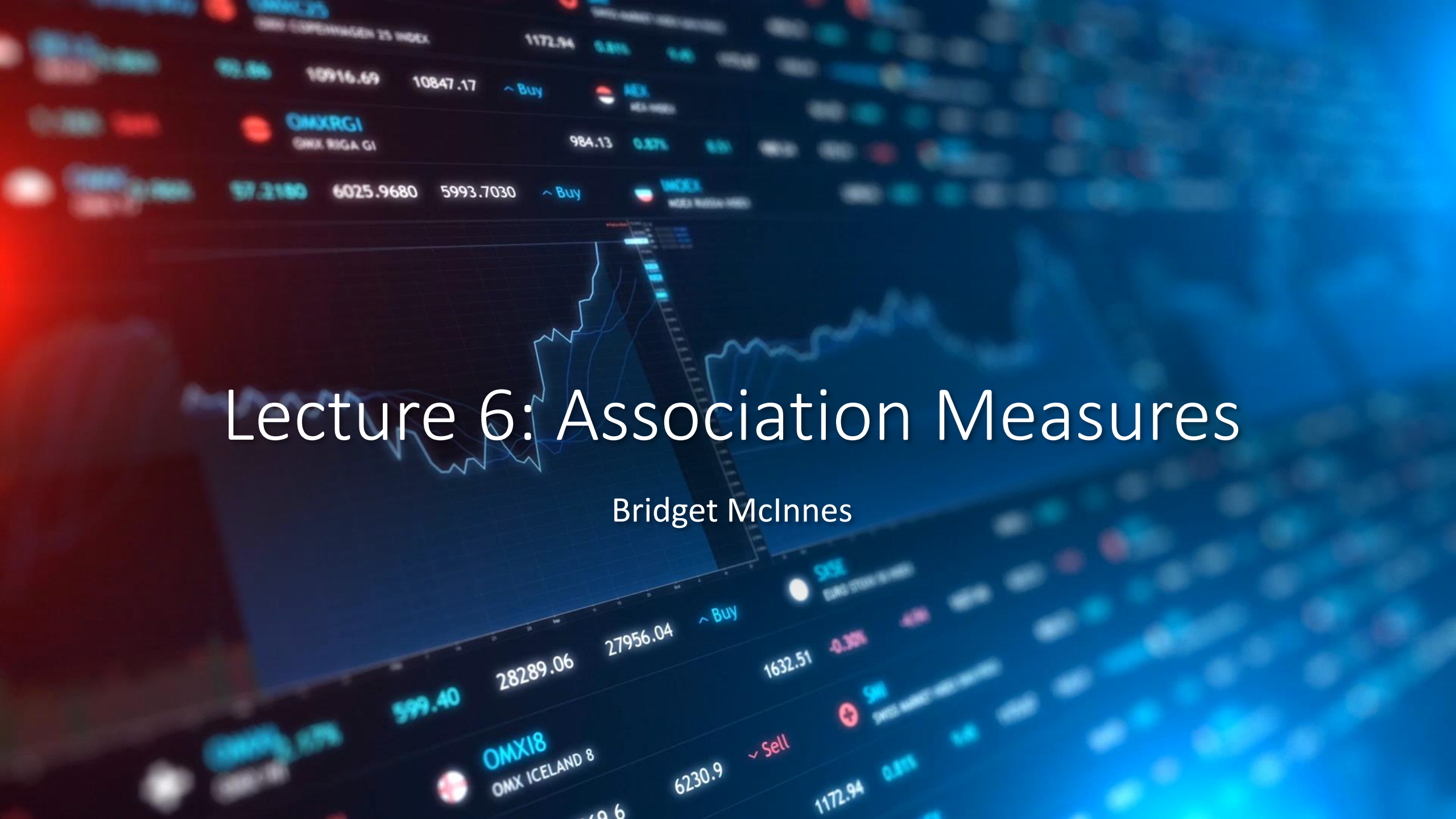
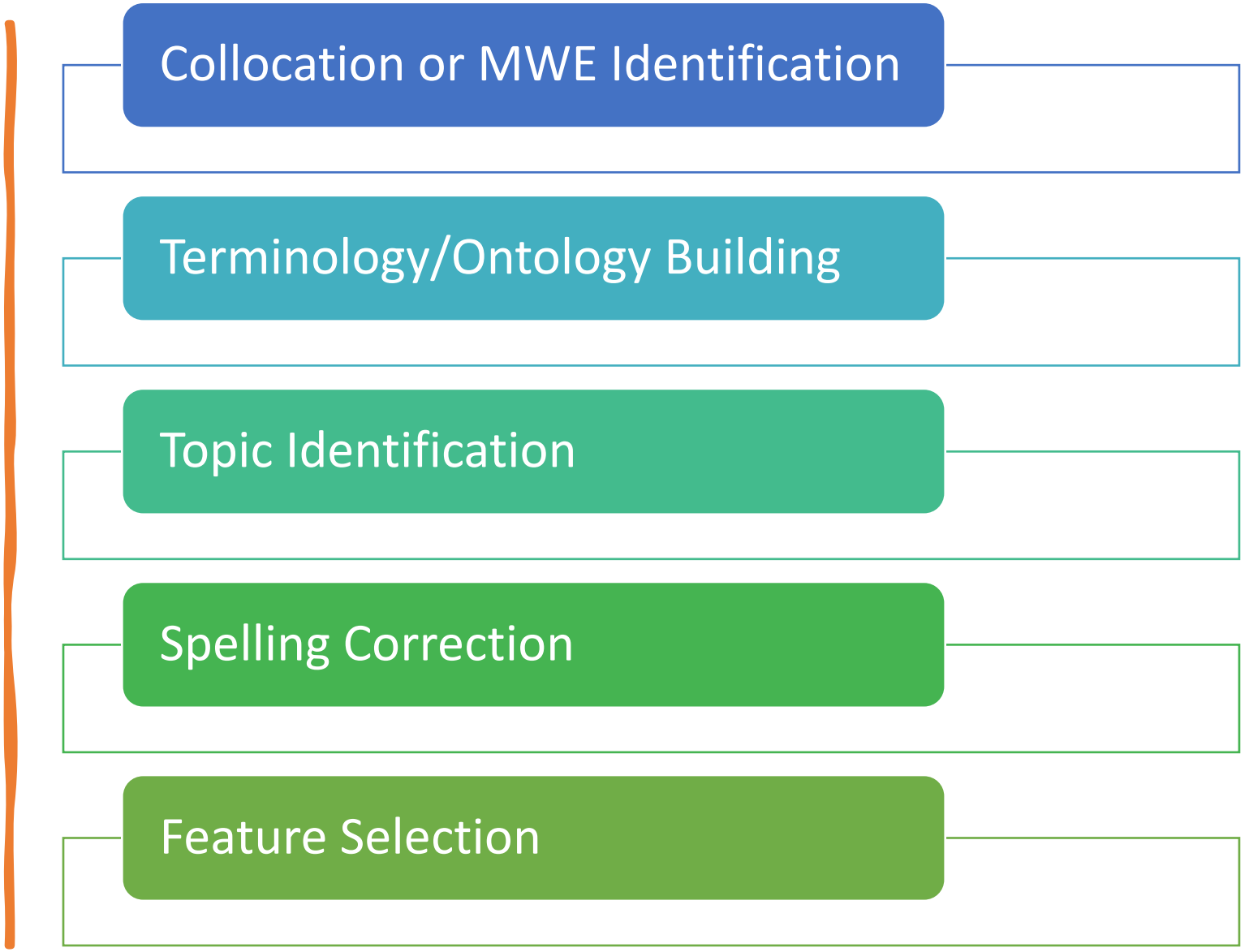


Lecture 6: Association Measures

Bridget McInnes

Hit record

Applications using Ngrams



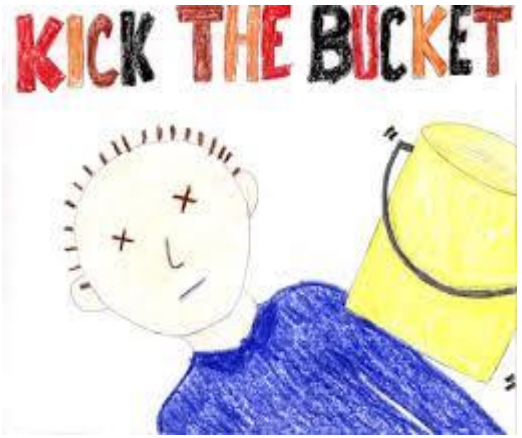
Collocations / Multi-word Expressions

Post Office

Nuclear family

spider crab

Collocations: Idioms

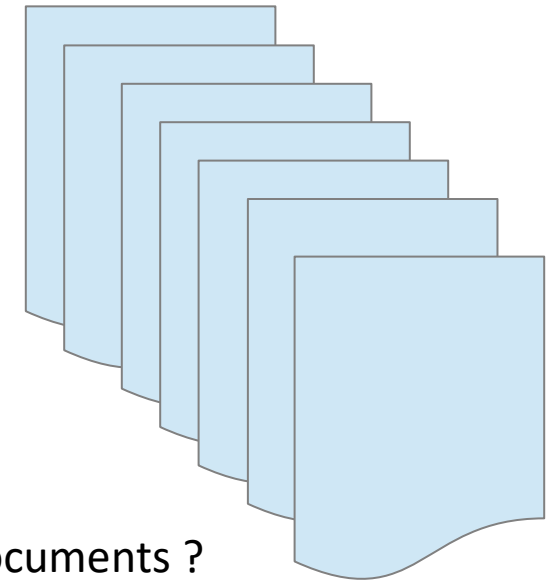


Kick the bucket

Pain in the neck

Black and blue

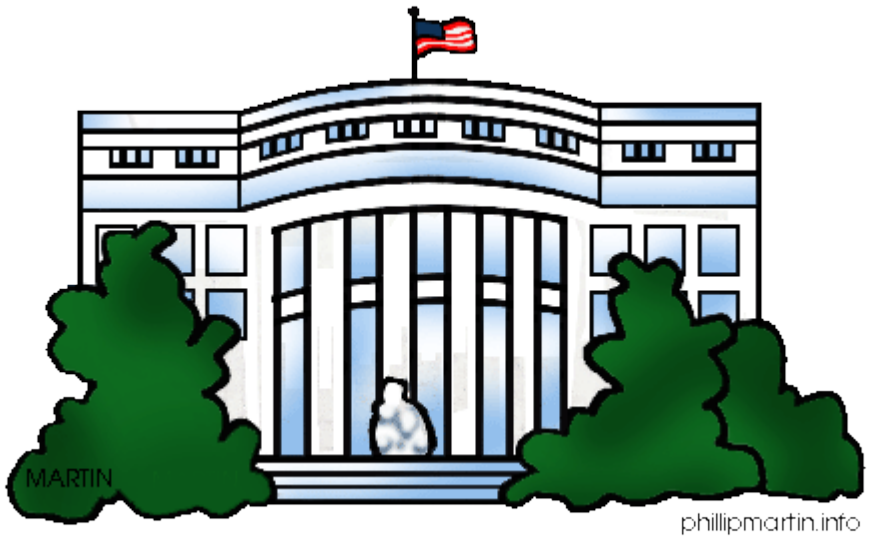
Collocations



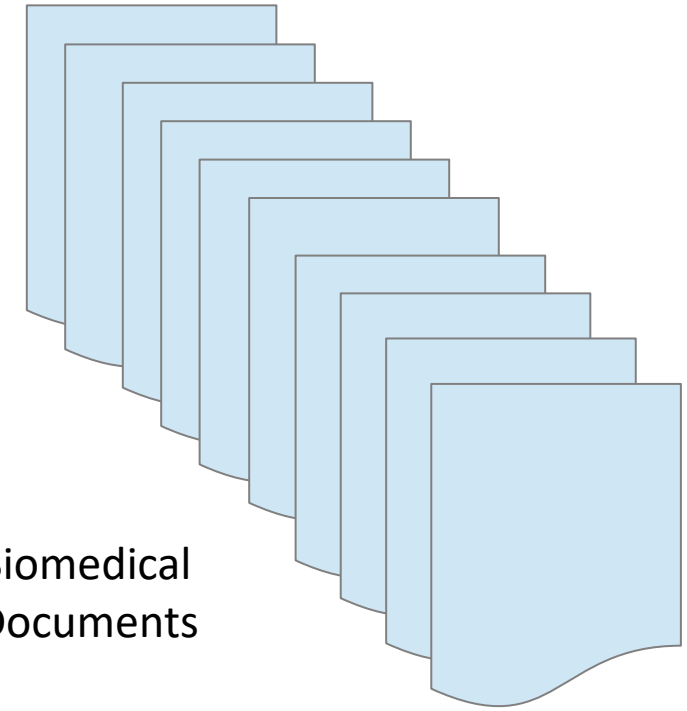
Documents ?
1. press releases
2. real estate notices

We walked past the white house.

Looking at the distribution of ngrams in the corpus can help us determine which is being referred to.

