UNIB20005

Language and Computation

Text Processing

Steven Bird, Department of Computing and Information Systems

Accessing Text

- Sources on the web and on disk
- NLTK corpora: special collections, already prepared
- What if you want to study some other text?
- Today: issues with processing raw text
- Wednesday: how to do it in Python

Text on the Web: Linguistic Value

Google hits	*adore*	*love*	*like*	*prefer*
absolutely	289,000	905,000	16,200	644
definitely	1,460	51,000	158,000	62,600
ratio	198:1	18:1	1:10	1:97

- Google queries, e.g.: “absolutely prefer”
- Benefits: coverage, ease of use
- Shortcomings of relying on a search engine: search patterns, inconsistencies, reproducibility, duplication
- Therefore: obtain texts and work with them directly (i.e. make your own corpus)

Text on the Web: additional material

http://news.bbc.co.uk/2/hi/health/2284783.stm

An urban legend published by BBC News

Text on the Web: HTML source

```
     !doctype html public "-//W3C//DTD HTML 4.0 Transitional//EN" "http://www.w3.org/TR/REC-html40/loose.dtd">
     <html>
<head>
<title>BBC NEWS | Health | Blondes 'to die out in 200 years'</title>

...650 lines...

<font face="sans-serif" size="1"><span class="date">Friday, 27 September, 2002, 11:51 GMT 12:51 UK
</span></font>
     <div class="headlinestory"><b>Blondes 'to die out in 200 years'</b><br></div>
     <div class="inlineimage">
          <img height="180" vspace="0" border="0" width="300" alt=" " src="/media/images/38280000/jpg/_38280456_blonde300.jpg">
               <div class="caption"><font size="1">Scientists believe the last blondes will be in Finland</font><br></div>
     </div>
     <font class="body" face="sans-serif" size="2">
     <div class="bodytext">
     The last natural blondes will die out within 200 years, scientists believe.
<P>
A study by experts in Germany suggests people with blonde hair are an endangered species and will become extinct by 2202.
<P>
Researchers predict the last truly natural blonde will be born in Finland - the country with the highest proportion of blondes.
<P>
<P>
<!-- GENInlineBOX -->
     <table bgcolor="#FFFFCC" class="boxbody" cellspacing="0" width="150" border="0" cellpadding="3" align="right">
<!-- GENInlineQUOTE -->
     <tr><td><img src="/nol/shared/img/startquote.gif" width="23" height="18" border="0" valign="TOP" alt=""><br><div
class="boxbody">
     The frequency of blondes may drop but they won't disappear
     </div><img align="RIGHT" src="/nol/shared/img/endquote.gif" width="23" height="18" border="0" valign="ABSBOTTOM" alt=""><br
clear="ALL"></td></tr>
<!-- GENInlineNAME -->
     <tr><td bgcolor="cccc99"><div class="boxhead">
     Prof Jonathan Rees, University of Edinburgh
     </div></td></tr>
     </table>
But they say too few people now carry the gene for blondes to last beyond the next two centuries.
<P>
The problem is that blonde hair is caused by a recessive gene.
<P>
```

