

NLTK Texts and Corpora

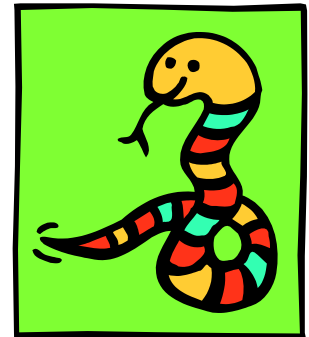
Assignment Project Exam Help

<https://powcoder.com>

LING 131A, Fall 2018

Marc Verhagen, Brandeis University

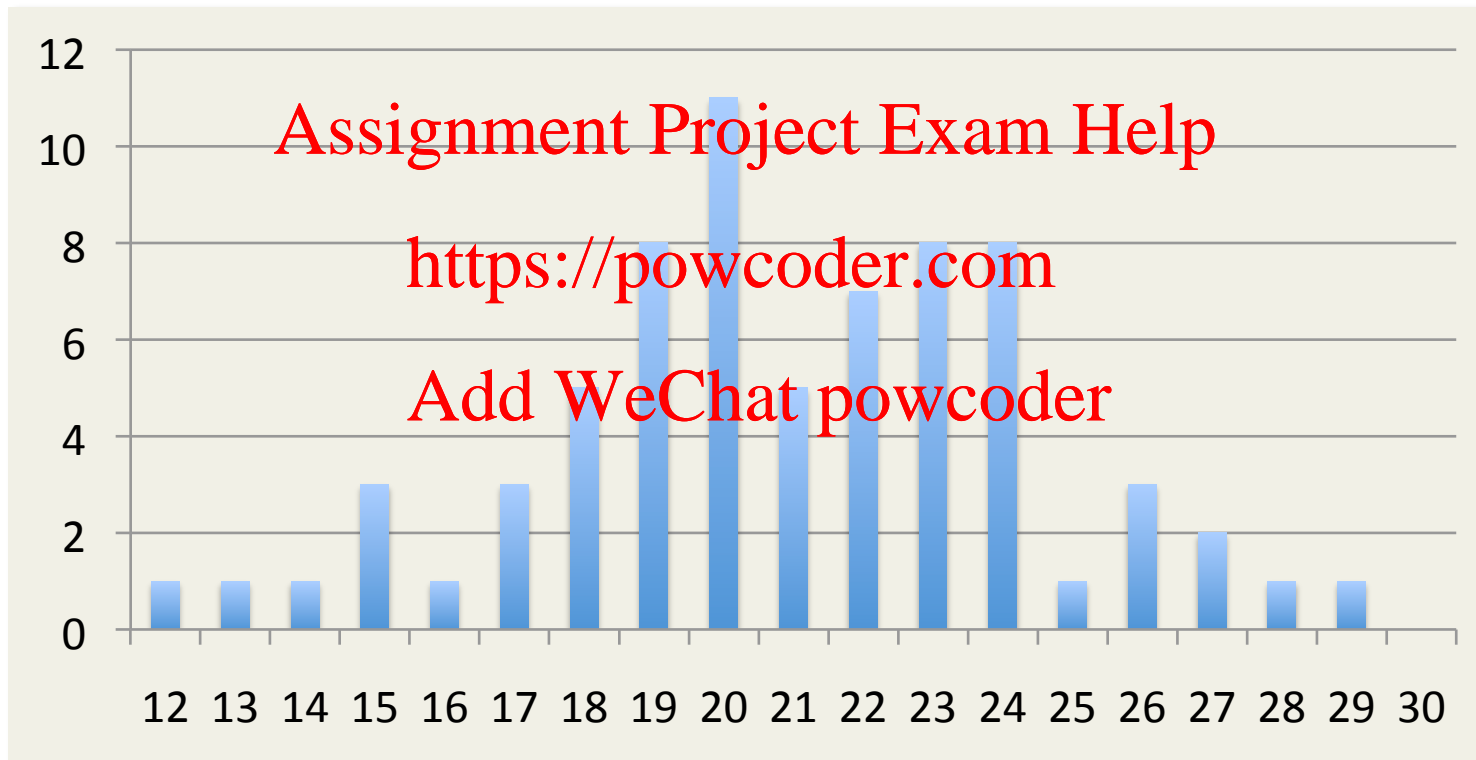
Add WeChat powcoder



Today

- Quiz 1 results
 - questions 8, 9, 11 and 12 updates
- Putting list comprehensions to bed
 - `[(x,y) for x in lst1 for y in lst2]`
- Animals, Dogs and a Zoo
 - doctest revisited
- Coding standards
- NLTK Texts and Corpora
- Assignment 3

