Part 1.1 -> Getting MLE:

Author used: Charles Dickens

(search was done case insensitive)

Corpus:

1. Book: Oliver twist
   Total words: 162850

   my: 512
   Cat: 2
   Likes: 6
   Dog: 75
   Food: 12

2. Book: David Copperfield
   Total words: 363544

   My: 5204
   Cat: 13
   Likes: 11
   Dog: 42
   Food: 3

3. Book: Nicholas Nickelby
   Total words: 328999

   My: 1314
   Cat: 5
   Likes: 9
   Dog: 27
   Food: 17

   (a) MLE of terms :
   i) my : (512+5204+1314)/ (162850+363544+328999) = 7030/855393 = 0.0082
   ii) cat : (2+13+5)/(162850+363544+328999) = 20/855393 = 0.000023
   iii) likes: (6+11+9)/(162850+363544+328999) = 26/855393 = 0.000030
   iv) dog: (75+42+27)/(162850+363544+328999) = 144/855393 = 0.00017
   v) food: (12+3+17)/(162850+363544+328999) = 0.000037

   b) Likelihood of phrase: My cat likes dog food
   p(my) * p(cat) * p(likes) * p(dog) * p(food)
   = 0.0082 * 0.000023 * 0.00003 * 0.00017 * 0.000037 = 3.5 * 10^-20

Part 1.2 -> Plotting Zipf's law

Pargraph:

It was into a place of this kind that Mr Ralph Nickleby gazed, as he sat with his hands in his pockets looking out of **the** window. He had fixed his eyes upon a distorted fir tree, planted by some former tenant in a tub that had once been green, and left there, years before, to rot away piecemeal. There was nothing very inviting in **the** object, but Mr Nickleby was wrapped in a brown study, and sat contemplating it with far greater attention than, in a more conscious mood, he would have deigned to bestow upon **the** rarest exotic. At length, his eyes wandered to a little dirty window on **the** left, through which **the** face of **the** clerk was dimly visible; that worthy chancing to look up, he beckoned him to attend.

