Lecture 23

408/508 *Computational Techniques for Linguists*

Today's Topics

- **Homework 10?**
  - Did you manage to install nltk and nltk_data?
- Some Python worked exercises today
  - inside nltk and in pure Python
- Python sets and lists
- Queues and Stacks
- Reading and writing files

Exercise 1

- Copy and past into Python
  - 1st and 2nd paragraphs in "*Alice's Adventures in Wonderland*" by Lewis Carroll:

```
text = 'Alice was beginning to get very tired of sitting by her sister
on the bank, and of having nothing to do. Once or twice she had peeped
into the book her sister was reading, but it had no pictures or
conversations in it, "and what is the use of a book," thought Alice,
"without pictures or conversations?"\nSo she was considering in her
own mind (as well as she could, for the hot day made her feel very
sleepy and stupid), whether the pleasure of making a daisy-chain would
be worth the trouble of getting up and picking the daisies, when
suddenly a White Rabbit with pink eyes ran close by her.'
```

1. `len(text)` counts what?
2. What does `text.split()` do?

- Store the result of split into a variable `words`

3. `len(words)` counts what?

Exercise 1

- Let's count the average word length in `text` three different ways:
  1. `len(text)` against `len(words)`
     - **note**: `len(text)` includes spaces
  2. use a variable `total=0`
     - then a for loop

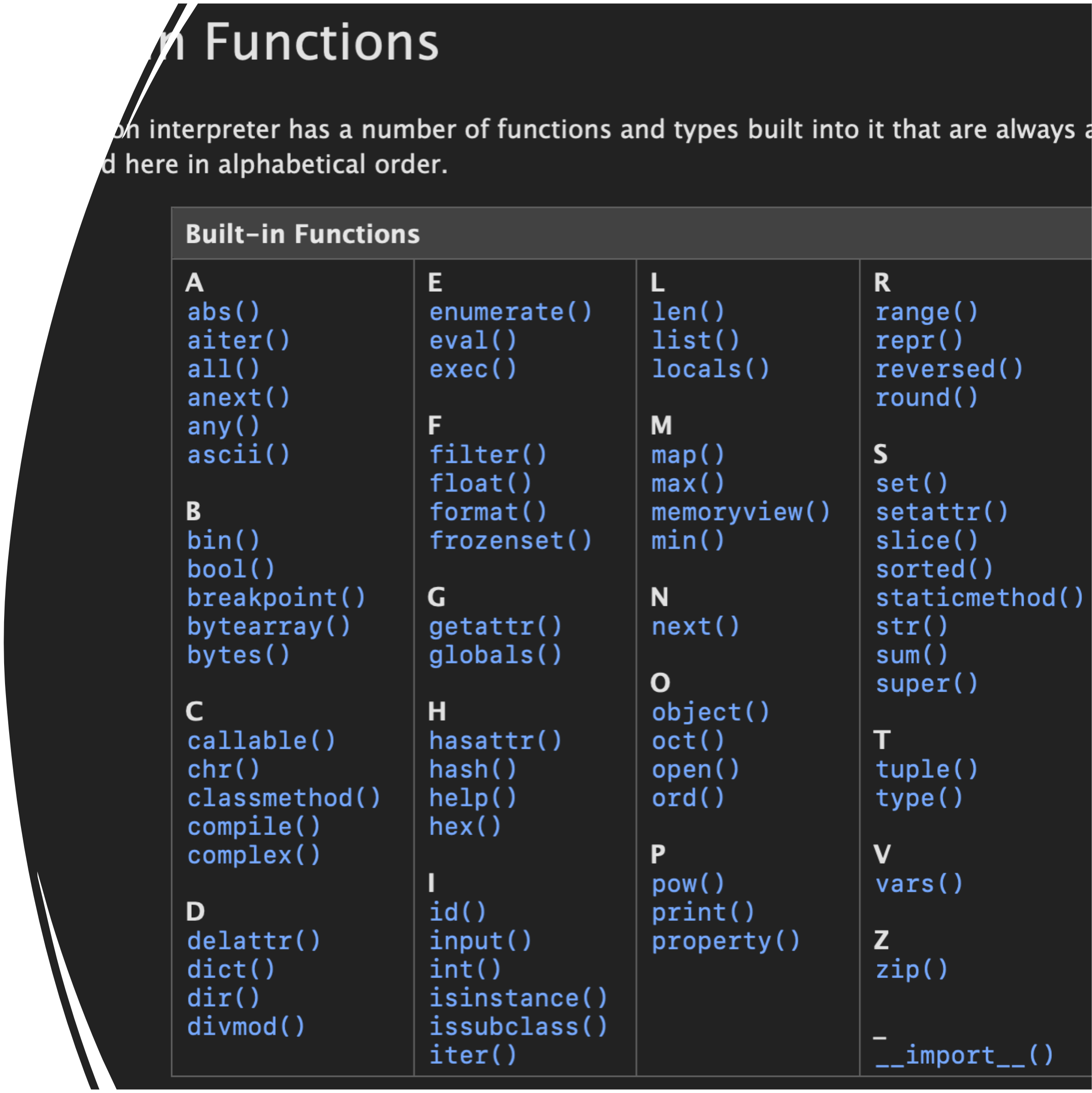
     ```
     for word in words:
             total += len(word)
     ```
  3. use a list comprehension
     - `sum([len(word) for word in words])`

`5.29` (*overestimate*)
`4.31` (*using for-loop*)

Python

Built-in functions

- https://docs.python.org/3/library/functions.html



n Functions

on interpreter has a number of functions and types built into it that are always a
d here in alphabetical order.

Built-in Functions

A
abs()
aiter()
all()
anext()
any()
ascii()

B
bin()
bool()
breakpoint()
bytearray()
bytes()

C
callable()
chr()
classmethod()
compile()
complex()

D
delattr()
dict()
dir()
divmod()

E
enumerate()
eval()
exec()

F
filter()
float()
format()
frozenset()

G
getattr()
globals()

H
hasattr()
hash()
help()
hex()

I
id()
input()
int()
isinstance()
issubclass()
iter()

L
len()
list()
locals()

M
map()
max()
memoryview()
min()

N
next()

O
object()
oct()
open()
ord()

P
pow()
print()
property()

R
range()
repr()
reversed()
round()

S
set()
setattr()
slice()
sorted()
staticmethod()
str()
sum()
super()