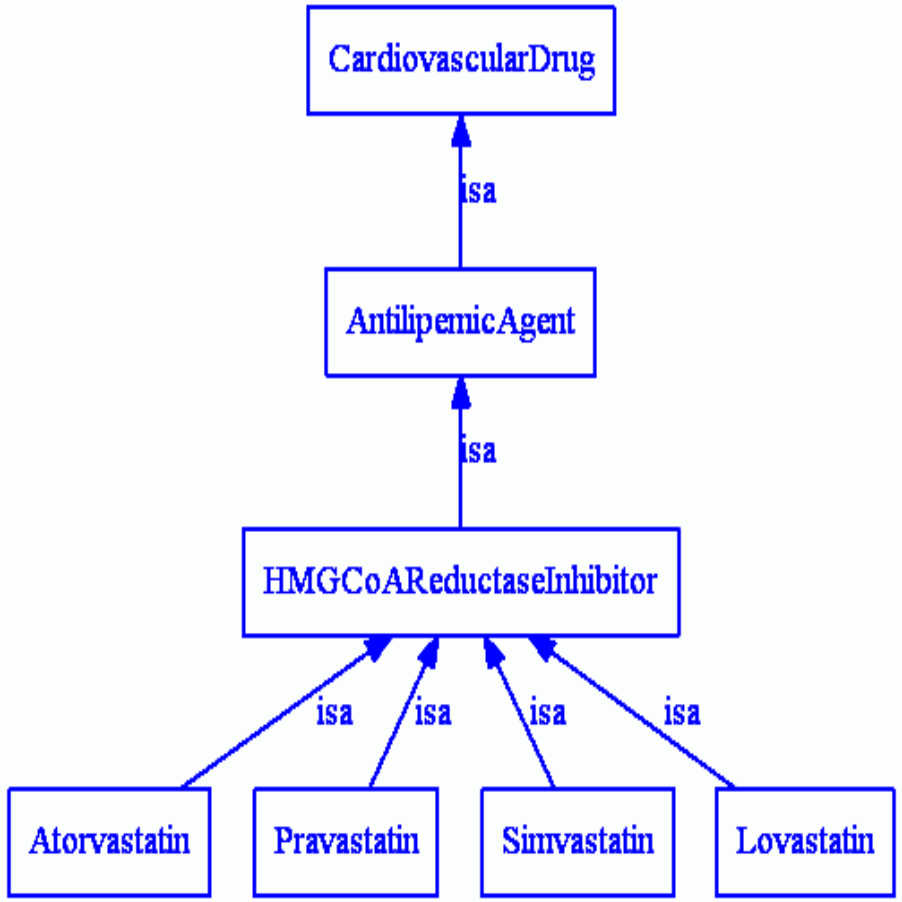


Terminology/Ontology Building

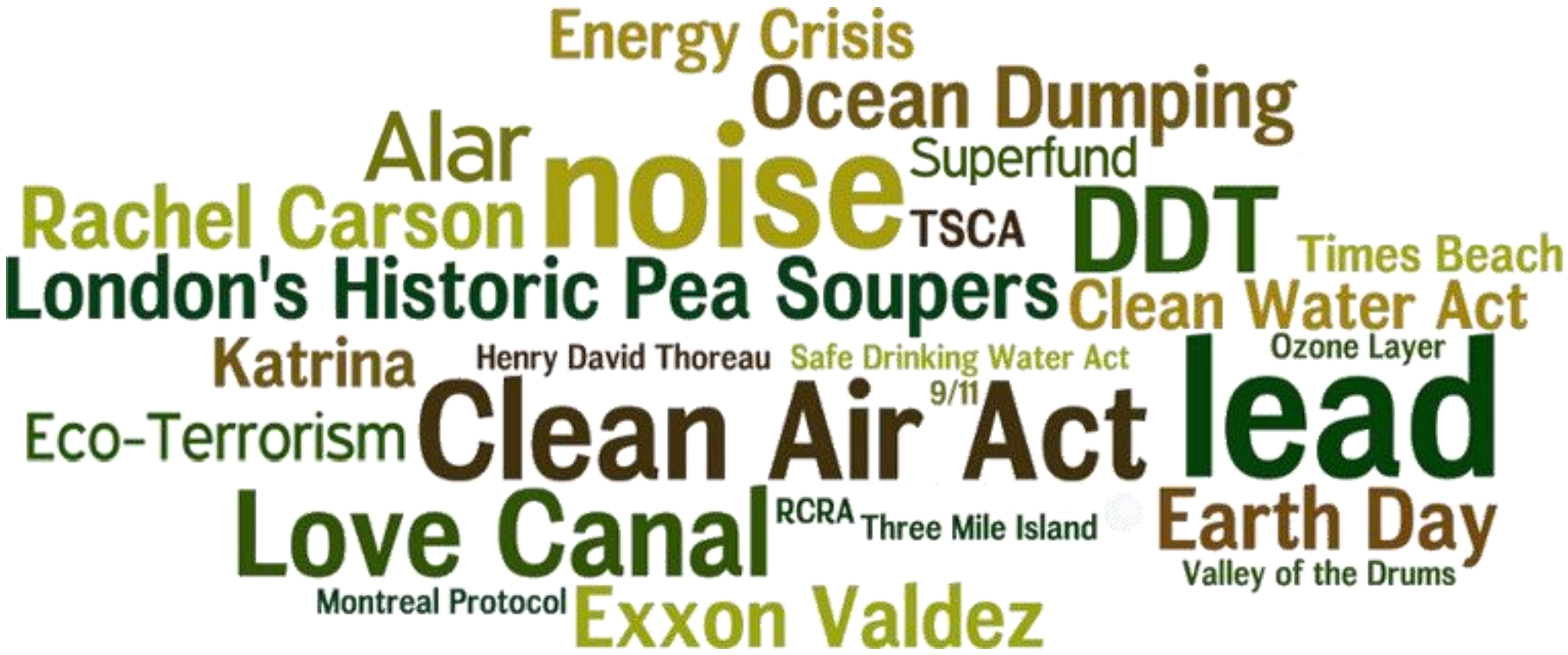


Biomedical
Documents

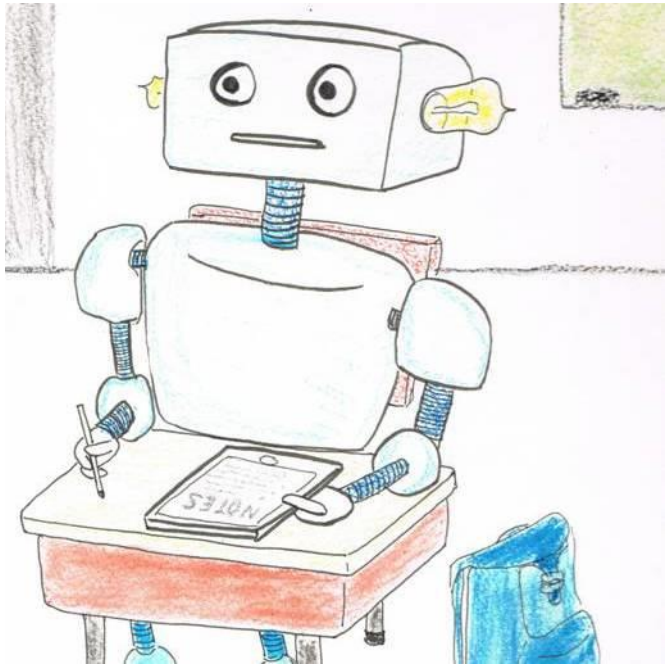
Prospective
Terms



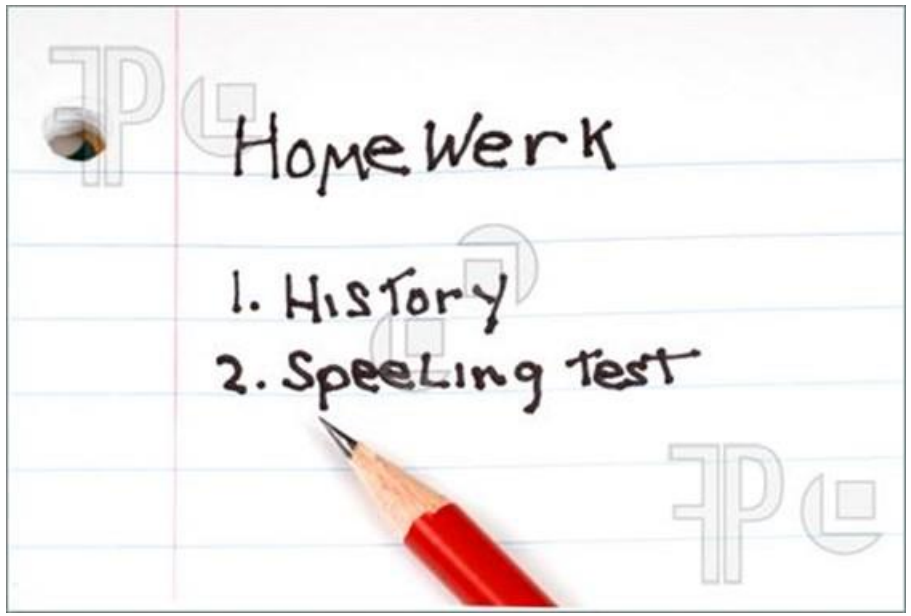
Topic Identification



Feature Selection



Spelling Correction

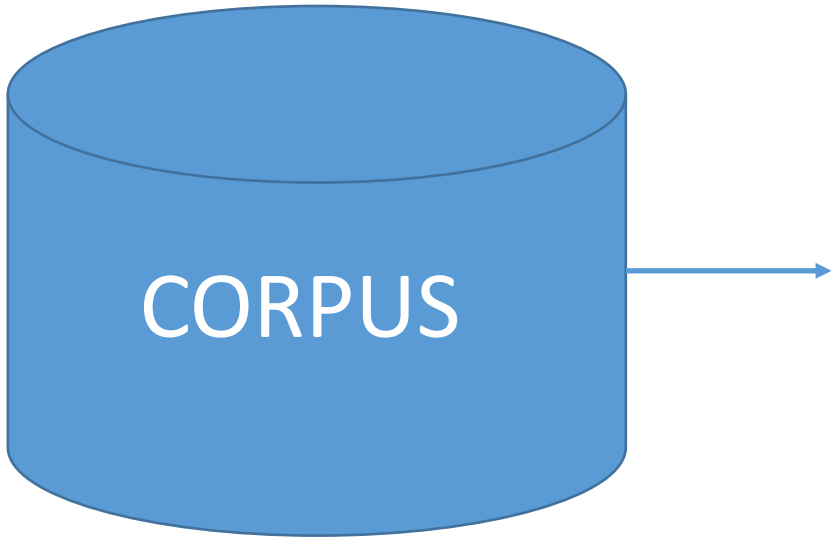


A lot of these tasks

Use the probability of an n-gram occurring
in the text:

we already know how to rank n-grams
based on their relative frequency

Ngram Probability – last lecture



	i	want	to	eat	chinese	food	lunch	spend	<start>	<end>
i	0.002	0.33	0	0.0036	0	0	0	0.00079	0	0
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0064	0.0011	0	0
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087	0	0
eat	0	0	0.0027	0	0.021	0.0027	0.056	0	0	0.011
chinese	0.0063	0	0	0	0	0.52	0.0063	0	0	0.008
food	0.014	0	0.014	0	0.00092	0.0037	0	0	0	0.004
lunch	0.0059	0	0	0	0	0.0029	0	0	0	0.003
spend	0.0036	0	0.0036	0	0	0	0	0	1	0.006
<start>	0.015	0	0.01	0	0.005	0.003	0.001	0	0	0
<end>	0	0	0	0	0.001	0.007	0.002	0.011	0	0

Relative Frequency Table

Another way to look at this

Likelihood the tokens in the n-gram tend
to occur together
more often than one would expect by chance

Ngram Statistics

Ngram statistics

Statistical measures are computed using various

co-occurrence

and

individual frequency counts

Contingency Table

Table 3: Contingency Table for Bigrams

	token2	¬ token2	Totals
token1	n_{11}	n_{12}	n_{1p}
¬ token1	n_{21}	n_{22}	n_{2p}
Totals	n_{p1}	n_{p2}	n_{pp}

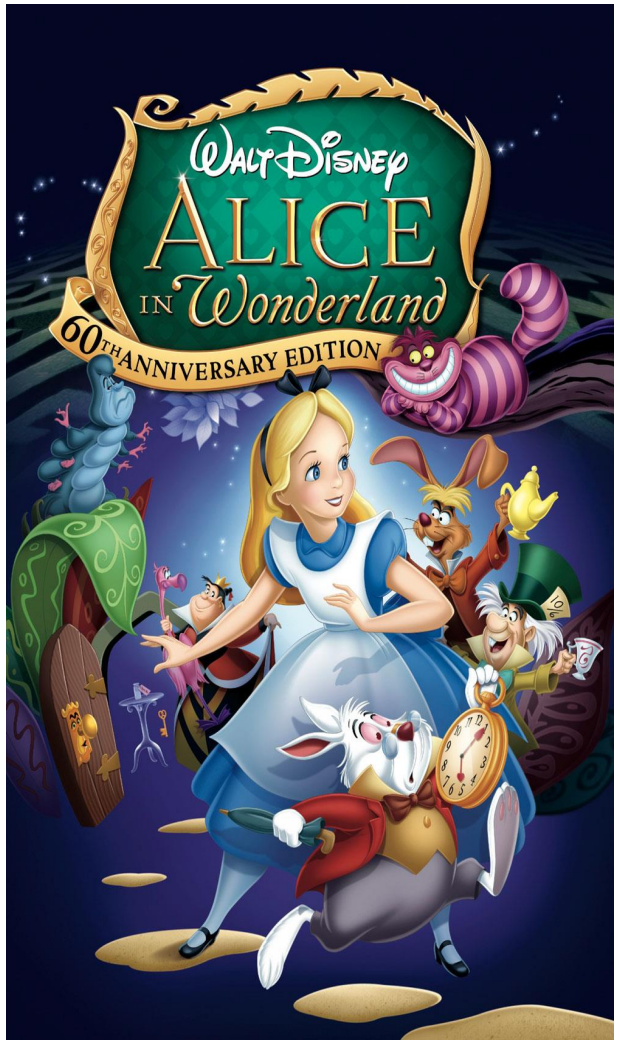
Contingency Table

Table 3: Contingency Table for Bigrams

	token2	¬ token2	Totals
token1	n_{11}	n_{12}	n_{1p}
¬ token1	n_{21}	n_{22}	n_{2p}
Totals	n_{p1}	n_{p2}	n_{pp}

$n_{11} = the\ number\ of\ times\ the\ bigram\ occurs$
$n_{12} = the\ number\ of\ token\ 1\ occurs\ without\ token\ 2$
$n_{21} = the\ number\ of\ token\ 2\ occurs\ without\ token\ 1$
$n_{22} = the\ number\ of\ times\ neither\ tokens\ are\ in\ the\ bigram$
$n_{p1} = the\ number\ of\ times\ token\ 2\ occurs$
$n_{p2} = the\ number\ of\ times\ token\ 2\ does\ not\ occur$
$n_{1p} = the\ number\ of\ times\ token\ 1\ occurs$
$n_{2p} = the\ number\ of\ times\ token\ 1\ does\ not\ occur$
$n_{2p} = the\ number\ of\ times\ token\ 1\ does\ not\ occur$
$n_{pp} = the\ total\ number\ of\ bigrams$

Example: white rabbit



	$\boldsymbol{rabbit}$	$\boldsymbol{\neg rabbit}$	
$\boldsymbol{white}$	$n_{11} = 21$	$n_{12} = 5$	$n_{1p} = 26$
$\boldsymbol{\neg white}$	$n_{21} = 2$	$n_{22} = 3757$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

Okay so we have our frequency counts …

Now what?