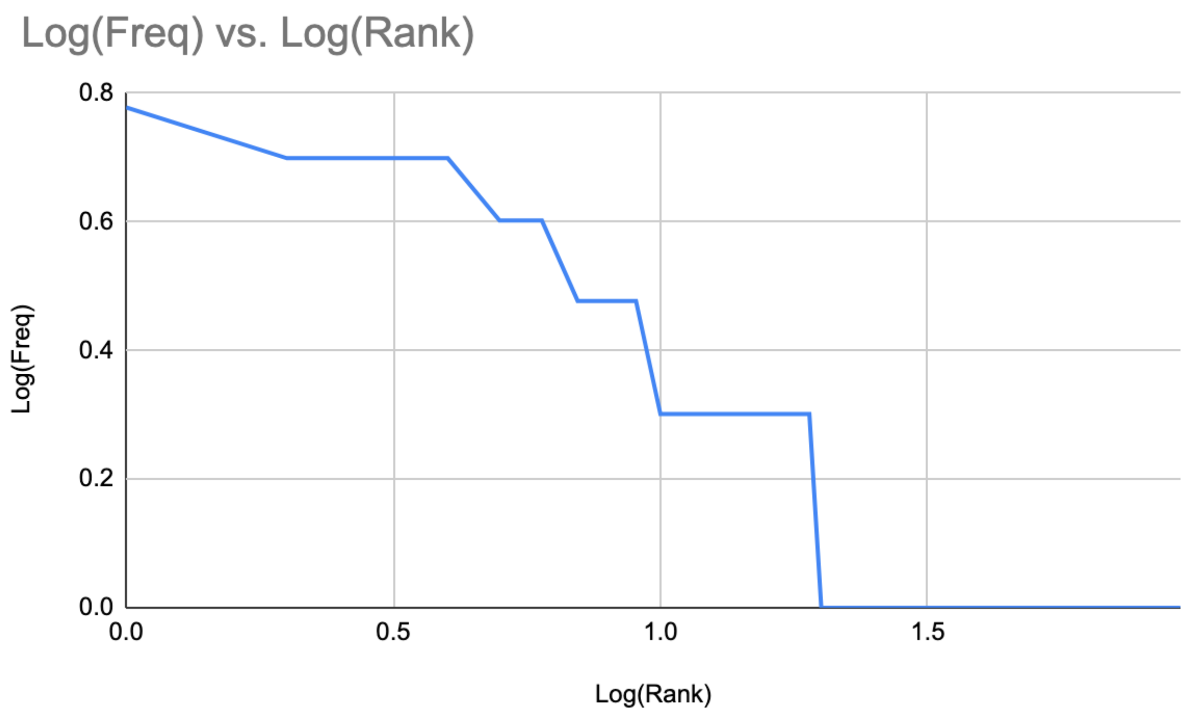
Term Frequency table

Term	Frequency	Rank	Log(Freq)	Log(Rank)
a	6	1	0.7781512504	0
the	5	2	0.6989700043	0.3010299957
in	5	3	0.6989700043	0.4771212547
to	5	4	0.6989700043	0.6020599913
was	4	5	0.6020599913	0.6989700043
his	4	6	0.6020599913	0.7781512504
of	3	7	0.4771212547	0.84509804
that	3	8	0.4771212547	0.903089987
he	3	9	0.4771212547	0.9542425094
Mr	2	10	0.3010299957	1
Nickleby	2	11	0.3010299957	1.041392685
sat	2	12	0.3010299957	1.079181246
with	2	13	0.3010299957	1.113943352
window	2	14	0.3010299957	1.146128036
had	2	15	0.3010299957	1.176091259
eyes	2	16	0.3010299957	1.204119983
upon	2	17	0.3010299957	1.230448921
and	2	18	0.3010299957	1.255272505
left	2	19	0.3010299957	1.278753601

It'	1	20	0	1.301029996
into'	1	21	0	1.322219295
place'	1	22	0	1.342422681
this'	1	23	0	1.361727836
kind'	1	24	0	1.380211242
Ralph'	1	25	0	1.397940009
gazed'	1	26	0	1.414973348
as'	1	27	0	1.431363764
hands'	1	28	0	1.447158031
pockets'	1	29	0	1.462397998
looking'	1	30	0	1.477121255
out'	1	31	0	1.491361694
He'	1	32	0	1.505149978
fixed'	1	33	0	1.51851394
distorted'	1	34	0	1.531478917
fir'	1	35	0	1.544068044
tree'	1	36	0	1.556302501
planted'	1	37	0	1.568201724
by'	1	38	0	1.579783597
some'	1	39	0	1.591064607
former'	1	40	0	1.602059991
tenant'	1	41	0	1.612783857
tub'	1	42	0	1.62324929
once'	1	43	0	1.633468456
been'	1	44	0	1.643452676
green'	1	45	0	1.653212514
there'	1	46	0	1.662757832
years'	1	47	0	1.672097858
before'	1	48	0	1.681241237
rot'	1	49	0	1.69019608
away'	1	50	0	1.698970004
piecemeal'	1	51	0	1.707570176
There'	1	52	0	1.716003344
nothing'	1	53	0	1.72427587
very'	1	54	0	1.73239376

inviting'	1	55	0	1.740362689
object'	1	56	0	1.748188027
but'	1	57	0	1.755874856
wrapt'	1	58	0	1.763427994
brown'	1	59	0	1.770852012
study'	1	60	0	1.77815125
contemplating'	1	61	0	1.785329835
it'	1	62	0	1.792391689
far'	1	63	0	1.799340549
greater'	1	64	0	1.806179974
attention'	1	65	0	1.812913357
than'	1	66	0	1.819543936
more'	1	67	0	1.826074803
conscious'	1	68	0	1.832508913
mood'	1	69	0	1.838849091
would'	1	70	0	1.84509804
have'	1	71	0	1.851258349
deigned'	1	72	0	1.857332496
bestow'	1	73	0	1.86332286
rarest'	1	74	0	1.86923172
exotic'	1	75	0	1.875061263
At'	1	76	0	1.880813592
length'	1	77	0	1.886490725
wandered'	1	78	0	1.892094603
little'	1	79	0	1.897627091
dirty'	1	80	0	1.903089987
on'	1	81	0	1.908485019
through'	1	82	0	1.913813852
which'	1	83	0	1.919078092
face'	1	84	0	1.924279286
clerk'	1	85	0	1.929418926
dimly'	1	86	0	1.934498451
visible'	1	87	0	1.939519253
worthy'	1	88	0	1.944482672
chancing'	1	89	0	1.949390007

