Assignment 1

Luigi Riz - Mat. 223823

To complete the first assignment, consisting in the creation of some functions based on the spaCy library, I worked on a Jupyter Notebook, that helped me in subdividing my work into chunks. For each method I provide both inline comments as documentation and an example of the use of them. While developing, I read spaCy online documentation and tried to understand how the provided datatypes and functions, could help me in maintaining the code simple and clear. To have a better visualization of the Dependency Trees provided by spaCy I used a submodule named displaCy, that plots a `Span` or `Document` in a more readable way.

The main methods I implemented in my code are the following:

1 dep_paths

The first method inside the Notebook has the following signature:

```
def dep_paths (sentence, nlp = spacy.load('en_core_web_sm'))
```

It takes as input a `string`, representing the sentence to be processed, and a spaCy `Language` object, the model used to perform dependency parsing.
It returns a `dict`, mapping every spaCy `Token` in the sentence, to a list of `string`s, constituting the path of dependency relations, starting from the `ROOT` and reaching the `token`. This key-value mapping is in the form:

```
'TOKEN'    ['ROOT', 'TOKEN_i'.dep_, ..., 'TOKEN_j'.dep_, 'TOKEN'.dep_]
```

meaning that to reach `Token` 'TOKEN' following the Dependency Tree, we have to start from the `ROOT` node and run through the sequence of archs 'TOKEN_i'.dep_, ..., 'TOKEN_j'.dep_, 'TOKEN'.dep_.

2 subtrees

This function is represented by the following signature:

```
def subtrees (sentence, nlp = spacy.load('en_core_web_sm'))
```

Similarly to the previous case, it requires a `string` sentence to be processed and a spaCy `Language` model to parse it. The output of the function is a `dict` in which:

- the *key* is a `Token` in the parsed sentence,
- the *value* is a `list` of `Token`s, representing the subtree of the related key. The `Token`s are inserted in the `list`, following their ordering in the original sentence.

So, every mapping in the output `dict` is in the form:

```
'TOKEN'    ['TOKEN_i', ..., 'TOKEN_j']
```

Note that the `list` is never empty since at least 'TOKEN' itself will always be contained in it.

3 check_subtree

The third method has the signature:

```
def check_subtree(sentence, subsentence, nlp = spacy.load('en_core_web_sm')):
```

In this case the required inputs are a `string` sentence to be parsed, an ordered `list` of words (strings) to be checked and, as previously, a spaCy `Language` model to perform parsing. The function's output is:

- `True`, if the ordered `list` of words represent at least a subtree for some of the `Token`s contained in the sentence;
- `False`, if the *subsentence* is not a subtree or if the words contained in it may form a subtree, but they are in the wrong order.

Note that the order in which the words appear in *subsentence* is meaningful.

4 span_root

This function is represented by the following signature:

```
def span_root (sentence, span_start, span_end, nlp = spacy.load('en_core_web_sm')):
```

It takes as input a `string` sentence, two `integers` representing the start and the end of the span we want to extract from the sentence and a spaCy `Language` model to perform parsing. The method provides as output a spaCy `Token`, representing the *root* of the `Span`. Following spaCy documentation, the returned object is *"The token with the shortest path to the root of the sentence (or the root itself). If multiple tokens are equally high in the tree, the first token is taken."*

Trying to develop a method which is independent from a sentence, also another function is provided:

```
def span_root_nosent (span, nlp = spacy.load('en_core_web_sm')):
```

In this case the user has only to provide a `string` span, out of the context of a more general sentence, and it will receive as output the `root` of it.

5 extract

The last method has the following definition:

```
def extract (sentence, nlp = spacy.load('en_core_web_sm')):
```

The inputs to provide are a `string` sentence and a spaCy `Language` model to parse it. The function will extract from the sentence three different `lists` of `Tokens`, each of them representing a span. The structure provided as output is a `dict`, associating the *keys* `'nsubj'`, `'dobj'` and `'iobj'` to the relative span. In this way the sentence subject, the direct object and the indirect object spans of the sentence are available. Note that if the sentence does not contain one of the spans, it will return a `dict` with an empty list to the related key.