Chapter 1

Appendix: Python and NLTK Cheat Sheet (Draft)

1.1 Python

1.1.1 Strings

```
>>> x = 'Python'; y = 'NLTK'; z = 'Natural Language Processing'
>>> x + '/' + y
'Python/NLTK'
>>> 'LT' in y
True
>>> x[2:]
'thon'
>>> x[::-1]
'nohtyP'
>>> len(x)
6
>>> z.count('a')
4
>>> z.endswith('ing')
True
>>> z.index('Language')
8
>>> '; '.join([x,y,z])
'Python; NLTK; Natural Language Processing'
>>> y.lower()
'nltk'
>>> z.replace(' ', '\n')
'Natural\nLanguage\nProcessing'
>>> print z.replace(' ', '\n')
Natural
Language
Processing
>>> z.split()
['Natural', 'Language', 'Processing']
```

For more information, type *help(str)* at the Python prompt.

1.1.2 Lists

```
>>> x = ['Natural', 'Language']; y = ['Processing']
>>> x[0]
'Natural'
>>> list(x[0])
['N', 'a', 't', 'u', 'r', 'a', 'l']
>>> x + y
['Natural', 'Language', 'Processing']
>>> 'Language' in x
True
>>> len(x)
2
>>> x.index('Language')
1
```

The following functions modify the list in-place:

```
>>> x.append('Toolkit')
>>> x
['Natural', 'Language', 'Toolkit']
>>> x.insert(0, 'Python')
>>> x
['Python', 'Natural', 'Language', 'Toolkit']
>>> x.reverse()
>>> x
['Toolkit', 'Language', 'Natural', 'Python']
>>> x.sort()
>>> x
['Language', 'Natural', 'Python', 'Toolkit']
```

For more information, type *help(list)* at the Python prompt.

1.1.3 Dictionaries

```
>>> d = {'natural': 'adj', 'language': 'noun'}
>>> d['natural']
'adj'
>>> d['toolkit'] = 'noun'
>>> d
{'natural': 'adj', 'toolkit': 'noun', 'language': 'noun'}
>>> 'language' in d
True
>>> d.items()
[('natural', 'adj'), ('toolkit', 'noun'), ('language', 'noun')]
>>> d.keys()
['natural', 'toolkit', 'language']
>>> d.values()
['adj', 'noun', 'noun']
```

For more information, type *help(dict)* at the Python prompt.

1.1.4 Regular Expressions

> **Note**
>
> to be written

1.2 NLTK

Many more examples can be found in the NLTK Guides, available at `http://nltk.org/doc/guides`.

1.2.1 Corpora

```
>>> import nltk
>>> dir(nltk.corpus)
```

1.2.2 Tokenization

```
>>> text = '''NLTK, the Natural Language Toolkit, is a suite of program
... modules, data sets and tutorials supporting research and teaching in
... computational linguistics and natural language processing.'''
>>> import nltk
>>> nltk.LineTokenizer().tokenize(text)
['NLTK, the Natural Language Toolkit, is a suite of program', 'modules,
data sets and tutorials supporting research and teaching in', 'computational
linguistics and natural language processing.']
>>> nltk.WhitespaceTokenizer().tokenize(text)
['NLTK,', 'the', 'Natural', 'Language', 'Toolkit,', 'is', 'a', 'suite',
 'of', 'program', 'modules,', 'data', 'sets', 'and', 'tutorials',
 'supporting', 'research', 'and', 'teaching', 'in', 'computational',
 'linguistics', 'and', 'natural', 'language', 'processing.']
>>> nltk.WordPunctTokenizer().tokenize(text)
['NLTK', ',', 'the', 'Natural', 'Language', 'Toolkit', ',', 'is', 'a',
 'suite', 'of', 'program', 'modules', ',', 'data', 'sets', 'and',
 'tutorials', 'supporting', 'research', 'and', 'teaching', 'in',
 'computational', 'linguistics', 'and', 'natural', 'language',
 'processing', '.']
>>> nltk.RegexpTokenizer(', ', gaps=True).tokenize(text)
['NLTK', 'the Natural Language Toolkit', 'is a suite of program\nmodules',
 'data sets and tutorials supporting research and teaching in\ncomputational
 linguistics and natural language processing.']
```

1.2.3 Stemming

```
>>> tokens = nltk.WordPunctTokenizer().tokenize(text)
>>> stemmer = nltk.RegexpStemmer('ing$|s$|e$')
>>> for token in tokens:
...     print stemmer.stem(token),
NLTK , th Natural Languag Toolkit , i a suit of program module ,
data set and tutorial support research and teach in computational
linguistic and natural languag process .
>>> stemmer = nltk.PorterStemmer()
```

```
>>> for token in tokens:
...     print stemmer.stem(token),
NLTK , the Natur Languag Toolkit , is a suit of program modul ,
data set and tutori support research and teach in comput linguist
and natur languag process .
```

1.2.4 Tagging

Note

to be written

About this document...

This chapter is a draft from *Introduction to Natural Language Processing* [http://nltk.org/book.html], by Steven Bird, Ewan Klein and Edward Loper,
This document is Revision: 6409 Mon Aug 11 04:34:08 EDT 2008