Quiz 1 Results



List comprehensions

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Aminals in the Zoo

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Python style guide

- <https://www.python.org/dev/peps/pep-0008/>

- Install and run

```
$ pip3 install pep8
```

```
$ pep8 some_script.py
```

Assignment Project Exam Help

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

<https://powcoder.com>

Add WeChat powcoder

- 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"

Assignment Project Exam Help

- Distribution

<https://powcoder.com>

- Frequency distribution

- Neighboring words

Add WeChat powcoder

- Concordance/KWIC

- Collocations

- Similar words

- Words that have the same neighbors

Corpora in NLTK

Assignment Project Exam Help
<https://powcoder.com>
 Add WeChat powcoder

NLTK Downloader

Collections Corpora Models All Packages

Identifier	Name	Size	Status
abc	Australian Broadcasting Commission 2006	1.4 MB	installed
alpino	Alpino Dutch Treebank	2.7 MB	installed
biocreative_ppi	BioCreative (Critical Assessment of Information Extraction Systems in Biology)	218.3 KB	installed
brown	Brown Corpus	3.2 MB	installed
brown_tei	Brown Corpus (TEI XML Version)	8.3 MB	installed
cess_cat	CESS-CAT Treebank	6.1 MB	installed
cess_esp	CESS-ESP Treebank	2.1 MB	installed
chat80	Chat80 Data Files	11.4 KB	installed
city_database	City Database	1.7 KB	installed
cmudict	The Carnegie Mellon Pronouncing Dictionary (0.6)	875.1 KB	installed
comparative_sentences	Comparative Sentence Dataset	272.6 KB	not installed
comtrans	ComTrans Corpus Sample	11.4 MB	installed
conll2000	CoNLL 2000 Chunking Corpus	738.9 KB	installed
conll2002	CoNLL 2002 Named Entity Recognition Corpus	1.8 MB	installed
conll2007	Dependency Treebanks from CoNLL 2007 (Catalan and Basque Subset)	1.2 MB	installed
crubadan	Crubadan Corpus	5.0 MB	installed
dependency_treebank	Dependency Parsed Treebank	446.7 KB	installed
dolch	Dolch Word List	2.1 KB	installed
europarl_raw	Sample European Parliament Proceedings Parallel Corpus	12.0 MB	not installed
floresta	Portuguese Treebank	4.8 MB	installed
framenet_v15	FrameNet 1.5	66.1 MB	installed
framenet_v17	FrameNet 1.7	94.6 MB	installed
gazetteers	Gazeteer Lists	8.1 KB	installed
genesis	Genesis Corpus	462.1 KB	installed
gutenberg	Project Gutenberg Selections	4.1 MB	installed
ieer	NIST IE-ER DATA SAMPLE	162.3 KB	installed
inaugural	C-Span Inaugural Address Corpus	313.8 KB	installed
indian	Indian Language POS-Tagged Corpus	194.5 KB	installed
jeita	JEITA Public Morphologically Tagged Corpus (in ChaSen format)	15.8 MB	installed
kimmo	PC-KIMMO Data Files	182.6 KB	installed
knbc	KNB Corpus (Annotated blog corpus)	8.4 MB	installed
lin_thesaurus	Lin's Dependency Thesaurus	85.0 MB	installed
mac_morpho	MAC-MORPHO: Brazilian Portuguese news text with part-of-speech tags	2.9 MB	installed
machado	Machado de Assis -- Obra Completa	5.9 MB	installed
masc_tagged	MASC Tagged Corpus	1.5 MB	installed
movie_reviews	Sentiment Polarity Dataset Version 2.0	3.8 MB	installed
mte_teip5	MULTEXT-East 1984 annotated corpus 4.0	14.1 MB	installed

Download Refresh

Server Index: https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml

Download Directory: /Users/marc/nltk_data

Package abc is up-to-date!

Corpus reader

- Define common interfaces for corpora that have different formats
- High-level tasks can then just use these common interfaces <https://powcoder.com>
- 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)
<https://powcoder.com>
Add WeChat powcoder
- 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"

Assignment Project Exam Help

Condition: News		Condition: Romance	
the		the	
cute		cute	
Monday		Monday	
could		could	
will		will	

NLTK

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

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder