Course

• Webpage for lecture slides and Panopto recordings:
  – http://elmo.sbs.arizona.edu/~sandiway/#courses
  – available from just before class time
    • (afterwards, look again for corrections/updates)
  – in .pptx (good for animations) and .pdf formats

• Meeting information
  – Tues/Thurs 3:30-4:45pm. McClellandPark, Room 102.
Course Objectives

• **Follow-on course to LING/C SC/PSYC 538 *Computational Linguistics***:
  - pre-requisite: 538
  - continue with selected material from the 538 textbook (J&M):
    • 25 chapters, a lot of material not covered in 438/538

• **And gain more extensive experience**
  - with new stuff **not in textbook**
  - dealing with natural language software packages
  - Installation, input data formatting
  - operation
  - project exercises
  - useful “real-world” computational experience
  - abilities gained will be of value to employers
Computational Facilities

- **Use your own laptop/desktop**
  - *can also make use of the computers in this lab*
    - but you don’t have installation rights on these computers

- **Platforms**
  - Windows is *maybe* possible but you really should run some variant of Unix...
  - Linux (e.g. Ubuntu, separate bootable partition or via virtualization software if you use Windows)
  - OSX
    - Not quite Linux, some porting issues, especially with C programs, can use Virtual Box (Linux under OSX)
    - Or Macports or Homebrew (no need for virtualization)
Grading

• Completion of all homework tasks will result in a satisfactory grade (A)

• Tasks typically should be completed before the corresponding class next week.
  – email me your work (sandiway@email.arizona.edu).
  – also be prepared to come up and present your work (if called upon).

• Office hours (by appointment):
  – also after class
Syllabus

• Homeworke
  – you may discuss questions with other students
  – however, you must write it up yourself (*in your own words, your own code etc.*)
  – cite (web) references and your classmates (*in the case of discussion*)
  – Student Code of Academic Integrity: *plagiarism* etc.
    • [http://deanofstudents.arizona.edu/codeofacademicintegrity](http://deanofstudents.arizona.edu/codeofacademicintegrity)

• Revisions to the syllabus
  – “the information contained in the course syllabus, other than the grade and absence policies, may be subject to change with reasonable advance notice, as deemed appropriate by the instructor.”
Python 3

Grab from python.org

- non-standard on MacOS
  - (Apple: Python 2.7.10)
- which python
  - /usr/bin/python
- which python3
  - /Library/Frameworks/Python.framework/Versions/3.5/bin/python3
- printenv PATH
Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The book is being updated for Python 3 and NLTK 3. (The original Python 2 version is still available at http://nltk.org/book_1ed.)
NLTK 3.2.5 Install

- See http://www.nltk.org/install.html
- Use pip (pip3 for python3) to install packages from the Python Package Index (PyPI)
  - sudo pip3 install -U nltk
NLTK Data Install

• See http://www.nltk.org/data.html
• python3
Homework 1

• Part 1:
  – Install (if you haven't already) a latest or recent version of Python 3, NLTK, and NLTK Data on your laptop

• Part 2:
  – The Penn Treebank is partially installed as a corpus in NLTK Data (Sections 00 and 01: wsj_0001.mrg to wsj_0199.mrg)
  – from nltk.corpus import treebank

```python
>>> treebank.words()
['Pierre', 'Vinker', '.', '.', '65', 'years', 'old', '.', '...']
>>> len(treebank.words())
108676
```
Homework 1

Part 2 (contd.)

- Submit relevant parts of your python console + a screenshot of a random tree from the treebank

- Parsed Sentence (example):
  - `t = treebankparsed_sents("wsj_0199.mrg")[0]`

- Tree drawing:
  - `t.draw()`

- Random number generation:
  - `import random`
  - `random.seed()`
  - `random.randrange(1, 200)` returns some x: `1 <= x < 200`
Part 3: Alice in Wonderland

- nltk.corpus.gutenberg.fileids()
- raw = nltk.corpus.gutenberg.raw('carroll-alice.txt')
- sents = nltk.sent_tokenize(raw)
- len(sents)
  1625
Homework 1

Part 3 (contd.): Alice in Wonderland

- Part of speech tag a random sentence from the book and comment on the accuracy of the tagging.
- Submit your console
- use
  - nltk.tokenize.word_tokenize(string)
  - nltk.pos_tag(list_of_words)

- Example:
  - [('The', 'DT'), ('Mouse', 'NNP'), ('only', 'RB'), ('growled', 'VBD'), ('in', 'IN'), ('reply', 'NN'), ('.', '.')]
Homework 1

• Due date:
  – next Monday (midnight)
  – by email (Subject: 581 Homework 1)
  – submit one PDF file (put your name at the top of the file)