UNIB20005

Language and Computation

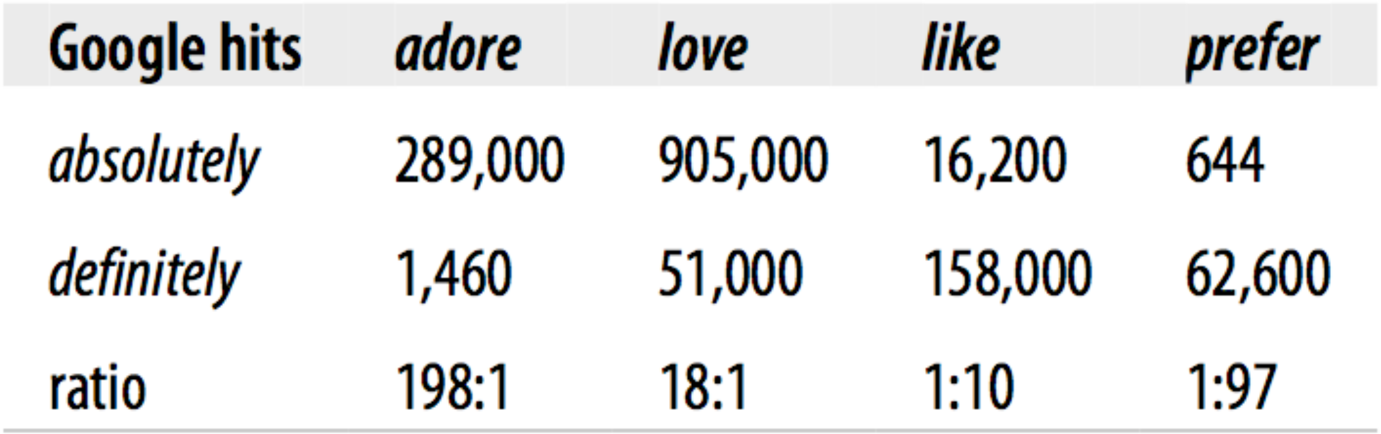
Text Processing

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

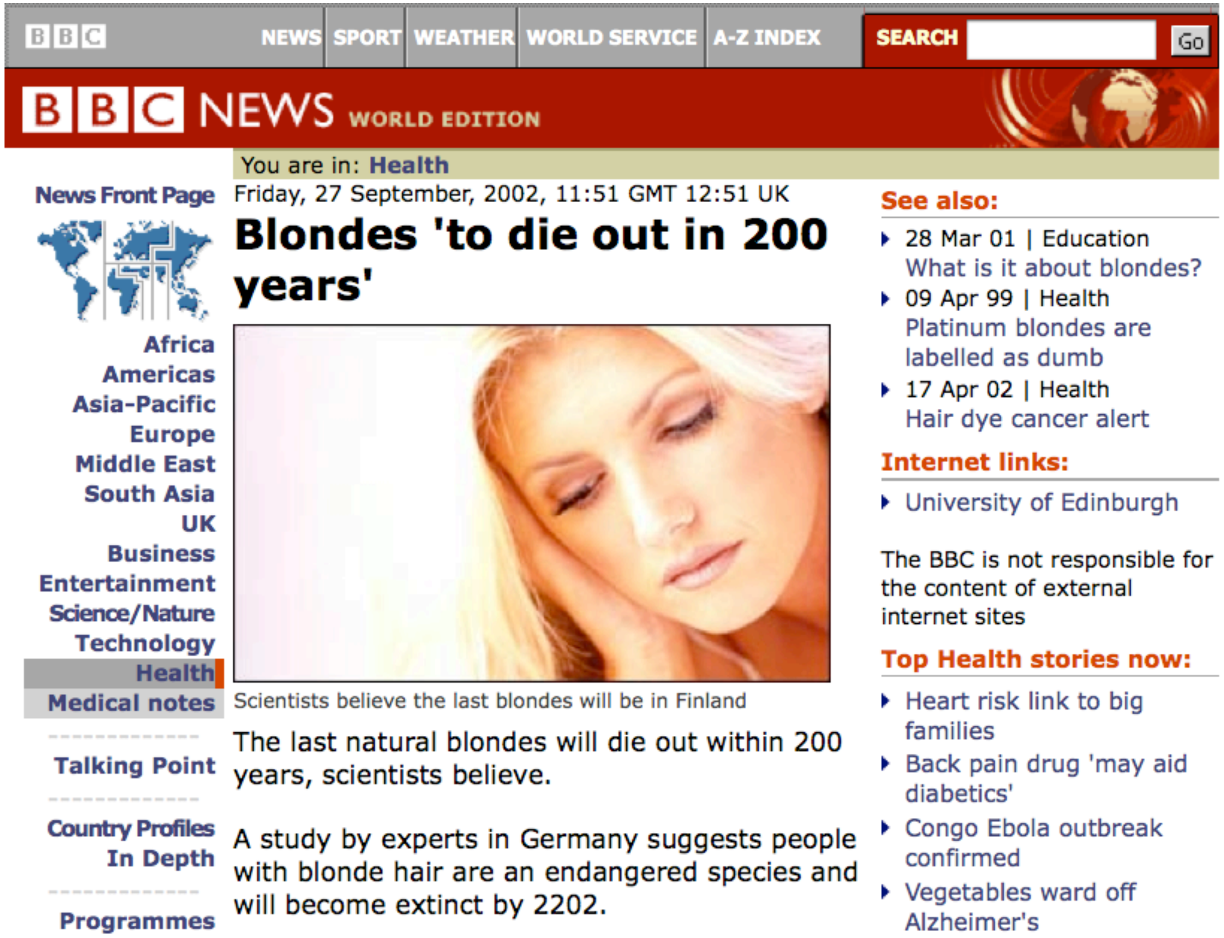
Text on the Web: Linguistic Value



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



You are in: Health

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.

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 stories now:

- Heart risk link to big families
- Back pain drug 'may aid diabetics'
- Congo Ebola outbreak confirmed
- Vegetables ward off Alzheimer's

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

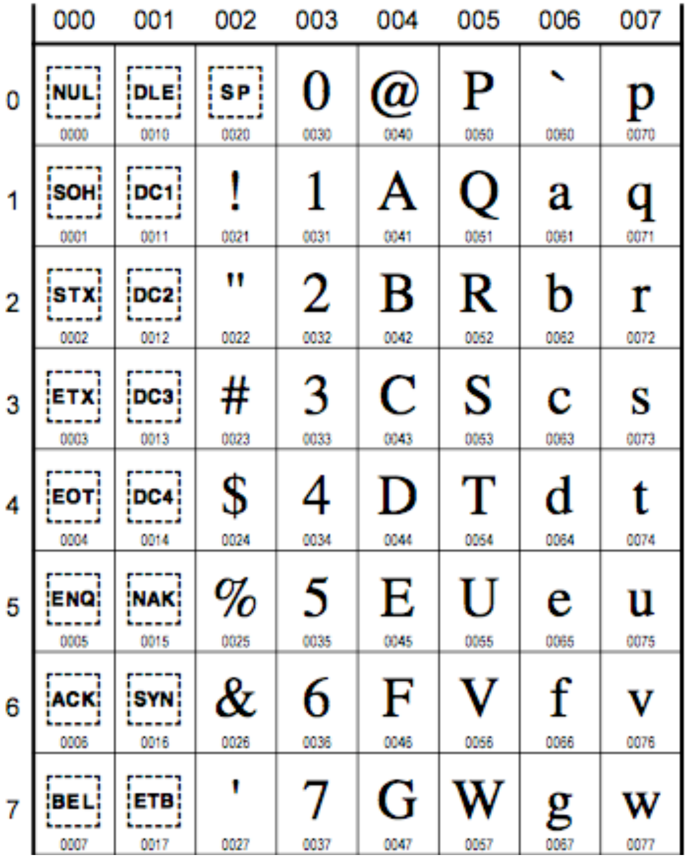
```
    BBC NEWS | Health | Blondes 'to die out in 200 years' NEWS &nbsp;&nbsp;SPORT
&nbsp;&nbsp;WEATHER &nbsp;&nbsp;WORLD SERVICE &nbsp;&nbsp;WHERE&nbsp;I&nbsp;LIVE -->
&nbsp;&nbsp;A-Z INDEX&nbsp; &nbsp;&nbsp;SEARCH&nbsp; &nbsp;You are in:&nbsp;Health
&nbsp; News Front Page Africa Americas Asia-Pacific 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&nbsp;|&nbsp;Education What is it about blondes?  09 Apr
99&nbsp;|&nbsp;Health Platinum blondes are labelled as dumb 17 Apr
02&nbsp;|&nbsp;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

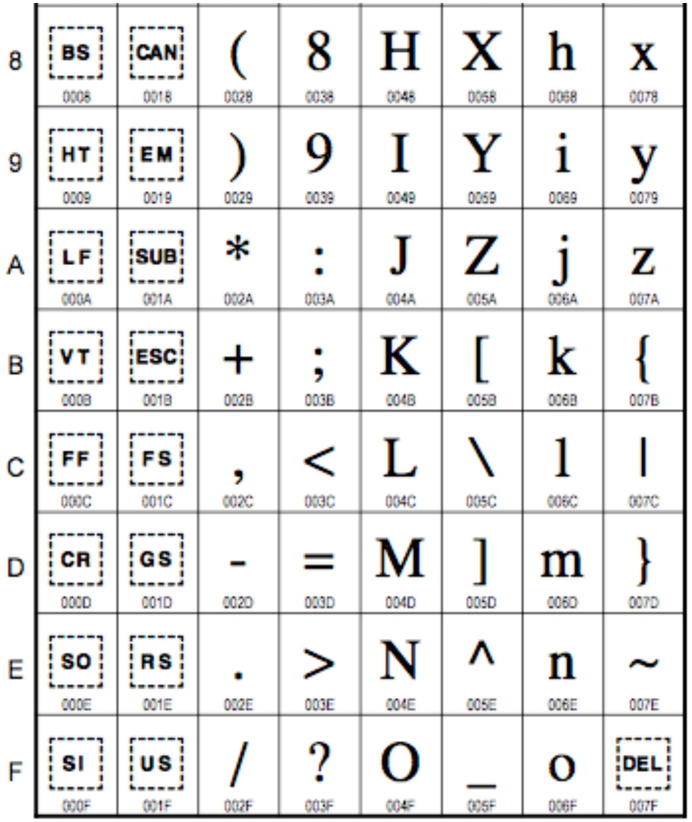
- This is a non-trivial task!
- We will use a built-in “library function” to do this

Unicode: Working at the level of Characters

Unicode: Code Points and Glyphs