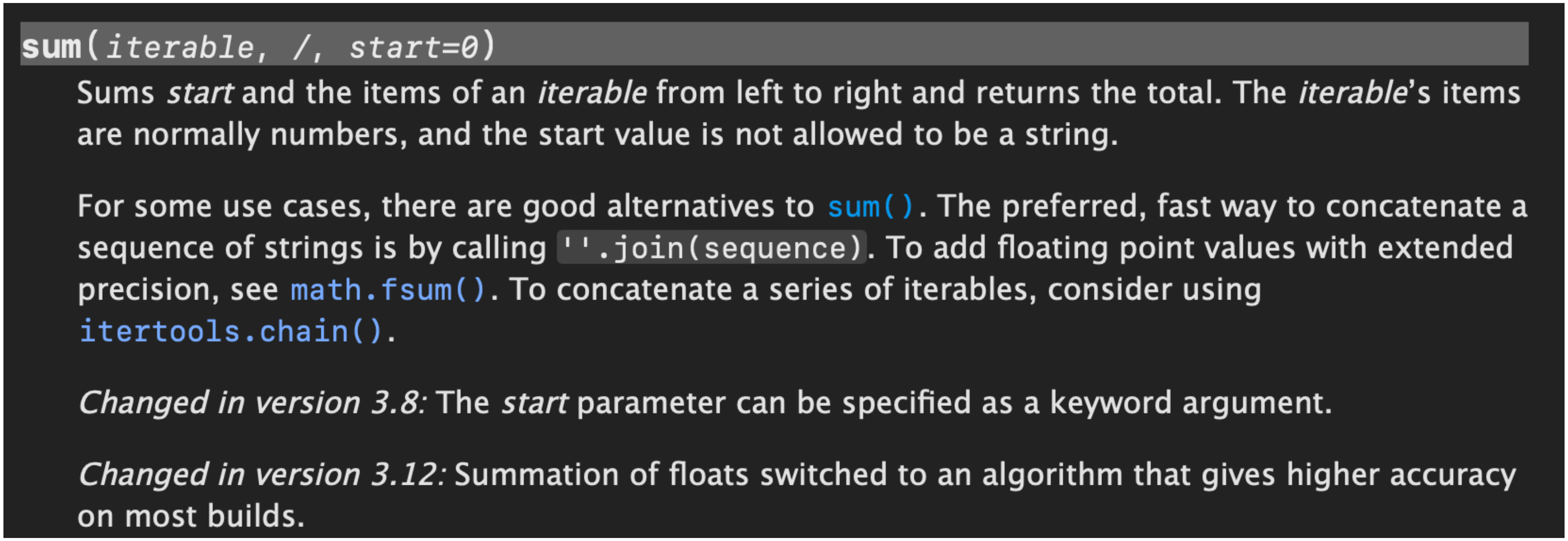
T
tuple()
type()

V
vars()

Z
zip()

_
__import__()

sum()



sum(*iterable*, /, *start=0*)

Sums *start* and the items of an *iterable* from left to right and returns the total. The *iterable*’s items are normally numbers, and the start value is not allowed to be a string.

For some use cases, there are good alternatives to `sum()`. The preferred, fast way to concatenate a sequence of strings is by calling `''.join(sequence)`. To add floating point values with extended precision, see `math.fsum()`. To concatenate a series of iterables, consider using `itertools.chain()`.

Changed in version 3.8: The *start* parameter can be specified as a keyword argument.

Changed in version 3.12: Summation of floats switched to an algorithm that gives higher accuracy on most builds.

nltk data: https://www.nltk.org/nltk_data/

1. *perluniprops: Index of Unicode Version 7.0.0 character properties in Perl* [download | source]
   id: perluniprops; size: 100266; author: ; copyright: ; license: ;

2. *The monolingual word aligner (Sultan et al. 2015) subset of the Paraphrase Database.* [download | source]
   id: mwa_ppdb; size: 1594711; author: ; copyright: ; license: Creative Commons Attribution 3.0 Unported (CC-BY);

3. *Punkt Tokenizer Models* [download | source]
   id: punkt; size: 13905355; author: Jan Strunk; copyright: ; license: ;

4. *RSLP Stemmer (Removedor de Sufixos da Lingua Portuguesa)* [download | source]
   id: rslp; size: 3805; author: Viviane Moreira Orengo (vmorengo@inf.ufrgs.br) and Christian Huyck; copyright: ; license: ;

5. *Porter Stemmer Test Files* [download | source]
   id: porter_test; size: 200510; author: ; copyright: ; license: ;

6. *Snowball Data* [download | source]
   id: snowball_data; size: 6785405; author: ; copyright: ; license: ;

7. *ACE Named Entity Chunker (Maximum entropy)* [download | source]
   id: maxent_ne_chunker; size: 13404747; author: ; copyright: ; license: ;

8. *Moses Sample Models* [download | source]
   id: moses_sample; size: 10961490; author: ; copyright: ; license: ;

9. *BLLIP Parser: WSJ Model* [download | source]
   id: bllip_wsj_no_aux; size: 24516205; author: ; copyright: ; license: ;

10. *Word2Vec Sample* [download | source]
    id: word2vec_sample; size: 49396025; author: ; copyright: ; license: ;

11. *Evaluation data from WMT15* [download | source]
    id: wmt15_eval; size: 383096; author: ; copyright: ; license: ;

12. *Grammars for Spanish* [download | source]
    id: spanish_grammars; size: 4047; author: Kepa Sarasola; copyright: ; license: ;

13. *Sample Grammars* [download | source]
    id: sample_grammars; size: 20293; author: ; copyright: ; license: ;

14. *Large context-free and feature-based grammars for parser comparison* [download | source]
    id: large_grammars; size: 283747; author: ; copyright: ; license: See the individual grammar files;

15. *Grammars from NLTK Book* [download | source]
    id: book_grammars; size: 9103; author: Ewan Klein; copyright: ; license: ;

16. *Grammars for Basque* [download | source]
    id: basque_grammars; size: 4704; author: Kepa Sarasola; copyright: ; license: ;

17. *Treebank Part of Speech Tagger (Maximum entropy)* [download | source]
    id: maxent_treebank_pos_tagger; size: 10156853; author: ; copyright: ; license: ;

18. *Averaged Perceptron Tagger* [download | source]
    id: averaged_perceptron_tagger; size: 2526731; author: ; copyright: ; license: ;

19. *Averaged Perceptron Tagger (Russian)* [download | source]
    id: averaged_perceptron_tagger_ru; size: 8628828; author: ; copyright: ; license: ;

20. *Mappings to the Universal Part-of-Speech Tagset* [download | source]
    id: universal_tagset; size: 19095; author: Slav Petrov; copyright: ; license: CC-BY-SA-4.0;

21. *VADER Sentiment Lexicon* [download | source]
    id: vader_lexicon; size: 90486; author: C.J. Hutto and Eric Gilbert; copyright: ; license: MIT License;

22. *Lin's Dependency Thesaurus* [download | source]
    id: lin_thesaurus; size: 89154019; author: Dekang Lin; copyright: ; license: Distributed with permission of Dekang Lin;

23. *Sentiment Polarity Dataset Version 2.0* [download | source]
    id: movie_reviews; size: 4004848; author: Bo Pang and Lillian Lee; copyright: Copyright (C) 2004 Bo Pang and Lillian Lee; license: Creative Commons Attribution 4.0 International;

24. *Problem Report Corpus* [download | source]
    id: problem_reports; size: 1032942; author: Andrew Ko, Carnegie Mellon University; copyright: ; license: ;

25. *Pros and Cons* [download | source]
    id: pros_cons; size: 746276; author: Bing Liu; copyright: Copyright (C) 2008 Bing Liu; license: Creative Commons Attribution 4.0 International;

26. *MASC Tagged Corpus* [download | source]
    id: masc_tagged; size: 1602143; author: Nancy Ide; copyright: Copyright (C) 2014 American National Corpus; license: This data may be used for the purposes of linguistic education, research, and development, including commercial development.;

27. *Sentence Polarity Dataset v1.0* [download | source]
    id: sentence_polarity; size: 490256; author: Bo Pang and Lillian Lee; copyright: Copyright (C) 2005 Bo Pang and Lillian Lee; license: Creative Commons Attribution 4.0 International;

28. *Web Text Corpus* [download | source]
    id: webtext; size: 646297; author: ; copyright: ; license: ;

29. *NPS Chat* [download | source]
    id: nps_chat; size: 301366; author: Craig Martell (cmartell@nps.edu); copyright: ; license: This corpus is distributed solely for non-commercial, non-profit educational and research use. It is a derivative compilation work of multiple works whose copyrights are held by the respective original authors.;

30. *City Database* [download | source]
    id: city_database; size: 1708; author: ; copyright: ; license: ;

31. *Sample European Parliament Proceedings Parallel Corpus* [download | source]
    id: europarl_raw; size: 12594977; author: Philipp Koehn, University of Edinburgh; copyright: ; license: ;

