

Applications of NLP

Hal Daumé III

School of Computing
University of Utah

me@hal3.name



The Basic Concepts

- Corpora, sentences, words, types/tokens
- Morphological analysis
- Syntactic analysis
- Word senses
- Phonemes

Parts of Speech

- Standard linguistic taxonomy:
 - Determiner the, a, an, that
 - Noun cat, ball, computer
 - Verb run, eat, digest
 - Adjective happy, loud, melancholy
 - Adverb quickly, very, not

- This is insufficient for our purposes

Parts of Speech, 3 Styles

See tags.jpg

Word Senses

- Standard examples:
 - bank
 - java
 - saturn

- Less obvious example: **run**

Word Senses

- Standard examples:
 - bank
 - java
 - saturn

- Less obvious example: **run** 42 senses of the verb!
 - travel rapidly, speed, hurry, zip
 - leave, go forth, go away
 - circulate, disperse, pass around
 - direct
 - move
 - function, work, operate, go
 - carry through, accomplish, execute
 - ...

Unix Tools

- You know become friends with:
 - wc
 - tr
 - sed
 - sort
 - uniq
 - cut

- (and maybe some others)

Unix Tools

- You know become friends with:
 - `wc` counts characters/words/lines
 - `tr` translates single characters
 - `sed` replaces regular expressions
 - `sort` sorts...duh!
 - `uniq` merges replicated lines
 - `cut` selects specific columns
 - `paste` generates multiple columns
 - `head/tail` list first/last lines
 - `grep` searching in a file

- (and maybe some others)

Unix Demos

- Lower-casing an entire files
- Finding all words not made entirely up of letters
- Splitting out hyphens
- Computing a vocabulary
 - Plus a frequency-sorted vocabulary
- Finding all words with non-unique capitalization
 - Finding all non-sentence initial words with n-u c
- Creating type-token curves (with xgraph)