 22, 24, '--NME--'), ('0990-014', 15, 16, 'Alexander_Lebed'), ('1058-001', 0, 1, '--NME--'), ('1103-020', 7, 9, '--NME--'), ('1070-014', 4, 5, 'Brazil'), ('0997-007', 0, 1, 'Boris_Yeltsin'), ('1128-011', 12, 13, '--NME--'), ('1096-034', 14, 15, 'Reserve_Bank_of_India'), ('0996-003', 1, 2, 'Dariusz_Rosati'), ('1112-004', 4, 5, 'Germany'), ('1102-003', 9, 11, 'Youri_Djorkaeff'), ('1143-003', 16, 17, 'France'), ('1042-003', 5, 7, '--NME--'), ('1016-007', 5, 6, 'Israel'), ('1039-014', 21, 23, 'Tanjung_Priok'), ('1045-003', 7, 8, 'Palestinian_territories'), ('1058-040', 3, 5, 'Marc_Newfield'), ('0966-021', 4, 5, '--NME--'), ('0957-019', 7, 9, 'Linda_Wild'), ('1045-009', 3, 4, 'Israel'), ('1094-055', 0, 2, 'San_Diego_Padres'), ('0966-036', 1, 3, '--NME--'), ('0980-014', 8, 9, '--NME--'), ('1033-013', 17, 18, '--NME--'), ('0949-004', 2, 3, '--NME--'), ('1056-028', 4, 6, 'New_Zealand'), ('1065-023', 0, 1, 'Brechin_City_F.C.'), ('1069-010', 15, 18, '--NM

E--'), ('1054-008', 4, 5, 'Netherlands'), ('0992-003', 25, 27, '--NME--'), ('0966-099', 1, 3, 'Dennis_Mitchell'), ('1036-013', 18, 19, 'Pyramiden'), ('1070-013', 1, 4, 'Al_Unser_Jr.'), ('0961-010', 0, 1, '--NME--'), ('1057-007', 31, 33, 'Waqar_Younis'), ('0946-008', 3, 4, 'Yorkshire'), ('1160-016', 9, 10, 'Dublin'), ('1072-019', 8, 10, '--NME--'), ('1090-007', 28, 29, 'Spain'), ('0954-004', 5, 6, 'Scotland'), ('0963-017', 32, 34, '--NME--'), ('0957-033', 15, 17, 'Andrea_Gaudenzi'), ('0946-010', 7, 9, 'Mark_Butcher'), ('0957-026', 0, 2, 'Andriy_Medvedev'), ('1072-013', 11, 13, 'Sean_Fitzpatrick'), ('1087-017', 13, 14, 'Australia'), ('0963-023', 8, 11, '--NME--'), ('0996-007', 0, 1, 'Poland'), ('1160-022', 16, 17, 'HarperCollins'), ('1010-002', 0, 1, 'Chicago'), ('0972-008', 3, 4, '--NME--'), ('1100-002', 1, 3, '--NME--'), ('1026-000', 5, 6, 'Islam'), ('0972-002', 10, 11, 'Australia'), ('1059-005', 8, 9, 'Surrey'), ('1155-009', 21, 23, 'Abolhassan_Banisadr'), ('1127-018', 0, 2, '--NME--'), ('1051-008', 14, 15, 'France'), ('0957-007', 0, 2, 'Tim_Henman'), ('0966-074', 1, 3, 'Rose_Cheruiyot'), ('0966-149', 1, 3, '--NME--'), ('0958-054', 2, 3, 'Pittsburgh_Pirates'), ('0974-000', 0, 3, '--NME--'), ('1011-022', 4, 5, 'NationsBank'), ('1035-010', 36, 37, 'Finland'), ('1116-014', 10, 13, '--NME--'), ('1160-007', 28, 30, 'Warner_Bros.'), ('1056-009', 1, 3, '--NME--'), ('1107-002', 6, 9, '--NME--'), ('1116-025', 4, 5, 'Baudouin_of_Belgium'), ('1094-025', 0, 1, 'Baltimore_Orioles'), ('1031-003', 0, 1, 'Italy'), ('0984-002', 11, 12, 'Europe'), ('0983-004', 1, 2, 'Iraq'), ('1055-020', 2, 4, '--NME--'), ('1116-024', 11, 12, '--NME--'), ('0999-003', 0, 1, 'Interfax'), ('1148-002', 7, 8, 'Tunisia'), ('0988-000', 0, 2, '--NME--'), ('0957-017', 3, 5, 'South_Africa'), ('1012-002', 0, 3, '--NME--'), ('1056-038', 3, 4, 'Australia'), ('1058-036', 1, 2, 'Detroit'), ('0983-002', 17, 18, 'London'), ('1087-018', 14, 15, 'Scotland'), ('1156-002', 3, 5, 'Lamberto_Dini'), ('1116-037', 9, 10, 'Belarus'), ('1035-006', 17, 18, '--NME--'), ('1027-023', 1, 3, '--NME--'), ('1057-007', 0, 1, 'Pakistan'), ('1057-000', 0, 1, '--NME--'), ('1056-026', 1, 2, '--NME--'), ('1152-000', 5, 6, 'Persian_Gulf'), ('1010-016', 7, 8, 'Italy'), ('1030-039', 2, 3, 'London'), ('1121-004', 20, 21, 'Erbil'), ('1014-009', 13, 14, 'Bob_Dole'), ('1141-007', 11, 12, 'Guangdong'), ('1006-009', 20, 23, '--NME--'), ('1056-040', 6, 8, 'Pavel_Buráň'), ('1068-001', 0, 1, 'London'), ('0958-052', 0, 2, 'San_Diego_Padres'), ('1062-002', 0, 1, 'Wales'), ('1096-030', 0, 1, 'Montreal'), ('1036-034', 0, 1, 'Norway'), ('1122-003', 5, 9, '--NME--'), ('1142-011', 6, 10, 'Patriotic_Union_of_Kurdistan'), ('0967-002', 0, 1, 'UEFA'), ('0951-005', 0, 4, '--NME--'), ('1096-020', 5, 7, 'Ryne_Sandberg'), ('1066-025', 0, 1, 'Oldham_Athletic_A.F.C.'), ('1091-003', 36, 38, 'Judith_Wiesner'), ('1072-015', 26, 27, '--NME--'), ('1116-029', 6, 8, 'Jackie_Stewart'), ('0949-001', 0, 1, 'London'), ('1127-003', 3, 4, 'Chechens'), ('0962-005', 4, 5, '--NME--'), ('1056-046', 0, 2, '--NME--'), ('1061-002', 4, 5, 'Scotland'), ('1029-002', 0, 1, 'Washington,_D.C.'), ('1069-013', 9, 12, '--NME--'), ('1152-012', 5, 6, 'Iran'), ('0949-003', 20, 22, '--NME--'), ('1012-007', 15, 16, 'Washington,_D.C.'), ('1121-005', 19, 20, 'Erbil'), ('1152-011', 18, 19, 'Kurds'), ('1033-005', 16, 17, '--NME--'), ('0946-009', 13, 14, 'England'), ('0947-002', 7, 8, 'England'), ('0952-002', 0, 2, 'FC_Rotor_Volgograd'), ('0959-007', 0, 1, 'Los_Angeles_Angels'), ('1033-005', 7, 8, '--NME--'), ('1148-008', 4, 5, 'Tunisia'), ('0952-006', 26, 27, 'Moscow'), ('1132-003', 0, 2, '--NME--'), ('1066-066', 0, 1, 'Lincoln,_Nebraska'), ('1060-018', 1, 3, 'Pádraig_Harrington'), ('1128-000', 5, 6, '--NME--'), ('0966-126', 1, 3, 'Inha_Babakova'), ('1122-002', 18, 19, 'Algeria'), ('1088-002', 21, 22, 'Pakistan'), ('1144-004', 25, 26, 'Reuters'), ('0984-003', 16, 17, '--NME--'), ('0960-019', 41, 42, 'Marcelo_Ríos'), ('1037-002', 3, 6, '--NME--'), ('0988-005', 17, 19, '--NME--'), ('1038-007', 6, 7, '--NME--'), ('1084-003', 7, 9, '--NME--'), ('0982-000', 5, 6, 'Chicago'), ('0997-004', 27, 28, 'Moscow'), ('1053-001', 0, 1, '--NME--'), ('1096-012', 28, 29, '--NME--'), ('1112-001', 0, 1, 'Bonn'), ('1128-006', 19, 20, '--NME--'), ('1027-017', 27, 29, '--NME--'), ('1088-002', 27, 28, 'Australia'), ('1156-007', 2, 3, 'Rome'), ('1045-007', 27, 30, '--NME--'), ('1058-008', 1, 2, 'Native_Americans_in_the_United_States'), ('0966-128', 4, 5, 'Russia'), ('0968-004', 5, 6, 'A._G._Edwards'), ('1033-010', 27, 28, 'Russia'), ('1036-011', 22, 23, 'Reuters'), ('1073-010', 0, 2, 'New_Zealand'), ('1089-004', 7, 9, 'Loren_Roberts'), ('1113-001', 0, 1, 'Bonn'), ('104

4-005', 2, 3, 'Arab_citizens_of_Israel'), ('1116-035', 2, 3, 'United_Kingdom
'), ('1142-002', 26, 27, 'Reuters'), ('1002-000', 4, 5, 'Colombia'), ('0996-
002', 0, 1, 'Poland'), ('1084-001', 0, 1, '--NME--'), ('1045-002', 8, 9, 'Is
rael'), ('1058-022', 3, 6, '--NME--'), ('0966-117', 1, 3, 'Igor_Trandenkov
'), ('0991-008', 21, 22, 'Chechnya'), ('1079-007', 8, 9, 'Yugoslavia'), ('10
76-031', 0, 1, '--NME--'), ('1051-010', 35, 36, 'Australia'), ('1142-005', 2
0, 21, 'Reuters'), ('1016-010', 8, 10, 'David_Levy_(Israeli_politician)'),
('0948-033', 3, 5, 'British_Universities_cricket_team'), ('0958-012', 0, 2,
'National_League_Central'), ('1072-014', 18, 19, 'South_Africa_national_rugb
y_union_team'), ('1044-006', 0, 1, 'Israel'), ('0999-003', 22, 24, '--NME
--'), ('1069-018', 1, 2, 'Bristol'), ('1111-009', 2, 4, '--NME--'), ('1160-0
07', 9, 10, 'United_Kingdom'), ('0963-017', 13, 15, 'Andrew_Magee'), ('0983-
005', 1, 2, 'England'), ('0984-010', 16, 17, 'United_States_Treasury_securit
y'), ('1079-005', 2, 3, 'Italy'), ('0946-007', 5, 6, 'England'), ('1017-004
', 5, 6, 'Africa'), ('1077-002', 20, 23, '--NME--'), ('1146-006', 31, 32, 'I
raq'), ('1155-014', 5, 6, 'Iran'), ('1138-007', 19, 22, '--NME--'), ('0984-0
17', 25, 26, '--NME--'), ('0950-004', 3, 4, 'Yugoslavia'), ('0997-000', 0,
1, 'Boris_Yeltsin'), ('1068-013', 2, 3, 'Birmingham'), ('0998-004', 3, 4, 'A
slan_Maskhadov'), ('1135-027', 6, 7, 'National_League_for_Democracy'), ('108
4-002', 10, 11, 'Europe'), ('1060-012', 1, 3, 'Wayne_Riley'), ('0966-113',
4, 5, 'Norway'), ('1037-000', 5, 7, '--NME--'), ('1085-003', 2, 4, '--NME
--'), ('1158-002', 17, 18, 'Belgium'), ('1051-010', 18, 19, 'Australia'), ('
1128-012', 9, 10, 'Poland'), ('0949-005', 22, 23, 'Blackburn_Rovers_F.C.'),
('0966-037', 1, 3, 'Iván_García'), ('1056-044', 0, 2, '--NME--'), ('1074-003
', 31, 34, '--NME--'), ('1012-005', 15, 16, 'Libya'), ('1029-017', 3, 7, '--
NME--'), ('1056-034', 11, 13, '--NME--'), ('1078-005', 7, 8, 'Russia'), ('10
90-014', 10, 12, 'Fed_Cup'), ('0990-014', 22, 23, 'Chechnya'), ('1126-009',
9, 10, 'Serbs'), ('0985-003', 6, 7, '--NME--'), ('1161-005', 0, 3, '--NME
--'), ('1056-017', 11, 13, 'Francis_Moreau'), ('1155-007', 0, 1, 'Iran'), ('
1056-010', 4, 5, 'Russia'), ('1116-010', 2, 3, 'Moscow'), ('0953-013', 8, 9,
'Panama'), ('1133-003', 0, 2, 'Xiao_Qiang'), ('1131-002', 24, 25, 'Chile'),
('1137-003', 7, 8, '--NME--'), ('1095-015', 0, 1, 'Miami_Marlins'), ('0974-0
11', 0, 2, 'Chaminda_Vaas'), ('0986-003', 0, 1, 'Birmingham'), ('1051-010',
25, 27, 'Mark_Roe'), ('0948-022', 7, 9, 'The_Oval'), ('1112-002', 3, 5, 'Jür
gen_Klinsmann'), ('1154-001', 0, 1, 'Paris'), ('0948-009', 0, 1, 'England'),
('0985-011', 1, 2, '--NME--'), ('0990-003', 11, 12, 'Interfax'), ('1130-006
', 3, 8, '--NME--'), ('0966-024', 4, 5, 'Cuba'), ('0967-003', 17, 19, 'Guy_H
ellers'), ('1007-010', 6, 7, 'Bob_Dole'), ('0961-007', 1, 2, '--NME--'), ('1
072-018', 26, 29, '--NME--'), ('1137-008', 32, 35, '--NME--'), ('1135-006',
18, 21, '--NME--'), ('1128-001', 0, 1, 'Warsaw'), ('1130-006', 17, 18, 'Unit
ed_States'), ('1126-010', 24, 25, 'NATO'), ('1116-011', 14, 15, 'Brazil'),
('0970-002', 30, 31, 'Atalanta_B.C.'), ('1006-009', 27, 28, 'Reuters'), ('11
51-001', 0, 1, 'Dallas'), ('1054-001', 2, 3, 'Netherlands'), ('0992-007', 0,
5, '--NME--'), ('0948-006', 0, 1, 'Australia'), ('0984-019', 0, 1, '--NME
--'), ('1103-016', 73, 75, '--NME--'), ('1058-031', 29, 33, '--NME--'), ('11
16-008', 4, 5, 'Strasbourg'), ('1121-000', 7, 8, 'Iraq'), ('1137-002', 6, 8,
'South_Korea'), ('1152-018', 22, 23, 'Persian_Gulf'), ('1137-003', 21, 23, '
International_Committee_of_the_Red_Cross'), ('1128-003', 37, 38, '--NME--'),
('1134-000', 2, 3, 'Indonesia'), ('1149-005', 31, 33, 'Interstate_95'), ('10
96-024', 8, 9, 'Atlanta'), ('1096-013', 34, 35, 'Philadelphia_Phillies'), ('
1131-007', 5, 6, 'Honduras'), ('0966-154', 4, 5, 'Cuba'), ('1072-011', 25, 2
6, 'South_Africa_national_rugby_union_team'), ('1116-006', 9, 10, 'Turin'),
('1046-006', 19, 21, '--NME--'), ('1116-006', 20, 21, 'France'), ('1072-012
', 36, 37, '--NME--'), ('0957-026', 6, 8, 'Ján_Krošlák'), ('1072-017', 5, 6,
'Andrew_Mehrtens'), ('1058-026', 5, 8, '--NME--'), ('1094-034', 0, 1, 'Montr
eal_Expos'), ('1128-011', 1, 2, 'Poland'), ('1121-003', 16, 17, 'Iraq'), ('0
957-009', 6, 8, 'David_Rikl'), ('0991-011', 14, 15, 'Alexander_Lebed'), ('09
55-013', 0, 1, 'Ulsan'), ('0982-002', 21, 22, 'Chicago'), ('1160-011', 17, 1
9, '--NME--'), ('1003-005', 5, 6, '--NME--'), ('1112-004', 18, 19, 'Europe
'), ('0963-017', 0, 1, '--NME--'), ('1059-002', 30, 32, 'England'), ('1103-0
06', 10, 12, '--NME--'), ('1121-003', 29, 30, 'Erbil'), ('1062-002', 2, 4, '
San_Marino'), ('0960-003', 0, 2, 'Andre_Agassi'), ('0982-002', 0, 1, 'United

```
_Kingdom'), ('1133-006', 18, 19, 'Metro_Manila'), ('1157-002', 0, 6, '--NME
--'), ('1027-006', 14, 15, 'Hebrew_language'), ('1057-007', 16, 18, 'Wasim_A
kram'), ('1075-001', 0, 1, '--NME--'), ('1001-000', 0, 1, 'Romania'), ('0958
-002', 0, 3, 'Major_League_Baseball'), ('0953-004', 26, 27, '--NME--'), ('10
73-007', 0, 4, '--NME--'), ('1037-002', 28, 29, 'China'), ('1153-003', 4, 5,
'--NME--'), ('0997-006', 5, 6, 'Moscow'), ('1022-002', 0, 1, 'Canada'), ('11
60-002', 2, 3, 'Italy'), ('1011-012', 18, 20, '--NME--'), ('1126-004', 4, 5,
'Islam'), ('1073-007', 11, 13, 'Justin_Marshall'), ('1097-000', 0, 7, '--NME
--'), ('1119-006', 29, 30, 'Kurds'), ('1000-018', 10, 11, '--NME--'), ('1022
-002', 21, 22, 'Canada'), ('1004-003', 22, 23, 'France'), ('1051-009', 1, 3,
'Colin_Montgomerie'), ('1014-003', 9, 10, 'Bill_Clinton'), ('0986-007', 0,
4, '--NME--'), ('0983-004', 39, 40, '--NME--'), ('1058-040', 7, 9, '--NME
--'), ('1069-017', 23, 26, '--NME--'), ('1087-004', 21, 22, 'Pakistan'), ('1
145-001', 0, 1, 'Manama'), ('1049-003', 29, 31, 'Niall_Quinn'), ('0986-006',
12, 13, '--NME--'), ('1026-002', 12, 17, '--NME--'), ('1097-003', 8, 10, 'St
efan_Edberg'), ('1022-000', 7, 8, '--NME--'), ('1039-007', 2, 3, 'Suharto'),
('1132-000', 0, 1, 'Mexico'), ('1124-002', 3, 4, 'Afrikaners'), ('1060-016',
1, 3, 'Frank_Nobilo'), ('1071-005', 0, 1, 'Panama'), ('1090-007', 10, 11, 'A
mericans'), ('1125-007', 10, 11, 'Rwanda'), ('0966-079', 4, 5, 'Kenya'), ('1
011-012', 13, 17, '--NME--'), ('0946-004', 2, 3, 'Somerset_County_Cricket_Cl
ub'), ('1126-005', 3, 4, 'NATO'), ('1143-003', 45, 46, 'United_Nations_Inter
im_Force_in_Lebanon'), ('1103-002', 0, 1, 'Brussels'), ('1076-040', 0, 1, 'F
K_Spartak_Subotica'), ('0970-004', 0, 1, '--NME--'), ('1094-045', 0, 2, 'San
_Diego_Padres'), ('1134-005', 0, 1, 'Jakarta'), ('0966-047', 1, 3, 'William_
Tanui'), ('1014-010', 50, 55, 'League_of_Women_Voters'), ('0990-005', 7, 8,
'Boris_Yeltsin'), ('0966-154', 1, 3, 'Isel_López'), ('0961-008', 30, 31, 'Na
tional_Football_League'), ('1003-003', 21, 22, '--NME--'), ('1066-042', 0,
2, 'Bristol_Rovers_F.C.'), ('0953-012', 0, 1, '--NME--'), ('1066-070', 0, 1,
'Darlington_F.C.'), ('1103-009', 28, 29, 'Turkey_national_football_team'),
('1117-008', 5, 6, 'Belgium'), ('1023-000', 0, 5, '--NME--'), ('1012-006', 3
3, 34, 'Islam'), ('0972-029', 0, 2, '--NME--'), ('1025-007', 15, 16, 'Israel
'), ('0984-021', 6, 9, 'Kansai_International_Airport'), ('1139-002', 16, 17,
'China'), ('1150-001', 2, 3, '--NME--'), ('1066-057', 0, 3, '--NME--'), ('11
52-012', 11, 12, 'Iraq'), ('1012-007', 0, 1, 'Louis_Farrakhan'), ('0966-015
', 1, 3, 'Yekaterina_Podkopayeva'), ('0953-002', 5, 8, '--NME--'), ('1015-00
6', 2, 3, 'RMS_Titanic'), ('1034-004', 1, 2, 'Netherlands'), ('1037-002', 2
5, 26, 'Guangdong'), ('1019-003', 9, 10, 'Groningen'), ('1130-001', 0, 1,
'--NME--'), ('1119-000', 5, 6, 'Erbil'), ('1096-019', 1, 2, 'Chicago'), ('11
60-011', 20, 21, 'China'), ('1116-038', 14, 15, '--NME--'), ('0983-004', 15,
16, 'Hussein_of_Jordan'), ('1055-037', 5, 7, '--NME--'), ('1069-018', 13, 1
5, '--NME--'), ('1118-002', 16, 17, 'Scotland'), ('1025-009', 18, 19, 'Israe
l'), ('1037-004', 27, 28, 'Ericsson'), ('1103-004', 0, 1, 'Turkish_Armed_For
ces'), ('1070-016', 8, 10, '--NME--'), ('1100-002', 0, 1, 'United_States'),
('1101-005', 13, 14, 'Mexico'), ('1012-006', 22, 23, 'Libya'), ('1116-037',
7, 8, 'Russia'), ('1027-002', 2, 4, 'Palestinian_territories'), ('1045-003',
```