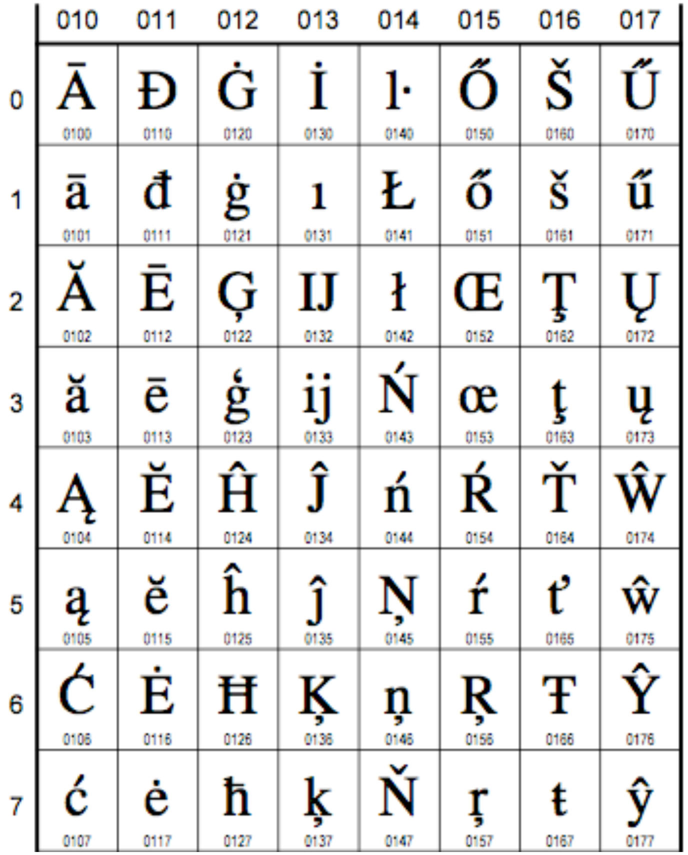


	000	001	002	003	004	005	006	007
0	NUL 0000	DLE 0010	SP 0020	0 0030	@ 0040	P 0050	` 0060	p 0070
1	SOH 0001	DC1 0011	! 0021	1 0031	A 0041	Q 0051	a 0061	q 0071
2	STX 0002	DC2 0012	" 0022	2 0032	B 0042	R 0052	b 0062	r 0072
3	ETX 0003	DC3 0013	# 0023	3 0033	C 0043	S 0053	c 0063	s 0073
4	EOT 0004	DC4 0014	$ 0024	4 0034	D 0044	T 0054	d 0064	t 0074
5	ENQ 0005	NAK 0015	% 0025	5 0035	E 0045	U 0055	e 0065	u 0075
6	ACK 0006	SYN 0016	& 0026	6 0036	F 0046	V 0056	f 0066	v 0076
7	BEL 0007	ETB 0017	' 0027	7 0037	G 0047	W 0057	g 0067	w 0077

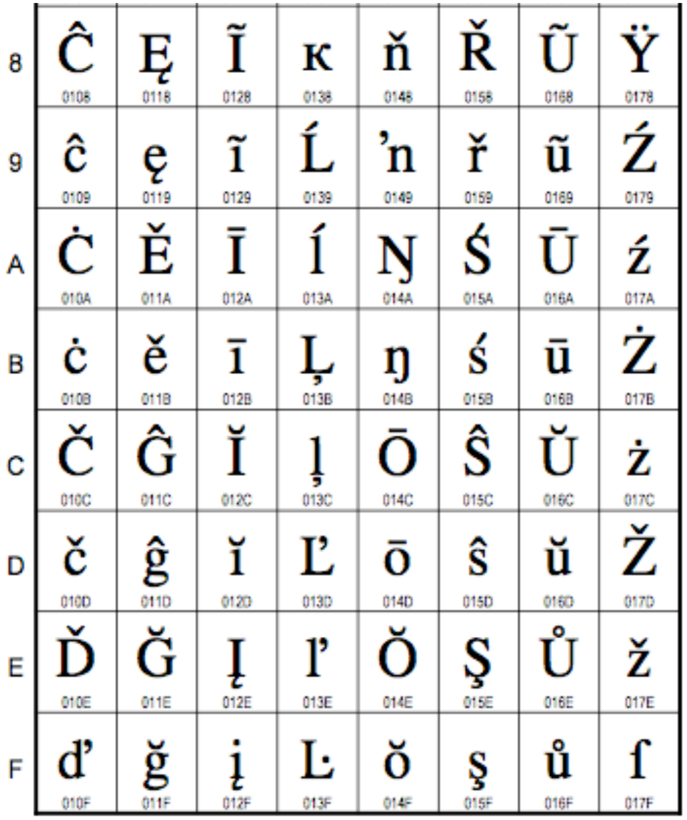


8	BS 0008	CAN 0018	(0028	8 0038	H 0048	X 0058	h 0068	x 0078
9	HT 0009	EM 0019	) 0029	9 0039	I 0049	Y 0059	i 0069	y 0079
A	LF 000A	SUB 001A	* 002A	: 003A	J 004A	Z 005A	j 006A	z 007A