32. *BioCreAtIvE (Critical Assessment of Information Extraction Systems in Biology)* [download | source]
    id: biocreative_ppi; size: 223566; author: ; copyright: Public Domain (not copyrighted); license: Public Domain;

33. *VerbNet Lexicon, Version 3.3* [download | source]
    id: verbnet3; size: 482025; author: Karin Kipper-Schuler; copyright: ; license: Distributed with permission of the author.;

34. *Cross-Framework and Cross-Domain Parser Evaluation Shared Task* [download | source]
    id: pe08; size: 80735; author: ; copyright: ; license: Distributed with permission;

35. *The Patient Information Leaflet (PIL) Corpus* [download | source]
    id: pil; size: 1510205; author: ; copyright: ; license: Distributed with permission;

36. *Crubadan Corpus* [download | source]
    id: crubadan; size: 5288655; author: Kevin Scannell; copyright: Copyright (C) 2010 Kevin Scannell; license: GPLv3;

37. *Project Gutenberg Selections* [download | source]
    id: gutenberg; size: 4251829; author: ; copyright: public domain; license: public domain;

38. *Proposition Bank Corpus 1.0* [download | source]
    id: propbank; size: 5323498; author: ; copyright: ; license: Distributed with permission;

39. *Machado de Assis -- Obra Completa* [download | source]
    id: machado; size: 6151774; author: Machado de Assis; copyright: ; license: Public Domain;

40. *C-Span State of the Union Address Corpus* [download | source]
    id: state_union; size: 808757; author: ; copyright: public domain; license: public domain;

41. *Twitter Samples* [download | source]
    id: twitter_samples; size: 16007673; author: ; copyright: Copyright (C) 2015 Twitter, Inc; license: Must be used subject to Twitter Developer Agreement (https://dev.twitter.com/overview/terms/agreement);

42. *SemCor 3.0 [download | source]*
    id: semcor; size: 4397021; author: Rada Mihalcea (rada@cs.unt.edu); copyright: ; license: You are granted permission to use, copy, modify and distribute this database for any purpose and without fee and royalty is hereby granted, provided that you agree to comply with the Princeton copyright notice and statements, including the disclaimer, and that the same appear on ALL copies of the database, including modifications that you make for internal use or for distribution. See semcor/README for more information.;

nltk data: https://www.nltk.org/nltk_data/

43. *Wordnet 3.1 [download | source]*
    id: wordnet31; size: 11058667; author: ; copyright: WordNet 3.1 Copyright 2011 by Princeton University. All rights reserved.; license: Permission to use, copy, modify and distribute this software and database and its documentation for any purpose and without fee or royalty is hereby granted, provided that you agree to comply with the following copyright notice and statements, including the disclaimer, and that the same appear on ALL copies of the software, database and documentation, including modifications that you make for internal use or for distribution.... [see webpage for full license];

44. *Extended Open Multilingual WordNet [download | source]*
    id: extended_omw; size: 11251284; author: ; copyright: Copyright (C) 2013 Francis Bond and Ryan Foster; license: CC by SA 3.0 Licence (for data from Wikitionary) and Unicode, Inc. Licence Agreement (for data from CLDR);

45. *Names Corpus, Version 1.3 (1994-03-29) [download | source]*
    id: names; size: 21326; author: Mark Kantrowitz and Bill Ross; copyright: Copyright (C) 1991 Mark Kantrowitz; license: You may use the lists of names for any purpose, so long as credit is given in any published work. You may also redistribute the list if you provide the recipients with a copy of this README file. The lists are not in the public domain (I retain the copyright on the lists) but are freely redistributable. If you have any additions to the lists of names, I would appreciate receiving them.;

46. *Penn Treebank [download | source]*
    id: ptb; size: 6289; author: ; copyright: Copyright (C) 1995 University of Pennsylvania; license: This is a stub for the full Penn Treebank Corpus version 3.;

47. *NomBank Corpus 1.0 [download | source]*
    id: nombank.1.0; size: 6728397; author: ; copyright: ; license: Distributed with permission;

48. *Portuguese Treebank [download | source]*
    id: floresta; size: 1882021; author: ; copyright: ; license: Non-commercial use only;

49. *ComTrans Corpus Sample [download | source]*
    id: comtrans; size: 11904518; author: Reinhard Rapp; copyright: ; license: ;

50. *KNB Corpus (Annotated blog corpus) [download | source]*
    id: knbc; size: 8760788; author: ; copyright: ; license: Freely re-distributable under the same license as the original KNB Corpus.;

51. *MAC-MORPHO: Brazilian Portuguese news text with part-of-speech tags [download | source]*
    id: mac_morpho; size: 3013904; author: ; copyright: ; license: Distributed with permission of Núcleo Interinstitucional de Lingüística Computacional (NILC), Universidade de São Paulo (USP) in São Carlos, Universidade Federal de São Carlos (UFSCar), Universidade Estadual Paulista (UNESP) of Araraquara.;

52. *Swadesh Wordlists [download | source]*
    id: swadesh; size: 22828; author: ; copyright: ; license: GNU Free Documentation License;

53. *PASCAL RTE Challenges 1, 2, and 3 [download | source]*
    id: rte; size: 386303; author: ; copyright: ; license: ;

54. *Toolbox Sample Files [download | source]*
    id: toolbox; size: 250616; author: ; copyright: ; license: ;

55. *JEITA Public Morphologically Tagged Corpus (in ChaSen format)* [download | source]
    id: jeita; size: 16531215; author: ; copyright: ; license: Freely re-distributable under the same license as the original JEITA corpus. Each document retains its own license from Aozora bunko and Project Sugita Genpaku.;

56. *Product Reviews (5 Products)* [download | source]
    id: product_reviews_1; size: 141287; author: Bing Liu; copyright: Copyright (C) 2004 Bing Liu; license: Creative Commons Attribution 4.0 International;

57. *Open Multilingual Wordnet* [download | source]
    id: omw; size: 12110409; author: Francis Bond; copyright: Please consult the copyright statements of the individual Wordnets; license: Please consult the LICENSE files included with the individual Wordnets. Note that all permit redistribution.;

58. *SentiWordNet* [download | source]
    id: sentiwordnet; size: 4686546; author: Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani; copyright: Copyright (C) 2013 SentiWordNet Project; license: Creative Commons Attribution ShareAlike 3.0 Unported license;

59. *Product Reviews (9 Products)* [download | source]
    id: product_reviews_2; size: 170698; author: Bing Liu; copyright: Copyright (C) 2007 Bing Liu; license: Creative Commons Attribution 4.0 International;

60. *Australian Broadcasting Commission 2006* [download | source]
    id: abc; size: 1487851; author: Australian Broadcasting Commission; copyright: ; license: ;

61. *Open English Wordnet 2021* [download | source]
    id: wordnet2021; size: 11332750; author: ; copyright: Open English Wordnet 2021 Copyright 2021 by the Open English Wordnet team. WordNet 3.1 Copyright 2011 by Princeton University. All rights reserved.; license: This resource is derived from Princeton WordNet under the WordNet License and further developed under the Creative Commons Attribution 4.0 International License. You may share and adapt this resource providing attribution is given to both Princeton WordNet and the Open English WordNet team.;

62. *Universal Declaration of Human Rights Corpus (Unicode Version)* [download | source]
    id: udhr2; size: 1653975; author: ; copyright: public domain; license: public domain;

63. *SENSEVAL 2 Corpus: Sense Tagged Text* [download | source]
    id: senseval; size: 2151350; author: ; copyright: ; license: Distributed with permission.;

64. *Word Lists* [download | source]
    id: words; size: 757777; author: ; copyright: public domain; license: public domain;

65. *FrameNet 1.5* [download | source]
    id: framenet_v15; size: 69337891; author: Collin F. Baker; copyright: ; license: May be used for non-commercial purposes.;

66. *Unicode Samples* [download | source]
    id: unicode_samples; size: 1212; author: ; copyright: ; license: ;

67. *PC-KIMMO Data Files* [download | source]
    id: kimmo; size: 186958; author: ; copyright: ; license: ;

68. *FrameNet 1.7* [download | source]
    id: framenet_v17; size: 99207152; author: Collin F. Baker; copyright: ; license: Creative Commons Attribution 3.0 Unported License;

69. *Chat-80 Data Files* [download | source]
    id: chat80; size: 19209; author: David Warren and Fernando Pereira; copyright: Copyright (C) 1982 David Warren and Fernando Pereira; license: This program may be used, copied, altered or included in other programs only for academic purposes and provided that the authorship of the initial program is aknowledged. Use for commercial purposes without the previous written agreement of the authors is forbidden.;

70. *Experimental Data for Question Classification* [download | source]
    id: qc; size: 125456; author: Xin Li and Dan Roth, UIUC; copyright: ; license: ;

71. *C-Span Inaugural Address Corpus* [download | source]
    id: inaugural; size: 346476; author: ; copyright: public domain; license: public domain;

72. *WordNet* [download | source]
    id: wordnet; size: 10775600; author: ; copyright: WordNet 3.0 Copyright 2006 by Princeton University. All rights reserved.; license: Permission to use, copy, modify and distribute this software and database and its documentation for any purpose and without fee or royalty is hereby granted, provided that you agree to comply with the following copyright notice and statements, including the disclaimer, and that the same appear on ALL copies of the software, database and documentation, including modifications that you make for internal use or for distribution.... [see webpage for full license];

73. *Stopwords Corpus* [download | source]
    id: stopwords; size: 34276; author: ; copyright: ; license: ;

74. *VerbNet Lexicon, Version 2.1* [download | source]
    id: verbnet; size: 323661; author: Karin Kipper-Schuler; copyright: ; license: Distributed with permission of the author.;

75. *Shakespeare XML Corpus Sample* [download | source]
    id: shakespeare; size: 475458; author: ; copyright: public domain; license: public domain;

76. *York-Toronto-Helsinki Parsed Corpus of Old English Prose* [download | source]
    id: ycoe; size: 477; author: ; copyright: ; license: ;

77. *NIST IE-ER DATA SAMPLE* [download | source]
    id: ieer; size: 166156; author: ; copyright: ; license: ;

78. *CESS-CAT Treebank* [download | source]
    id: cess_cat; size: 5396688; author: ; copyright: ; license: If you use these corpora for research, please cite thusly: CESS-Cat project (M. Antonia Martí, MarionaTaulé, Lluís Márquez, Manuel Bertran (2007) ?CESS-ECE: A Multilingual and Multilevel Annotated Corpus? in http://www.lsi.upc.edu/~mbertran/cess-ece/publications).;

nltk data: https://www.nltk.org/nltk_data/

79. *Switchboard Corpus Sample* [download | source]
id: switchboard; size: 791161; author: ; copyright: ; license: Permission is granted for use of this material in accordance with the Open Content License [http://opencontent.org/opl.shtml]. This corpus contains transcripts and annotations for 36 calls from the Switchboard Corpus [http://www.ldc.upenn.edu/Catalog/LDC93S7.html].;

80. *Comparative Sentence Dataset* [download | source]
id: comparative_sentences; size: 279121; author: Nitin Jindal and Bing Liu; copyright: Copyright (C) 2006 Nitin Jindal and Bing Liu; license: Creative Commons Attribution 4.0 International;

81. *Subjectivity Dataset v1.0* [download | source]
id: subjectivity; size: 521628; author: Bo Pang and Lillian Lee; copyright: Copyright (C) 2004 Bo Pang and Lillian Lee; license: Creative Commons Attribution 4.0 International;

82. *Universal Declaration of Human Rights Corpus* [download | source]
id: udhr; size: 1170177; author: ; copyright: public domain; license: public domain;

83. *Polish language of the XX century sixties* [download | source]
id: pl196x; size: 7051453; author: I. Kurcz, A. Lewicki, J. Sambor, K. Szafran, J. Woronczak; copyright: ; license: GNU General Public License;

84. *Paradigm Corpus* [download | source]
id: paradigms; size: 24902; author: Cathy Bow, University of Melbourne; copyright: ; license: Distributed with the permission of the author;

85. *Gazeteer Lists* [download | source]
id: gazetteers; size: 8265; author: ; copyright: ; license: GNU Free Documentation License; or public domain (depending on the file);

86. *TIMIT Corpus Sample* [download | source]
id: timit; size: 22251869; author: ; copyright: ; license: This corpus sample is Copyright 1993 Linguistic Data Consortium, and is distributed under the terms of the Creative Commons Attribution, Non-Commercial, ShareAlike license. http://creativecommons.org/;

87. *Penn Treebank Sample* [download | source]
id: treebank; size: 1740034; author: ; copyright: Copyright (C) 1995 University of Pennsylvania; license: This is a 10% fragment of Penn Treebank, (C) LDC 1995. It is made available under fair use for the purposes of illustrating NLTK tools for tokenizing, tagging, chunking and parsing. This data is for non-commercial use only.;

88. *Sinica Treebank Corpus Sample* [download | source]
id: sinica_treebank; size: 906706; author: ; copyright: ; license: Distributed with the Natural Language Toolkit under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike License [http://creativecommons.org/licenses/by-nc-sa/2.5/].;

89. *Opinion Lexicon* [download | source]
id: opinion_lexicon; size: 24947; author: Bing Liu; copyright: Copyright (C) 2011 Bing Liu; license: Creative Commons Attribution 4.0 International;

90. *Prepositional Phrase Attachment Corpus* [download | source]
id: ppattach; size: 781714; author: Adwait Ratnaparkhi; copyright: (C) 1994 Adwait Ratnaparkhi; license: Distributed with the permission of the author.;

91. *Dependency Parsed Treebank* [download | source]
id: dependency_treebank; size: 457429; author: ; copyright: Copyright (C) 1995 University of Pennsylvania; license: This is a 10% fragment of Penn Treebank, (C) LDC 1995, which has been dependency parsed. It is made available under fair use for the purposes of illustrating NLTK tools for tokenizing, tagging, chunking and parsing. This data is for non-commercial use only.;

92. *The Reuters-21578 benchmark corpus, ApteMod version* [download | source]
id: reuters; size: 6378691; author: ; copyright: ; license: The copyright for the text of newswire articles and Reuters annotations in the Reuters-21578 collection resides with Reuters Ltd. Reuters Ltd. and Carnegie Group, Inc. have agreed to allow the free distribution of this data *for research purposes only*. If you publish results based on this data set, please acknowledge its use, refer to the data set by the name 'Reuters-21578, Distribution 1.0', and inform your readers of the current location of the data set.;

93. *Genesis Corpus* [download | source]
id: genesis; size: 473239; author: ; copyright: public domain; license: public domain;

94. *CESS-ESP Treebank* [download | source]
id: cess_esp; size: 2220392; author: ; copyright: ; license: If you use these corpora for research, please cite thusly: CESS-Cat project (M. Antonia Martí, MarionaTaulé, Lluís Márquez, Manuel Bertran (2007) ?CESS-ECE: A Multilingual and Multilevel Annotated Corpus? in http://www.lsi.upc.edu/~mbertran/cess-ece/publications).;

95. *Dependency Treebanks from CoNLL 2007 (Catalan and Basque Subset)* [download | source]
id: conll2007; size: 1242958; author: ; copyright: Copyright (C) 2007 The University of the Basque Country; license: Creative Commons Attribution-NonCommercial-NoDerivativeWorks license;

96. *Non-Breaking Prefixes (Moses Decoder)* [download | source]
id: nonbreaking_prefixes; size: 25437; author: ; copyright: ; license: Gnu LGPL;

97. *Dolch Word List* [download | source]
id: dolch; size: 2116; author: ; copyright: ; license: ;

98. *SMULTRON Corpus Sample* [download | source]
id: smultron; size: 166207; author: Sofia Gustafson-Capkova, Yvonne Samuelsson, and Martin Volk; copyright: ; license: ;

99. *Alpino Dutch Treebank* [download | source]
id: alpino; size: 2797255; author: ; copyright: ; license: Distributed with permission of Gertjan van Noord;

100. *WordNet-InfoContent* [download | source]
id: wordnet_ic; size: 12056682; author: ; copyright: ; license: ;

101. *Brown Corpus* [download | source]
id: brown; size: 3314357; author: W. N. Francis and H. Kucera; copyright: ; license: May be used for non-commercial purposes.;

102. *PanLex Swadesh Corpora* [download | source]
id: panlex_swadesh; size: 2861668; author: Jonathan Pool (editor); copyright: ; license: CC0 1.0 Universal;

103. *CONLL 2000 Chunking Corpus* [download | source]
id: conll2000; size: 756607; author: ; copyright: ; license: ;

104. *Universal Treebanks Version 2.0* [download | source]
id: universal_treebanks_v20; size: 25908853; author: ; copyright: ; license: Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States;

105. *Brown Corpus (TEI XML Version)* [download | source]
id: brown_tei; size: 8737738; author: W. N. Francis and H. Kucera; copyright: ; license: May be used for non-commercial purposes.;

106. *The Carnegie Mellon Pronouncing Dictionary (0.6)* [download | source]
id: cmudict; size: 896069; author: ; copyright: Copyright 1998 Carnegie Mellon University; license: Use of this dictionary, for any research or commercial purpose, is completely unrestricted. If you use or redistribute this material, we would appreciate acknowlegement of its origin.;

107. *Open Multilingual Wordnet* [download | source]
id: omw-1.4; size: 26634772; author: Francis Bond; copyright: Please consult the copyright statements of the individual Wordnets; license: Please consult the LICENSE files included with the individual Wordnets. Note that all permit redistribution.;

108. *MULTEXT-East 1984 annotated corpus 4.0* [download | source]
id: mte_teip5; size: 14800561; author: Erjavec, Tomaž; Barbu, Ana-Maria; Derzhanski, Ivan; Dimitrova, Ludmila; Garabík, Radovan; Ide, Nancy; Kaalep, Heiki-Jaan; Kotsyba, Natalia; Krstev, Cvetana; Oravecz, Csaba; Petkevič, Vladimír; Priest-Dorman, Greg; QasemiZadeh, Behrang; Radziszewski, Adam; Simov, Kiril; Tufiş, Dan and Zdravkova, Katerina; copyright: ; license: Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0);

109. *Indian Language POS-Tagged Corpus* [download | source]
id: indian; size: 199187; author: A Kumaran; copyright: ; license: Distributed with permission;

110. *CONLL 2002 Named Entity Recognition Corpus* [download | source]
id: conll2002; size: 1867449; author: ; copyright: ; license: ;

111. *Help on Tagsets* [download | source]
id: tagsets; size: 34531; author: UCREL, Lancaster University; copyright: ; license: ;

Exercise 2

- The full text of "*Alice's Adventures in Wonderland*" is available in nltk data.
- Example:

```
>>> import nltk
>>> alice = nltk.corpus.gutenberg.words('carroll-alice.txt')
>>> len(alice)
34110
>>> alice[:100]
['[', 'Alice', "'", 's', 'Adventures', 'in', 'Wonderland', 'by', 'Lewis', 'Carroll',
'1865', ']', 'CHAPTER', 'I', '.', 'Down', 'the', 'Rabbit', '-', 'Hole', 'Alice',
'was', 'beginning', 'to', 'get', 'very', 'tired', 'of', 'sitting', 'by', 'her',
'sister', 'on', 'the', 'bank', ',', 'and', 'of', 'having', 'nothing', 'to', 'do',
':', 'once', 'or', 'twice', 'she', 'had', 'peeped', 'into', 'the', 'book', 'her',
'sister', 'was', 'reading', ',', 'but', 'it', 'had', 'no', 'pictures', 'or',
'conversations', 'in', 'it', ',', "'", 'and', 'what', 'is', 'the', 'use', 'of', 'a',
'book', ",'", 'thought', 'Alice', "'", 'without', 'pictures', 'or', 'conversation',
"?'", 'So', 'she', 'was', 'considering', 'in', 'her', 'own', 'mind', '(', 'as',
'well', 'as', 'she', 'could', ',']
>>>
```

.words() already tokenized the text

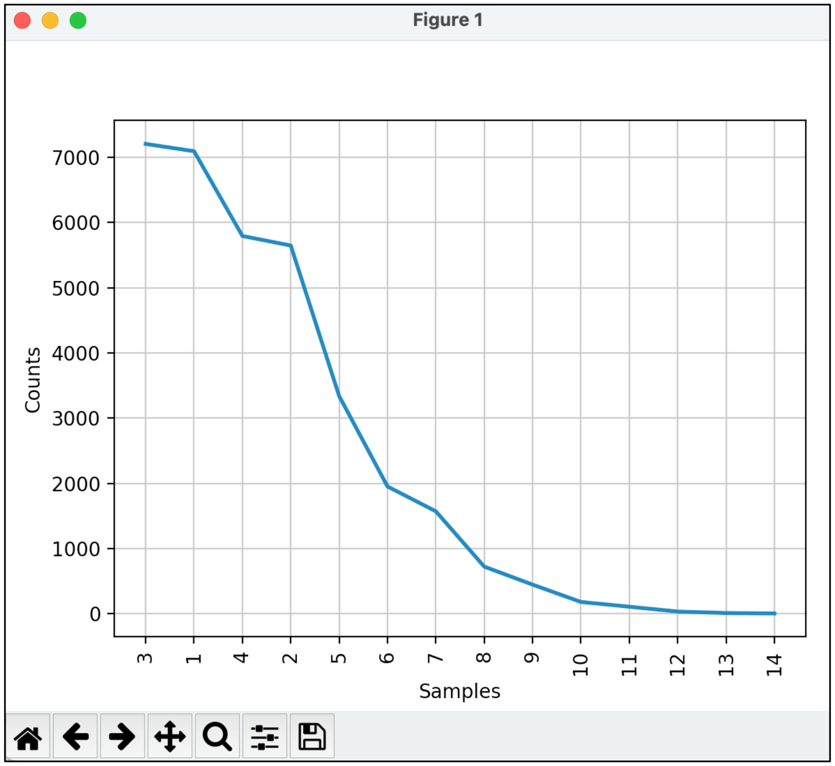
Exercise 2

- Calculate the average number of characters per word (cf. exercise 1 answer: 4.31)

1. for-loop method
2. list comprehension method
   - `sum()` or
   - `from statistics import mean`
   - `mean()`

Exercise 2

- word length distribution
- `>>> fd = nltk.FreqDist(wlen)`
- `>>> fd`
- `FreqDist({3: 7205, 1: 7093, 4: 5793, 2: 5647, 5: 3340, 6: 1952, 7: 1571, 8: 723, 9: 447, 10: 181, ...})`
- `>>> fd.plot()`
- `<AxesSubplot:xlabel='Samples', ylabel='Counts'>`



Python

- https://docs.python.org/3/tutorial/introduction.html
- Numbers
- Strings
- Lists
- Dictionaries

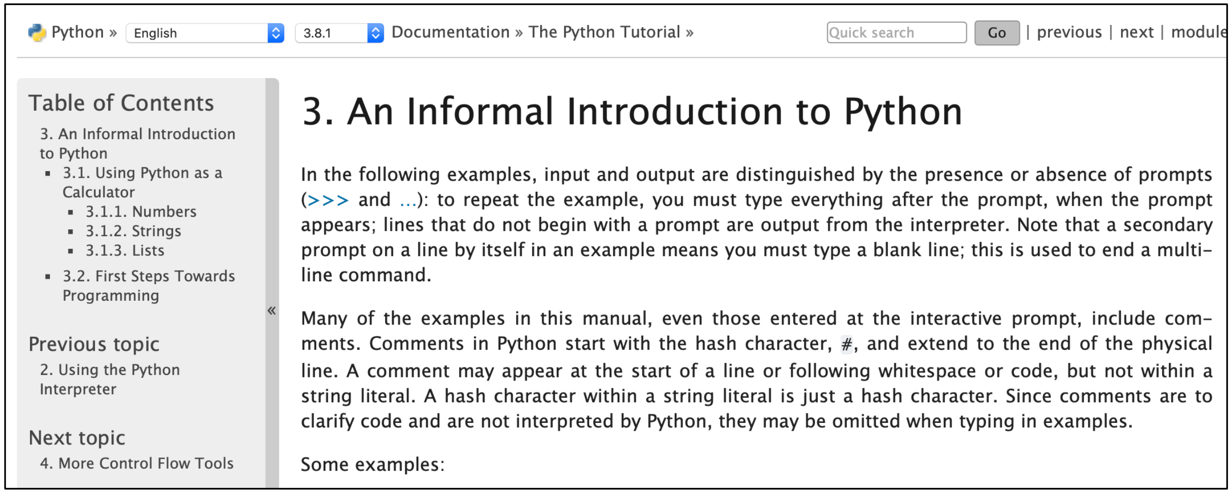

Python »
English
3.8.1
Documentation » The Python Tutorial »
Quick search
Go
| previous | next | module
«

Table of Contents

Previous topic

2. Using the Python Interpreter

Next topic

4. More Control Flow Tools

3. An Informal Introduction to Python

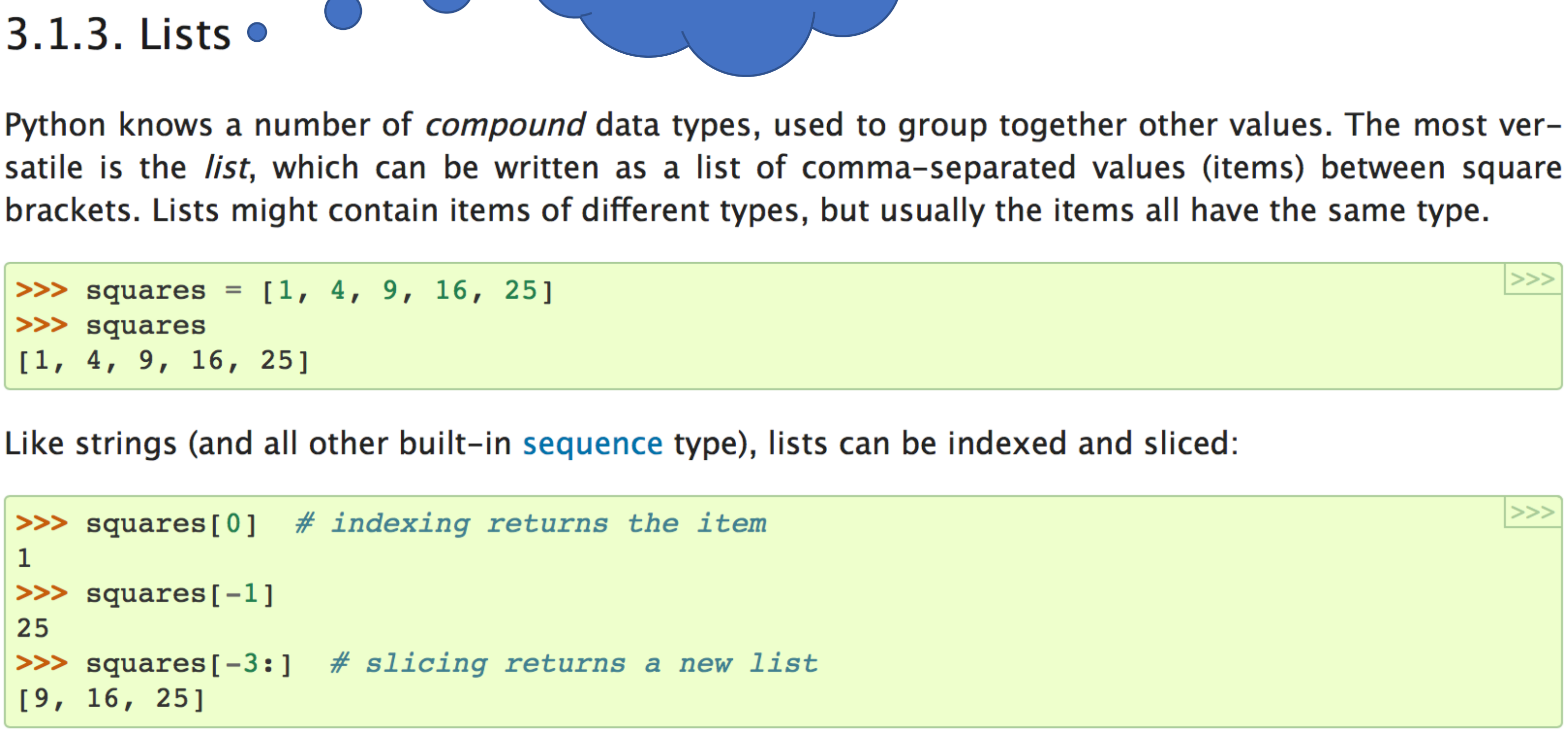
In the following examples, input and output are distinguished by the presence or absence of prompts (`>>>` and `...`): to repeat the example, you must type everything after the prompt, when the prompt appears; lines that do not begin with a prompt are output from the interpreter. Note that a secondary prompt on a line by itself in an example means you must type a blank line; this is used to end a multi-line command.

Many of the examples in this manual, even those entered at the interactive prompt, include comments. Comments in Python start with the hash character, `#`, and extend to the end of the physical line. A comment may appear at the start of a line or following whitespace or code, but not within a string literal. A hash character within a string literal is just a hash character. Since comments are to clarify code and are not interpreted by Python, they may be omitted when typing in examples.