In [95]:
```
evaluation_report(set(gold_mentions(df_dev)),set(wiki_pred))
```

```
66.01456815816857 53.608247422680414 59.1680656593919
```

Problem 6: Context-sensitive disambiguation

Consider the entity mention ‘Lincoln’. The most probable entity for this mention turns out to be Lincoln, Nebraska; but in pages about American history, we would be better off to predict Abraham Lincoln. This suggests that we should try to disambiguate between different entity references based on the textual context on the page from which the mention was taken. Your task in this last problem is to implement this idea.

Set up a dictionary that contains, for each mention m that can refer to more than one entity e, a

separate Naive Bayes classifier that is trained to predict the correct entity e, given the textual context of the mention. As the prior probabilities of the classifier, choose the probabilities $P(e|m)$ that you used in Problem 5. To let you estimate the context-specific probabilities, we have compiled a data set with mention contexts:

In [96]:
```
with bz2.open('contexts.tsv.bz2') as source:
    df_contexts = pd.read_csv(source, sep='\t', quoting=csv.QUOTE_NONE)
```

This data frame contains, for each ambiguous mention m and each knowledge base entity e to which this mention can refer, up to 100 randomly selected contexts in which m is used to refer to e. For this data, a **context** is defined as the 5 tokens to the left and the 5 tokens to the right of the mention. Here are a few examples:

In [97]:
```
df_contexts.head()
```

Out[97]:

	mention	entity	context
0	1970	UEFA_Champions_League	Cup twice the first in @ and the second in 1983
1	1970	FIFA_World_Cup	America 1975 and during the @ and 1978 World C...
2	1990 World Cup	1990_FIFA_World_Cup	Manolo represented Spain at the @
3	1990 World Cup	1990_FIFA_World_Cup	Hašek represented Czechoslovakia at the @ and ...
4	1990 World Cup	1990_FIFA_World_Cup	renovations in 1989 for the @ The present capa...

Note that, in each context, the position of the mention is indicated by the `@` symbol.

From this data frame, it is easy to select the data that you need to train the classifiers – the contexts and corresponding entities for all mentions. To illustrate this, the following cell shows how to select all contexts that belong to the mention 'Lincoln':

In [98]:
```
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from collections import Counter

classifiers={}
for w in set(df_contexts.mention):
    X=df_contexts.context[df_contexts.mention==w]
    y=df_contexts.entity[df_contexts.mention==w]

    c=Counter(y)
    prior = sorted(c.items(), key=lambda pair: (pair[0]))
    prior= [p[1]/sum(c.values()) for p in prior]
    pipe=Pipeline([("vectorize",CountVectorizer()),("MultinomialNB",Multinor

    pipe.fit(X,y)
    classifiers[w]=pipe
```

Implement the context-sensitive disambiguation method and evaluate its performance. Here are

some more hints that may help you along the way:

Hint 1: The prior probabilities for a Naive Bayes classifier can be specified using the `class_prior` option. You will have to provide the probabilities in the same order as the alphabetically sorted class (entity) names.

Hint 2: Not all mentions in the knowledge base are ambiguous, and therefore not all mentions have context data. If a mention has only one possible entity, pick that one. If a mention has no entity at all, predict the `--NME--` label.

In [99]:

```
# TODO: Write code here to implement the context-sensitive disambiguation me
cs_pred=[]


for x in pred_df.itertuples():
    doc=nlp(x[2])
    tmp_str=str(doc[x[3]:x[4]])
    tmp_df=df_kb.loc[df_kb.mention==tmp_str]
    try:
        cs_pred.append((x[1],x[3],x[4],classifiers[tmp_str].predict([x[2]])[
    except KeyError:
        if len(tmp_df)!=0:
            new_ent=(x[1],x[3],x[4],tmp_df.iloc[0].entity)
            cs_pred.append(new_ent)
        else:
            new_ent=(x[1],x[3],x[4],"--NME--")
            cs_pred.append(new_ent)
print(cs_pred)


# for x in pred_df.itertuples():
#     doc = nlp(x[2])

#     for ent in doc.ents:
#         tmp_str=[]
#         for i in range(ent.start,ent.end):
#             tmp_str.append(doc[i].text)
#         tmp_str=" ".join(tmp_str)
#         tmp_df=df_kb.loc[df_kb.mention == tmp_str]
#         try:
#             cs_pred.append((x[1],ent.start,ent.end,classifiers[tmp_str].pr
#         except KeyError:
#             if len(tmp_df)!=0:
#                 cs_pred.append((x[1],ent.start,ent.end,tmp_df.iloc[0].enti
#             else:
#                 cs_pred.append((x[1],ent.start,ent.end,"--NME--"))



# cs_pred=[]
# for x in pred_df.itertuples():
#     doc = nlp(x[2])
#     for ent in doc.ents:
#         tmp_str=[]
#         for i in range(ent.start,ent.end):
#             tmp_str.append(doc[i].text)
#     tmp_str=" ".join(tmp_str)
#     try:
#         cs_pred.append((x[1],ent.start,ent.end,classifiers[tmp_str].predic
#     except KeyError:
#         cs_pred.append((x[1],ent.start,ent.end,"--NME--"))

# print(cs_pred)
```

```
s_(baseball)'), ('0946-006', 0, 1, 'Essex_County_Cricket_Club'), ('1155-011
', 3, 4, 'Germany'), ('0957-006', 11, 12, 'Australia_national_cricket_team
'), ('1089-001', 2, 3, 'Wisconsin'), ('0966-087', 1, 3, 'Rohan_Robinson'),
('0948-003', 4, 5, 'England_cricket_team'), ('1056-009', 4, 5, 'United_Kingd
```