Text on the Web: Extracting Text from HTML

Simple method: delete all markup and collapse whitespace

BBC NEWS | Health | Blondes 'to die out in 200 years' NEWS   SPORT   WEATHER   WORLD SERVICE   WHERE I LIVE -->   A-Z INDEX   SEARCH   You are in: Health   News Front Page Africa Americas Asia-Pacific Europe Middle East South Asia UK Business Entertainment Science/Nature Technology Health Medical notes ---------- Talking Point ---------- Country Profiles In Depth ---------- Programmes ---------- SERVICES Daily E-mail News Ticker Mobile/PDAs ---------- Text Only Feedback Help EDITIONS Change to UK Friday, 27 September, 2002, 11:51 GMT 12:51 UK Blondes 'to die out in 200 years' Scientists believe the last blondes will be in Finland **The last natural blondes will die out within 200 years, scientists believe. A study by experts in Germany suggests people with blonde hair are an endangered species and will become extinct by 2202. Researchers predict the last truly natural blonde will be born in Finland - the country with the highest proportion of blondes.** The frequency of blondes may drop but they won't disappear Prof Jonathan Rees, University of Edinburgh **But they say too few people now carry the gene for blondes to last beyond the next two centuries. The problem is that blonde hair is caused by a recessive gene. In order for a child to have blonde hair, it must have the gene on both sides of the family in the grandparents' generation. Dyed rivals The researchers also believe that so-called bottle blondes may be to blame for the demise of their natural rivals. They suggest that dyed-blondes are more attractive to men who choose them as partners over true blondes.** Bottle-blondes like Ann Widdecombe may be to blame **But Jonathan Rees, professor of dermatology at the University of Edinburgh said it was unlikely blondes would die out completely "Genes don't die out unless there is a disadvantage of having that gene or by chance. They don't disappear," he told BBC News Online. "The only reason blondes would disappear is if having the gene was a disadvantage and I do not think that is the case. "The frequency of blondes may drop but they won't disappear."** See also: 28 Mar 01 | Education What is it about blondes? 09 Apr 99 | Health Platinum blondes are labelled as dumb 17 Apr 02 | Health Hair dye cancer alert Internet links: University of Edinburgh The BBC is not responsible for the content of external internet sites Top Health

Extracting text from HTML (3.1)

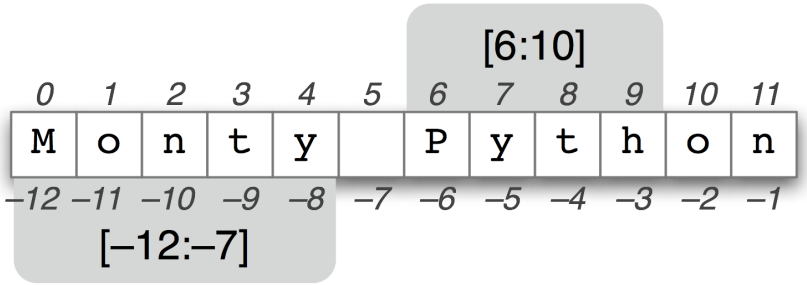
- This is a non-trivial task!

- We will use a built-in “library function” to do this

```
>>> from urllib import urlopen
>>> url = ‘http://...’
>>> html = urlopen(url).read()
>>> raw = clean_html(html)
```

String processing (3.2)

- Basic data type: sequence of characters
  - its what we get when we read from a file or URL
  - we cannot process a text until we split it into tokens
- Basic string operations
  - indexing
  - slicing
  - concatenation
- From strings to lists to strings
  - splitting
  - joining



Regular Expressions (3.4)

- Metacharacters
  - start ^, end $, wildcard .
  - slicing
  - concatenation
- Ranges and closures
  - [] [^] (| |)
  - +, *, ?
  - \w, \d, \s (plus inverses)
- Applications
  - [w for w in wordlist if re.search('...', w)]

Tokenization (3.7)

- Simple approach: split on whitespace:

```
The last natural blondes will die out within 200 years, scientists believe.
```

- Split off punctuation as well:

```
"The frequency of blondes may drop but they won't disappear."
```

- Harder case (*Alice in Wonderland*)

```
'When I'M a Duchess,' she said to herself, (not in a very hopeful tone though), 'I
won't have any pepper in my kitchen AT ALL. Soup does very well without--Maybe it's
always pepper that makes people hot-tempered,'
```

- Sentence tokenization (aka “sentence segmentation”):

```
But Jonathan Rees, professor of dermatology at the University of Edinburgh said it
was unlikely blondes would die out completely "Genes don't die out unless there is
a disadvantage of having that gene or by chance. They don't disappear," he told BBC
News Online.  The only reason blondes would disappear is if having the gene was a
disadvantage and I do not think that is the case.  "The frequency of blondes may
drop but they won't disappear."
```

Aside:
Word segmentation and language learning

a. doyouseethekitty
b. seethedoggy
c. doyoulikethekitty
d. likethedoggy