B	VT 000B	ESC 001B	+ 002B	; 003B	K 004B	[005B	k 006B	{ 007B
C	FF 000C	FS 001C	, 002C	< 003C	L 004C	\ 005C	l 006C	| 007C
D	CR 000D	GS 001D	- 002D	= 003D	M 004D	] 005D	m 006D	} 007D
E	SO 000E	RS 001E	. 002E	> 003E	N 004E	^ 005E	n 006E	~ 007E
F	SI 000F	US 001F	/ 002F	? 003F	O 004F	_ 005F	o 006F	DEL 007F

Unicode (cont)



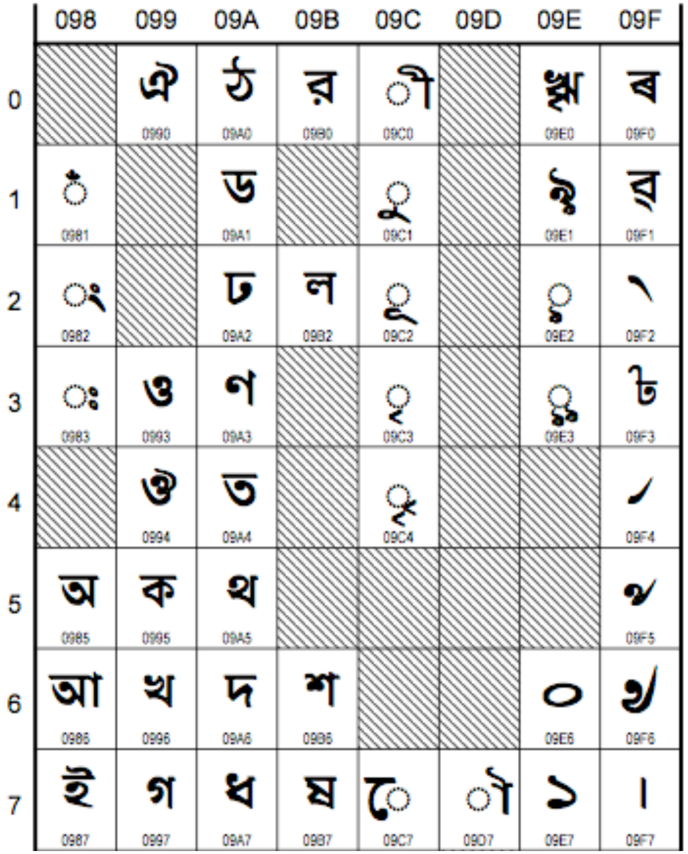
	010	011	012	013	014	015	016	017
0	Ā 0100	Đ 0110	Ġ 0120	İ 0130	ŀ 0140	Ő 0150	Š 0160	Ű 0170
1	ā 0101	đ 0111	ġ 0121	ı 0131	Ł 0141	ő 0151	š 0161	ű 0171
2	Ă 0102	Ē 0112	Ģ 0122	Ĳ 0132	ł 0142	Œ 0152	Ţ 0162	Ų 0172
3	ă 0103	ē 0113	ģ 0123	ĳ 0133	Ń 0143	œ 0153	ţ 0163	ų 0173
4	Ą 0104	Ĕ 0114	Ĥ 0124	Ĵ 0134	ń 0144	Ŕ 0154	Ť 0164	Ŵ 0174
5	ą 0105	ĕ 0115	ĥ 0125	ĵ 0135	Ņ 0145	ŕ 0155	ť 0165	ŵ 0175
6	Ć 0106	Ė 0116	Ħ 0126	Ķ 0136	ņ 0146	Ŗ 0156	Ŧ 0166	Ŷ 0176
7	ć 0107	ė 0117	ħ 0127	ķ 0137	Ň 0147	ŗ 0157	ŧ 0167	ŷ 0177



