

HOW TO PERFORM SUPER SIMPLE TEXT ANALYSIS

DR. ALVIN ANG

	A	B	C	D	E
1	RANK	FREQ	WORD	% OF TOTAL	VIZ
2	1	102	THE	6.59767141	
3	2	57	TO	3.686934023	
4	3	49	THAT	3.169469599	
5	4	36	IN	2.328589909	
6	5	30	OF	1.940491591	
7	6	29	AND	1.875808538	
8	7	29	IS	1.875808538	
9	8	24	I	1.552393273	
10	9	23	A	1.48771022	
11	10	22	IT	1.423027167	
12	11	20	CHINA	1.293661061	
13	12	19	HAVE	1.228978008	
14	13	19	NUMBER	1.228978008	
15	14	17	NOT	1.099611902	
16	15	16	WE	1.034928849	
17	16	15	WILL	0.9702457956	
18	17	14	WOULD	0.9055627426	
19	18	13	THIS	0.8408796895	
20	20	12	THEY	0.7761966365	

CONTENTS

<i>Introduction</i>	3
<i>Web Frequency Indexer</i>	4
Step 1: Copy Paste Text	4
Step 2: Web Frequency Indexer Output.....	5
<i>Excel Analysis</i>	6
Step 1: Pasting the Data	6
Step 2: Creating the Count “ “	9
<i>Conclusion</i>	10
<i>References</i>	11
<i>About the Authors</i>	12

INTRODUCTION

- This article follows Bajak (2015).
- <https://www.storybench.org/how-to-do-super-simple-textual-analysis/>
- It makes use of a Web Frequency Indexer found here:
- <https://www.lex tutor.ca/freq/eng/>
- We may expect an Excel output like this:

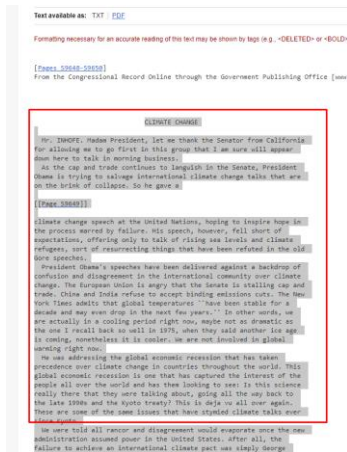
	A	B	C	D	E
1	RANK	FREQ	WORD	% OF TOTAL	VIZ
2	1	102	THE	6.59767141	
3	2	57	TO	3.686934023	
4	3	49	THAT	3.169469599	
5	4	36	IN	2.328589909	
6	5	30	OF	1.940491591	
7	6	29	AND	1.875808538	
8	7	29	IS	1.875808538	
9	8	24	I	1.552393273	
10	9	23	A	1.48771022	
11	10	22	IT	1.423027167	
12	11	20	CHINA	1.293661061	
13	12	19	HAVE	1.228978008	
14	13	19	NUMBER	1.228978008	
15	14	17	NOT	1.099611902	
16	15	16	WE	1.034928849	
17	16	15	WILL	0.9702457956	
18	17	14	WOULD	0.9055627426	
19	18	13	THIS	0.8408796895	
20	20	12	THEY	0.7761966365	

WEB FREQUENCY INDEXER

STEP 1: COPY PASTE TEXT



- Go to <https://www.lexutor.ca/freq/eng/>



- Go to <https://www.congress.gov/congressional-record/2009/09/22/senate-section/article/s9648-2/>
- Select the text and copy.
- Paste it into the 'Web Frequency Indexer' box → Click 'Submit Window'.

STEP 2: WEB FREQUENCY INDEXER OUTPUT

st Builder > Frequency Text Input > Freq. List Output

Text: Untitled
 Date: 3/16/2020 2:46
 Tokens: 1514
 Types: 523
 Ratio: 0.3454
 Sort: descending

RANK	FREQ	COVERAGE		WORD
		indivd	cumulative	
1.	98	6.47%	6.47%	THE
2.	57	3.76%	10.23%	TO
3.	48	3.17%	13.40%	THAT
4.	36	2.38%	15.78%	IN
5.	30	1.98%	17.76%	OF
6.	29	1.92%	19.68%	AND
7.	28	1.85%	21.53%	IS
8.	23	1.52%	23.05%	A
9.	23	1.52%	24.57%	I
10.	21	1.39%	25.96%	IT
11.	20	1.32%	27.28%	CHINA
12.	19	1.25%	28.53%	HAVE
13.	19	1.25%	29.78%	NUMBER
14.	17	1.12%	30.90%	NOT
15.	16	1.06%	31.96%	WE
16.	15	0.99%	32.95%	WILL
17.	14	0.92%	33.87%	WOULD
18.	13	0.86%	34.73%	THIS
19.	12	0.79%	35.52%	CLIMATE
20.	12	0.79%	36.31%	THEY
21.	11	0.73%	37.04%	ARE
22.	11	0.73%	37.77%	EMISSIONS
23.	10	0.66%	38.43%	CHANGE
24.	10	0.66%	39.09%	FOR
25.	10	0.66%	39.75%	SENATE

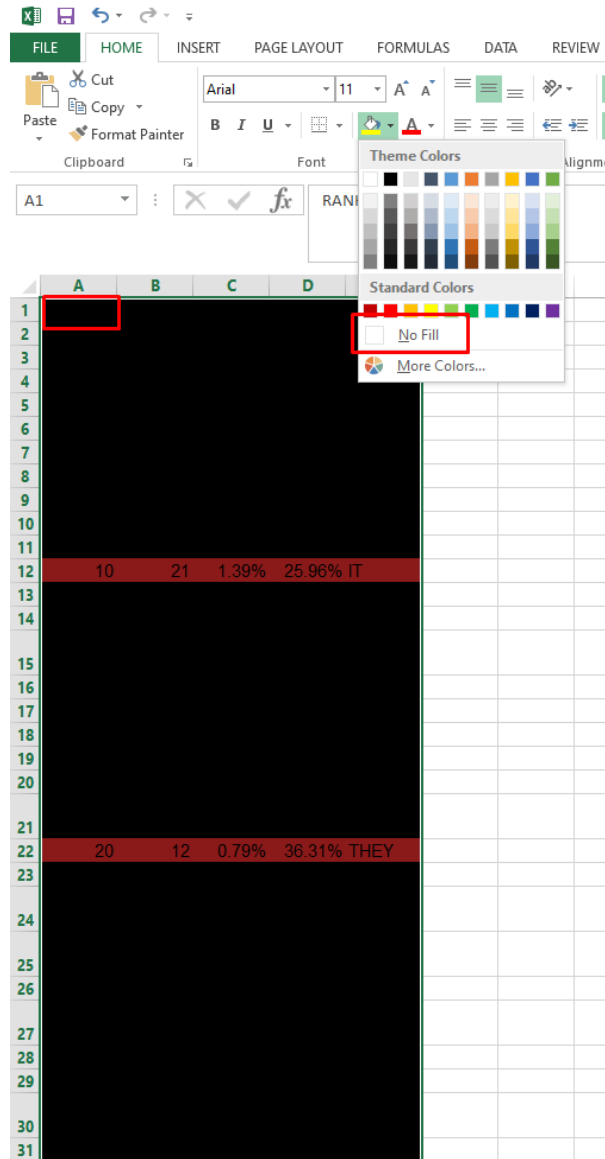
Same list but with 'selectable' word column
(for extracting list of freq>x)

1.	98	THE
2.	57	TO
3.	48	THAT
4.	36	IN
5.	30	OF
6.	29	AND
7.	28	IS
8.	23	A
9.	23	I
10.	21	IT
11.	20	CHINA
12.	19	HAVE
13.	19	NUMBER
14.	17	NOT
15.	16	WE
16.	15	WILL
17.	14	WOULD
18.	13	THIS
19.	12	CLIMATE
20.	12	THEY
21.	11	ARE
22.	11	EMISSIONS
23.	10	CHANGE
24.	10	FOR
25.	10	SENATE

- You will come to this new page.
- Select and copy the LHS of the table i.e. Rank / Freq / Coverage / Word.

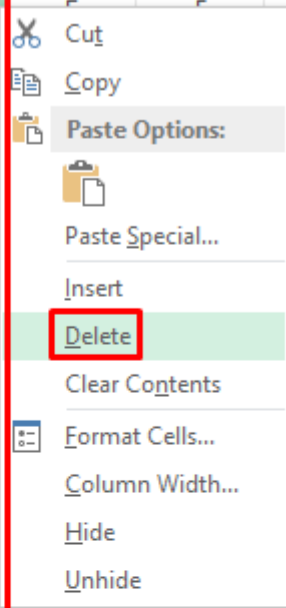
EXCEL ANALYSIS

STEP 1: PASTING THE DATA



- Open Excel
- Paste everything into cell A1.
- Remember to change the background color to → No Fill.

	A	B	C	D	E	F	G
1			COVERAGE				
2	RANK	FREQ	individ	cumulative			
3	1	98	6.47%	6.47%			
4	2	57	3.76%	10.23%			
5	3	48	3.17%	13.40%			
6	4	36	2.38%	15.78%			
7	5	30	1.98%	17.76%			
8	6	29	1.92%	19.68%			
9	7	28	1.85%	21.53%			
10	8	23	1.52%	23.05%			
11	9	23	1.52%	24.57%			
12	10	21	1.39%	25.96%			
13	11	20	1.32%	27.28%			
14	12	19	1.25%	28.53%			
15	13	19	1.25%	29.78%	NUMBER		
16	14	17	1.12%	30.90%	NOT		
17	15	16	1.06%	31.96%	WE		
18	16	15	0.99%	32.95%	WILL		
19	17	14	0.92%	33.87%	WOULD		
20	18	13	0.86%	34.73%	THIS		
21	19	12	0.79%	35.52%	CLIMATE		
22	20	12	0.79%	36.31%	THEY		
23	21	11	0.73%	37.04%	ARE		
24	22	11	0.73%	37.77%	EMISSIONS		
25	23	10	0.66%	38.43%	CHANGE		



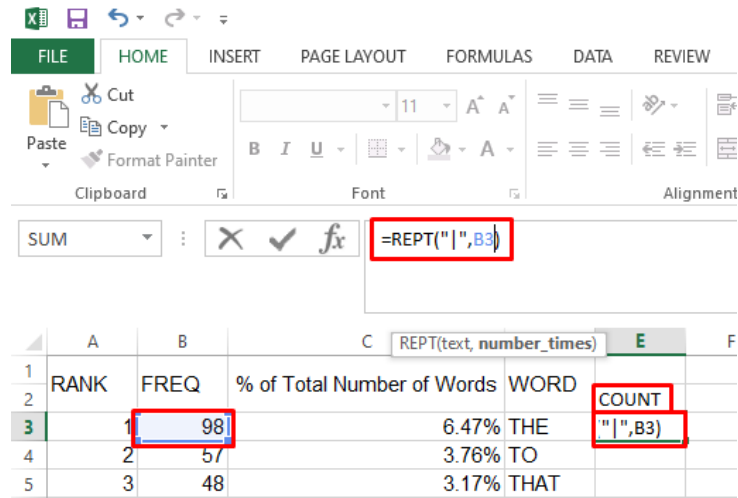
- Delete column D (for simplicity).