Some examples:

Python Lists

Recall strings from last time? Think of those as lists too



3.1.3. Lists

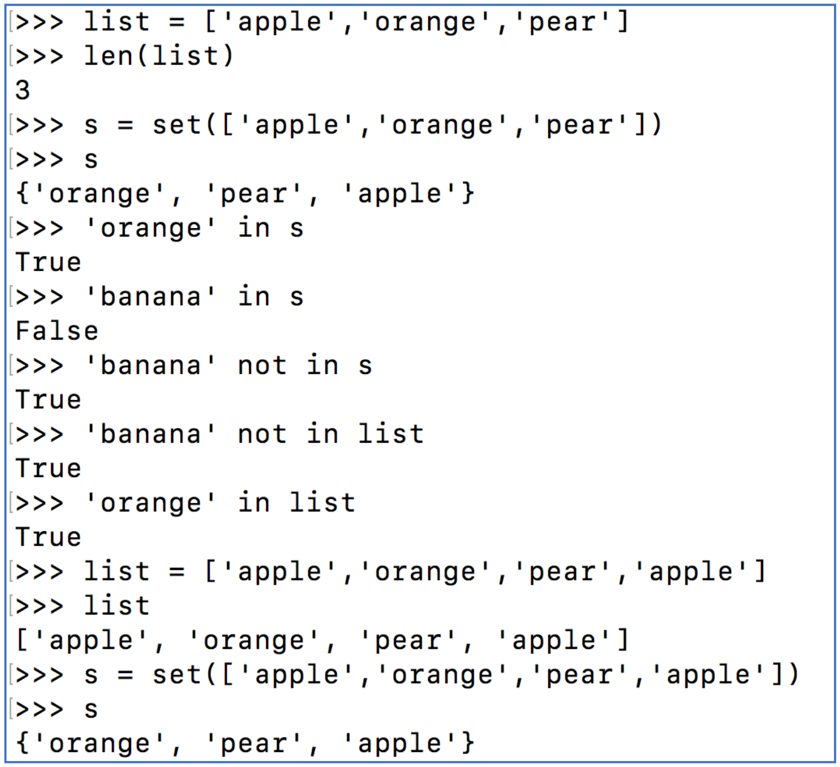
Python knows a number of *compound* data types, used to group together other values. The most versatile is the *list*, which can be written as a list of comma-separated values (items) between square brackets. Lists might contain items of different types, but usually the items all have the same type.

```
>>> squares = [1, 4, 9, 16, 25]
>>> squares
[1, 4, 9, 16, 25]
```

Like strings (and all other built-in sequence type), lists can be indexed and sliced:

```
>>> squares[0]  # indexing returns the item
1
>>> squares[-1]
25
>>> squares[-3:]  # slicing returns a new list
[9, 16, 25]
```

Python Lists vs. Sets



```
[>>> list = ['apple','orange','pear']
[>>> len(list)
3
[>>> s = set(['apple','orange','pear'])
[>>> s
{'orange', 'pear', 'apple'}
[>>> 'orange' in s
True
[>>> 'banana' in s
False
[>>> 'banana' not in s
True
[>>> 'banana' not in list
True
[>>> 'orange' in list
True
[>>> list = ['apple','orange','pear','apple']
[>>> list
['apple', 'orange', 'pear', 'apple']
[>>> s = set(['apple','orange','pear','apple'])
[>>> s
{'orange', 'pear', 'apple'}
```

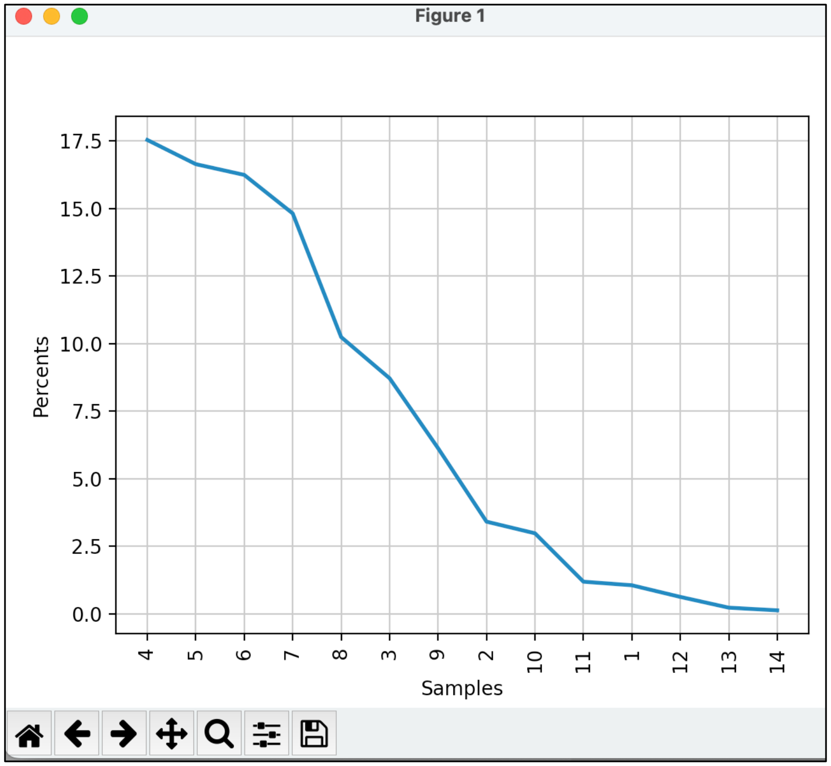
- `in` – membership
- `not in` – not a member of
- `set(`*`List`*`)` - produces a set, **no duplicates permitted**
- set is faster than list for lookup (hashtable)

Exercise 2 extended

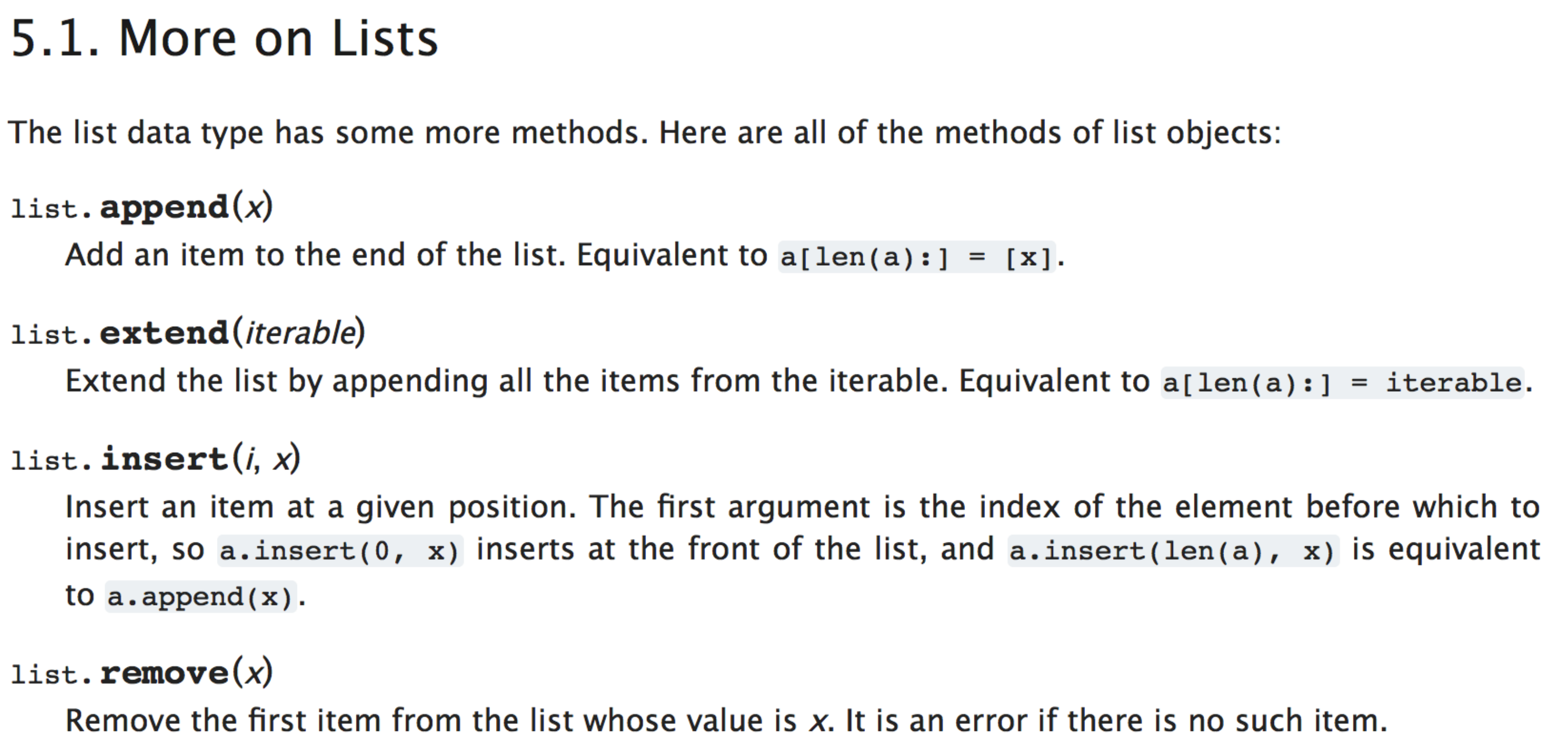
- word length distribution for the vocab of *Alice* (using set)

```
>>> vocab = set(alice)
>>> vwlen = [len(word) for word
in vocab]
>>> vfd = nltk.FreqDist(vwlen)
>>> vfd
FreqDist({4: 529, 5: 502, 6:
490, 7: 447, 8: 309, 3: 263, 9:
185, 2: 103, 10: 90, 11: 36,
...})
>>> vfd.plot(percents=True)
<AxesSubplot:xlabel='Samples',
ylabel='Percents'>
```