The next step is:

to approximate the expected value
of the n-gram if it
occurred by chance in the corpus

Expected Values

Table 5: Contingency Table for Expected Values

	token2	¬ tokens2	Totals
token1	$m_{11} = \frac{n_{1p} * n_{p1}}{n_{pp}}$	$m_{12} = \frac{n_{1p} * n_{p2}}{n_{pp}}$	n_{1p}
¬ token1	$m_{21} = \frac{np1 * n_{2p}}{n_{pp}}$	$m_{22} = \frac{n_{p2} * n_{2p}}{n_{pp}}$	n_{2p}
Totals	n_{p1}	n_{p2}	n_{pp}

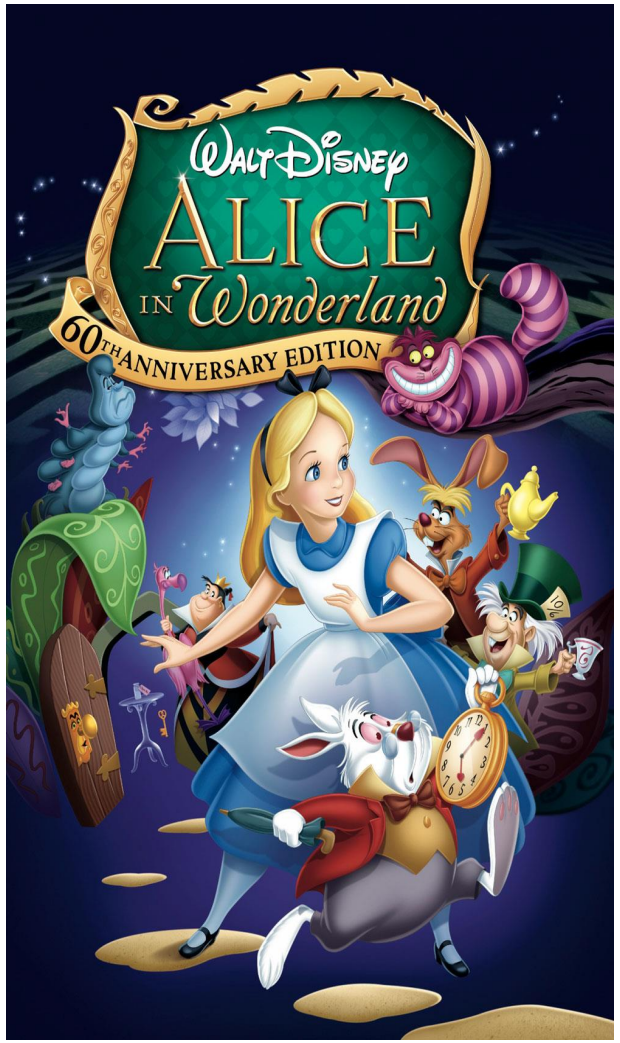
Expected Values

Table 5: Contingency Table for Expected Values

	token2	¬ tokens2	Totals
token1	$m_{11} = \frac{n_{1p} * n_{p1}}{n_{pp}}$	$m_{12} = \frac{n_{1p} * n_{p2}}{n_{pp}}$	n_{1p}
¬ token1	$m_{21} = \frac{np1 * n_{2p}}{n_{pp}}$	$m_{22} = \frac{n_{p2} * n_{2p}}{n_{pp}}$	n_{2p}
Totals	n_{p1}	n_{p2}	n_{pp}

	rabbit	***¬rabbit***	
white	$n_{11} = 21$	$n_{12} = 5$	$n_{1p} = 26$
¬white	$n_{21} = 2$	$n_{22} = 3757$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

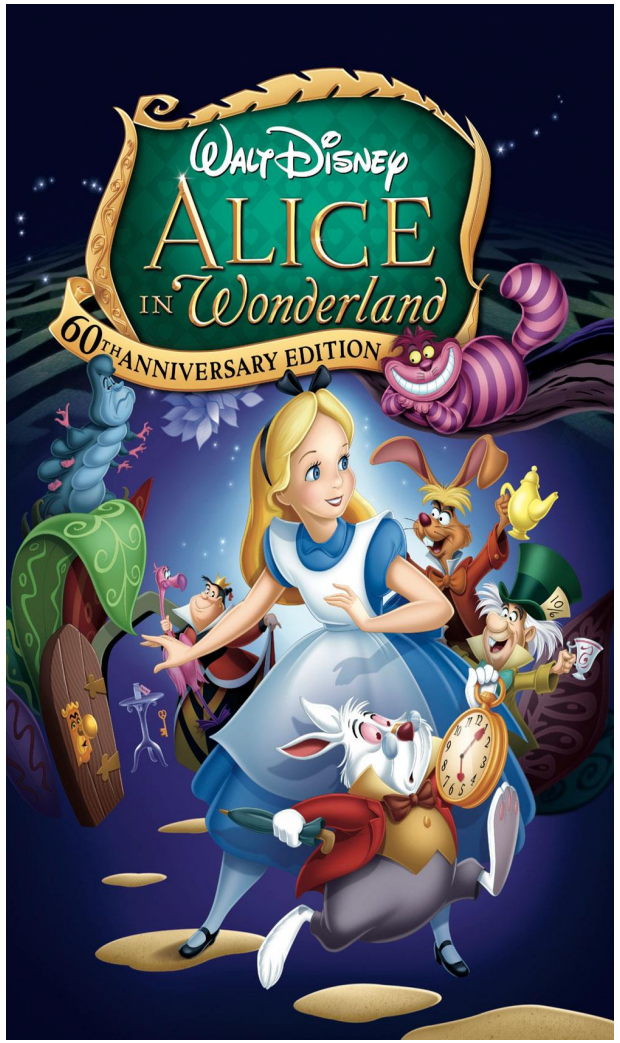
Expected values



Seen values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$n_{11} = 21$	$n_{12} = 5$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$n_{21} = 2$	$n_{22} = 3757$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

Expected values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$m_{11} =?$	$m_{12} =?$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$m_{21} =?$	$m_{22} =?$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

Expected values

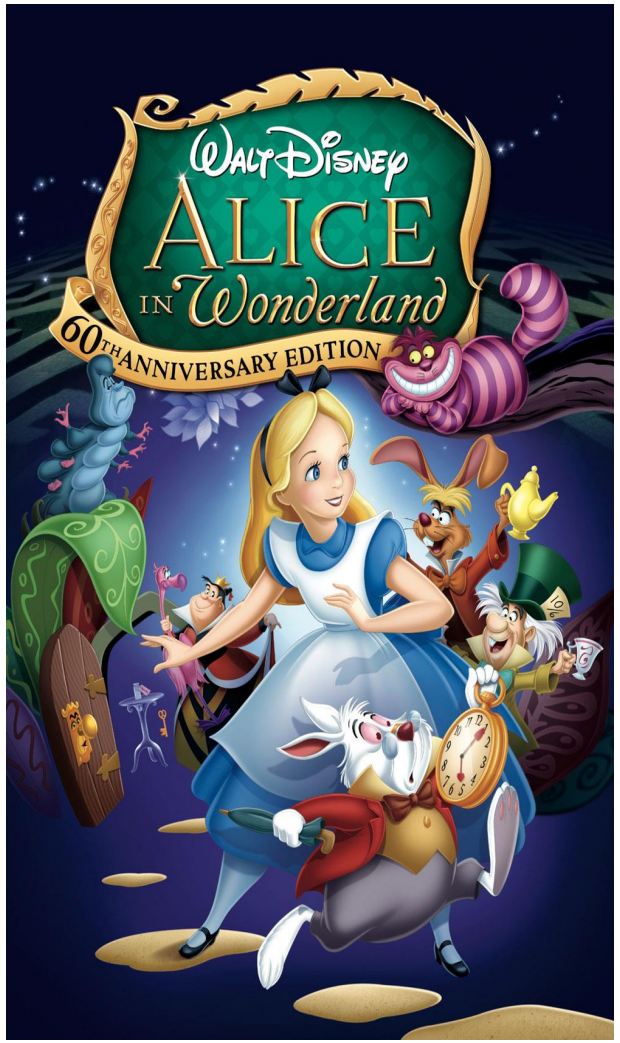


Seen values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$n_{11} = 21$	$n_{12} = 5$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$n_{21} = 2$	$n_{22} = 3757$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

Expected values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$m_{11} =?$	$m_{12} =?$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$m_{21} =?$	$m_{22} =?$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

$$m_{11} = \frac{n_{1p} * n_{p1}}{n_{pp}} = \frac{23 * 26}{3785} = 0.1579$$

Expected values

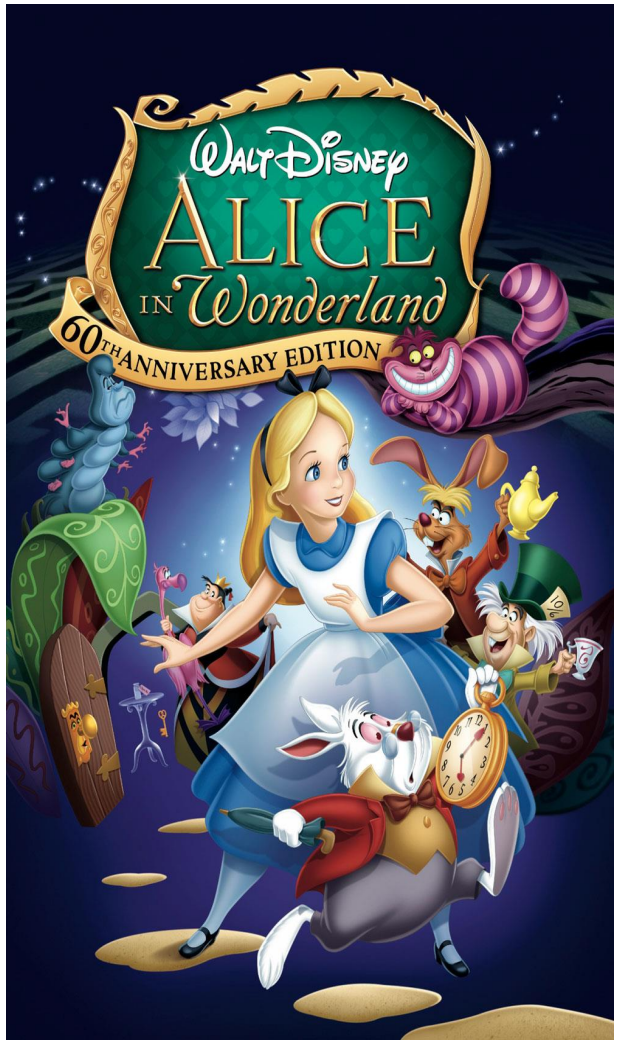


Seen values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$n_{11} = 21$	$n_{12} = 5$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$n_{21} = 2$	$n_{22} = 3757$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

Expected values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$m_{11} = 0.1579$	$m_{12} =?$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$m_{21} =?$	$m_{22} =?$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

$$m_{11} = \frac{n_{1p} * n_{p1}}{n_{pp}} = \frac{23 * 26}{3785} = 0.1579$$

Expected values

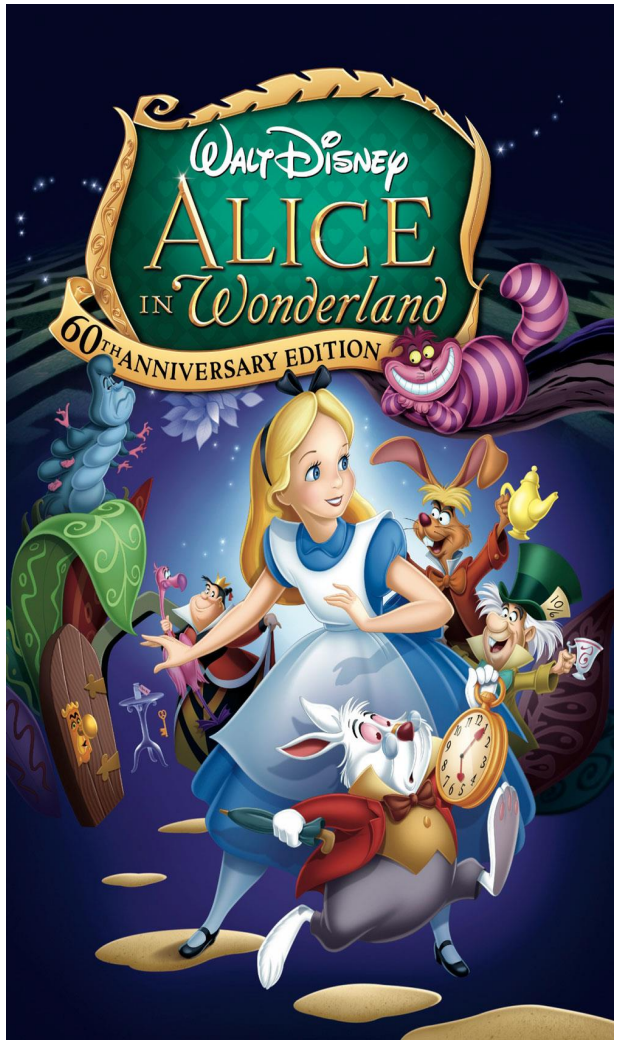


Seen values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$n_{11} = 21$	$n_{12} = 5$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$n_{21} = 2$	$n_{22} = 3757$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

Expected values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$m_{11} = 0.1579$	$m_{12} =?$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$m_{21} =?$	$m_{22} =?$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

$$m_{12} = \frac{n_{1p} * n_{p2}}{n_{pp}} = \frac{23 * 3762}{3785} = 22.8602$$

Expected values

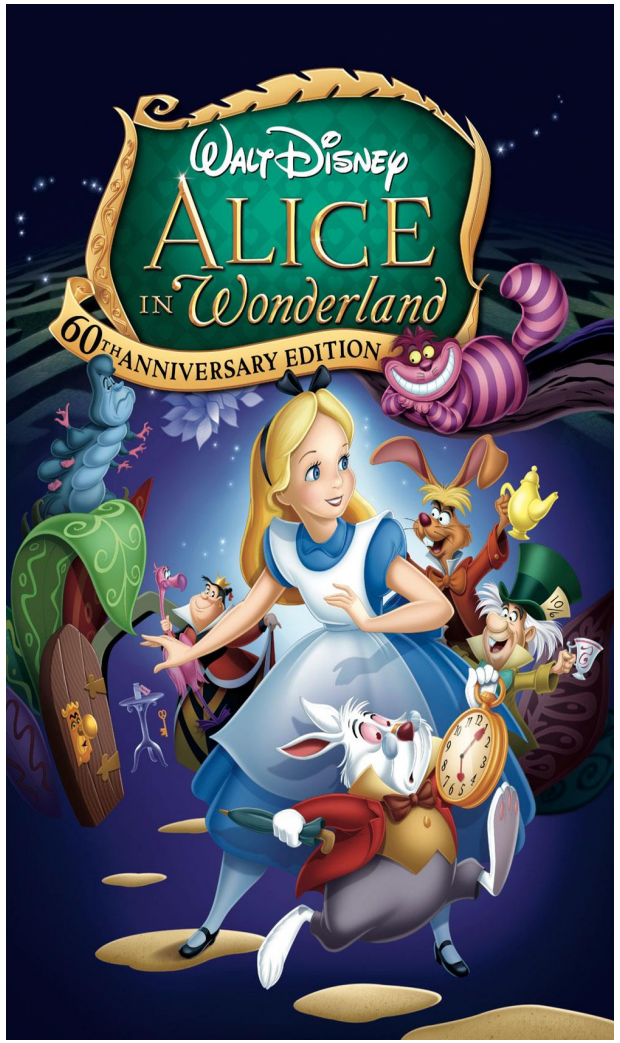


Seen values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$n_{11} = 21$	$n_{12} = 5$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$n_{21} = 2$	$n_{22} = 3757$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

Expected values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$m_{11} = 0.1579$	$m_{12} = 25.8602$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$m_{21} =?$	$m_{22} =?$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

$$m_{12} = \frac{n_{1p} * n_{p2}}{n_{pp}} = \frac{23 * 3762}{3785} = 25.8602$$

Expected values

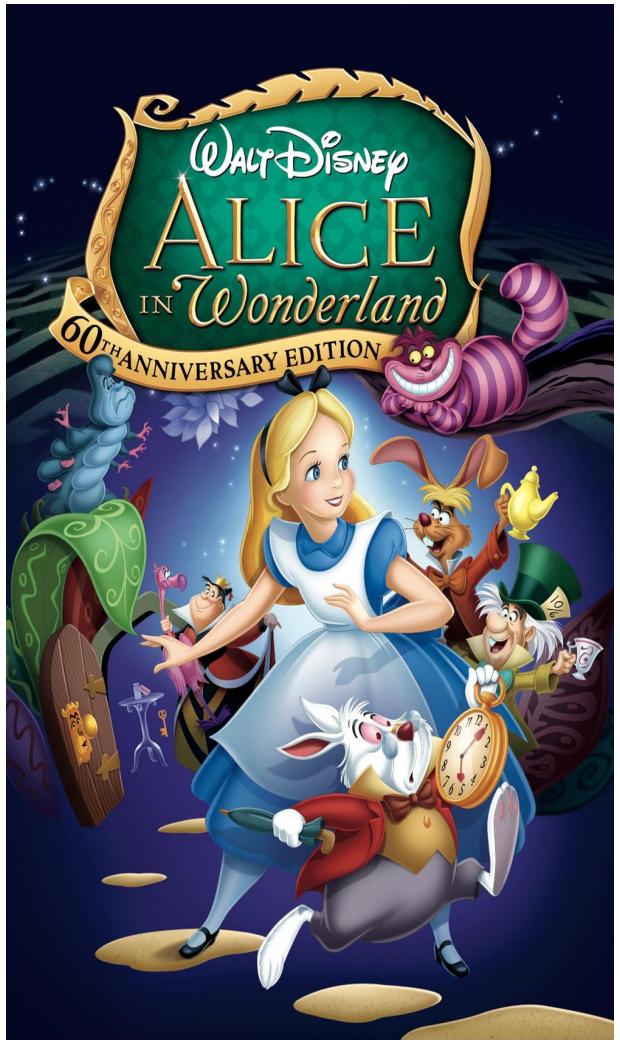


Seen values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$n_{11} = 21$	$n_{12} = 5$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$n_{21} = 2$	$n_{22} = 3757$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

Expected values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$m_{11} = 0.1579$	$m_{12} = 25.8602$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$m_{21} =?$	$m_{22} =?$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

$$m_{21} = \frac{n_{2p} * n_{p1}}{n_{pp}} = \frac{3759 * 23}{3785} = 25.8420$$

Expected values

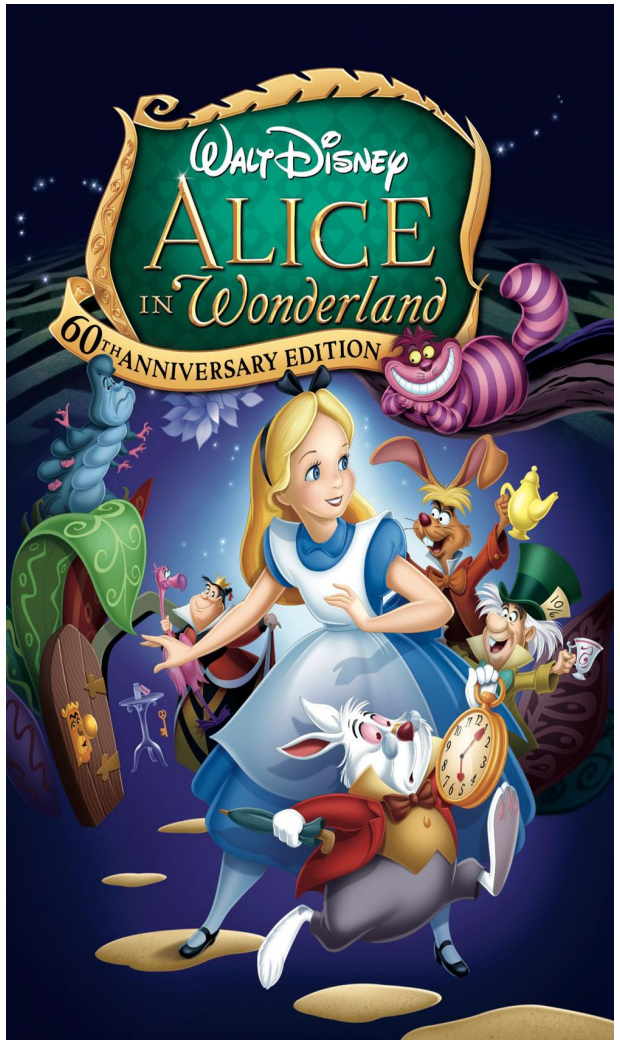


Seen values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$n_{11} = 21$	$n_{12} = 5$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$n_{21} = 2$	$n_{22} = 3757$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

Expected values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$m_{11} = 0.1579$	$m_{12} = 22.8602$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$m_{21} = 22.8420$	$m_{22} =?$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

$$m_{21} = \frac{n_{2p} * n_{p1}}{n_{pp}} = \frac{3759 * 23}{3785} = 22.8420$$

Expected values

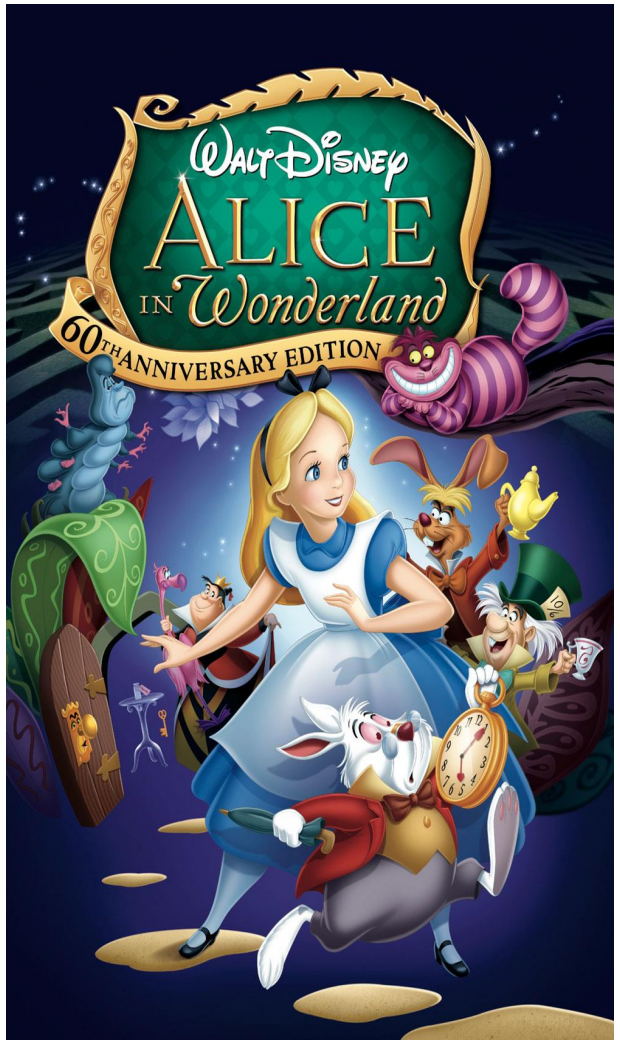


Seen values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$n_{11} = 21$	$n_{12} = 5$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$n_{21} = 2$	$n_{22} = 3757$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

Expected values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$m_{11} = 0.1579$	$m_{12} = 22.8602$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$m_{21} = 22.8420$	$m_{22} =?$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

$$m_{22} = \frac{n_{2p} * n_{p2}}{n_{pp}} = \frac{3759 * 3762}{3785} = 3736.1579$$

Expected values



Seen values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$n_{11} = 21$	$n_{12} = 5$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$n_{21} = 2$	$n_{22} = 3757$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

Expected values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$m_{11} = 0.1579$	$m_{12} = 22.8602$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$m_{21} = 22.8420$	$m_{22} = 3736.1579$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

$$m_{22} = \frac{n_{2p} * n_{p2}}{n_{pp}} = \frac{3759 * 3762}{3785} = 3736.1579$$

Likelihood

Now we have our known values and we have our expected values

Next: we can calculate the likelihood the tokens in the n-gram have occurred together by chance

Measures of Association

Number of measures of association

- Mutual Information
- Log Likelihood
- Chi Squared
- Dice co-efficient

…. the list goes on

Log Likelihood Ratio (G^2)

$$G^2 = 2 * \sum_{i}^{j} n_{ij} * \log(\frac{n_{ij}}{m_{ij}})$$

Basic idea

We are taking the log ratio of our known values over our expected values

Seen values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$n_{11} = 21$	$n_{12} = 5$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$n_{21} = 2$	$n_{22} = 3757$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

Expected values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$m_{11} = 0.1579$	$m_{12} = 22.8602$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$m_{21} = 22.8420$	$m_{22} =?$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

$$G^2 = 2 * \sum_{i}^{j} n_{ij} * \log(\frac{n_{ij}}{m_{ij}})$$

i, j = { 1, 2}

We have:

Seen values

Expected values

Seen values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$n_{11} = 21$	$n_{12} = 5$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$n_{21} = 2$	$n_{22} = 3757$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

Expected values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$m_{11} = 0.1579$	$m_{12} = 22.8602$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$m_{21} = 22.8420$	$m_{22} = ?$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

$$G^2 = 2 * \sum_{i}^{j} n_{ij} * \log(\frac{n_{ij}}{m_{ij}})$$

i, j = { 1, 2}

We have:

Seen values

Expected values

$$G^2 = 2 * \left(\left(n_{11} * \log\left(\frac{n_{11}}{m_{11}}\right)\right) + \left(n_{12} * \log\left(\frac{n_{12}}{m_{12}}\right)\right) + \left(n_{21} * \log\left(\frac{n_{21}}{m_{21}}\right)\right) + \left(n_{22} * \log\left(\frac{n_{22}}{m_{22}}\right)\right) \right)$$

Seen values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$n_{11} = 21$	$n_{12} = 5$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$n_{21} = 2$	$n_{22} = 3757$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

Expected values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$m_{11} = 0.1579$	$m_{12} = 22.8602$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$m_{21} = 22.8420$	$m_{22} = ?$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

$$G^2 = 2 * \sum_{i}^{j} n_{ij} * \log(\frac{n_{ij}}{m_{ij}})$$

i, j = { 1, 2}

We have:

Seen values

Expected values

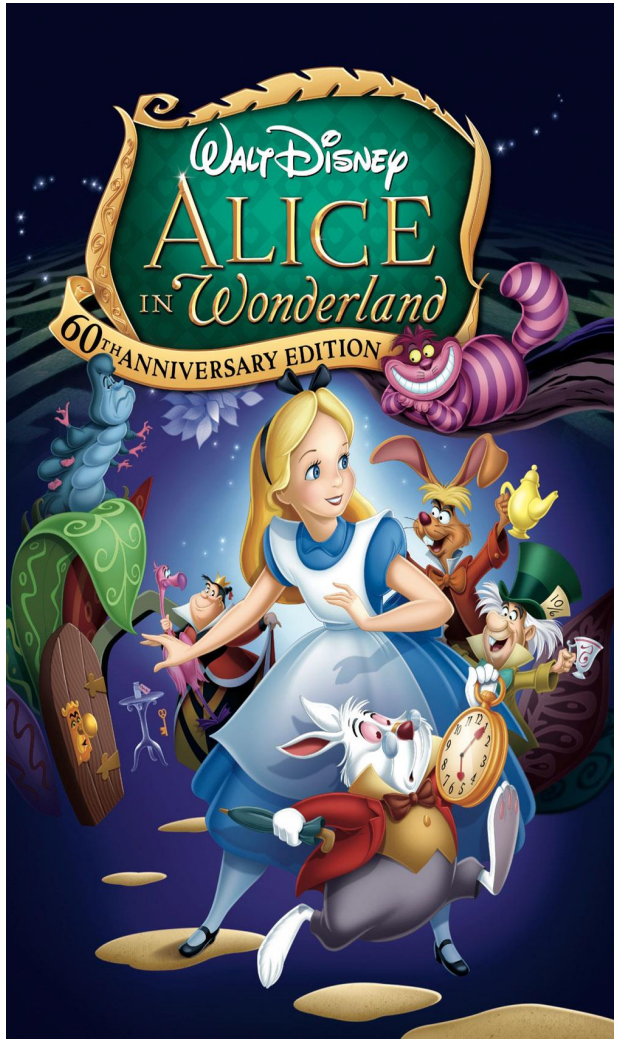
$$G^2 = 2 * \left(\left(n_{11} * \log\left(\frac{n_{11}}{m_{11}}\right)\right) + \left(n_{12} * \log\left(\frac{n_{12}}{m_{12}}\right)\right) + \left(n_{21} * \log\left(\frac{n_{21}}{m_{21}}\right)\right) + \left(n_{22} * \log\left(\frac{n_{22}}{m_{22}}\right)\right) \right)$$

