INF5830, 2015, Group 1 – Getting started with Python and NLTK

NLTK is a Python library for natural language processing. We will use it in INF5830. Hence we will also need some knowledge of Python itself. Later on we will use other Python libraries, like Scipy.

I know your background is varied. Some of you have no experience with NLTK, nor with Python. Others have met them both in INF1820 or INF2820, and some of you may have more Python experience from other courses or work. This variation is a challenge. But it shouldn't be too difficult for them with no prior experience with these tools to catch up. It might require some extra efforts the first weeks, though. You should aslo try to learn from each other, from the experiences of your fellow students.

Choice of Python

We start by discussing various Python interfaces: (python, ipython, idle, anaconda, spyder, canopy, others?), various editors and their relative advantages.

Initialization

To get access to the nltk data you should first add the following line to the end of your .bashrc-file.
```
export NLTK_DATA=/projects/nlp/nltk_data
```
Then logout and login again, and you're all set.

Part 1 (for those with no background in Python, or those who need a fresh up)

The NLTK book teaches NLTK and Python simultaneously. We mainly follow the book. While reading the book, you should sit on the terminal and type the examples from the book.

- Start with section 1.1 and section 1.2.
- Then do exercises : 1, 2, 3, 8, 16, 19 from section 1.8
- Work through section 3.2 Strings
- Do exercises: 9, 10, 13 from section 1.8
- And exercises 2, 4, 5, 10 from section 3.12

Python

We assume you know how to program in some language or other, but you would at some time have to learn the quirks and quiddities of Python. Where to look?

- Of course, the NLTK book, e.g. sec.1.4, 2.3, (and eventually parts of) ch. 4
- Sooner or later you will have to consult the excellent official Python documentation, in particular the tutorial and library reference. (We are using Python 2.x – not 3.x)
- The background we assume is covered by INF1820 and. Some of the slides from this page may be useful (in particular for those who have taken INF1820 but started to forget…)
- INF1820 also recommends How to think like a computer scientist: Learning with Python as an easy introduction to Python.

Part 2 – For everybody

Exercise 1: Work through the NLTK.book sec 1.3 where you meet the NLTK favorite tool: FreqDist and make a first encounter with two key concepts we will meet again: Bigram and Collocation

Exercise 2: NLTK makes a lot out of the FreqDist class. To understand it better, it may be useful to see that the core is a Python dictionary. Make a Python function which takes a list, *j*, as an argument and returns a dictionary, *d*. The dictionary *d* should take the members of *j* as keys and to each key, *k*, return the number of occurrences of *k* in *j*. Compare *d* to *d2*, constructed by d2 = FreqDist(j). Do they have the same keys? To they get the same value for each key?

Exercises 3-5: Section 1.8: exercises 22, 26, 28

Exercise 6: Make a Python function which takes a list of numbers and returns the median? (Hint: sort the list)

Exercise 7: Make a Python function which takes a list of numbers and returns the mean. (Hint: cf exercise 1.8.26 above)

Exercise 8: Make a function which takes a set of numbers and returns the variance. (Hint: You may import sqrt from maths).

Exercise 9: Work through NLTK book sec. 2.1 up to "Annotated Text corpora"

Exercise 10: NLTK book sec. 2.2

Exercises 11-14: NLTk-book sec.2.8: exercises 4, 8, 15

If your already an expert

and need some challenges, do exercise 23 from section 2.8