om'), ('0955-009', 0, 1, '--NME--'), ('1030-040', 1, 2, 'Tokyo'), ('1058-003
', 6, 8, '--NME--'), ('1142-005', 17, 18, 'Kurdistan_Democratic_Party'), ('1
152-017', 3, 4, 'United_States_Marine_Corps'), ('0984-006', 29, 30, 'Missour
i'), ('1056-043', 3, 4, 'France_national_football_team'), ('1058-009', 15, 1
6, 'Texas'), ('0981-000', 0, 3, '--NME--'), ('1017-002', 19, 21, 'Air_France
'), ('0957-012', 0, 2, 'Alexander_Volkov_(tennis)'), ('0957-020', 35, 36, 'A
ustralia_national_cricket_team'), ('1075-002', 0, 1, 'Mauritania_national_fo
otball_team'), ('0955-014', 0, 1, 'FC_Seoul'), ('1018-004', 14, 15, 'Berlin
'), ('1086-002', 0, 1, 'England_national_football_team'), ('1099-008', 17, 1
8, 'Duncan_Ferguson'), ('1116-035', 8, 9, '--NME--'), ('1152-003', 29, 30, '
Iraq'), ('0977-008', 0, 1, 'Canada'), ('1139-002', 9, 11, '--NME--'), ('1142
-002', 16, 17, 'Baghdad'), ('0968-004', 24, 25, 'Bosnia_and_Herzegovina'),
('0992-002', 3, 4, 'Islam'), ('1056-040', 0, 2, 'Marty_Nothstein'), ('1057-0
03', 13, 15, 'Graham_Lloyd'), ('1090-010', 21, 22, 'Nagoya'), ('1016-014',
7, 9, 'Yitzhak_Mordechai'), ('1051-009', 32, 34, 'Klas_Eriksson'), ('0997-00
4', 14, 15, 'Russian_Empire'), ('1031-007', 35, 37, '--NME--'), ('0972-005',
0, 3, '--NME--'), ('1117-003', 0, 2, '--NME--'), ('1139-003', 2, 3, '--NME
--'), ('0992-000', 1, 2, 'Islam'), ('0991-002', 10, 12, 'Aslan_Maskhadov'),
('1009-014', 0, 1, 'Mexico'), ('1056-017', 16, 17, 'Germany_national_footbal
l_team'), ('1109-007', 7, 8, '--NME--'), ('1125-005', 19, 20, 'Rwanda'), ('1
047-000', 8, 9, '--NME--'), ('1099-009', 9, 11, 'Craig_Brown'), ('1146-007',
17, 18, 'Iraq'), ('0972-015', 0, 2, '--NME--'), ('0999-002', 1, 2, 'Moscow
'), ('1099-018', 14, 16, 'Colin_Hendry'), ('0994-007', 0, 1, 'Alexander_Luka
shenko'), ('0966-068', 4, 5, 'The_Bahamas'), ('1160-012', 1, 2, 'HarperColli
ns'), ('0966-146', 4, 5, '--NME--'), ('1145-002', 10, 11, 'Bahrain'), ('1156
-002', 0, 1, 'Italy'), ('0956-002', 2, 4, 'South_Korea'), ('1066-047', 0, 2,
'Notts_County_F.C.'), ('0963-003', 12, 14, 'Billy_Andrade'), ('1012-009', 1
0, 11, 'Libya'), ('1154-006', 16, 18, '--NME--'), ('1027-009', 9, 12, '--NME
--'), ('1092-004', 7, 8, 'Germany_national_football_team'), ('1084-002', 0,
1, 'Sweden_national_football_team'), ('1101-008', 33, 35, '--NME--'), ('1033
-007', 10, 11, 'Switzerland'), ('1009-001', 0, 2, '--NME--'), ('1009-011',
0, 1, 'Mexico'), ('0947-011', 19, 21, '--NME--'), ('0949-005', 24, 25, 'Sout
hampton'), ('1058-007', 10, 13, '--NME--'), ('0966-081', 1, 3, 'Torrance_Zel
lner'), ('1033-010', 6, 7, 'Chechnya'), ('0972-014', 0, 3, '--NME--'), ('097
6-002', 0, 5, '--NME--'), ('1033-009', 0, 1, '--NME--'), ('1044-004', 30, 3
1, 'Arabs'), ('1068-026', 4, 5, 'Wrexham_A.F.C.'), ('1089-005', 0, 2, '--NME
--'), ('1058-019', 5, 6, 'New_York_Yankees'), ('1072-019', 67, 69, 'Robin_Br
ooke'), ('1044-004', 11, 12, 'Palestine'), ('1117-004', 7, 8, 'Netherlands
'), ('0957-026', 3, 4, 'Ukraine_national_football_team'), ('1108-000', 0, 3,
'--NME--'), ('1077-003', 16, 17, 'England_national_football_team'), ('1155-0
10', 31, 32, 'Berlin'), ('0959-012', 0, 1, 'Cincinnati_Reds'), ('0960-021',
11, 12, 'Jeff_Tarango'), ('0960-039', 7, 8, 'Jeff_Tarango'), ('1008-018', 1,
3, '--NME--'), ('1056-034', 21, 23, 'Philippe_Ermenault'), ('1076-010', 0,
4, '--NME--'), ('1099-017', 35, 38, '--NME--'), ('1010-021', 20, 22, '--NME
--'), ('1056-030', 1, 3, '--NME--'), ('1103-017', 0, 1, 'Belgium'), ('1126-0
02', 33, 34, 'Serbs'), ('1095-005', 0, 1, 'Detroit_Tigers'), ('0955-012', 0,
1, '--NME--'), ('1056-043', 7, 9, '--NME--'), ('1136-002', 19, 20, 'Tianjin
'), ('0961-004', 6, 7, '--NME--'), ('1099-018', 23, 25, 'Stuart_McCall'), ('
1006-007', 5, 6, 'Niger'), ('0963-017', 23, 25, 'PGA_Tour'), ('1036-014', 3,
4, 'Norway'), ('0966-084', 1, 3, 'Fabrizio_Mori'), ('1056-031', 1, 3, '--NME
--'), ('1121-001', 0, 1, 'London'), ('1142-006', 20, 21, 'Kurdistan_Democrat
ic_Party'), ('1068-013', 6, 8, 'Manchester_City_F.C.'), ('0993-006', 14, 19,
'--NME--'), ('1064-005', 28, 29, 'Gloucester_Rugby'), ('1056-021', 1, 3, 'An
nett_Neumann'), ('1099-004', 0, 1, 'Scotland_national_football_team'), ('104
9-004', 5, 6, 'Ireland'), ('0946-001', 0, 1, 'London'), ('0962-010', 10, 12,
'Ken_Green_(golfer)'), ('0966-034', 1, 3, 'Patrick_Stevens'), ('0966-143',
4, 5, 'United_Kingdom'), ('1051-008', 1, 3, 'Joakim_Haeggman'), ('1094-036',
0, 2, 'New_York_Yankees'), ('1072-018', 49, 51, 'Gary_Teichmann'), ('0960-00
7', 9, 10, 'Andre_Agassi'), ('1051-011', 15, 17, 'Peter_Hedblom'), ('0949-01
0', 2, 4, 'UEFA_Euro_1996'), ('1133-004', 33, 35, 'Deng_Xiaoping'), ('1133-0
11', 0, 1, '--NME--'), ('1116-006', 2, 3, 'France'), ('1028-002', 8, 9, 'Mic
higan'), ('0962-009', 25, 27, '--NME--'), ('1115-004', 6, 8, '--NME--'), ('1

035-008', 12, 13, 'Armenia'), ('0966-048', 4, 5, 'Kenya_national_cricket_tea
m'), ('0980-013', 19, 22, '--NME--'), ('1054-023', 5, 6, 'Netherlands'), ('1
068-021', 0, 4, '--NME--'), ('0962-010', 22, 24, '--NME--'), ('0990-014', 1
5, 16, 'Alexander_Lebed'), ('1058-001', 0, 1, '--NME--'), ('1103-020', 7, 9,
'--NME--'), ('1070-014', 4, 5, 'Brazil_national_football_team'), ('0997-007
', 0, 1, 'Boris_Yeltsin'), ('1128-011', 12, 13, '--NME--'), ('1096-034', 14,
15, 'Run_(baseball)'), ('0996-003', 1, 2, 'Dariusz_Rosati'), ('1112-004', 4,
5, 'Germany_national_football_team'), ('1102-003', 9, 11, 'Youri_Djorkaeff
'), ('1143-003', 16, 17, 'France_national_football_team'), ('1042-003', 5,
7, '--NME--'), ('1016-007', 5, 6, 'Israel'), ('1039-014', 21, 23, 'Tanjung_P
riok'), ('1045-003', 7, 8, 'State_of_Palestine'), ('1058-040', 3, 5, 'Marc_N
ewfield'), ('0966-021', 4, 5, '--NME--'), ('0957-019', 7, 9, 'Linda_Wild'),
('1045-009', 3, 4, 'Israel'), ('1094-055', 0, 2, 'San_Diego_Padres'), ('0966
-036', 1, 3, '--NME--'), ('0980-014', 8, 9, '--NME--'), ('1033-013', 17, 18,
'--NME--'), ('0949-004', 2, 3, '--NME--'), ('1056-028', 4, 6, 'New_Zealander
s'), ('1065-023', 0, 1, 'Brechin_City_F.C.'), ('1069-010', 15, 18, '--NME
--'), ('1054-008', 4, 5, 'Netherlands'), ('0992-003', 25, 27, '--NME--'), ('
0966-099', 1, 3, 'Dennis_Mitchell'), ('1036-013', 18, 19, 'Pyramiden'), ('10
70-013', 1, 4, 'Al_Unser_Jr.'), ('0961-010', 0, 1, '--NME--'), ('1057-007',
31, 33, 'Waqar_Younis'), ('0946-008', 3, 4, 'Yorkshire_County_Cricket_Club
'), ('1160-016', 9, 10, 'Dublin'), ('1072-019', 8, 10, '--NME--'), ('1090-00
7', 28, 29, 'Spain_Fed_Cup_team'), ('0954-004', 5, 6, 'Scotland_national_rug
by_union_team'), ('0963-017', 32, 34, '--NME--'), ('0957-033', 15, 17, 'Andr
ea_Gaudenzi'), ('0946-010', 7, 9, 'Mark_Butcher'), ('0957-026', 0, 2, 'Andri
y_Medvedev'), ('1072-013', 11, 13, 'Sean_Fitzpatrick'), ('1087-017', 13, 14,
'Australia'), ('0963-023', 8, 11, '--NME--'), ('0996-007', 0, 1, 'Poland'),
('1160-022', 16, 17, 'HarperCollins'), ('1010-002', 0, 1, 'Chicago'), ('0972
-008', 3, 4, '--NME--'), ('1100-002', 1, 3, '--NME--'), ('1026-000', 5, 6, '
Islam'), ('0972-002', 10, 11, 'Australia_national_cricket_team'), ('1059-005
', 8, 9, 'Surrey_County_Cricket_Club'), ('1155-009', 21, 23, 'Abolhassan_Ban
isadr'), ('1127-018', 0, 2, '--NME--'), ('1051-008', 14, 15, 'France_nationa
l_football_team'), ('0957-007', 0, 2, 'Tim_Henman'), ('0966-074', 1, 3, 'Ros
e_Cheruiyot'), ('0966-149', 1, 3, '--NME--'), ('0958-054', 2, 3, 'Pittsburgh
_Pirates'), ('0974-000', 0, 3, '--NME--'), ('1011-022', 4, 5, 'NationsBank
'), ('1035-010', 36, 37, 'Finland'), ('1116-014', 10, 13, '--NME--'), ('1160
-007', 28, 30, 'Warner_Bros.'), ('1056-009', 1, 3, '--NME--'), ('1107-002',
6, 9, '--NME--'), ('1116-025', 4, 5, 'Baudouin_of_Belgium'), ('1094-025', 0,
1, 'Baltimore_Orioles'), ('1031-003', 0, 1, 'Italy'), ('0984-002', 11, 12, '
Europe'), ('0983-004', 1, 2, 'Iraq'), ('1055-020', 2, 4, '--NME--'), ('1116-
024', 11, 12, '--NME--'), ('0999-003', 0, 1, 'Interfax'), ('1148-002', 7, 8,
'Tunisia'), ('0988-000', 0, 2, '--NME--'), ('0957-017', 3, 5, 'South_Africa
'), ('1012-002', 0, 3, '--NME--'), ('1056-038', 3, 4, 'Australia_national_cr
icket_team'), ('1058-036', 1, 2, 'Detroit'), ('0983-002', 17, 18, 'London'),
('1087-018', 14, 15, 'Scotland_national_rugby_union_team'), ('1156-002', 3,
5, 'Lamberto_Dini'), ('1116-037', 9, 10, 'Belarus'), ('1035-006', 17, 18,
'--NME--'), ('1027-023', 1, 3, '--NME--'), ('1057-007', 0, 1, 'Pakistan_nati
onal_cricket_team'), ('1057-000', 0, 1, '--NME--'), ('1056-026', 1, 2, '--NM
E--'), ('1152-000', 5, 6, 'Persian_Gulf'), ('1010-016', 7, 8, 'Italy'), ('10
30-039', 2, 3, 'London'), ('1121-004', 20, 21, 'Erbil'), ('1014-009', 13, 1
4, 'Bob_Dole'), ('1141-007', 11, 12, 'Guangdong'), ('1006-009', 20, 23, '--N
ME--'), ('1056-040', 6, 8, 'Pavel_Buráň'), ('1068-001', 0, 1, 'London'), ('0
958-052', 0, 2, 'San_Diego_Padres'), ('1062-002', 0, 1, 'Wales_national_unde
r-21_football_team'), ('1096-030', 0, 1, 'Montreal'), ('1036-034', 0, 1, 'No
rway'), ('1122-003', 5, 9, '--NME--'), ('1142-011', 6, 10, 'Patriotic_Union_
of_Kurdistan'), ('0967-002', 0, 1, 'UEFA'), ('0951-005', 0, 4, '--NME--'),
('1096-020', 5, 7, 'Ryne_Sandberg'), ('1066-025', 0, 1, 'Oldham_Athletic_A.
F.C.'), ('1091-003', 36, 38, 'Judith_Wiesner'), ('1072-015', 26, 27, '--NME
--'), ('1116-029', 6, 8, 'Jackie_Stewart'), ('0949-001', 0, 1, 'London'), ('
1127-003', 3, 4, 'Chechens'), ('0962-005', 4, 5, '--NME--'), ('1056-046', 0,
2, '--NME--'), ('1061-002', 4, 5, 'Scotland'), ('1029-002', 0, 1, 'Washingto
n,_D.C.'), ('1069-013', 9, 12, '--NME--'), ('1152-012', 5, 6, 'Iranian_Kurdi
stan'), ('0949-003', 20, 22, '--NME--'), ('1012-007', 15, 16, 'Washington,_

