CSC 5930/9010: Text Mining
Lab 2

Dr. Paula Matuszek
Paula.Matuszek@villanova.edu
Paula.Matuszek@gmail.com
(610) 647-9789
Goals

• Goals for this lab are:
  • Get started with Python
  • Get started with NLTK
• If you have a laptop with you, get them installed on it.
What is NLTK?

- Natural Language ToolKit
- Set of modules for Python
- A large number of methods for importing and manipulating text
- A large set of relevant data
- Starting point is http://www.nltk.org/.
Why NLTK?

- Easy to get started
- Powerful set of tools useful for preparing documents for mining
- Several text mining tools built in
- Very well documented
- However: this is not a NLP course, and we will be ignoring much of NLTK which isn’t important for our topics
Okay, What is Python

- Python is an object-oriented, interpreted language.
- Unlike Java, it has very little required structure.
- IDLE is a simple integrated development for Python and the easiest way to use it
- Can also be run at command line

We will cover just some basics in class, more or less the minimum to use NLTK effectively
Why Python? (besides NLTK)

- Built-in types for strings, lists, dictionaries
- Strong numeric processing capabilities
- Clean syntax, powerful extensions
Getting Started: Python

- Do you already have Python installed?
- If you have a mac, Python is bundled with it. To check which version, open a terminal window and type “python”. You should get something like

  Paladin:$ python
  Python 2.7.2 (v2.7.2:8527427914a2, Jun 11 2011, 14:13:39)
  [GCC 4.0.1 (Apple Inc. build 5493)] on darwin
  Type "help", "copyright", "credits" or "license" for more information.

  >>>

- If you have Windows you should be able to open a command window, type “python”, and get something similar if you already have it.

- If you don’t already have Python it’s easiest to install from the NLTK page (next slide)

Getting Started: NLTK

  • Starting URLs for the systems we will be using: downloads, documentation, etc.

• This is what we are actually interested in.

• [http://nltk.org/install.html](http://nltk.org/install.html) has detailed installation instructions for macs/windows/unix.
Version Nightmares...

• Biggest problem you are likely to run into is version incompatibilities.
  • Latest version of Python is 3.2, but Python 3.x does not preserve backward compatibility with 2.x.
  • Latest stable version of NLTK is 2.0.4
    • It is not compatible with Python 3.x.
    • Docs say works with Python 2.6.x, 2.7.x. Worked for me with 2.7.2.
    • There is an alpha release of NLTK 3.0 which is compatible with Python 3; don’t do it unless you’re exceptionally brave.
    • I recommend using Python 2.7 and NLTK 2.0.4.
NLTK Data

- http://www.nltk.org/data

- Data include corpora, dictionaries, gazetteers, trained models

- http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml gives a list, with downloadable versions

- We looked at these last week.
Go! Step 1

- Start at http://nltk.org/install.html. This gives instructions for installing everything you will need and an initial test.

- You will need to be sure you have
  - Python and IDLE, PyYAML, NLTK, numpy

- If you didn’t bring a laptop either work with someone who has the same system or go on to the next slide with the PCs in G87

- If you run into trouble, there is more information at https://sites.google.com/site/naturallanguagetoolkit/download
Step 2

• Go to https://sites.google.com/site/naturallanguagetoolkit/getting-started

• Follow the instructions to make sure you can run Python, IDLE and NLTK.

• Test some of the demos at the bottom of the page.
Step 3

• Look through the preface of the NLTK book by Bird et al: http://www.nltk.org/book

• Work through chapter 1, through section 1.4. This will include both some Python and some NLTK methods.

• More version nightmares... The book assumes that you are running Python 2.4 or 2.5 and NLTK 2.0. I haven’t found any significant differences in the early tutorials, but there may be some.
If Python is new to you and you feel like you need more practice, work through the exercises at the end of Ch. 1. The ☼ exercises are easy; moderately difficult exercises involving more Python are marked with ☉.

If you’re already familiar with Python and NLTK, you have a head start! See if you can help some of your fellow students.

Start assignment 2. Finish it as homework, if necessary.