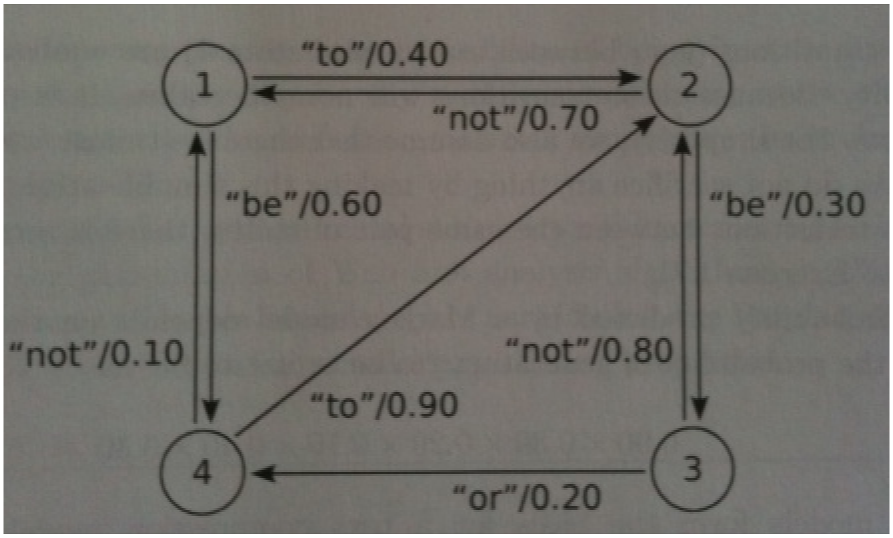
look'	1	90	0	1.954242509
up'	1	91	0	1.959041392
beckoned'	1	92	0	1.963787827
him'	1	93	0	1.968482949
attend'	1	94	0	1.973127854



So we can conclude from the above graph that $F_i = 1/i^{\alpha}$, hence proving Zipf's law

As the sample size is small, we don't see the smooth graph but as we can see from a frequency distribution, it is inversely proportional to rank.

Part 1.3 -> Hidden Markov Model example diagram from class, state to reach phrase of length 3



(a) All the 3 length phrases that can be computed from the Markov’s models are:
(Assuming we are starting from state 1)

1. “to be or” = 1-> 2 -> 3 -> 4 = 0.40 *0.30 * 0.20 = 0.024
2. “be to be” = 1 -> 4 -> 2-> 3 = 0.60 * 0.90 * 0.30 = 0.162
3. “be to not” = 1 -> 4 -> 2 -> 1 = 0.60 * 0.90 * 0.70 = 0.378
4. “to not to” = 1-> 2-> 1-> 2 = 0.40 * 0.70 * 0.40 = 0.112
5. “to be not” = 1->2-> 3->2 = 0.40 * 0.30 * 0.80 = 0.096
6. “be not be” = 1-> 4 -> 1 -> 4 = 0.60 * 0.10 * 0.60 =0.036
7. “to not be” = 1-> 2 -> 1 -> 4 = 0.40 * 0.70 *0.60 = 0.168
8. “be not to” = 1-> 4 -> 1 -> 2 = 0.60 * 0.10 * 0.40 = 0.024

(b) To reach to the final probability using matrix multiplication, we compute the multiplication 3 times as we need phrase of length 3.

Transition matrix as per example in slide:

0.00 0.40 0.00 0.60
0.70 0.00 0.30 0.00
0.00 0.80 0.00 0.20
0.10 0.90 0.00 0.00

State 1: (1 0 0 0)

On multiplying the transition matrix with state 1 : we get

$$(1\ 0\ 0\ 0) \quad * \quad \begin{matrix} 0.00 & 0.40 & 0.00 & 0.60 \\ 0.70 & 0.00 & 0.30 & 0.00 \\ 0.00 & 0.80 & 0.00 & 0.20 \\ 0.10 & 0.90 & 0.00 & 0.00 \end{matrix} \quad = (0\ 0.4\ 0\ 0.6)$$

Meaning the probability from state 1 to:
State 2 is 0.4
State 3 is 0
state 4 is 0.6

Multiplying again :

$$(0\ 0.4\ 0\ 0.6) \quad * \quad \begin{matrix} 0.00 & 0.40 & 0.00 & 0.60 \\ 0.70 & 0.00 & 0.30 & 0.00 \\ 0.00 & 0.80 & 0.00 & 0.20 \\ 0.10 & 0.90 & 0.00 & 0.00 \end{matrix} \quad = (0.34\ 0.54\ 0.12\ 0.00)$$

Meaning the probability to reach from state 1 to:
state 1 is 0.34
state 2 is 0.54
State 3 is 0.12
State 4 is 0.0

$$(0.34\ 0.54\ 0.12\ 0.00) \quad * \quad \begin{matrix} 0.00 & 0.40 & 0.00 & 0.60 \\ 0.70 & 0.00 & 0.30 & 0.00 \\ 0.00 & 0.80 & 0.00 & 0.20 \\ 0.10 & 0.90 & 0.00 & 0.00 \end{matrix} \quad = (0.378\ 0.232\ 0.162\ 0.228)$$

Meaning the probability to reach from state 1 to :
state 1 is 0.378
state 2 is 0.232
State 3 is 0.162
State 4 is 0.228

To reach the state of 1 from 1 after 3 multiplication :0.378
To reach the state of 2 from 1 after 3 multiplication : 0.232 (0.112 + 0.096 + 0.024)
To reach the state of 3 from 1 after 3 multiplication : 0.162
To reach the state of 4 from 1 after 3 multiplication : 0.228 (0.024+ 0.036+ 0.168)