$$G^2 = 2 * \left(\left(21 * \log\left(\frac{21}{0.1579}\right)\right) + \left(5 * \log\left(\frac{5}{22.8602}\right)\right) + \left(2 * \log\left(\frac{2}{22.8420}\right)\right) + \left(3767 * \log\left(\frac{3757}{3762}\right)\right) \right) = 221.0014$$

```
bridget@caterpillar: ~/cla
File  Edit  View  Search  Terminal  Help
3785
mock<>turtle<>59 59 61
march<>hare<>31 31 31
thought<>alice<>26 30 82
white<>rabbit<>21 26 23
alice<>thought<>12 143 13
poor<>alice<>11 28 82
my<>dear<>11 54 20
sat<>down<>9 14 85
alice<>did<>9 143 22
poor<>little<>9 28 30
alice<>replied<>9 143 27
any<>rate<>8 25 8
tell<>me<>8 11 45
let<>me<>8 10 45
alice<>looked<>8 143 14
beautiful<>soup<>8 10 10
little<>thing<>8 113 36
same<>thing<>7 18 36
soo<>oop<>7 7 7
looked<>down<>7 16 85
alice<>began<>7 143 22
cried<>alice<>7 7 82
oh<>dear<>7 10 20
few<>minutes<>6 9 7
found<>herself<>6 9 33
down<>here<>6 23 19
three<>gardeners<>6 21 7
great<>hurry<>6 34 6
golden<>key<>6 8 7
little<>door<>6 113 8
mary<>ann<>6 6 6
cheshire<>cat<>5 7 8
another<>moment<>5 21 8
should<>think<>5 21 10
came<>upon<>5 21 27
should<>like<>5 21 33
alice<>went<>5 143 29
little<>golden<>5 113 7
white<>kid<>5 26 5
yer<>honour<>5 5 5
right<>size<>5 17 10
play<>croquet<>5 5 6
next<>witness<>5 18 6
kid<>gloves<>5 5 6
trembling<>voice<>5 6 27
alice<>felt<>5 143 5
feet<>high<>5 5 14
--More--(1%)
```

```
bridget@caterpillar: ~/class
File  Edit  View  Search  Terminal  Help
3785
mock<>turtle<>1 590.5000 59 59 61
march<>hare<>2 359.6439 31 31 31
white<>rabbit<>3 221.0014 21 26 23
thought<>alice<>4 184.9358 26 30 82
soo<>oop<>5 102.0875 7 7 7
beautiful<>soup<>6 94.5189 8 10 10
mary<>ann<>7 89.3550 6 6 6
any<>rate<>8 83.1894 8 25 8
yer<>honour<>9 76.2870 5 5 5
golden<>key<>10 74.6171 6 8 7
alice<>thought<>11 72.6410 12 143 13
few<>minutes<>12 72.1577 6 9 7
play<>croquet<>13 70.8803 5 5 6
kid<>gloves<>13 70.8803 5 5 6
my<>dear<>14 68.5555 11 54 20
oh<>dear<>15 63.9931 7 10 20
ootiful<>soo<>16 62.8158 4 4 4
beau<>ootiful<>16 62.8158 4 4 4
caucus<>race<>16 62.8158 4 4 4
let<>me<>17 62.4435 8 10 45
tell<>me<>18 59.5803 8 11 45
three<>gardeners<>19 58.4940 6 21 7
feet<>high<>20 58.0378 5 5 14
rose<>tree<>21 57.8118 4 4 5
great<>hurry<>22 57.6670 6 34 6
cheshire<>cat<>23 57.3294 5 7 8
guinea<>pigs<>24 55.1777 4 6 4
lobster<>quadrille<>24 55.1777 4 6 4
poor<>little<>25 55.1228 9 28 30
cried<>alice<>26 54.2515 7 7 82
sat<>down<>27 51.2520 9 14 85
white<>kid<>28 50.8303 5 26 5
next<>witness<>29 49.6169 5 18 6
poor<>alice<>30 48.9469 11 28 82
saucepan<>flew<>31 48.8388 3 3 3
dead<>silence<>32 48.2527 4 5 7
old<>fellow<>33 46.7675 4 13 4
found<>herself<>34 46.6476 6 9 33
whole<>pack<>35 46.0643 4 14 4
trembling<>voice<>36 45.0170 5 6 27
fast<>asleep<>37 44.3401 3 3 4
_that<>s_<>37 44.3401 3 3 4
croquet<>ground<>37 44.3401 3 4 3
hand<>bit<>38 44.0915 4 8 6
jury<>box<>38 44.0915 4 8 6
crowded<>round<>39 43.5023 5 6 31
same<>thing<>40 42.7329 7 18 36
--More--(1%)
```

Left hand column: frequency
Right hand column: Log Likelihood



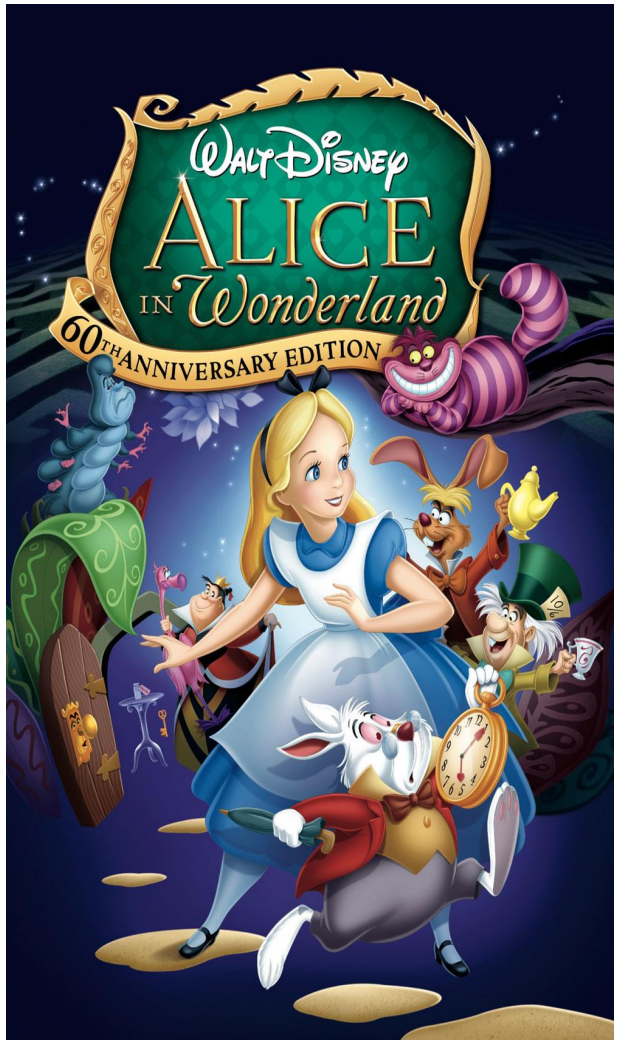
4

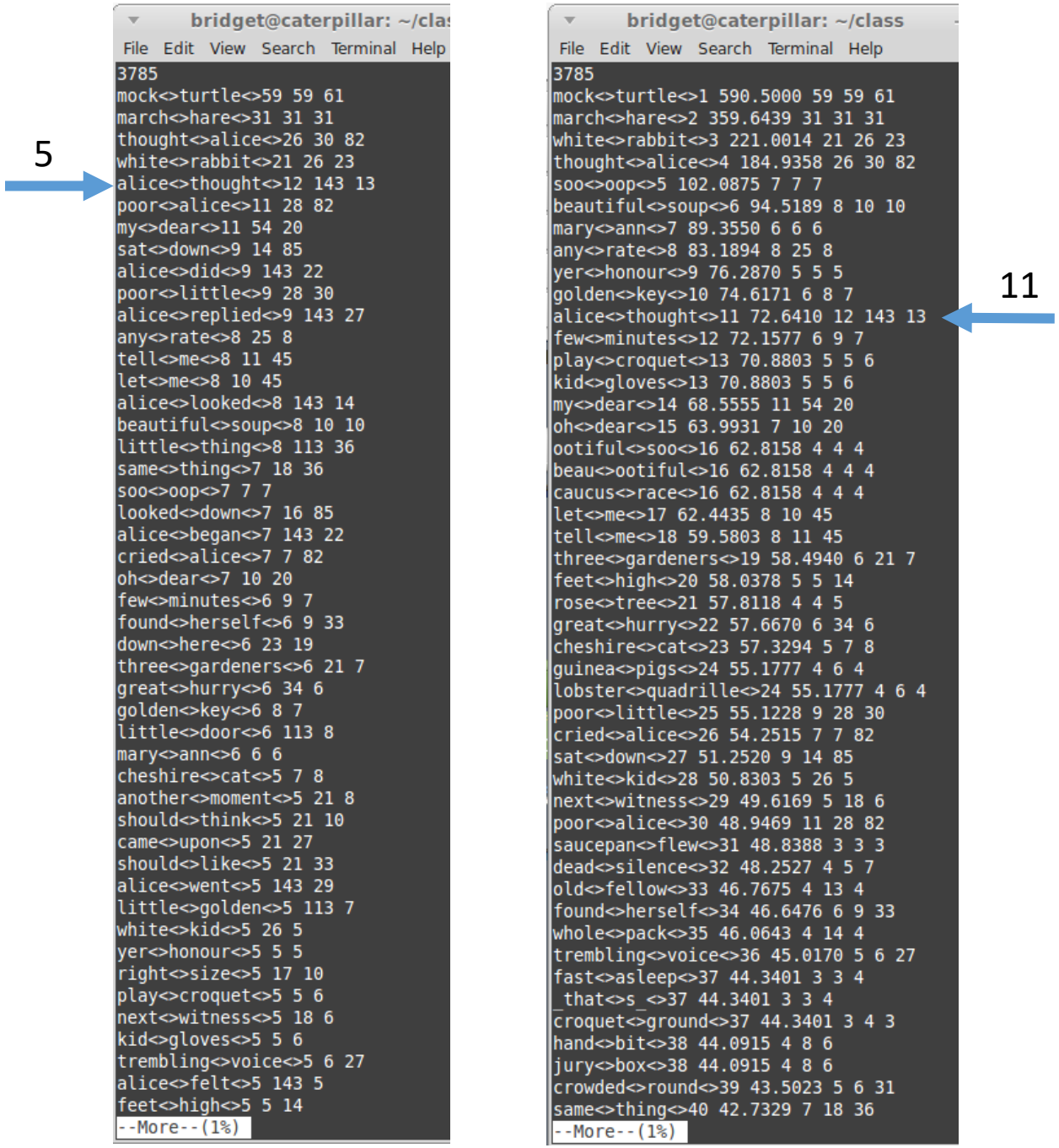
```
bridget@caterpillar: ~/cla
File  Edit  View  Search  Terminal  Help
3785
mock<>turtle<>59 59 61
march<>hare<>31 31 31
thought<>alice<>26 30 82
white<>rabbit<>21 26 23
alice<>thought<>12 143 13
poor<>alice<>11 28 82
my<>dear<>11 54 20
sat<>down<>9 14 85
alice<>did<>9 143 22
poor<>little<>9 28 30
alice<>replied<>9 143 27
any<>rate<>8 25 8
tell<>me<>8 11 45
let<>me<>8 10 45
alice<>looked<>8 143 14
beautiful<>soup<>8 10 10
little<>thing<>8 113 36
same<>thing<>7 18 36
soo<>oop<>7 7 7
looked<>down<>7 16 85
alice<>began<>7 143 22
cried<>alice<>7 7 82
oh<>dear<>7 10 20
few<>minutes<>6 9 7
found<>herself<>6 9 33
down<>here<>6 23 19
three<>gardeners<>6 21 7
great<>hurry<>6 34 6
golden<>key<>6 8 7
little<>door<>6 113 8
mary<>ann<>6 6 6
cheshire<>cat<>5 7 8
another<>moment<>5 21 8
should<>think<>5 21 10
came<>upon<>5 21 27
should<>like<>5 21 33
alice<>went<>5 143 29
little<>golden<>5 113 7
white<>kid<>5 26 5
yer<>honour<>5 5 5
right<>size<>5 17 10
play<>croquet<>5 5 6
next<>witness<>5 18 6
kid<>gloves<>5 5 6
trembling<>voice<>5 6 27
alice<>felt<>5 143 5
feet<>high<>5 5 14
--More--(1%)
```

```
bridget@caterpillar: ~/class
File  Edit  View  Search  Terminal  Help
3785
mock<>turtle<>1 590.5000 59 59 61
march<>hare<>2 359.6439 31 31 31
white<>rabbit<>3 221.0014 21 26 23
thought<>alice<>4 184.9358 26 30 82
soo<>oop<>5 102.0875 7 7 7
beautiful<>soup<>6 94.5189 8 10 10
mary<>ann<>7 89.3550 6 6 6
any<>rate<>8 83.1894 8 25 8
yer<>honour<>9 76.2870 5 5 5
golden<>key<>10 74.6171 6 8 7
alice<>thought<>11 72.6410 12 143 13
few<>minutes<>12 72.1577 6 9 7
play<>croquet<>13 70.8803 5 5 6
kid<>gloves<>13 70.8803 5 5 6
my<>dear<>14 68.5555 11 54 20
oh<>dear<>15 63.9931 7 10 20
ootiful<>soo<>16 62.8158 4 4 4
beau<>ootiful<>16 62.8158 4 4 4
caucus<>race<>16 62.8158 4 4 4
let<>me<>17 62.4435 8 10 45
tell<>me<>18 59.5803 8 11 45
three<>gardeners<>19 58.4940 6 21 7
feet<>high<>20 58.0378 5 5 14
rose<>tree<>21 57.8118 4 4 5
great<>hurry<>22 57.6670 6 34 6
cheshire<>cat<>23 57.3294 5 7 8
guinea<>pigs<>24 55.1777 4 6 4
lobster<>quadrille<>24 55.1777 4 6 4
poor<>little<>25 55.1228 9 28 30
cried<>alice<>26 54.2515 7 7 82
sat<>down<>27 51.2520 9 14 85
white<>kid<>28 50.8303 5 26 5
next<>witness<>29 49.6169 5 18 6
poor<>alice<>30 48.9469 11 28 82
saucepan<>flew<>31 48.8388 3 3 3
dead<>silence<>32 48.2527 4 5 7
old<>fellow<>33 46.7675 4 13 4
found<>herself<>34 46.6476 6 9 33
whole<>pack<>35 46.0643 4 14 4
trembling<>voice<>36 45.0170 5 6 27
fast<>asleep<>37 44.3401 3 3 4
_that<>s_<>37 44.3401 3 3 4
croquet<>ground<>37 44.3401 3 4 3
hand<>bit<>38 44.0915 4 8 6
jury<>box<>38 44.0915 4 8 6
crowded<>round<>39 43.5023 5 6 31
same<>thing<>40 42.7329 7 18 36
--More--(1%)
```

