PIC 16, Winter 2018 – Preparation 8F

Assigned 2/26/2018. To be completed by class 3/2/2018.

Intended Learning Outcomes

By the end of this preparatory assignment, students should be able to:

- download data for practice analysis from the NLTK repository,
- explore word usage with NLTK's `concordance`, `common_contexts`, and `dispersion_plot` functions,
- calculate a crude metric of the lexical diversity of a text by comparing the ratio of word (token) *types* to total words,
- calculate the frequency of each word type in a text, just as in 1F, and other word metrics using the `FreqDist` function, and
- learn from the NLTK book as needed.

Tasks

- ☐ Read the introduction to Chapter 1 of the NLTK Book. Note that this version of the book is for Python 3, but it can be used (with a few changes to the code) with Python 2. If you are having trouble with version issues, you can find the first edition of the book here.
- ☐ Follow 1.2, 1.3, and 1.4
- ☐ You should be able to skip Chapter 1, section 2 ("A Closer Look at Python: Texts as Lists of Words")
- ☐ Follow all of Chapter 1, section 3 ("Computing with Language: Simple Statistics")
- ☐ You should be able to skip Chapter 1, section 4 ("Automatic Natural Language Understanding")
- ☐ Skim Chapter 1, section 5. This will give you a good overview of the issues in natural language processing.
- ☐ Try this Track B Final Exam option from last year (one of three that were offered). (We've been working with Jupyter notebook all quarter, but if you need a refresher, see the "More on Jupyter" video posted 1M).