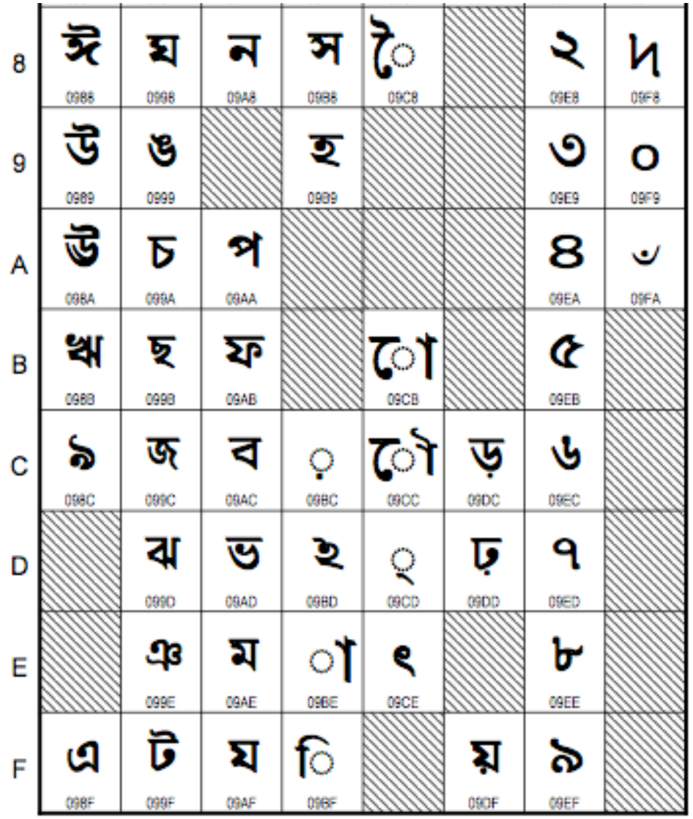
8	Ĉ 0108	Ę 0118	Ĩ 0128	ĸ 0138	ň 0148	Ř 0158	Ũ 0168	Ÿ 0178
9	ĉ 0109	ę 0119	ĩ 0129	Ĺ 0139	ŉ 0149	ř 0159	ũ 0169	Ź 0179
A	Ċ 010A	Ě 011A	Ī 012A	ĺ 013A	Ŋ 014A	Ś 015A	Ū 016A	ź 017A
B	ċ 010B	ě 011B	ī 012B	Ļ 013B	ŋ 014B	ś 015B	ū 016B	Ż 017B
C	Č 010C	Ĝ 011C	Ĭ 012C	ļ 013C	Ō 014C	Ŝ 015C	Ŭ 016C	ż 017C
D	č 010D	ĝ 011D	ĭ 012D	Ľ 013D	ō 014D	ŝ 015D	ŭ 016D	Ž 017D
E	Ď 010E	Ğ 011E	Į 012E	ľ 013E	Ŏ 014E	Ş 015E	Ů 016E	ž 017E
F	ď 010F	ğ 011F	į 012F	Ŀ 013F	ŏ 014F	ş 015F	ů 016F	ſ 017F

- Code pages: http://www.unicode.org/charts/

E.g. Bengali



	098	099	09A	09B	09C	09D	09E	09F
0		ঐ 0990	ঠ 09A0	র 09B0	◌ী 09C0		ৠ 09E0	ৰ 09F0
1	◌ঁ 0981		ড 09A1		◌ু 09C1		ৡ 09E1	ৱ 09F1
2	◌ং 0982		ঢ 09A2	ল 09B2	◌ূ 09C2		◌ৢ 09E2	৲ 09F2
3	◌ঃ 0983	ও 0993	ণ 09A3		◌ৃ 09C3		◌ৣ 09E3	৳ 09F3
4		ঔ 0994	ত 09A4		◌ৄ 09C4			৴ 09F4
