Due Monday, March 25, 2019

POS-Tagging with N-Grams

PRP	VBD	VBN	IN	PRP	NN	NNP	NNP	IN	DT
I	had	called	upon	my	friend	Sherlock	Holmes	upon	the

JJ	NN	IN	NNP	,	IN	DT	NN	IN	VBG
second	morning	after	Christmas	,	with	the	intention	of	wishing

PRP	DT	NNS	IN	DT	NN	.	PRP	VBD	VBG	IN
him	the	compliments	of	the	season	.	He	was	lounging	upon

DT	NN	IN	DT	JJ	NN	,	DT	NN	IN
the	sofa	in	a	purple	dressing-gown	,	a	pipe-rack	within

PRP	NN	IN	DT	NN	,	CC	DT	NN	IN	JJ
his	reach	upon	the	right	,	and	a	pile	of	crumpled

NN	NNS	,	RB	RB	VBN	,	IN	IN	NN	.
morning	papers	,	evidently	newly	studied	,	near	at	hand	.

IN	DT	NN	VBD	DT	JJ	NN	,	CC	IN	DT
Beside	the	couch	was	a	wooden	chair	,	and	on	the

NN	IN	DT	NN	VBD	DT	RB	JJ	CC	JJ	NN
angle	of	the	back	hung	a	very	seedy	and	disreputable	hard-felt

NN	,	RB	DT	NN	IN	NN	,	CC	VBD	IN
hat	,	much	the	worse	for	wear	,	and	cracked	in

JJ	NNS	.	DT	NN	CC	DT	NN	VBG	IN	DT
several	places	.	A	lens	and	a	forceps	lying	upon	the

NN	IN	DT	NN	VBD	IN	DT	NN	VBD	VBN
seat	of	the	chair	suggested	that	the	hat	had	been

VBN	IN	DT	NN	IN	DT	NN	IN	NN	.
suspended	in	this	manner	for	the	purpose	of	examination	.

Answer the following questions about the passage above.

a) How many bigrams are there?

b) What is P (DT | .)?

c) What is P (JJ | DT)?

d) What is P (NN | DT)?

The bigram probability can be estimated as:

$$P(t_n \mid t_{n-1}) = \frac{Count(t_{n-1}\ t_n)}{Count(t_{n-1})}$$

e) Rewrite this equation to estimate trigrams.

f) What is P (NN | IN DT)?

Estimate using interpolation:

g) $\hat{P}(JJ \mid VBD\ DT)$

h) $\hat{P}(DT \mid PRP\ VBD)$

$$\hat{P}(t_n \mid t_{n-2}\ t_{n-1}) = \lambda_1 P(t_n) + \lambda_2 P(t_n \mid t_{n-1}) + \lambda_3 P(t_n \mid t_{n-2}\ t_{n-1})$$

$$\lambda_1 = 0.1\ \lambda_2 = 0.7\ \lambda_3 = 0.2$$