3

Left hand column: frequency
Right hand column: Log Likelihood

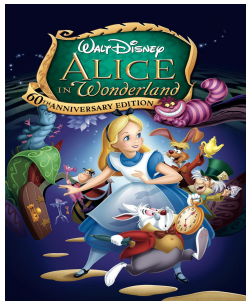




Left hand column: frequency
Right hand column: Log Likelihood

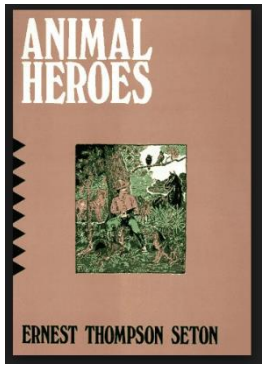
So it is weighting the ngram based on its distribution with respect to the other tokens in the ngram

white rabbit from two different corpora



Seen values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$n_{11} = 21$	$n_{12} = 5$	$n_{1p} = 26$
$\neg\boldsymbol{white}$	$n_{21} = 2$	$n_{22} = 3757$	$n_{2p} = 3759$
	$n_{p1} = 23$	$n_{p2} = 3762$	$n_{pp} = 3785$

$G^2 = 221.0014$

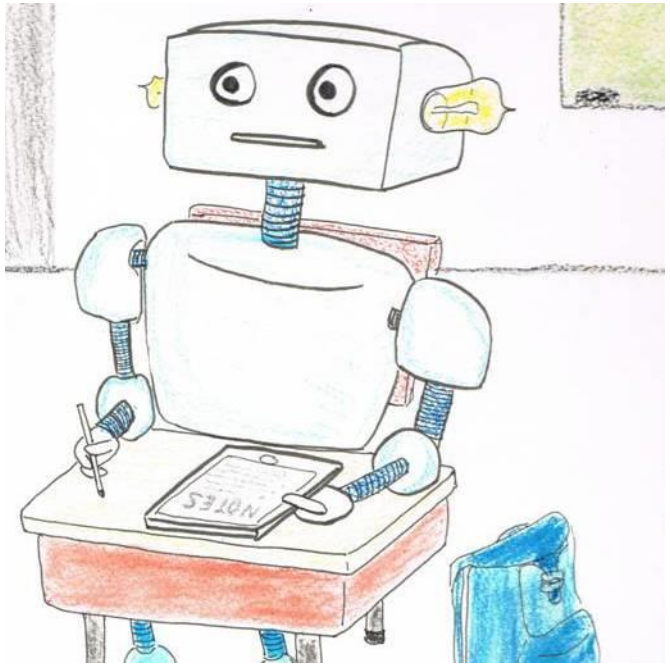


Seen values	$\boldsymbol{rabbit}$	$\neg\boldsymbol{rabbit}$	
$\boldsymbol{white}$	$n_{11} = 1$	$n_{12} = 82$	$n_{1p} = 83$
$\neg\boldsymbol{white}$	$n_{21} = 23$	$n_{22} = 9289$	$n_{2p} = 9312$
	$n_{p1} = 24$	$n_{p2} = 9371$	$n_{pp} = 9395$

$G^2 = 1.5601$

Application: Feature Selection

```
bridget@caterpillar: ~/class
File  Edit  View  Search  Terminal  Help
3785
mock<>turtle<>1 590.5000 59 59 61
march<>hare<>2 359.6439 31 31 31
white<>rabbit<>3 221.0014 21 26 23
thought<>alice<>4 184.9358 26 30 82
soo<>oop<>5 102.0875 7 7 7
beautiful<>soup<>6 94.5189 8 10 10
mary<>ann<>7 89.3550 6 6 6
any<>rate<>8 83.1894 8 25 8
yer<>honour<>9 76.2870 5 5 5
golden<>key<>10 74.6171 6 8 7
alice<>thought<>11 72.6410 12 143 13
few<>minutes<>12 72.1577 6 9 7
play<>croquet<>13 70.8803 5 5 6
kid<>gloves<>13 70.8803 5 5 6
my<>dear<>14 68.5555 11 54 20
oh<>dear<>15 63.9931 7 10 20
ootiful<>soo<>16 62.8158 4 4 4
beau<>ootiful<>16 62.8158 4 4 4
caucus<>race<>16 62.8158 4 4 4
let<>me<>17 62.4435 8 10 45
tell<>me<>18 59.5803 8 11 45
three<>gardeners<>19 58.4940 6 21 7
feet<>high<>20 58.0378 5 5 14
rose<>tree<>21 57.8118 4 4 5
great<>hurry<>22 57.6670 6 34 6
cheshire<>cat<>23 57.3294 5 7 8
guinea<>pigs<>24 55.1777 4 6 4
lobster<>quadrille<>24 55.1777 4 6 4
poor<>little<>25 55.1228 9 28 30
cried<>alice<>26 54.2515 7 7 82
sat<>down<>27 51.2520 9 14 85
white<>kid<>28 50.8303 5 26 5
next<>witness<>29 49.6169 5 18 6
poor<>alice<>30 48.9469 11 28 82
saucepan<>flew<>31 48.8388 3 3 3
dead<>silence<>32 48.2527 4 5 7
old<>fellow<>33 46.7675 4 13 4
found<>herself<>34 46.6476 6 9 33
whole<>pack<>35 46.0643 4 14 4
trembling<>voice<>36 45.0170 5 6 27
fast<>asleep<>37 44.3401 3 3 4
_that<>s_<>37 44.3401 3 3 4
croquet<>ground<>37 44.3401 3 4 3
hand<>bit<>38 44.0915 4 8 6
jury<>box<>38 44.0915 4 8 6
crowded<>round<>39 43.5023 5 6 31
same<>thing<>40 42.7329 7 18 36
--More--(1%)
```



Threshold
Cutoff

Application: Context Sensitive Spelling Correction

We went to Paris and stayed (their|there)? eleven days

P(stayed their eleven) vs P(stayed there eleven)

Backoff model

Avg(P(stayed their), P(their eleven)) vs Avg(P(stayed there), P(there eleven))

Application: Context Sensitive Spelling Correction

We went to Paris and stayed (their|there)? eleven days

P(stayed their eleven) vs P(stayed there eleven)

Backoff model

Avg(P(stayed their), P(their eleven)) vs Avg(P(stayed there), P(there eleven))

	i	want	to	eat	chinese	food	lunch	spend	<start>	<end>
i	0.002	0.33	0	0.0036	0	0	0	0.00079	0	0
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0064	0.0011	0	0
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087	0	0
eat	0	0	0.0027	0	0.021	0.0027	0.056	0	0	0.011
chinese	0.0063	0	0	0	0	0.52	0.0063	0	0	0.008
food	0.014	0	0.014	0	0.00092	0.0037	0	0	0	0.004
lunch	0.0059	0	0	0	0	0.0029	0	0	0	0.003
spend	0.0036	0	0.0036	0	0	0	0	0	1	0.006
<start>	0.015	0	0.01	0	0.005	0.003	0.001	0	0	0
<end>	0	0	0	0	0.001	0.007	0.002	0.011	0	0

Relative Frequency Table

Application: Context Sensitive Spelling Correction

$G^2 Table$

bridget@caterpillar: ~/class

File Edit View Search Terminal Help

```
3785
mock<>turtle<>1 590.5000 59 59 61
march<>hare<>2 359.6439 31 31 31
white<>rabbit<>3 221.0014 21 26 23
thought<>alice<>4 184.9358 26 30 82
soo<>oop<>5 102.0875 7 7 7
beautiful<>soup<>6 94.5189 8 10 10
mary<>ann<>7 89.3550 6 6 6
any<>rate<>8 83.1894 8 25 8
yer<>honour<>9 76.2870 5 5 5
golden<>key<>10 74.6171 6 8 7
alice<>thought<>11 72.6410 12 143 13
few<>minutes<>12 72.1577 6 9 7
play<>croquet<>13 70.8803 5 5 6
kid<>gloves<>13 70.8803 5 5 6
my<>dear<>14 68.5555 11 54 20
oh<>dear<>15 63.9931 7 10 20
ootiful<>soo<>16 62.8158 4 4 4
beau<>ootiful<>16 62.8158 4 4 4
caucus<>race<>16 62.8158 4 4 4
let<>me<>17 62.4435 8 10 45
tell<>me<>18 59.5803 8 11 45
three<>gardeners<>19 58.4940 6 21 7
feet<>high<>20 58.0378 5 5 14
rose<>tree<>21 57.8118 4 4 5
great<>hurry<>22 57.6670 6 34 6
cheshire<>cat<>23 57.3294 5 7 8
guinea<>pigs<>24 55.1777 4 6 4
lobster<>quadrille<>24 55.1777 4 6 4
poor<>little<>25 55.1228 9 28 30
cried<>alice<>26 54.2515 7 7 82
sat<>down<>27 51.2520 9 14 85
white<>kid<>28 50.8303 5 26 5
next<>witness<>29 49.6169 5 18 6
poor<>alice<>30 48.9469 11 28 82
saucepan<>flew<>31 48.8388 3 3 3
dead<>silence<>32 48.2527 4 5 7
old<>fellow<>33 46.7675 4 13 4
found<>herself<>34 46.6476 6 9 33
whole<>pack<>35 46.0643 4 14 4
trembling<>voice<>36 45.0170 5 6 27
fast<>asleep<>37 44.3401 3 3 4
_that<>s_<>37 44.3401 3 3 4
croquet<>ground<>37 44.3401 3 4 3
hand<>bit<>38 44.0915 4 8 6
jury<>box<>38 44.0915 4 8 6
crowded<>round<>39 43.5023 5 6 31
same<>thing<>40 42.7329 7 18 36
--More--(1%)
```

We went to Paris and stayed (their|there)? eleven days

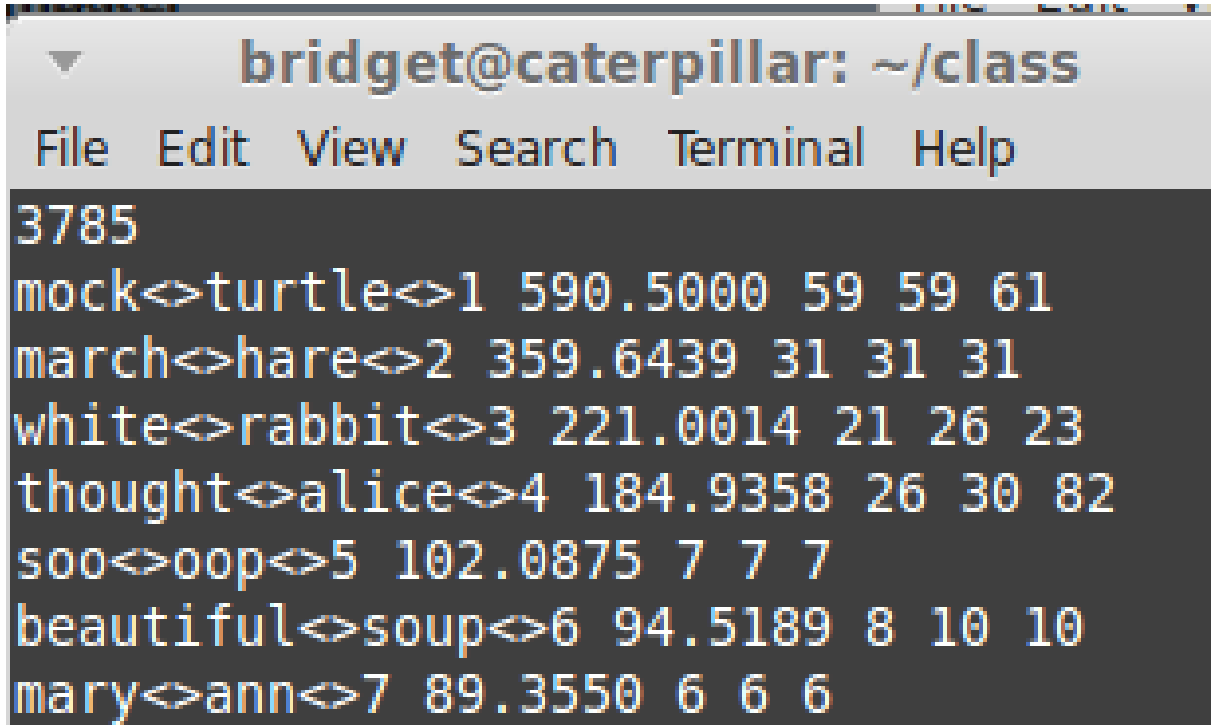
Backoff model

Avg(P(stayed their), P(their eleven)) vs Avg(P(stayed there), P(there eleven))

Avg(G^2(stayed their), G^2(their eleven)) vs Avg(G^2(stayed there), G^2(there eleven))

Application: Collocations, Terms and Multiword Expressions?

