Tools and Techniques for Speech and Language Processing

University of Illinois at Urbana-Champaign

Week 02 of 16 — Day 03 of 29

Quiz (∼10 minutes)

Review HW01 (~10 minutes)

Review HW02 (∼5 minutes)

Present HW03 (∼5 minutes)

Practice piping commands (∼20 minutes)

Use shell commands and pipe operators to identify the most frequent tokens in the English Europarl corpus:

```
/usr/share/nltk_data/corpora/europarl_raw/english/*
```

Practice piping commands (∼20 minutes)

Use shell commands and pipe operators to identify the most frequent word lengths in the English Europarl corpus

```
/usr/share/nltk_data/corpora/europarl_raw/english/*
```

In other words, before doing other sorting, etc, convert all English tokens to placeholders based on their length. There is more than one way to do this.

Advanced option

If you get through the previous work quickly,

- Put the previous commands in a shell script
- Allow the user to specify which language via a command line argument