NLTK Tutorial

CSC485/2501 – Fall 2017

UNIVERSITY OF TORONTO

Where to code things

- Computing Disciplines Facility (CDF)
  - Teaching (computer) labs
    - BA2200, BA2210, BA2220, BA2230, BA2240, BA2270, BA3175, BA3185, BA3195, BA3200, BA3219
    - Accessible by T-Card
  - Computing environments (UNIX)
- Administrative office located in BA3224
- **You must must must make sure all your code runs on the CDF computing environments!**
- https://www.cdf.toronto.edu

UNIVERSITY OF TORONTO

CDF account

- Must be enrolled in a CS course (this one counts)
- Account name most likely your UTORid
- Password initially your student number
- Here's a guide: http://www.cdf.toronto.edu/resources/intro_for_new_students.html

UNIVERSITY OF TORONTO

Connecting from outside

- Desktop: NX (Windows, Mac, Linux)
  - Surprisingly fast, but buggy
  - https://www.teach.cs.toronto.edu/nx/
- Command line: ssh
  - Can open GUIs, but slow

    ```
    $ ssh -X <cdf_id>@cdf.toronto.edu
    ```

  - Can set up public key/private key for password-free access
- File transfer: stfp, rsync

UNIVERSITY OF TORONTO

Electronic submission

- When you're asked to submit code
- You can do so either…
  - From the command line:

```
$ submit -c csc{485,2501}h -a <assignment_name>
-f <file_1> [<file_2> [...]]
```

  - From the CDF Student Secure Website:
    - https://www.teach.cs.toronto.edu/students/

UNIVERSITY OF TORONTO

Python

- High-level scripting language
- "Readable" code
- Duck typing
- Auto garbage collection and memory management
- Large standard library
- Great documentation!
- Centralized package repositories
  - Conda, PyPI (cool stuff!)
  - Not on CDF though ☹
- We are using python 3.5!

Python 2 vs. 3 in a nutshell

```
1 / 2 # py2: 0, py3: .5
'é' # py2: '\xc3\xa9', py3: 'é'
print 'foo' # py2: prints "foo" to stdout, py3: syntax error
print('foo') # same
try:
    raise IOError, 'foo' # py2: IOError, py3: syntax error
    raise IOError('foo') # same
except IOError as e: # same
    pass
type(range(3)) # py2: list, py3: range (a generator)
```

- Won't matter much for you
  - But for more info: http://sebastianraschka.com/Articles/2014_python_2_3_key_diff.html
  - And if you like cross-compatibility: http://python-future.org/compatible_idioms.html

UNIVERSITY OF TORONTO

Python Shell vs. Script

- Python can both be *compiled* and *interpreted*
- Interpreted: python shell
  - One line at a time, see the return values

```
$ python3.5
```

- Compiled: python scripts
  - Or make a file with the suffix .py and call

```
$ python3.5 <file_name>
```

  - Runs until completion

IDEs + editors

- Editors are computationally cheap and often have syntax highlighting
  - On CDF: nano, vim, emacs, gedit, kate
- Interactive Development Environments (IDEs) have all sorts of bells and whistles
  - On CDF: idle, wing-101
  - Other: PyCharm
    - https://www.jetbrains.com/pycharm/

BIRDS!

UNIVERSITY OF TORONTO

Natural Language Toolkit (NLTK)

- Python package that implements many standard NLP data structures and algorithms
- First developed in 2001 as a part of a CL course at the University of Pennsylvania
- We are using NLTK 3
  - Reference for all things NLTK: http://www.nltk.org/book/
- Simple; modular; not optimized with tricks

UNIVERSITY OF TORONTO

Modules in NLTK

Task	Module	Functionality	Chpt
Accessing corpora	nltk.corpus	Standardized interface for accessing corpora	2
Finding resources	nltk.data	Convenient way to find text resources. File, URL, etc	2, 3
String processing	nltk.stem, nltk.tokenize	Stemming/lemmatizing, tokenizing	3
Collocation discovery	nltk.collocations	Methods for finding collocations (bigrams, etc.)	3
POS tagging	nltk.tag	Various part-of-speech taggers	5
Classification	nltk.classify	Various statistical classifiers + NLP feature sets	6
Chunking	nltk.chunk	Non-overlapping parsing of text	7
Parsing	nltk.parse	Various text parsers	8
Grammars	nltk.grammar	Grammar definitions for parsing	8, 9
Semantics	nltk.sem	First order logic and sets. Agents being agents.	10
Understanding	nltk.translate, nltk.wsd	Machine translation; word sense disambiguation	1-ish

UNIVERSITY OF TORONTO

Installing NLTK elsewhere

- **It still has to run on CDF machines!**
- Various python distributions
  - Standard
    - Install Python 3.5: https://www.python.org/downloads/
    - Then:
      ```
      $ pip install nltk
      ```
  - Anaconda
    - Install Python 3.5: https://www.continuum.io/downloads
    - Then:
      ```
      $ conda install nltk
      ```
- Side note: iPython is a really nice interpreter

```
$ pip install ipython
$ conda install ipython
```

MORE DEMOS!!!!

UNIVERSITY OF TORONTO