

AntConc

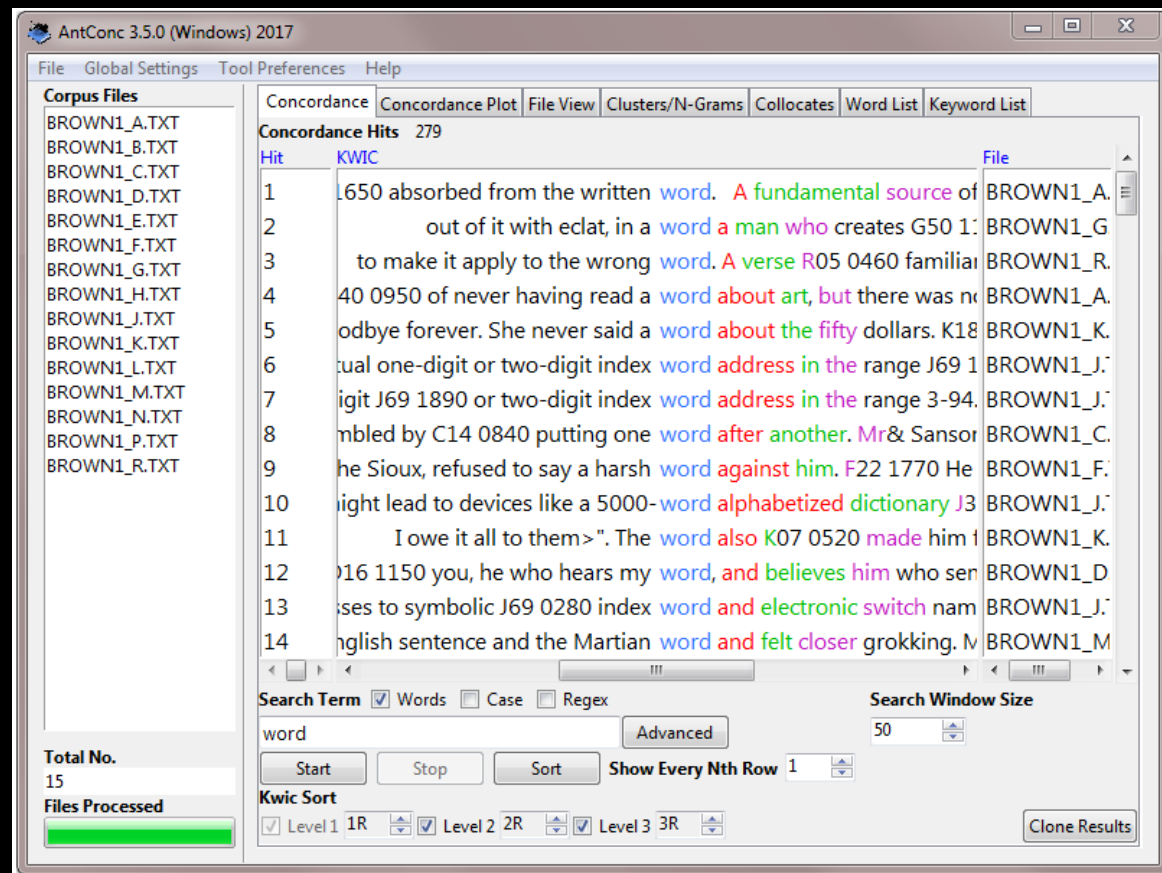
Big Data and Language

Seonmin Park

seonminpark@kaist.ac.kr

AntConc

- AntConc
 - Laurence Anthony, Ph.D
- A freeware corpus analysis toolkit for concordancing and text analysis.



Test with your own text

- <http://www.glozman.com/textpages.html>
 - Harry Potter
 - Lord of the Rings
 - Star Wars
 - ...
- You can also crawl the transcripts from web pages by programming.

What is the frequent words in Harry Potter?

