Regular Expressions

Solutions to the exercises

Solutions are in the *3-Regex_exercises_solutions* file
You can download it at https://github.com/SimoneBarandoni/nlp-python

Natural Language Processing

Text Pre-Processing

Text Pre-Processing

- Preparing data is one of the most important step in a data analysis process
- Raw texts contain a lot of noise (typographic errors, colloquialisms, etc.) and many other elements which are meaningless for a machine
- Text Pre-Processing is usually fundamental to make a text suitable for machine interpretation

Text Pre-Processing – main steps

1. Text cleaning:
   Useless or noisy elements are usually removed or modified to produce a cleaner text. Some examples are:

   - Unicode characters: punctuation, Emoji’s, URL’s
   - Numbers
   - Extra spaces
   - Stopwords: words which do not add meaning to a sentence (articles, prepositions, etc.)
   - Uppercase letters: a machine do not know that “Hey” and “hey” are the same thing

- Which elements should be removed depend on the kind of text we have and on the kind of analysis to be done

Text Pre-Processing – main steps

2. Tokenization:

   Identification of the basic elements composing a text: **tokens**. A text can be divided into sentences or words.

```
I saw a dog
```

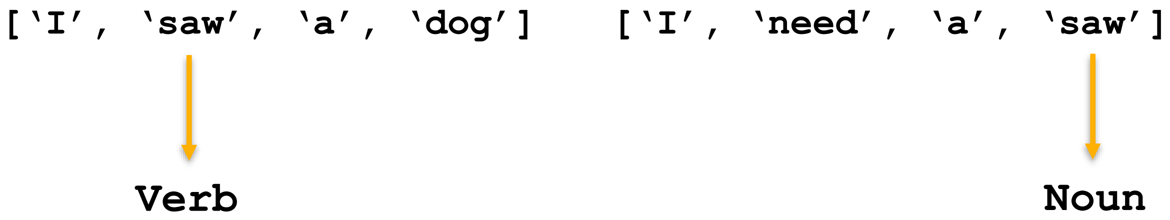
↓

```
['I', 'saw', 'a', 'dog']
```

Text Pre-Processing – main steps

3. Part of Speech (PoS) tagging:

   Assignment of the corresponding Part of Speech (noun, verb, adjective, etc.) to each term.
   This can help with ambiguity.



```
['I', 'saw', 'a', 'dog']      ['I', 'need', 'a', 'saw']
        Verb                                      Noun
```

Text Pre-Processing – main steps

3. Part of Speech (PoS) tagging:

NLTK uses many different and specific POS labels:

NN : common noun (singular)
NNP : proper noun
NNS : common noun (plural)
…

Nouns

VB : verb, base form
VBD : verb, past tense
VBN : verb, past participle
VBP : verb, present tense
…

Verbs

Text Pre-Processing – main steps

4. Lemmatization

   Texts contain different forms of a word (e.g. *organise, organises, organising*) or derivationally related words with similar meanings (e.g. *democracy, democratic, democratization*). It is often useful to reduce inflectional and derivationally related forms to a common base (**lemma**)

```
[‘I’, ‘saw’, ‘two’, ‘dogs’]
          ⬇
[‘I’, ‘see’, ‘two’, ‘dog’]
```

Text Pre-Processing – main steps

5. Stemming

   As for lemmatization, stemming reduces each word to a root form (**stem**). But, differently to the lemma, this can result in a lexically incorrect or non-meaningful word

```
[‘I’, ‘was’, ‘eating’]
```

⇩

```
[‘I’, ‘wa’, ‘eat’]
```

Text-Preprocessing

With Python

Open Jupyter Notebook and *4-Introduction_to_NLTK* file
You can download it at https://github.com/SimoneBarandoni/nlp-python