5	অ 0985	ক 0995	থ 09A5					৵ 09F5
6	আ 0986	খ 0996	দ 09A6	শ 09B6			০ 09E6	৶ 09F6
7	ই 0987	গ 0997	ধ 09A7	ষ 09B7	◌ে 09C7	◌ৗ 09D7	১ 09E7	৷ 09F7

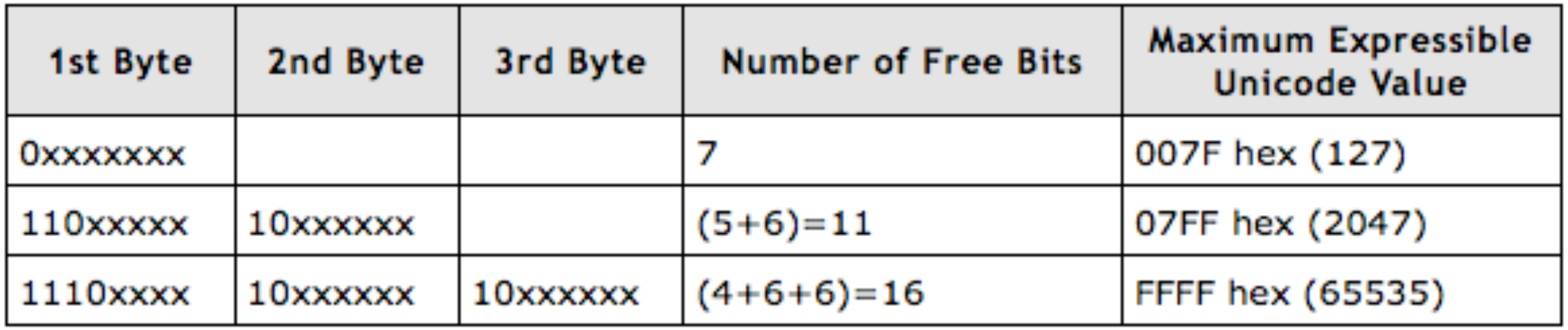


8	ঈ 0988	ঘ 0998	ন 09A8	স 09B8	◌ৈ 09C8		২ 09E8	৸ 09F8
9	উ 0989	ঙ 0999		হ 09B9			৩ 09E9	৹ 09F9
A	ঊ 098A	চ 099A	প 09AA				৪ 09EA	৺ 09FA
B	ঋ 098B	ছ 099B	ফ 09AB		◌ো 09CB		৫ 09EB	
C	ঌ 098C	জ 099C	ব 09AC	◌় 09BC	◌ৌ 09CC	ড় 09DC	৬ 09EC	
D		ঝ 099D	ভ 09AD	ঽ 09BD	◌্ 09CD	ঢ় 09DD	৭ 09ED	
E		ঞ 099E	ম 09AE	◌া 09BE	ৎ 09CE		৮ 09EE	
F	এ 098F	ট 099F	য 09AF	◌ি 09BF		য় 09DF	৯ 09EF	

Unicode Normalization

- typographic ligatures, e.g. ﬃ
- precomposed diacritics, e.g. o + ̈ = ö
- U+006F U+0308 = U+00F6

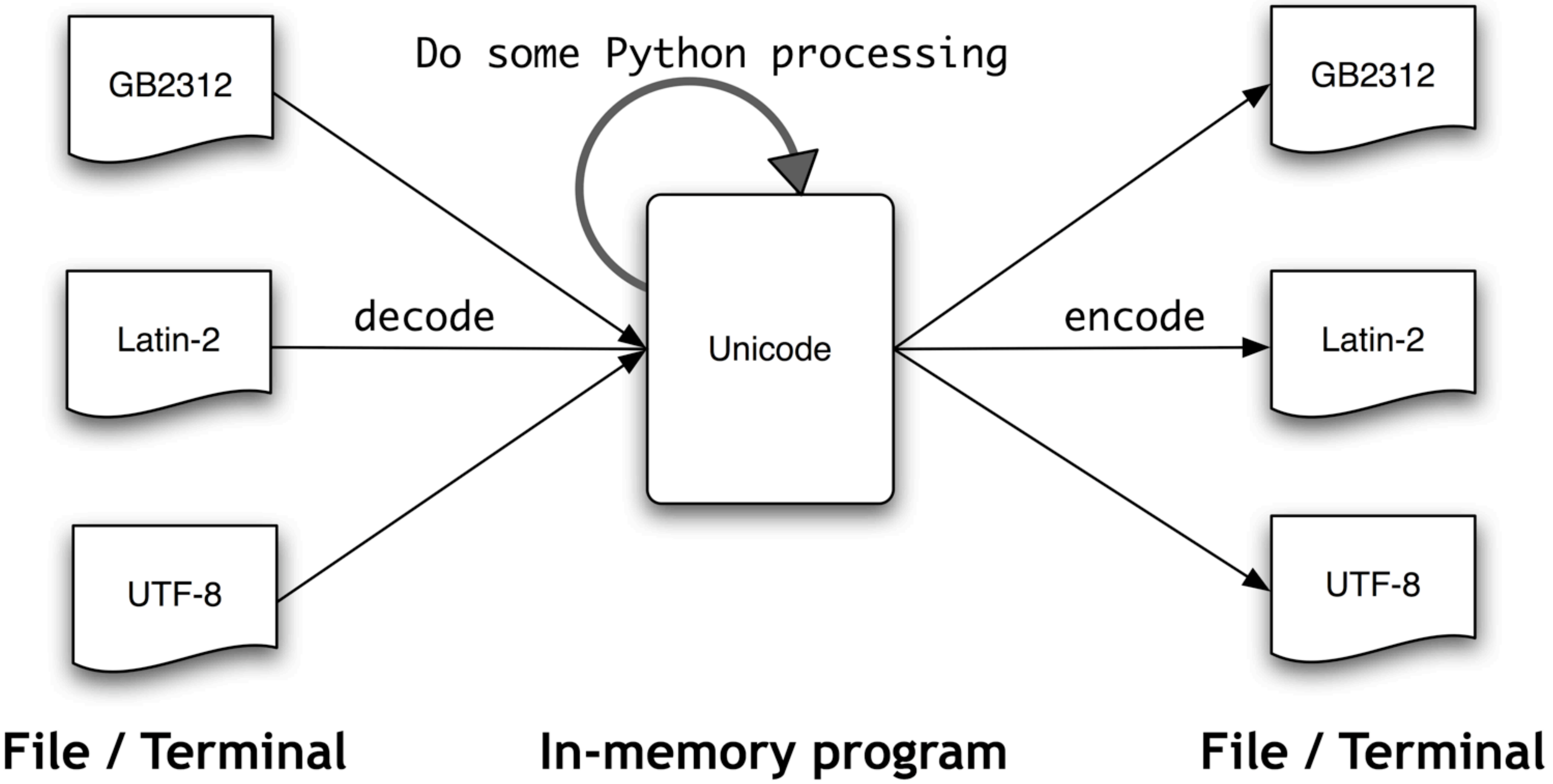
UTF-8 Encoding



1st Byte	2nd Byte	3rd Byte	Number of Free Bits	Maximum Expressible Unicode Value
0xxxxxxx			7	007F hex (127)
110xxxxx	10xxxxxx		(5+6)=11	07FF hex (2047)
1110xxxx	10xxxxxx	10xxxxxx	(4+6+6)=16	FFFF hex (65535)

- UTF = "Unicode Transformation Format", e.g. UTF-8, UTF-16, UTF-32
- An *encoding* is how we represent a *codepoint* as a unique sequence of *bytes*
- For codepoints 0..127 we use one byte
- For codepoints 128..2047 we use two bytes
- For codepoints 2048..65535 we use three bytes
- This is called a "variable length" encoding

Text Processing with Unicode



Extracting characters from bytes

- This is a non-trivial task!
- We will use a built-in “library function” to do this

Tokenization: From Strings to Tokens

Strings

- Basic data type: sequence of characters
- what we get when we read from a file or URL
- we cannot process a text till we split it into tokens

Tokenization

- 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."
  ```