Ngrams that are collocations, terms and mwes
are more likely to occur together than by chance

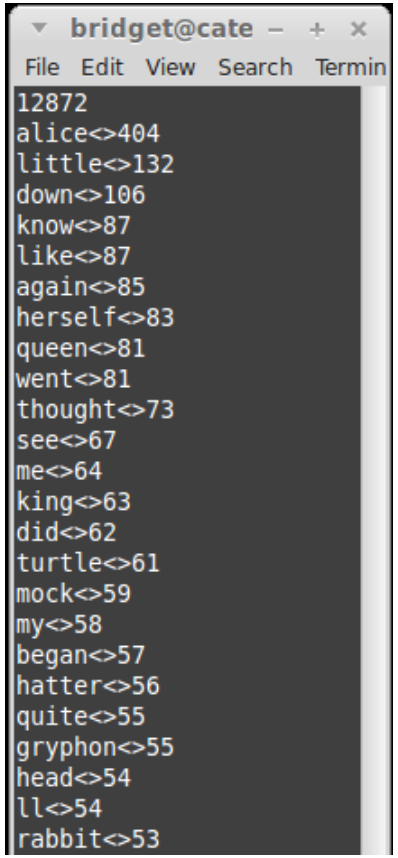


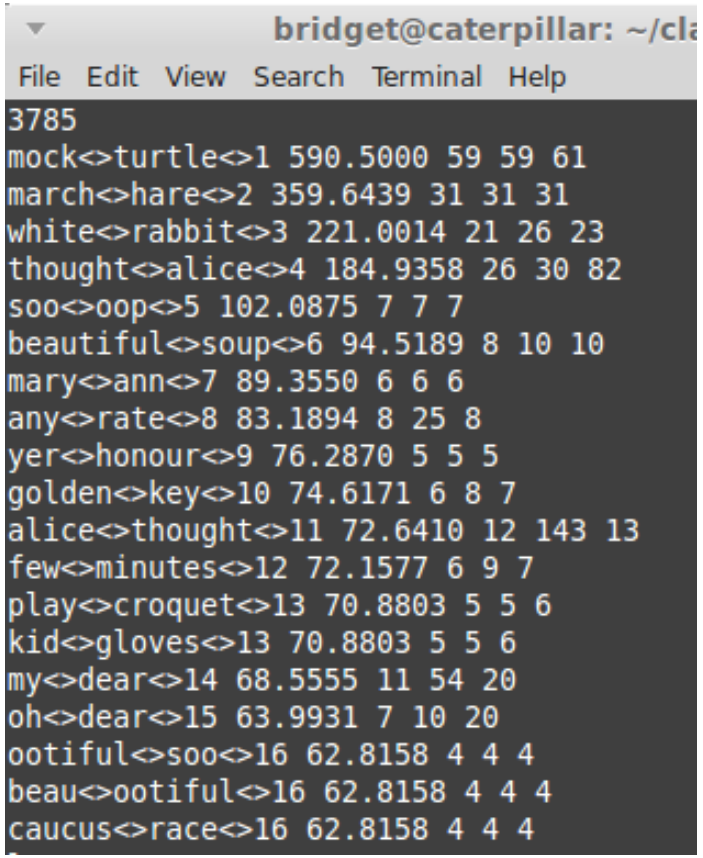
Application: Topic Identification

Terms

provide more specific contextual information

than individual words





Questions?

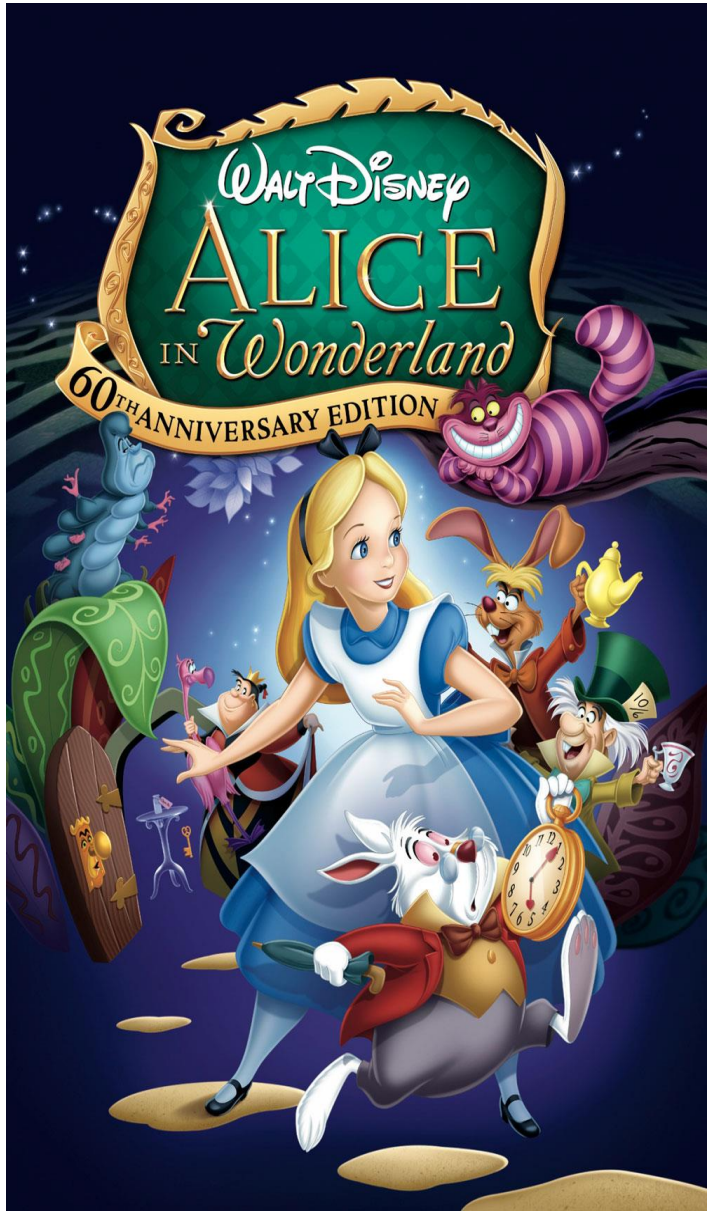
Ngram: Tokenization

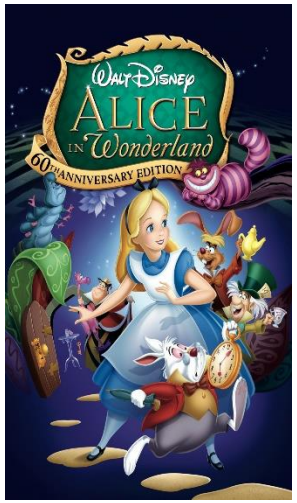
Extracting Relevant Contextual information from the text

Alice in Wonderland

and

Bigrams

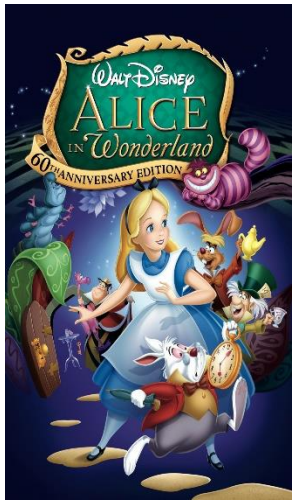




N-gram Extractor

```
bridget@caterpillar: ~/class
File  Edit  View  Search  Terminal  Help
32867
,<>and<>1 1457.3545 490 2444 899
mock<>turtle<>2 846.3665 59 59 61
said<>the<>3 663.9441 211 466 1676
don<>t<>4 620.7031 60 60 217
,<>said<>5 577.9117 225 2444 466
said<>alice<>6 533.8303 115 466 403
march<>hare<>7 473.5838 31 34 31
i<>m<>8 443.9968 57 522 62
the<>queen<>9 407.4693 75 1676 81
went<>on<>10 386.5964 47 81 191
in<>a<>11 364.6575 99 383 649
the<>king<>12 361.0673 62 1676 63
a<>little<>13 304.9718 61 649 132
the<>gryphon<>14 300.0905 53 1676 55
.<>i<>15 289.4790 116 1017 522
the<>mock<>16 289.3467 54 1676 59
she<>had<>17 284.6907 62 556 180
the<>hatter<>18 282.6445 52 1676 56
you<>know<>19 282.6369 45 399 87
of<>the<>20 250.5836 134 523 1676
it<>was<>21 247.6654 74 603 353
;<>and<>22 246.5375 68 198 899
white<>rabbit<>23 243.1178 21 30 53
the<>duchess<>24 241.5998 42 1676 43
to<>herself<>25 237.2455 46 757 83
as<>she<>26 231.2743 62 264 556
i<>ve<>27 226.3519 33 522 44
it<>s<>28 220.8583 57 603 210
won<>t<>29 217.1130 24 29 217
did<>not<>30 215.6460 27 62 140
to<>be<>31 211.1520 53 757 152
!<>said<>32 198.8397 65 456 466
there<>was<>33 192.6939 35 100 353
out<>of<>34 184.7397 40 119 523
of<>course<>35 181.1068 24 523 27
the<>dormouse<>36 179.3941 35 1676 40
i<>ll<>37 176.8172 30 522 54
doesn<>t<>38 172.0503 17 17 217
the<>the<>39 169.0912 1 1676 1676
;<>but<>40 168.3603 32 198 173
can<>t<>41 166.5136 24 57 217
such<>a<>42 161.9726 28 45 649
she<>could<>43 161.9636 32 556 75
?<>said<>44 161.0593 43 211 466
,<>but<>45 160.9052 73 2444 173
--More--(0%)
```

These ngrams are a little different than we saw before



N-gram Extractor

```
bridget@caterpillar: ~/class
File  Edit  View  Search  Terminal  Help
32867
,<>and<>1 1457.3545 490 2444 899
mock<>turtle<>2 846.3665 59 59 61
said<>the<>3 663.9441 211 466 1676
don<>t<>4 620.7031 60 60 217
,<>said<>5 577.9117 225 2444 466
said<>alice<>6 533.8303 115 466 403
march<>hare<>7 473.5838 31 34 31
i<>m<>8 443.9968 57 522 62
the<>queen<>9 407.4693 75 1676 81
went<>on<>10 386.5964 47 81 191
in<>a<>11 364.6575 99 383 649
the<>king<>12 361.0673 62 1676 63
a<>little<>13 304.9718 61 649 132
the<>gryphon<>14 300.0905 53 1676 55
.<>i<>15 289.4790 116 1017 522
the<>mock<>16 289.3467 54 1676 59
she<>had<>17 284.6907 62 556 180
the<>hatter<>18 282.6445 52 1676 56
you<>know<>19 282.6369 45 399 87
of<>the<>20 250.5836 134 523 1676
it<>was<>21 247.6654 74 603 353
;<>and<>22 246.5375 68 198 899
white<>rabbit<>23 243.1178 21 30 53
the<>duchess<>24 241.5998 42 1676 43
to<>herself<>25 237.2455 46 757 83
as<>she<>26 231.2743 62 264 556
i<>ve<>27 226.3519 33 522 44
it<>s<>28 220.8583 57 603 210
won<>t<>29 217.1130 24 29 217
did<>not<>30 215.6460 27 62 140
to<>be<>31 211.1520 53 757 152
!<>said<>32 198.8397 65 456 466
there<>was<>33 192.6939 35 100 353
out<>of<>34 184.7397 40 119 523
of<>course<>35 181.1068 24 523 27
the<>dormouse<>36 179.3941 35 1676 40
i<>ll<>37 176.8172 30 522 54
doesn<>t<>38 172.0503 17 17 217
the<>the<>39 169.0912 1 1676 1676
;<>but<>40 168.3603 32 198 173
can<>t<>41 166.5136 24 57 217
such<>a<>42 161.9726 28 45 649
she<>could<>43 161.9636 32 556 75
?<>said<>44 161.0593 43 211 466
,<>but<>45 160.9052 73 2444 173
--More--(0%)
```

Uh oh – what is the Problem?

Lots of punctuation
And
Non-content tokens

Stopwords

These are often words
that do not provide any contextual information for the task at hand

e.g. if we were analyzing clinical notes the token *patient* is often removed because it provides no contextual information – all the notes are about patients.

Punctuation

Is often included in the stopword list

Stop words

From Wikipedia, the free encyclopedia

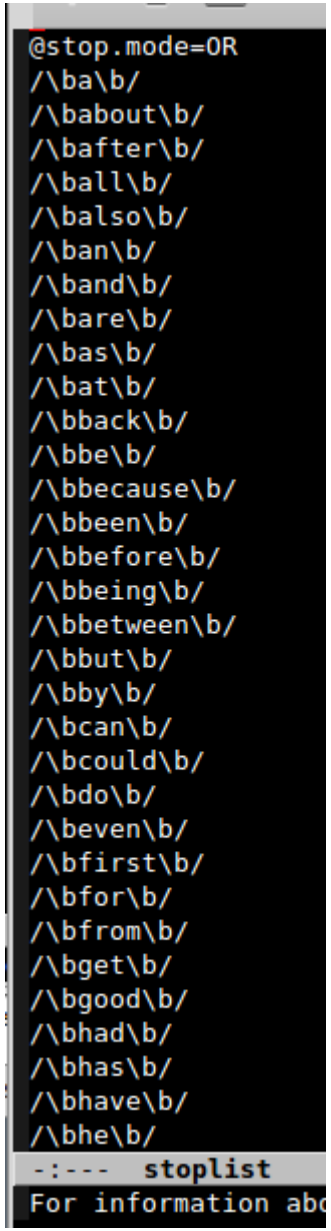
Not to be confused with Safeword.

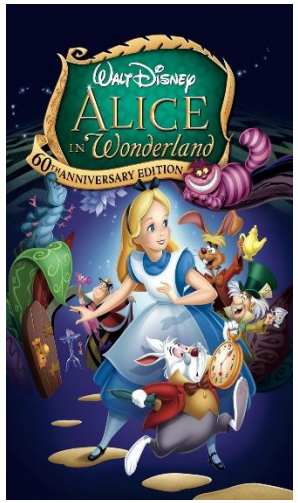
In computing, **stop words** are words which are filtered out before or after processing of natural language data (text).[1] There is no single universal list of stop words used by all processing of natural language tools, and indeed not all tools even use such a list. Some tools specifically avoid removing these **stop words** to support phrase search.

