

## Informatics 2A: Tutorial Sheet 5 (Week 7)

### Part-of-speech tagging

SHAY COHEN

1. Suppose you are given a text corpus of 100,000 tokens involving 10,000 distinct word types. Assume for the purpose of this question that the corpus perfectly obeys Zipf's law. (This wouldn't be possible in practice, since the number of occurrences of each word type must be a whole number.)
  - (a) Estimate how many word types are needed to account for *half* the tokens in the corpus. (Your calculation need not be too precise.)
  - (b) The word *about* is the 60th most common word in the corpus. Estimate how many times it occurs.
2. The following are some of the POS tags from the Penn Treebank tagset, illustrated with sample words

NN	Noun, singular ( <i>cat</i> )
NNS	Noun, plural ( <i>cats</i> )
VBD	Verb, past tense ( <i>sang</i> )
VBG	Verb, gerund form ( <i>singing</i> )
JJ	Adjective ( <i>huge</i> )
RB	Adverb ( <i>hugely, soon, very</i> )
PRP	Personal pronoun ( <i>you</i> )
PRP\$	Possessive pronoun ( <i>your</i> )
IN	Preposition ( <i>on, up</i> )
DT	Determiner ( <i>the, some</i> )
CC	Coordinating conjunction ( <i>and, but</i> )
WRB	Wh-adverb ( <i>where, why</i> )

Tag the following text, ignoring punctuation. You may assume the above tags are sufficient for this purpose. Highlight any points at which you think difficult or debatable tagging decisions arise.

I was walking down the high street yesterday when I noticed an old man acting suspiciously. He was peering into various shop windows and writing things in a notebook. When he spotted me, he stuffed the notebook into his pocket and wandered off.

(For this question, getting the 'right answer' is less important than having the experience of trying.)

3. Consider the following (artificial) sentence:

The old man the lifeboats

Use the version of *bigram tagging* described in the lectures to tag this sentence, using the tags DT, N, V, Adj and the following frequency data. (Rows correspond to potential POS tags for the word in question; columns

correspond to the POS tag of the preceding word.) You may assume *the* and *lifeboats* can only be tagged as DT and N respectively.

<b>old</b>	DT	N	V	Adj	<b>man</b>	DT	N	V	Adj
N	8	2	3	2	N	102	45	15	86
V	0	0	0	0	V	0	11	4	4
Adj	34	5	13	17	Adj	0	0	0	0

4. Now use the *Viterbi algorithm* to tag the sentence

The old man the lifeboats

using the following transition and emission probabilities. Include explicit backtrace pointers in your Viterbi matrix.

	DT	N	V	Adj		lifeboat	man	old	the
start	0.4	0.3	0.1	0.2	DT	0	0	0	0.5
DT	0	0.6	0	0.4	N	0.2	0.3	0.2	0
N	0.05	0.3	0.4	0.25	V	0	0.1	0	0
V	0.4	0.3	0.1	0.2	Adj	0	0	0.4	0
Adj	0.1	0.5	0.2	0.2					

Transitions

Emissions

You may want to use a calculator to help with the arithmetic. Note too that in the transition matrix, rows represent the ‘previous state’, and columns represent the ‘next state’ (the opposite of the convention in question 3).