https://www.nltk.org/api/nltk.probability.html



Python Lists



5.1. More on Lists

The list data type has some more methods. Here are all of the methods of list objects:

`list.`**`append`**(*x*)

Add an item to the end of the list. Equivalent to `a[len(a):] = [x]`.

`list.`**`extend`**(*iterable*)

Extend the list by appending all the items from the iterable. Equivalent to `a[len(a):] = iterable`.

`list.`**`insert`**(*i*, *x*)

Insert an item at a given position. The first argument is the index of the element before which to insert, so `a.insert(0, x)` inserts at the front of the list, and `a.insert(len(a), x)` is equivalent to `a.append(x)`.

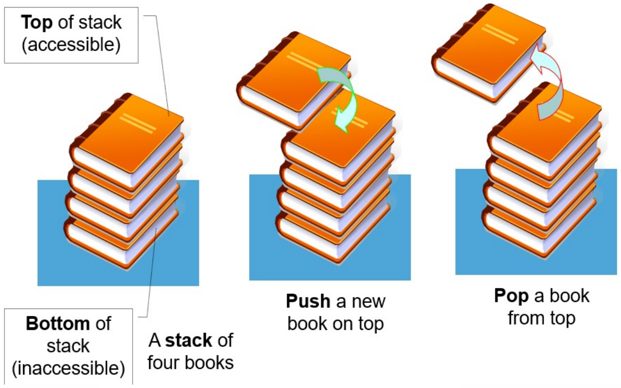
`list.`**`remove`**(*x*)

Remove the first item from the list whose value is *x*. It is an error if there is no such item.

Python Lists

- Lists as stacks
- Lists as queues
- List Comprehensions (*see Exercise 2*)

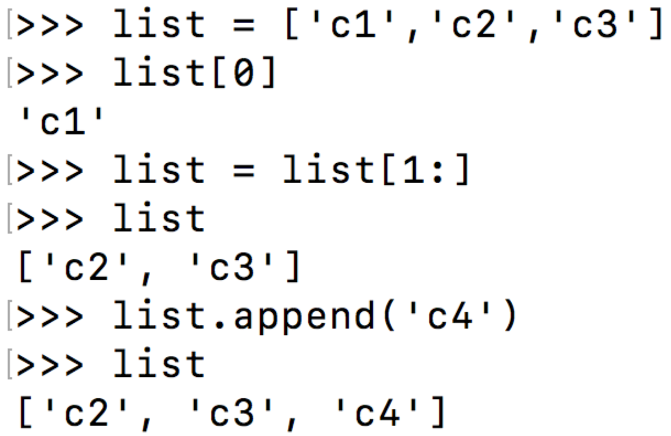
https://visualgo.net/en/list?slide=4



https://www.appcoda.com/ios-concurrency/

Python List as a Queue

EXAMPLE:

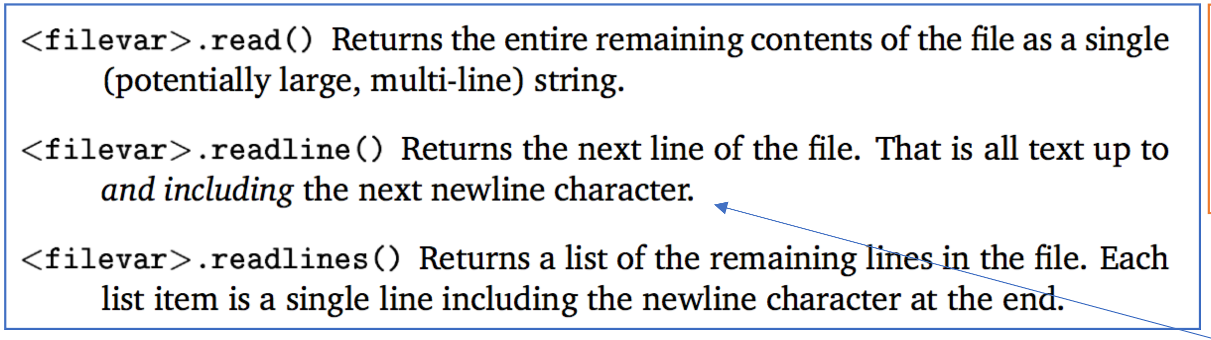


```
[>>> list = ['c1','c2','c3']
[>>> list[0]
'c1'
[>>> list = list[1:]
[>>> list
['c2', 'c3']
[>>> list.append('c4')
[>>> list
['c2', 'c3', 'c4']
```

- Method append() to add to the queue
- list[0] gives us the head of the queue
- list = list[1:] deletes the head of the queue from the queue

Python: Files

- Like all other programming languages, uses a file handle, called **file variable**: open()
- `infile = open("file.txt","r")`   `outfile = open("results.txt","w")`



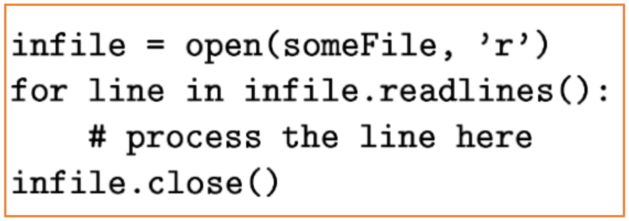
<filevar>.read() Returns the entire remaining contents of the file as a single (potentially large, multi-line) string.

<filevar>.readline() Returns the next line of the file. That is all text up to *and including* the next newline character.

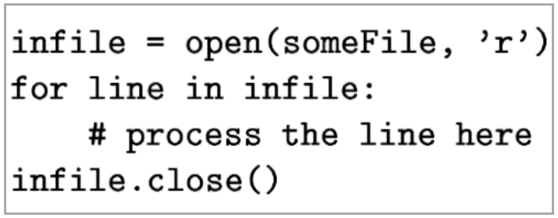
<filevar>.readlines() Returns a list of the remaining lines in the file. Each list item is a single line including the newline character at the end.

f = open(fname)
f = open(fname, encoding="utf-8")
f = open(fname, encoding="latin-1")
f = open(fname, encoding="ascii")

Removing the newline:
`.strip()`
`.rstrip()`
`.lstrip()`

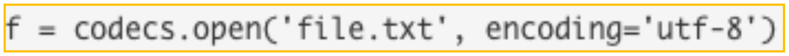


```
infile = open(someFile, ’r’)
for line in infile.readlines():
    # process the line here
infile.close()
```



```
infile = open(someFile, ’r’)
for line in infile:
    # process the line here
infile.close()
```

Python 2.7



```
f = codecs.open('file.txt', encoding='utf-8')
```