Informatics 2A: Tutorial Sheet 5 Solutions

Shay Cohen

1. (a) For the purpose of a rough calculation, we can approximate the frequency graph by the curve $y = c/x$ for a suitable constant c. The total number of tokens will then be

$$\int_1^{10000} c/x \; dx \;=\; c[\ln x]_1^{10000} \;=\; 4c \ln 10 \;=\; 100000$$

   (and this fixes the value of c). So to obtain half the total number of tokens, we clearly want to take the 100 most common word types:

$$\int_1^{100} c/x \; dx \;=\; c[\ln x]_1^{100} \;=\; 2c \ln 10$$

   (b) From the above, we have $c \approx 10857$. So the frequency of *about* is roughly $10857/60 \approx 181$.

2. Here's one way to tag the text, based on the Penn Treebank tagging guidelines:

   I/PRP was/VBD walking/VBG down/IN the/DT high/JJ street/NN yesterday/NN when/CC I/PRP noticed/VBD an/DT old/JJ man/NN acting/VBG suspiciously/RB . He/PRP was/VBD peering/VBG into/IN various/JJ shop/NN windows/NN and/CC writing/VBG things/NNS in/IN a/DT notebook/NN . When/WRB he/PRP spotted/VBD me/PRP, he/PRP stuffed/VBD the/DT notebook/NN into/IN his/PRP$ pocket/NN and/CC wandered/VBD off/RP ./.

   Here's how the Stanford tagger tags it:

   I/PRP was/VBD walking/VBG down/RP the/DT high/JJ street/NN yesterday/NN when/WRB I/PRP noticed/VBD an/DT old/JJ man/NN acting/VBG suspiciously/RB ./. He/PRP was/VBD peering/VBG into/IN various/JJ shop/NN windows/NNS and/CC writing/VBG things/NNS in/IN a/DT notebook/NN ./. When/WRB he/PRP spotted/VBD me/PRP ,/, he/PRP stuffed/VBD the/DT notebook/NN into/IN his/PRP$ pocket/NN and/CC wandered/VBD off/RP ./.

   You can see it sometimes makes mistakes, for example, denoting "down" as a particle.

   Here is the Penn treebank POS tagset if needed for discussion:

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential there
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	to
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

For most words here the tagging is straightforward, but the following points might be discussed:

- If *high street* were regarded as a compound noun, the tagging would be high/NN street/NN.
- One might very reasonably want to tag *yesterday* as a temporal adverb (RB). The Penn guidelines, however, say that it should be treated as a noun (even in contexts like the above), pointing out e.g. that it admits a possessive form *yesterday's news*.
- We have tagged the first *when* as a coordinating conjunction, and the second as a Wh-adverb, though it is not entirely clear whether this accords with Penn Treebank policy.

3. We only have to tag the words *old* and *man*, since the tagging of the other words is fixed. Proceeding from left to right, we see that if *old* is preceded by a DT, its most likely POS is Adj, while if *man* is preceded by Adj, its most likely POS is N.

   (This is admittedly a rather weak example, in that the tagging of *man* would be the same whatever preceded it!)

4. The Viterbi matrix is as follows:

	the	old	man	the	lifeboats
DT	.4x.5 = .2	0	0	.00096x.4x.5 = .000192	0
N	0	.2x.6x.2 = .024	.032x.5x.3 = .0048	0	etc.
V	0	0	.024x.4x.1 = .00096	0	0
Adj	0	.2x.4x.4 = .032	0	0	0

   Thus the most probable tagging is:

   The/DT old/N man/V the/DT lifeboats/N

   (The backtrace pointers can be read off from the above matrix in an ad hoc fashion: e.g. in the cell for (man,N), the first factor is .032 which comes from the cell for (the,Adj).)