NLTK Cheatsheet

For workshops and meetups go to: https://research.unimelb.edu.au/infrastructure/research-platform-services/training/nltk (https://research.unimelb.edu.au/infrastructure/research-platform-services/training/nltk)

In [1]:

```
import nltk # import the library
from nltk import word_tokenize
```

1. Access a corpus and preprocess

In [ ]:

```
file = open('datasets/...', "r", encoding='UTF-8') # open and read a file
text = file.read()
words=nltk.tokenize.word_tokenize(text) # tokenize the text
```

2. Filter by part-of-speech (nouns)

In [ ]:

```
tagged_words=nltk.pos_tag(words) # part of speech
nouns=[]
for tagged_word in tagged_words: # filter by nouns
    word=tagged_word[0].lower()
    tag=tagged_word[1]
    if tag.startswith('NN'):
        nouns.append(word)
```

3. Analise the text

In [ ]:

```
frequency_distribution=nltk.FreqDist(words) # count words and make interpretation
frequency_distribution.most_common(10)
```