D.C.'), ('1121-005', 19, 20, 'Erbil'), ('1152-011', 18, 19, 'Kurds'), ('1033
-005', 16, 17, '--NME--'), ('0946-009', 13, 14, 'England_cricket_team'), ('0
947-002', 7, 8, 'England'), ('0952-002', 0, 2, 'FC_Rotor_Volgograd'), ('0959
-007', 0, 1, 'Los_Angeles_Angels'), ('1033-005', 7, 8, '--NME--'), ('1148-00
8', 4, 5, 'Tunisia_national_football_team'), ('0952-006', 26, 27, 'Moscow'),
('1132-003', 0, 2, '--NME--'), ('1066-066', 0, 1, 'Lincoln,_Nebraska'), ('10
60-018', 1, 3, 'Pádraig_Harrington'), ('1128-000', 5, 6, '--NME--'), ('0966-
126', 1, 3, 'Inha_Babakova'), ('1122-002', 18, 19, 'Algeria'), ('1088-002',
21, 22, 'Pakistan_national_cricket_team'), ('1144-004', 25, 26, 'Reuters'),
('0984-003', 16, 17, '--NME--'), ('0960-019', 41, 42, 'Marcelo_Ríos'), ('103
7-002', 3, 6, '--NME--'), ('0988-005', 17, 19, '--NME--'), ('1038-007', 6,
7, '--NME--'), ('1084-003', 7, 9, '--NME--'), ('0982-000', 5, 6, 'Chicago'),
('0997-004', 27, 28, 'Moscow'), ('1053-001', 0, 1, '--NME--'), ('1096-012',
28, 29, '--NME--'), ('1112-001', 0, 1, 'Bonn'), ('1128-006', 19, 20, '--NME
--'), ('1027-017', 27, 29, '--NME--'), ('1088-002', 27, 28, 'Australia_natio
nal_rugby_union_team'), ('1156-007', 2, 3, 'Rome'), ('1045-007', 27, 30, '--
NME--'), ('1058-008', 1, 2, 'Indigenous_peoples_of_the_Americas'), ('0966-12
8', 4, 5, 'Russia_national_football_team'), ('0968-004', 5, 6, 'A._G._Edward
s'), ('1033-010', 27, 28, 'Russia'), ('1036-011', 22, 23, 'Reuters'), ('1073
-010', 0, 2, 'New_Zealand_national_rugby_union_team'), ('1089-004', 7, 9, 'L
oren_Roberts'), ('1113-001', 0, 1, 'Bonn'), ('1044-005', 2, 3, 'Arab_citizen
s_of_Israel'), ('1116-035', 2, 3, 'England'), ('1142-002', 26, 27, 'Reuters
'), ('1002-000', 4, 5, 'Colombia'), ('0996-002', 0, 1, 'Poland_national_foot
ball_team'), ('1084-001', 0, 1, '--NME--'), ('1045-002', 8, 9, 'Israel'), ('
1058-022', 3, 6, '--NME--'), ('0966-117', 1, 3, 'Igor_Trandenkov'), ('0991-0
08', 21, 22, 'Chechnya'), ('1079-007', 8, 9, 'Socialist_Federal_Republic_of_
Yugoslavia'), ('1076-031', 0, 1, '--NME--'), ('1051-010', 35, 36, 'Australia
_national_cricket_team'), ('1142-005', 20, 21, 'Reuters'), ('1016-010', 8, 1
0, 'David_Levy_(Israeli_politician)'), ('0948-033', 3, 5, 'British_Universit
ies_cricket_team'), ('0958-012', 0, 2, 'National_League_Central'), ('1072-01
4', 18, 19, 'South_Africa_national_rugby_union_team'), ('1044-006', 0, 1, 'I
srael'), ('0999-003', 22, 24, '--NME--'), ('1069-018', 1, 2, 'Bristol'), ('1
111-009', 2, 4, '--NME--'), ('1160-007', 9, 10, 'England'), ('0963-017', 13,
15, 'Andrew_Magee'), ('0983-005', 1, 2, 'English_language'), ('0984-010', 1
6, 17, 'United_States_Treasury_security'), ('1079-005', 2, 3, 'Italy'), ('09
46-007', 5, 6, 'England_cricket_team'), ('1017-004', 5, 6, 'Africa'), ('1077
-002', 20, 23, '--NME--'), ('1146-006', 31, 32, 'Iraq'), ('1155-014', 5, 6,
'Iran'), ('1138-007', 19, 22, '--NME--'), ('0984-017', 25, 26, '--NME--'),
('0950-004', 3, 4, 'Socialist_Federal_Republic_of_Yugoslavia'), ('0997-000',
0, 1, 'Boris_Yeltsin'), ('1068-013', 2, 3, 'Birmingham_City_F.C.'), ('0998-0
04', 3, 4, 'Aslan_Maskhadov'), ('1135-027', 6, 7, 'National_League_for_Democ
racy'), ('1084-002', 10, 11, 'Europe'), ('1060-012', 1, 3, 'Wayne_Riley'),
('0966-113', 4, 5, 'Norway'), ('1037-000', 5, 7, '--NME--'), ('1085-003', 2,
4, '--NME--'), ('1158-002', 17, 18, 'Belgium'), ('1051-010', 18, 19, 'Austra
lia_national_cricket_team'), ('1128-012', 9, 10, 'Poles'), ('0949-005', 22,
23, 'Blackburn_Rovers_F.C.'), ('0966-037', 1, 3, 'Iván_García'), ('1056-044
', 0, 2, '--NME--'), ('1074-003', 31, 34, '--NME--'), ('1012-005', 15, 16, '
Libya'), ('1029-017', 3, 7, '--NME--'), ('1056-034', 11, 13, '--NME--'), ('1
078-005', 7, 8, 'Russia_national_football_team'), ('1090-014', 10, 12, 'Fed_
Cup'), ('0990-014', 22, 23, 'Chechnya'), ('1126-009', 9, 10, 'Serbia'), ('09
85-003', 6, 7, '--NME--'), ('1161-005', 0, 3, '--NME--'), ('1056-017', 11, 1
3, 'Francis_Moreau'), ('1155-007', 0, 1, 'Iran'), ('1056-010', 4, 5, 'Russia
_national_football_team'), ('1116-010', 2, 3, 'Moscow'), ('0953-013', 8, 9,
'Panama'), ('1133-003', 0, 2, 'Xiao_Qiang'), ('1131-002', 24, 25, 'Chile'),
('1137-003', 7, 8, '--NME--'), ('1095-015', 0, 1, 'Miami_Marlins'), ('0974-0
11', 0, 2, 'Chaminda_Vaas'), ('0986-003', 0, 1, 'Birmingham'), ('1051-010',
25, 27, 'Mark_Roe'), ('0948-022', 7, 9, 'The_Oval'), ('1112-002', 3, 5, 'Jür
gen_Klinsmann'), ('1154-001', 0, 1, 'Paris'), ('0948-009', 0, 1, 'England'),
('0985-011', 1, 2, '--NME--'), ('0990-003', 11, 12, 'Interfax'), ('1130-006
', 3, 8, '--NME--'), ('0966-024', 4, 5, 'Cuba_national_football_team'), ('09
67-003', 17, 19, 'Guy_Hellers'), ('1007-010', 6, 7, 'Bob_Dole'), ('0961-007
', 1, 2, '--NME--'), ('1072-018', 26, 29, '--NME--'), ('1137-008', 32, 35,

