Today's Topics

• Penn Treebank 3 (on USB stick)
• Homework 1 Review
• Homework 2
Homework 1: Part 2

• The Penn Treebank is partially installed as a corpus in NLTK Data (Sections 00 and 01: wsj_0001.mrg to wsj_0199.mrg)

• from nltk.corpus import treebank

```python
>>> treebank.words()
['Pierre', 'Vinker', ',', '.', '61', 'years', 'old', ',', '...']
>>> len(treebank.words())
108676
```
Homework 1: Part 2

Part 2 (contd.)

• Submit relevant parts of your python console + a screenshot of a random tree from the treebank

• Parsed Sentence:
  • treebank.parsed_sents("wsj_0199.mrg")[0]

• Random number generation:
  • import random
  • random.seed()
  • random.randrange(1,200)
Homework 1: Part 2

- import random
- from nltk.corpus import treebank
- random.seed()
- treebank.parsed_sents("wsj_0"+str(random.randrange(1,200)).zfill(3)+".mrg")[0].draw()

**note**: zfill used to avoid this:

```python
>>> "wsj_0"+str(random.randrange(1,200))+.mrg"
'wsj_054.mrg'
>>> "wsj_0"+str(random.randrange(1,200))+".mrg"
'wsj_0139.mrg'
```
Homework 1: Part 3

Alice in Wonderland by Lewis Carroll

- nltk.corpus.gutenberg.fileids()

- raw = nltk.corpus.gutenberg.raw('carroll-alice.txt')
- sents = nltk.sent_tokenize(raw)
- len(sents)
  1625
Homework 1: Part 3

• Part of speech tag a random sentence from the book and comment on the accuracy of the tagging.

• Submit your console

• use
  • `nltk.tokenize.word_tokenize(string)`
  • `nltk.pos_tag(list_of_words)`

https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
Homework 1: Part 3

- ```from nltk.tokenize import word_tokenize```
- ```from nltk import pos_tag, sent_tokenize```
- ```from nltk.corpus import gutenberg```
- ```from random import seed, randrange```
- ```seed()```
- ```sents = sent_tokenize(gutenberg.raw('carroll-alice.txt'))```
- ```pos_tag(word_tokenize(sents[randrange(200)]))```
- ```[('How', 'WRB'), ('queer', 'NN'), ('everything', 'NN'), ('is', 'VBZ'), ('today', 'JJ'), ('!', '.')]```
Another example (more difficult)

- "Jabberwocky" is a nonsense poem written by Lewis Carroll about the killing of a creature named "the Jabberwock". It was included in his 1871 novel Through the Looking-Glass, and What Alice Found There, the sequel to Alice's Adventures in Wonderland. (Wikipedia)
CMU Pronouncing Dictionary (CMUdict)

• http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Look up the pronunciation for a word or phrase in CMUdict (version 0.7b)

PIZZA
P IY1 T S AH0 .

ABSOLUTELY
AE2 B S AH0 L UW1 T L IY0 .

ROUTE
R UW1 T .
### CMU Pronouncing Dictionary (CMUdict)

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Example</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>odd</td>
<td>AA D</td>
</tr>
<tr>
<td>AE</td>
<td>at</td>
<td>AE T</td>
</tr>
<tr>
<td>AH</td>
<td>hut</td>
<td>HH AH T</td>
</tr>
<tr>
<td>AO</td>
<td>ought</td>
<td>AO T</td>
</tr>
<tr>
<td>AW</td>
<td>cow</td>
<td>K AW</td>
</tr>
<tr>
<td>AW</td>
<td>hide</td>
<td>HH AY D</td>
</tr>
<tr>
<td>B</td>
<td>be</td>
<td>B IY</td>
</tr>
<tr>
<td>CH</td>
<td>cheese</td>
<td>CH IY Z</td>
</tr>
<tr>
<td>D</td>
<td>dee</td>
<td>D IY</td>
</tr>
<tr>
<td>DH</td>
<td>thee</td>
<td>DH IY</td>
</tr>
<tr>
<td>EH</td>
<td>Ed</td>
<td>EH D</td>
</tr>
<tr>
<td>ER</td>
<td>hurt</td>
<td>HH ER T</td>
</tr>
<tr>
<td>EY</td>
<td>ate</td>
<td>EY T</td>
</tr>
<tr>
<td>F</td>
<td>fee</td>
<td>F IY</td>
</tr>
<tr>
<td>G</td>
<td>green</td>
<td>G R IY N</td>
</tr>
<tr>
<td>HH</td>
<td>he</td>
<td>HH IY</td>
</tr>
<tr>
<td>IH</td>
<td>it</td>
<td>IH T</td>
</tr>
<tr>
<td>IY</td>
<td>eat</td>
<td>IY T</td>
</tr>
<tr>
<td>JH</td>
<td>gee</td>
<td>JH IY</td>
</tr>
<tr>
<td>K</td>
<td>key</td>
<td>K IY</td>
</tr>
<tr>
<td>L</td>
<td>lee</td>
<td>L IY</td>
</tr>
<tr>
<td>M</td>
<td>me</td>
<td>M IY</td>
</tr>
<tr>
<td>N</td>
<td>knee</td>
<td>N IY</td>
</tr>
<tr>
<td>NG</td>
<td>ping</td>
<td>P IH NG</td>
</tr>
<tr>
<td>OW</td>
<td>oat</td>
<td>OW T</td>
</tr>
<tr>
<td>OY</td>
<td>toy</td>
<td>T OY</td>
</tr>
<tr>
<td>P</td>
<td>pee</td>
<td>P IY</td>
</tr>
<tr>
<td>R</td>
<td>read</td>
<td>R IY D</td>
</tr>
<tr>
<td>S</td>
<td>sea</td>
<td>S IY</td>
</tr>
<tr>
<td>SH</td>
<td>she</td>
<td>SH IY</td>
</tr>
<tr>
<td>T</td>
<td>tea</td>
<td>T IY</td>
</tr>
<tr>
<td>TH</td>
<td>theta</td>
<td>TH EY T AH</td>
</tr>
<tr>
<td>UH</td>
<td>hood</td>
<td>HH UH D</td>
</tr>
<tr>
<td>UW</td>
<td>two</td>
<td>T UW</td>
</tr>
<tr>
<td>V</td>
<td>vee</td>
<td>V IY</td>
</tr>
<tr>
<td>W</td>
<td>we</td>
<td>W IY</td>
</tr>
<tr>
<td>Y</td>
<td>yield</td>
<td>Y IY L D</td>
</tr>
<tr>
<td>Z</td>
<td>zee</td>
<td>Z IY</td>
</tr>
<tr>
<td>ZH</td>
<td>seizure</td>
<td>S IY ZH ER</td>
</tr>
</tbody>
</table>

**Vowels**

- 0 — No stress
- 1 — Primary stress
- 2 — Secondary stress
CMUDict for Perl

- CPAN (Comprehensive Perl Archive Network)
  - [https://www.cpan.org/modules/INSTALL.html](https://www.cpan.org/modules/INSTALL.html)
- cpan App::cpanminus
- cpanm Module::Name
- CMUDict
  - [https://metacpan.org/pod/Lingua::EN::CMUDict](https://metacpan.org/pod/Lingua::EN::CMUDict)
  - cpanm Lingua::EN::CMUDict
cpanm install

cpan App::cpanminus
CPAN.pm requires configuration, but most of it can be done automatically. If you answer 'no' below, you will enter an interactive dialog for each configuration option instead.

Would you like to configure as much as possible automatically? [yes]

Use of uninitialized value $what in concatenation ($) or string at /System/Library/Perl/5.18/App/Cpan.pm line 553, <STDIN> line 1.

Warning: You do not have write permission for Perl library directories. To install modules, you need to configure a local Perl library directory or escalate your privileges. CPAN can help you by bootstrapping the local::lib module or by configuring itself to use 'sudo' (if available). You may also resolve this problem manually if you need to customize your setup. What approach do you want? (Choose 'local::lib', 'sudo' or 'manual')

[local::lib] sudo

Would you like me to automatically choose some CPAN mirror sites for you? (This means connecting to the Internet) [yes]

Fetching with LWP: http://mirrors-usa.go-parts.com/cpan/authors/01mailrc.txt.gz

....

... 
Running make install
Password:
Installing /Library/Perl/5.18/App/cpanminus.pm
Installing /Library/Perl/5.18/App/cpanminus/fatscript.pm
Installing /usr/local/share/man/man1/cpanm.1
Installing /usr/local/share/man/man3/App::cpanminus.3pm
Installing /usr/local/share/man/man3/App::cpanminus::fatscript.3pm
Installing /usr/local/bin/cpanm
Appending installation info to /Library/Perl/Updates/5.18.2/darwin-thread-multi-2level/perllocal.pod
MIYAGAWA/App-cpanminus-1.7043.tar.gz
sudo /usr/bin/make install -- OK
$ which cpanm
/usr/local/bin/cpanm
Module::Install install

```
sudo cpanm Module::Install
--> Working on Module::Install
Fetching http://www.cpan.org/authors/id/E/ET/ETHER/Module-Install-1.19.tar.gz ... OK
==> Found dependencies: YAML::Tiny
--> Working on YAML::Tiny
Fetching http://www.cpan.org/authors/id/E/ET/ETHER/YAML-Tiny-1.70.tar.gz ... OK
Configuring YAML-Tiny-1.70 ... OK
Building and testing YAML-Tiny-1.70 ... OK
Successfully installed YAML-Tiny-1.70
Configuring Module-Install-1.19 ... OK
==> Found dependencies: Parse::CPAN::Meta
--> Working on Parse::CPAN::Meta
Fetching http://www.cpan.org/authors/id/D/DA/DAGOLDEN/CPAN-Meta-2.150010.tar.gz ... OK
Configuring CPAN-Meta-2.150010 ... OK
Building and testing CPAN-Meta-2.150010 ... OK
Successfully installed CPAN-Meta-2.150010 (upgraded from 2.150010)
```
TAP::Harness::Env and File::ShareDir::Install install

```bash
sudo cpanm TAP::Harness::Env
--> Working on TAP::Harness::Env
Fetching http://www.cpan.org/authors/id/L/LE/LEONT/Test-Harness-3.39.tar.gz ... OK
Configuring Test-Harness-3.39 ... OK
Building and testing Test-Harness-3.39 ... OK
Successfully installed Test-Harness-3.39
1 distribution installed
```

```bash
sudo cpanm File::ShareDir::Install
--> Working on File::ShareDir::Install
Fetching http://www.cpan.org/authors/id/E/ET/ETHER/File-ShareDir-Install-0.11.tar.gz ... OK
Configuring File-ShareDir-Install-0.11 ... OK
Building and testing File-ShareDir-Install-0.11 ... OK
Successfully installed File-ShareDir-Install-0.11
1 distribution installed
```
sudo cpanm Lingua::EN::CMUDict

--> Working on Lingua::EN::CMUDict
Fetching http://www.cpan.org/authors/id/L/LM/LMETCALF/Lingua-EN-CMUDict-0.05.tar.gz ... OK
Configuring Lingua-EN-CMUDict-0.05 ... OK

==> Found dependencies: File::ShareDir

--> Working on File::ShareDir
Fetching http://www.cpan.org/authors/id/R/RE/REHSACK/File-ShareDir-1.104.tar.gz ... OK
Configuring File-ShareDir-1.104 ... OK
Building and testing File-ShareDir-1.104 ... OK
Successfully installed File-ShareDir-1.104
Building and testing Lingua-EN-CMUDict-0.05 ... OK
Successfully installed Lingua-EN-CMUDict-0.05
2 distributions installed
Using cmudict

```perl
use Lingua::EN::CMUDict;
my $obj = new Lingua::EN::CMUDict;
my $n = $obj->number_of_syllables($ARGV[0]);
if ($n) {
  print "$ARGV[0]: $n\n";
} else {
  print "Not in cmudict\n";
}
```

$ perl cmudict.perl pizza
pizza: 2
$ perl cmudict.perl absolutely
absolutely: 4
$ perl cmudict.perl sprezzatura
Not in cmudict
$
Perl @INC

- perl -e 'print "@INC\n"'
- /Library/Perl/5.18/darwin-thread-multi-2level /Library/Perl/5.18
  /Network/Library/Perl/5.18/darwin-thread-multi-2level
  /Network/Library/Perl/5.18 /Library/Perl/Updates/5.18.2/darwin-thread-multi-2level /Library/Perl/Updates/5.18.2
  /System/Library/Perl/5.18/darwin-thread-multi-2level
  /System/Library/Perl/5.18 /System/Library/Perl/Extras/5.18/darwin-thread-multi-2level /System/Library/Perl/Extras/5.18.
cmudict from nltk.corpus

```python
from nltk.corpus import cmudict

cmudict.dict()['pizza']
[['P', 'IY1', 'T', 'S', 'AH0']]

cmudict.dict()['absolutely']
[['AE2', 'B', 'S', 'AH0', 'L', 'UW1', 'T', 'L', 'IY0']]

cmudict.dict()['sprezzatura']
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
KeyError: 'sprezzatura'

cmudict.dict()['route']
[['R', 'UW1', 'T'], ['R', 'AW1', 'T']]
```
Homework 2

• Write a Python program to print out the number of syllables in a word in CMUdict
• Print an error message if not found
• Show it working
• Optional (100% Extra Credit):
  • add a "fallback" algorithm to calculate the number of syllables, e.g.

1 The "Written Method" Rules
   a Count the number of vowels (A, E, I, O, U) in the word.
      • Add 1 every time the letter 'y' makes the sound of a vowel (A, E, I, O, U).
      • Subtract 1 for each silent vowel (like the silent 'e' at the end of a word).
   b Subtract 1 for each diphthong or triphthong in the word.
      • Diphthong: when 2 vowels make only 1 sound (au, oy, oo)
      • Triphthong: when 3 vowels make only 1 sound (iou)
   c Does the word end with "le" or "les"? Add 1 only if the letter before the "le" is a consonant.
   d The number you get is the number of syllables in your word.
Homework 2

• due next Monday (by midnight)
• email to sandiway@email.arizona.edu
• usual rules: one PDF file etc.