The screenshot shows the Microsoft Excel interface. The 'HOME' tab is active, and the 'Merge & Center' button in the Alignment group is highlighted with a red box. Below the ribbon, the formula bar shows 'C1' and the text 'COVERAGE'. The main grid displays a table with the following data:

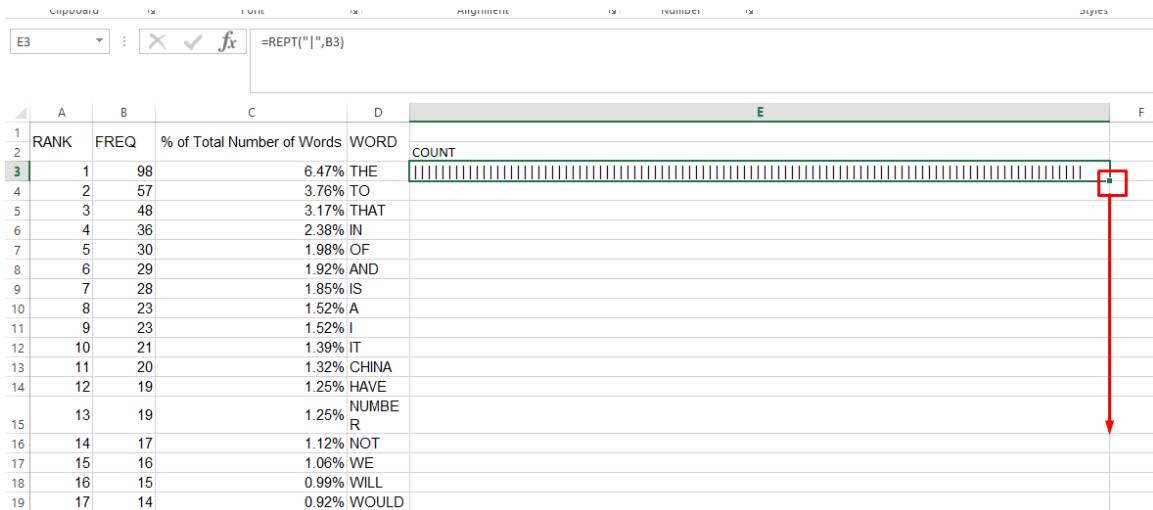
	A	B	C	D	E	F	G	H	I
1			COVERAGE						
2	RANK	FREQ	individ cumulative	WORD					
3	1	98	6.47%	THE					
4	2	57	3.76%	TO					
5	3	48	3.17%	THAT					
6	4	36	2.38%	IN					
7	5	30	1.98%	OF					
8	6	29	1.92%	AND					
9	7	20	1.05%	IS					

- Merge and Center the cell 'Coverage' and 'individ cumulative'.
- Change the title to “% of Total Number of Words”.

STEP 2: CREATING THE COUNT “ | “



- Create a new column called “Count”
- In the first row, type in
 - = REPT (“|”, B3)
- Press Enter.



- It will draw many |||||
- Drag it all the way down.

CONCLUSION

	A	B	C	D	E
1	RANK	FREQ	% of Total Number of Words	WORD	COUNT
2					
3	1	98	6.47%	THE	
4	2	57	3.76%	TO	
5	3	48	3.17%	THAT	
6	4	36	2.38%	IN	
7	5	30	1.98%	OF	
8	6	29	1.92%	AND	
9	7	28	1.85%	IS	
10	8	23	1.52%	A	
11	9	23	1.52%	I	
12	10	21	1.39%	IT	
13	11	20	1.32%	CHINA	
14	12	19	1.25%	HAVE	
15	13	19	1.25%	NUMBER	
16	14	17	1.12%	NOT	
17	15	16	1.06%	WE	
18	16	15	0.99%	WILL	
19	17	14	0.92%	WOULD	
20	18	13	0.86%	THIS	
21	19	12	0.79%	CLIMATE	
22	20	12	0.79%	THEY	
23	21	11	0.73%	ARE	
24	22	11	0.73%	EMISSIONS	

- You will finish with what you see above.

REFERENCES

Bajak, A. (2015). "How to do super simple textual analysis." from <https://www.storybench.org/how-to-do-super-simple-textual-analysis/>.

ABOUT THE AUTHORS

Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at www.AlvinAng.sg.