'--NME--'), ('1135-006', 18, 21, '--NME--'), ('1128-001', 0, 1, 'Warsaw'),
('1130-006', 17, 18, 'United_States'), ('1126-010', 24, 25, 'NATO'), ('1116-
011', 14, 15, 'Brazil_national_football_team'), ('0970-002', 30, 31, 'Atalan
ta_B.C.'), ('1006-009', 27, 28, 'Reuters'), ('1151-001', 0, 1, 'Dallas'), ('
1054-001', 2, 3, 'Netherlands'), ('0992-007', 0, 5, '--NME--'), ('0948-006',
0, 1, 'Australia_national_cricket_team'), ('0984-019', 0, 1, '--NME--'), ('1
103-016', 73, 75, '--NME--'), ('1058-031', 29, 33, '--NME--'), ('1116-008',
4, 5, 'Strasbourg'), ('1121-000', 7, 8, 'Iraq'), ('1137-002', 6, 8, 'South_K
orea'), ('1152-018', 22, 23, 'Persian_Gulf'), ('1137-003', 21, 23, 'Internat
ional_Committee_of_the_Red_Cross'), ('1128-003', 37, 38, '--NME--'), ('1134-
000', 2, 3, 'Indonesia'), ('1149-005', 31, 33, 'Interstate_95'), ('1096-024
', 8, 9, 'Atlanta'), ('1096-013', 34, 35, 'Philadelphia_Phillies'), ('1131-0
07', 5, 6, 'Honduras'), ('0966-154', 4, 5, 'Cuba_national_football_team'),
('1072-011', 25, 26, 'South_Africa_national_rugby_union_team'), ('1116-006',
9, 10, 'Turin'), ('1046-006', 19, 21, '--NME--'), ('1116-006', 20, 21, 'Fran
ce'), ('1072-012', 36, 37, '--NME--'), ('0957-026', 6, 8, 'Ján_Krošlák'), ('
1072-017', 5, 6, 'Andrew_Mehrtens'), ('1058-026', 5, 8, '--NME--'), ('1094-0
34', 0, 1, 'Montreal_Expos'), ('1128-011', 1, 2, 'Poland'), ('1121-003', 16,
17, 'Iraq'), ('0957-009', 6, 8, 'David_Rikl'), ('0991-011', 14, 15, 'Alexand
er_Lebed'), ('0955-013', 0, 1, 'Ulsan'), ('0982-002', 21, 22, 'Chicago'), ('
1160-011', 17, 19, '--NME--'), ('1003-005', 5, 6, '--NME--'), ('1112-004', 1
8, 19, 'Europe'), ('0963-017', 0, 1, '--NME--'), ('1059-002', 30, 32, 'Engla
nd'), ('1103-006', 10, 12, '--NME--'), ('1121-003', 29, 30, 'Erbil'), ('1062
-002', 2, 4, 'San_Marino_national_football_team'), ('0960-003', 0, 2, 'Andre
_Agassi'), ('0982-002', 0, 1, 'United_Kingdom'), ('1133-006', 18, 19, 'Metro
_Manila'), ('1157-002', 0, 6, '--NME--'), ('1027-006', 14, 15, 'Hebrew_langu
age'), ('1057-007', 16, 18, 'Wasim_Akram'), ('1075-001', 0, 1, '--NME--'),
('1001-000', 0, 1, 'Romania_national_football_team'), ('0958-002', 0, 3, 'Ma
jor_League_Baseball'), ('0953-004', 26, 27, '--NME--'), ('1073-007', 0, 4,
'--NME--'), ('1037-002', 28, 29, 'China'), ('1153-003', 4, 5, '--NME--'), ('
0997-006', 5, 6, 'Moscow'), ('1022-002', 0, 1, 'Canada'), ('1160-002', 2, 3,
'Italy_national_football_team'), ('1011-012', 18, 20, '--NME--'), ('1126-004
', 4, 5, 'Islam'), ('1073-007', 11, 13, 'Justin_Marshall'), ('1097-000', 0,
7, '--NME--'), ('1119-006', 29, 30, 'Kurds'), ('1000-018', 10, 11, '--NME
--'), ('1022-002', 21, 22, 'Canada'), ('1004-003', 22, 23, 'France_national_
football_team'), ('1051-009', 1, 3, 'Colin_Montgomerie'), ('1014-003', 9, 1
0, 'Bill_Clinton'), ('0986-007', 0, 4, '--NME--'), ('0983-004', 39, 40, '--N
ME--'), ('1058-040', 7, 9, '--NME--'), ('1069-017', 23, 26, '--NME--'), ('10
87-004', 21, 22, 'Pakistan'), ('1145-001', 0, 1, 'Manama'), ('1049-003', 29,
31, 'Niall_Quinn'), ('0986-006', 12, 13, '--NME--'), ('1026-002', 12, 17,
'--NME--'), ('1097-003', 8, 10, 'Stefan_Edberg'), ('1022-000', 7, 8, '--NME
--'), ('1039-007', 2, 3, 'Suharto'), ('1132-000', 0, 1, 'Mexico'), ('1124-00
2', 3, 4, 'Afrikaners'), ('1060-016', 1, 3, 'Frank_Nobilo'), ('1071-005', 0,
1, 'Panama'), ('1090-007', 10, 11, 'United_States'), ('1125-007', 10, 11, 'R
wanda'), ('0966-079', 4, 5, 'Kenya_national_cricket_team'), ('1011-012', 13,
17, '--NME--'), ('0946-004', 2, 3, 'Somerset_County_Cricket_Club'), ('1126-0
05', 3, 4, 'NATO'), ('1143-003', 45, 46, 'United_Nations_Interim_Force_in_Le
banon'), ('1103-002', 0, 1, 'Brussels'), ('1076-040', 0, 1, 'FC_Spartak_Mosc
ow'), ('0970-004', 0, 1, '--NME--'), ('1094-045', 0, 2, 'San_Diego_Padres'),
('1134-005', 0, 1, 'Jakarta'), ('0966-047', 1, 3, 'William_Tanui'), ('1014-0
10', 50, 55, 'League_of_Women_Voters'), ('0990-005', 7, 8, 'Boris_Yeltsin'),
('0966-154', 1, 3, 'Isel_López'), ('0961-008', 30, 31, 'National_Football_Le
ague'), ('1003-003', 21, 22, '--NME--'), ('1066-042', 0, 2, 'Bristol_Rovers_
F.C.'), ('0953-012', 0, 1, '--NME--'), ('1066-070', 0, 1, 'Darlington_F.
C.'), ('1103-009', 28, 29, 'Turkey_national_football_team'), ('1117-008', 5,
6, 'Belgium_national_football_team'), ('1023-000', 0, 5, '--NME--'), ('1012-
006', 33, 34, 'Islam'), ('0972-029', 0, 2, '--NME--'), ('1025-007', 15, 16,
'Israel'), ('0984-021', 6, 9, 'Kansai_International_Airport'), ('1139-002',
16, 17, 'China'), ('1150-001', 2, 3, '--NME--'), ('1066-057', 0, 3, '--NME
--'), ('1152-012', 11, 12, 'Iraq'), ('1012-007', 0, 1, 'Louis_Farrakhan'),
('0966-015', 1, 3, 'Yekaterina_Podkopayeva'), ('0953-002', 5, 8, '--NME--'),
('1015-006', 2, 3, 'RMS_Titanic'), ('1034-004', 1, 2, 'Netherlands'), ('1037

```
-002', 25, 26, 'Guangdong'), ('1019-003', 9, 10, 'Groningen'), ('1130-001',
0, 1, '--NME--'), ('1119-000', 5, 6, 'Erbil'), ('1096-019', 1, 2, 'Chicago
'), ('1160-011', 20, 21, 'China'), ('1116-038', 14, 15, '--NME--'), ('0983-0
04', 15, 16, 'Hussein_of_Jordan'), ('1055-037', 5, 7, '--NME--'), ('1069-018
', 13, 15, '--NME--'), ('1118-002', 16, 17, 'Scotland_national_football_team
'), ('1025-009', 18, 19, 'Israelis'), ('1037-004', 27, 28, 'Ericsson'), ('11
03-004', 0, 1, 'Turkish_Cypriots'), ('1070-016', 8, 10, '--NME--'), ('1100-0
02', 0, 1, 'United_States'), ('1101-005', 13, 14, 'Mexico'), ('1012-006', 2
2, 23, 'Libya'), ('1116-037', 7, 8, 'Russia_national_football_team'), ('1027
-002', 2, 4, 'Palestinian territories'), ('1045-003', 8, 9, '--NME--'), ('11
```

In [100…

```
evaluation_report(set(gold_mentions(df_dev)),set(cs_pred))
```

```
62.70551508844954 50.92107486902147 56.202201081887715
```

You should expect to see a small (around 1 unit) increase in both precision, recall, and F1.

This was the last lab in the Text Mining course. Congratulations!

Please read the section ‘General information’ on the ‘Labs’ page of the course website before submitting this notebook!