The butterflies of the Florentine Codex

[21]:

```
import pandas as pd
import re
import spacy

from spacy.matcher import PhraseMatcher
from spacy.matcher import Matcher
from spacy import displacy

from spacy.tokens import Span
```

[22]:

```
papalotl = open('papalotl.txt')
butterflies = papalotl.read()
```

[23]:

```
from spacy.lang.en import English

raw_text = butterflies
nlp = English()
nlp.add_pipe(nlp.create_pipe('sentencizer'))
doc = nlp(raw_text)
sentences = [sent.string.strip() for sent in doc.sents]
```

[24]:

```
# Get rid of newlines
sentences = [item.replace('\n', " ") for item in sentences]
```

[25]:

```
df = pd.DataFrame(sentences)
df.rename(columns={0: "Butterflies"})[1:5]
```

[25]:

```
                                          Butterflies
1                     lt is fuzzy, like fat; winged.
2                             Its wings are twofold.
3          It has arms, it has legs, it has antennae.
4  lt is a flyer, a constant flyer, a flutterer, …
```

[26]:

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

[27]:

```
doc = nlp(butterflies)
```

[28]:

```
matcher = PhraseMatcher(nlp.vocab)
papalotl = nlp(butterflies)
phrase_list = ['abdomen', 'neck', 'wings', 'arms', 'legs', 'antennae']
phrase_patterns = [nlp(text) for text in phrase_list]
matcher.add('mariposa', None, *phrase_patterns)
found_matches = matcher(papalotl)
```

[29]:

```
for match_id, start, end in found_matches:
    string_id = nlp.vocab.strings[match_id]
    span = papalotl[start:end]
    print(start, end, span.text)
```

```
15 16 abdomen
20 21 neck
34 35 wings
40 41 arms
44 45 legs
48 49 antennae
85 86 wings
234 235 wings
369 370 wings
516 517 wings
```

[30]:
```python
def show_ents(doc):
    if doc.ents:
        for ent in doc.ents:
            print(ent.text+' - '+ent.label_)
    else:
        print('No entity found')
```

[31]:
```python
ORGAN = doc.vocab.strings['ORGAN']
```

[32]:
```python
new_ent = Span(doc, 15, 16, label=ORGAN)
new_ent1 = Span(doc, 20, 21, label=ORGAN)
new_ent2 = Span(doc, 34, 35, label=ORGAN)
new_ent3 = Span(doc, 40, 41, label=ORGAN)
new_ent4 = Span(doc, 44, 45, label=ORGAN)
new_ent5 = Span(doc, 48, 49, label=ORGAN)
new_ent6 = Span(doc, 85, 86, label=ORGAN)
new_ent7 = Span(doc, 234, 235, label=ORGAN)
new_ent8 = Span(doc, 369, 370, label=ORGAN)
new_ent9 = Span(doc, 516, 517, label=ORGAN)
```

[33]:
```python
doc.ents = list(doc.
 ↪ents)+[new_ent]+[new_ent1]+[new_ent2]+[new_ent3]+[new_ent4]+[new_ent5]+[new_ent6]+[new_ent7]+[n
```

[34]:
```python
show_ents(doc)
```

```
abdomen - ORGAN
neck - ORGAN
wings - ORGAN
arms - ORGAN
legs - ORGAN
antennae - ORGAN
wings - ORGAN
XICALPAPALOTL - ORG
XICALTECONPAPALOTL - ORG
XICALTECON - PERSON
xicalli - PERSON
TLILPAPALOTL - PERSON
wings - ORGAN
TLECOCOZPAPALOTL - ORG
quappachpapalotl - PERSON
lts - PERSON
```

```
IZTAC PAPALOTL
 - PRODUCT
CHIAN PAPALOTL - PRODUCT
wings - ORGAN
TEXOPAPALOTL
 - LAW
XOCHIPAPALOTL - PERSON
UAPPAPALOTL - PERSON
wings - ORGAN
```

[41]:
```
colors = {'ORGAN': 'purple'}
options = {'ents': ['ORGAN'], 'colors': colors}
```

[42]:
```
displacy.render(doc, style='ent', options=options)
```

```
<IPython.core.display.HTML object>
```

[44]:
```
pwd
```

[44]:
```
'C:\\Users\\User'
```

[ ]:

[ ]:

[ ]:

[ ]:

[ ]: