Tools and Techniques for Speech and Language Processing

University of Illinois at Urbana-Champaign

Week 03 of 16 — Day 05 of 29

Quiz (~15 minutes)

Practice time

The following exercise expects you to:

- Write a bash script that reads from stdin (w/o `read` command)
- Make use of bash variables and expansions
- Get the output of a single python command from the command line (w/o opening the python interpreter)

Measuring the size of presidents' vocabularies

Write a bash script which accepts text from stdin, and then uses shell commands and variables as well as python commands to:

- remove punctuation (,.;:) from the text
- output the total number of **types** (unique tokens) in the text
- output the total number of **tokens** in the text
- output as a float the **lexical diversity** (types/tokens) of the text

Then, use your script to find the types, tokens, and lexical diversity of three U.S. presidential inaugural addresses of your choice from:

`/usr/share/nltk_data/corpora/inaugural/`

Measuring the size of presidents' vocabularies

What does your numerical value of lexical diversity mean?

What are some ways you might improve on preprocessing the text before calculating lexical diversity? Just think about what you would do, even if you don't know how to do it yet.