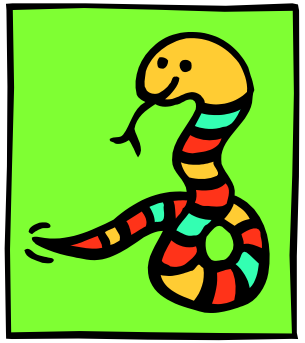
NLTK Texts and Corpora

LING 131A, Fall 2018

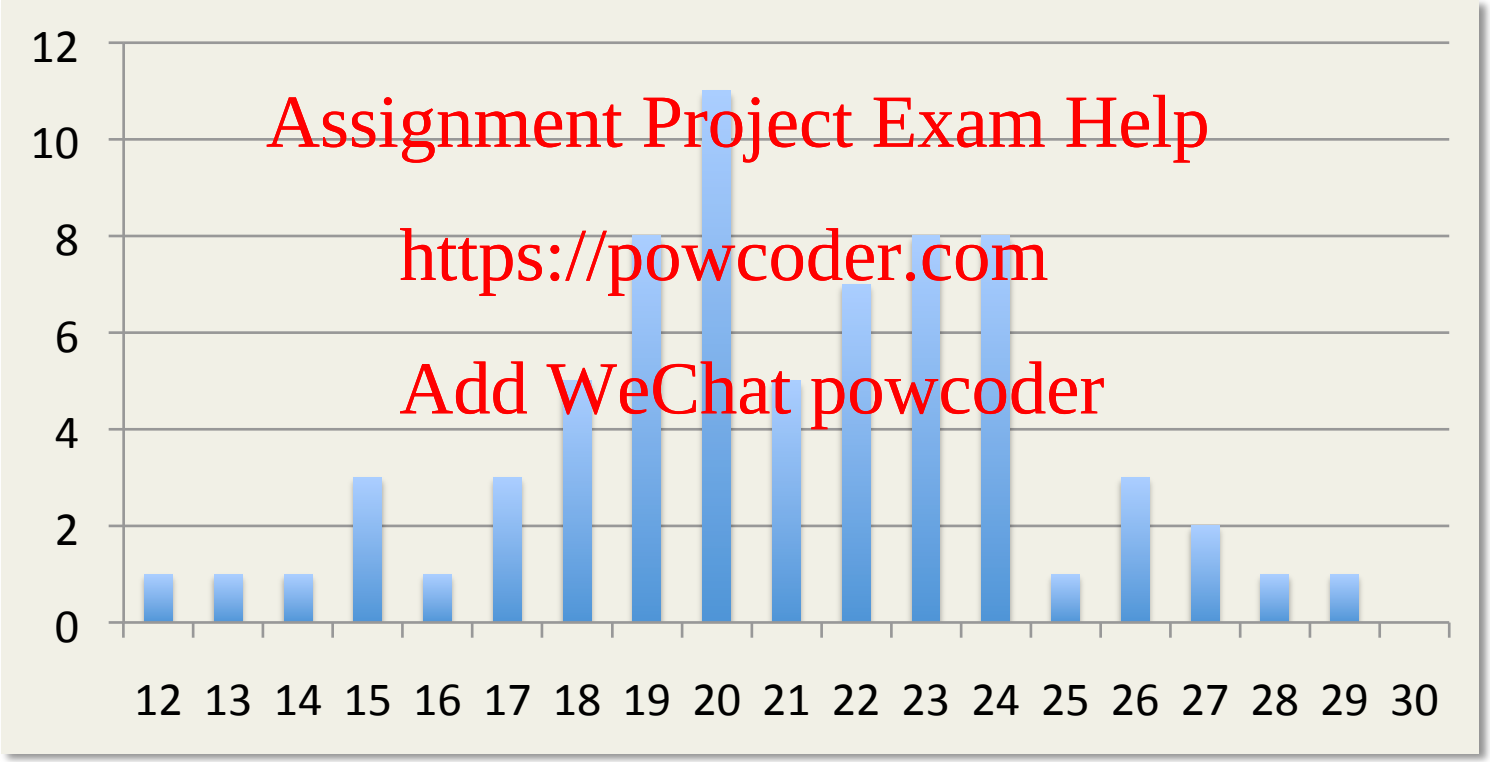
Marc Verhagen, Brandeis University



Today

- Quiz 1 results
  - questions 8, 9, 11 and 12 updates
- Putting list comprehensions to bed
  - [[(x,y) for x in s1] for y in s2]
- Animals, Dogs and a Zoo
  - doctest revisited
- Coding standards
- NLTK Texts and Corpora
- Assignment 3

Quiz 1 Results



List comprehensions

Aminals in the Zoo

Python style guide

- https://www.python.org/dev/peps/pep-0008/
- Install and run

```
$ pip3 install pep8
$ pep8 some_script.py
```


- Key insight: code is read much more often than it is written. Style guides are about consistency and readability

- But… always:
  - Guides are just guides
  - There is danger in getting distracted too much by the guide
  - A Foolish Consistency is the Hobgoblin of Little Minds

Corpora

- Large set of text
  - Structured (has levels of annotation)
  - Balanced
  - Samples of real world text
    - competence versus performance
- Used for
  - Hypothesis testing
  - Statistical analysis
  - Training examples for supervised machine learning