Example Stopword

- They are in regular expression form
  - Makes them powerful
  - E.g. /\b\d+\b/ -> removes a series of digits without having
  - to enumerate over all the digits

- @stop.mode
  - OR
  - Do not include ngram if either word is a stopword
  - AND
  - Do not include ngram if both words are a stopword





stop list

N-gram Modeler

```
bridget@caterpillar: ~/class
File  Edit  View  Search  Terminal  Help
3785
mock<>turtle<>1 590.5000 59 59 61
march<>hare<>2 359.6439 31 31 31
white<>rabbit<>3 221.0014 21 26 23
thought<>alice<>4 184.9358 26 30 82
soo<>oop<>5 102.0875 7 7 7
beautiful<>soup<>6 94.5189 8 10 10
mary<>ann<>7 89.3550 6 6 6
any<>rate<>8 83.1894 8 25 8
yer<>honour<>9 76.2870 5 5 5
golden<>key<>10 74.6171 6 8 7
alice<>thought<>11 72.6410 12 143 13
few<>minutes<>12 72.1577 6 9 7
play<>croquet<>13 70.8803 5 5 6
kid<>gloves<>13 70.8803 5 5 6
my<>dear<>14 68.5555 11 54 20
oh<>dear<>15 63.9931 7 10 20
ootiful<>soo<>16 62.8158 4 4 4
beau<>ootiful<>16 62.8158 4 4 4
caucus<>race<>16 62.8158 4 4 4
let<>me<>17 62.4435 8 10 45
tell<>me<>18 59.5803 8 11 45
three<>gardeners<>19 58.4940 6 21 7
feet<>high<>20 58.0378 5 5 14
rose<>tree<>21 57.8118 4 4 5
great<>hurry<>22 57.6670 6 34 6
cheshire<>cat<>23 57.3294 5 7 8
guinea<>pigs<>24 55.1777 4 6 4
lobster<>quadrille<>24 55.1777 4 6 4
poor<>little<>25 55.1228 9 28 30
cried<>alice<>26 54.2515 7 7 82
sat<>down<>27 51.2520 9 14 85
white<>kid<>28 50.8303 5 26 5
next<>witness<>29 49.6169 5 18 6
poor<>alice<>30 48.9469 11 28 82
saucepan<>flew<>31 48.8388 3 3 3
dead<>silence<>32 48.2527 4 5 7
old<>fellow<>33 46.7675 4 13 4
found<>herself<>34 46.6476 6 9 33
whole<>pack<>35 46.0643 4 14 4
trembling<>voice<>36 45.0170 5 6 27
fast<>asleep<>37 44.3401 3 3 4
croquet<>ground<>37 44.3401 3 4 3
_that<>s_<>37 44.3401 3 3 4
jury<>box<>38 44.0915 4 8 6
hand<>bit<>38 44.0915 4 8 6
--More--(1%)
```

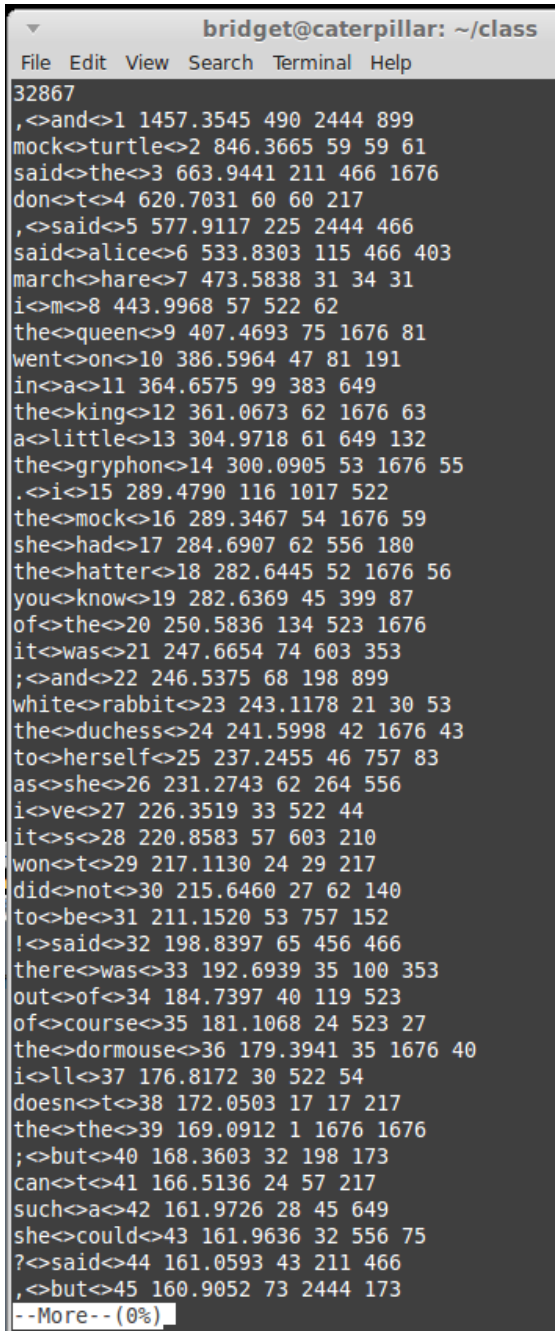
Compare the outputs

- What can you see

  The likelihood score of the top bigrams has decreased

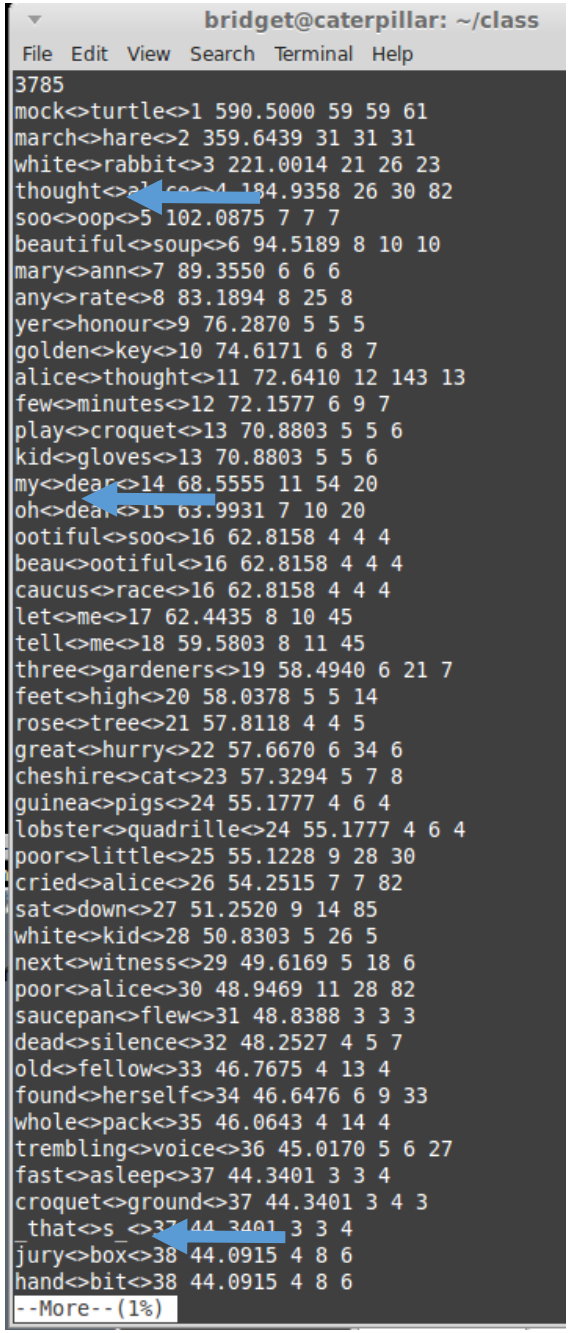
But

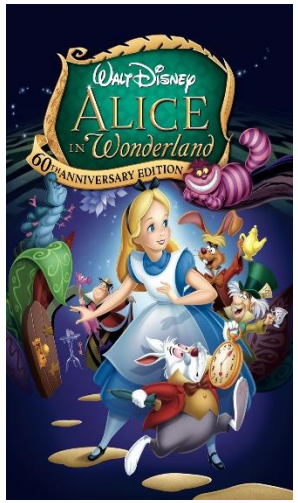
- he contextual information of the top bigrams has increased

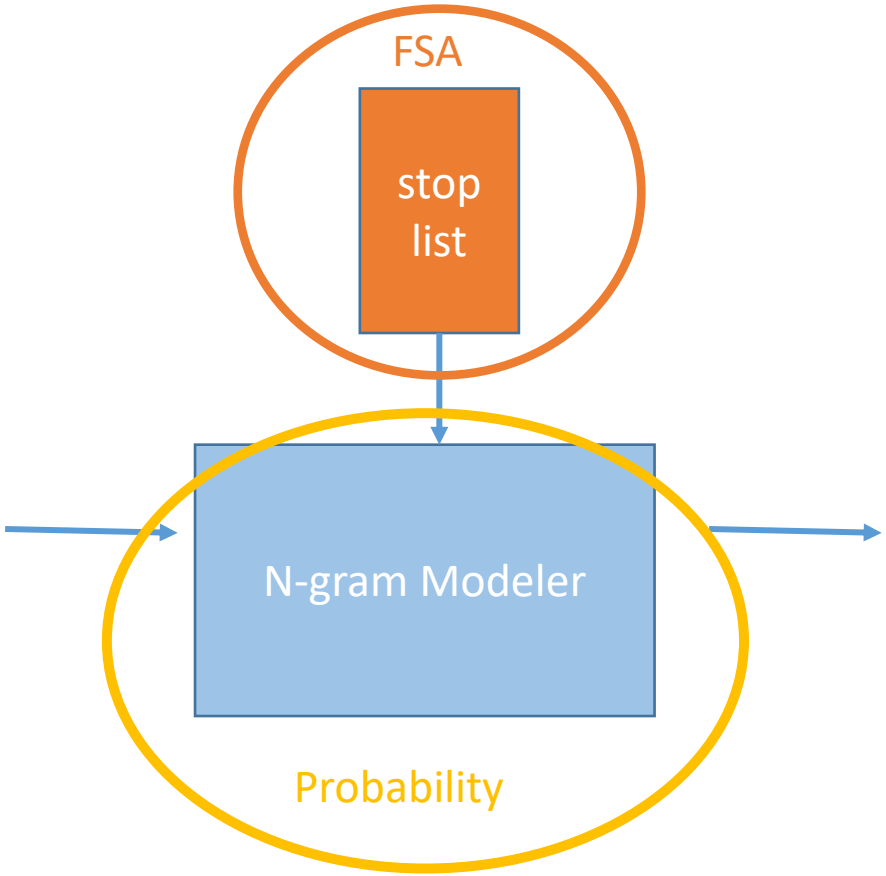


Creating a Stoplist

A bit of art … with trial and error.








```
3785
mock<>turtle<>59
march<>hare<>31
thought<>alice<>26
white<>rabbit<>21
alice<>thought<>12
poor<>alice<>11
my<>dear<>11
poor<>little<>9
alice<>did<>9
sat<>down<>9
alice<>replied<>9
any<>rate<>8
let<>me<>8
little<>thing<>8
tell<>me<>8
alice<>looked<>8
beautiful<>soup<>8
looked<>down<>7
oh<>dear<>7
alice<>began<>7
soo<>oop<>7
same<>thing<>7
cried<>alice<>7
golden<>key<>6
few<>minutes<>6
great<>hurry<>6
little<>door<>6
mary<>ann<>6
down<>here<>6
three<>gardeners<>6
found<>herself<>6
play<>croquet<>5
-:---  alice.2          Top L
Wrote /home/bridget/alice.2
```

NLP Application that is using both:
Rules and Probability
Information

Stoplists are commonly used

But … would we necessarily want to use
one with our Spelling Correction Application?

We went to Paris and stayed (their|there)? eleven days

Avg(G^2(stayed their), G^2(their eleven)) vs Avg(G^2(stayed there), G^2(there eleven))

Look a little more at Ngrams

- Ngrams
  - we are going to modify our definition slightly
  - a continguous or non-continguous sequence of tokens that occur in some proximity to each other in a corpus

Look a little more at Ngrams

- Ngrams
  - We are going to modify our definition slightly
  - a continguous or non-continguous sequence of tokens that occur in some proximity to each other in a corpus

to be or not to be

Our unique tokens: to be or not to be

Look a little more at Ngrams

- Ngrams
  - We are going to modify our definition slightly
  - a continguous or non-continguous sequence of tokens that occur in some proximity to each other in a corpus

to be or not to be

Our unigrams are: to be or not to be

Look a little more at Ngrams

- Ngrams
  - We are going to modify our definition slightly
  - a continguous or non-continguous sequence of tokens that occur in some proximity to each other in a corpus

to be or not to be

Our bigrams are: to be be or or not not to to be

Look a little more at Ngrams

- Ngrams
  - we are going to modify our definition slightly
  - a contiguous or non-contiguous sequence of tokens that occur in some proximity to each other in a corpus

These are contiguous ngrams – the are situated next to each other in the text

to be or not to be

Our bigrams are: to be be or or not not to to be

Non-contiguous n-grams

We open up the window under which an n-gram can occur while still maintaining order

to be or not to be

Our bigrams with a *window* size of three are:

to be to or

Non-contiguous n-grams

We open up the window under which an n-gram can occur while still maintaining order

to be or not to be

Our bigrams with a *window* size of three are:

to be to or

Non-contiguous n-grams

We open up the window under which an n-gram can occur while still maintaining order

to be or not to be

Our bigrams with a *window* size of three are:

to be to or be or be not

Non-contiguous n-grams

We open up the window under which an n-gram can occur while still maintaining order

to be or not to be

Our bigrams with a *window* size of three are:

to be to or be or be not

Non-contiguous n-grams

We open up the window under which an n-gram can occur while still maintaining order

to be or not to be

Our bigrams with a *window* size of three are:

to be to or be or be not or not or to

Non-contiguous n-grams

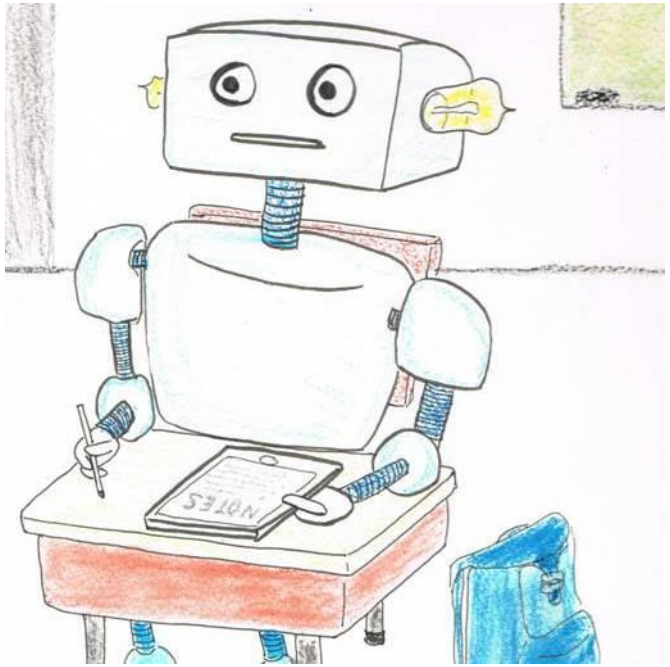
We open up the window under which an n-gram can occur while still maintaining order

to be or not to be

Our bigrams with a *window* size of three are:

to be to or be or be not or not or to not to not be to be

We may think about this with feature selection



Why do this?

- Sparsity
  - Increases the number of tokens seen in the data set while still making them reasonable

- Expands the context

Term, Collocation or MWE Identification?

Term, Collocation or MWE Identification?

Probably Not

Spelling Correction

We went to Paris and stayed (their|there)? eleven days

Avg(G^2(stayed their), G^2(their eleven)) vs Avg(G^2(stayed there), G^2(there eleven))

Spelling Correction

We went to Paris and stayed (their|there)? eleven days

Avg(G^2(stayed their), G^2(their eleven)) vs Avg(G^2(stayed there), G^2(there eleven))

? Maybe?

Spelling Correction

How about smoothing?

We went to Paris and stayed (their|there)? eleven days

Avg(G^2(stayed their), G^2(their eleven)) vs Avg(G^2(stayed there), G^2(there eleven))

Questions?