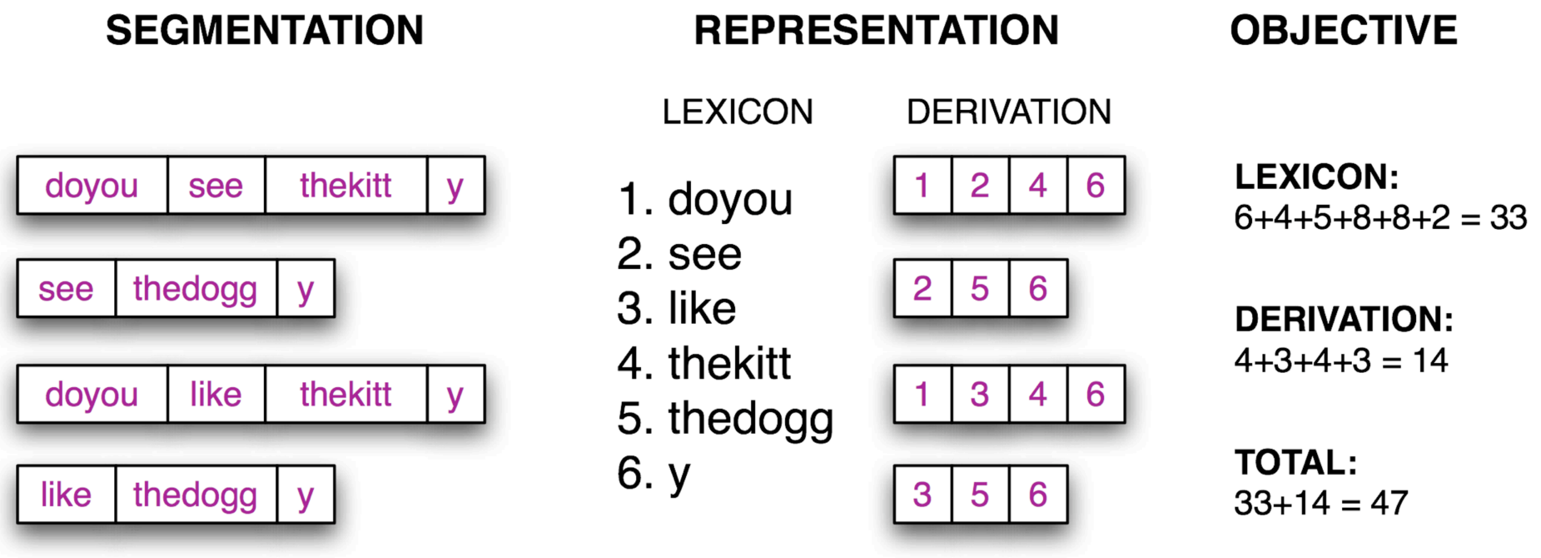
Aside:
Word segmentation and language learning

a. doyouseethekitty

b. seethedoggy

c. doyoulikethekitty

d. likethedoggy



SEGMENTATION

doyou see thekitt y

see thedogg y

doyou like thekitt y

like thedogg y

REPRESENTATION

LEXICON

1. doyou
2. see
3. like
4. thekitt
5. thedogg
6. y

DERIVATION

1 2 4 6

2 5 6

1 3 4 6

3 5 6

OBJECTIVE

LEXICON:
6+4+5+8+8+2 = 33

DERIVATION:
4+3+4+3 = 14

TOTAL:
33+14 = 47

Regular Expressions

- Motivations: tokenization, morphology
- Metacharacters: . ^ $
- Ranges: [abcdefg] [a-g] [^aeiou]
  \w (word character) [a-zA-Z0-9]
  \d (digit) [0-9]
  \s (space) [\t\n\r\f\v]
  \W, \D, \S
- Closures: a* a+ a? a{3,7}
- Alternatives: (...|...|...)
- Demonstration (nltk.app.nemo)