Distribution

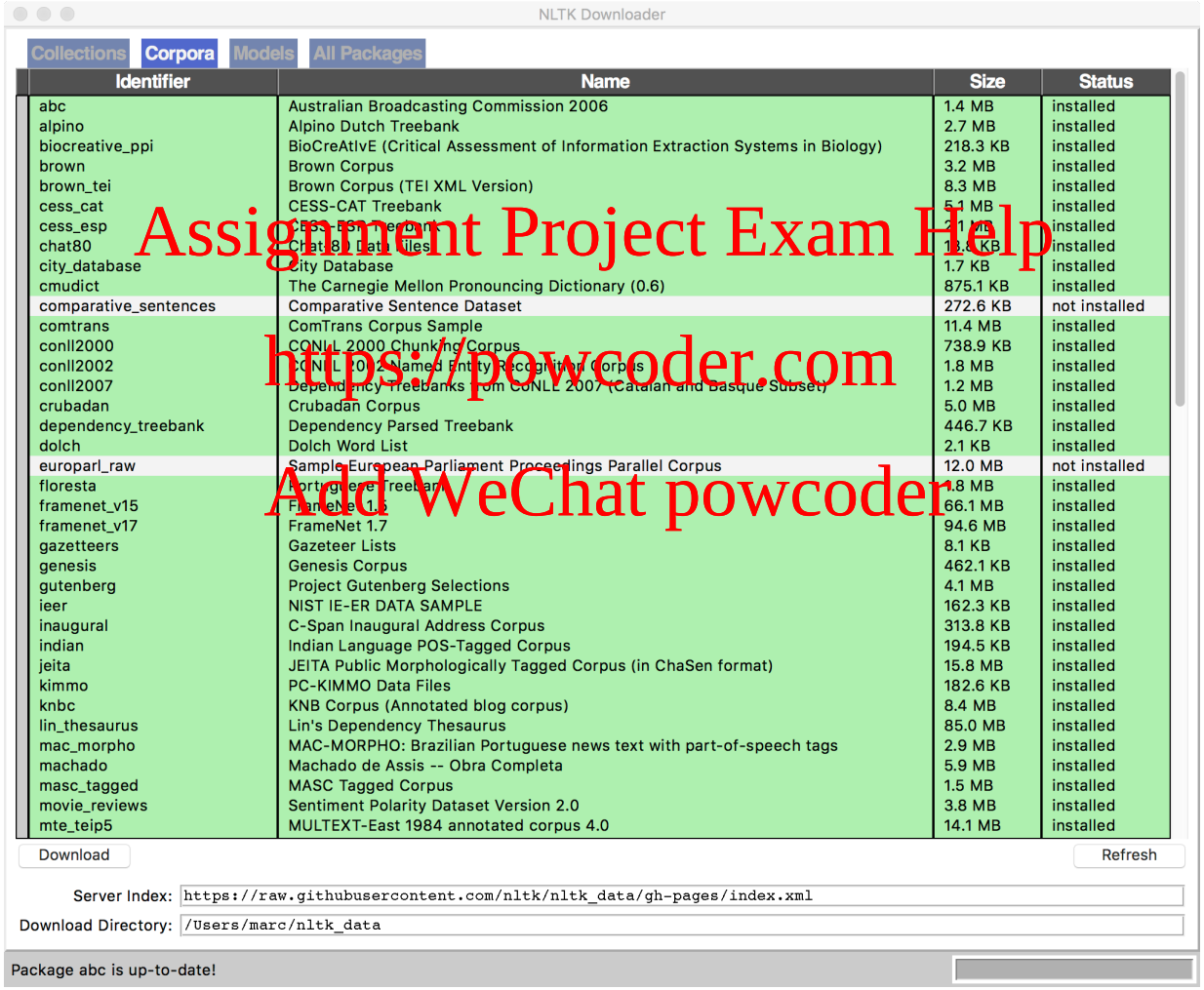
- "You know a word by the company it keeps"

- Distribution

  - Frequency distribution
  - Neighboring words

    - Concordance/KWIC
    - Collocations
  - Similar words
    - Words that have the same neighbors

Corpora in NLTK



Corpus reader

- Define common interfaces for corpora that have different formats
- High-level tasks can then just use these common interfaces
- Sample corpus readers implemented in NLTK
  - PlaintextCorpusReader, CategorizedPlaintextReader, CategorizedTaggedCorpusReader, BracketParseCorpusReader, DependencyCorpusReader, WordlistCorpusReader …

Frequency Distributions and Conditional Frequency Distributions

- Taking text provided by corpus readers, we can build frequency distributions and conditional frequency distributions (or to compute probabilities and conditional probabilities)
- Using these distributions, we can inspect data, or build language models or classifiers (See Chapter 6 for building classifiers)

Conditional Frequency Distributions

a collection of frequency distributions, each one for a different "condition"

Condition: News

the	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |||
cute	
Monday	𝍸 𝍸 |||
could	|
will	𝍸 |||

Condition: Romance

the	𝍸 𝍸 |||
cute	|||
Monday	|
could	𝍸 𝍸 𝍸
will	||||

NLTK

- Text
- FreqDist
- CorpusReader
  - PlainTextCorpusReader
  - CategorizedTaggedCorpusReader
- ConcatenatedCorpusView
- StreamBackedCorpusView