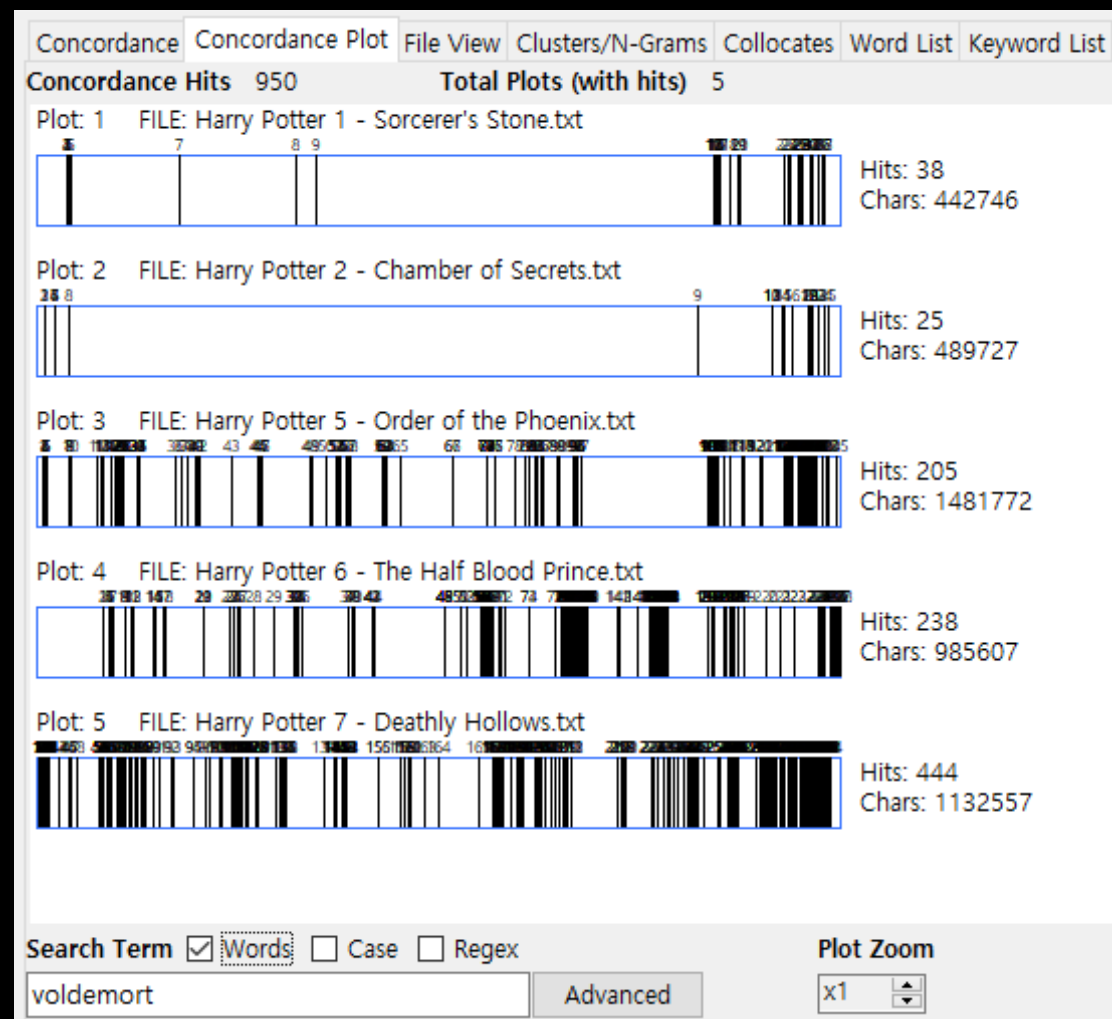
AntConc 3.5.0 (Windows) 2017

File Global Settings Tool Preferences Help

Corpus Files		Concordance	Concordance Plot	File View	Clusters/N
		Word Types: 19780	Word Tokens: 814835		
Rank	Freq	Word			
1	37500	the			
2	19917	and			
3	19634	to			
4	16157	he			
5	15878	of			
6	15345	a			
7	12963	harry			
8	11239	was			
9	10945	you			
10	10734	it			
11	10497	s			
12	10334	said			
13	10224	his			
14	10066	i			
15	9281	in			
16	8142	that			

But we can't find any meaningful result...

Concordance Plot



- Plots where the word “Voldemort” appears.

Collocations with the word “magic”

Concordance		Concordance Plot		File View		Clusters/N-Grams		Collocates		Word List		Keyword List	
Total No. of Collocate Types: 215						Total No. of Collocate Tokens: 934							
Rank	Freq	Freq(L)	Freq(R)	Stat	Collocate								
1	215	215	0	0	of								
2	103	103	0	7.28766	ministry								
3	51	51	0	7.25761	minister								
4	36	36	0	0	use								
5	35	35	0	8.51865	history								
6	24	24	0	0	for								
7	23	23	0	0	to								
8	18	18	0	4.00234	dark								
9	16	16	0	0	do								
10	15	15	0	0	by								
11	13	13	0	0	the								
12	8	8	0	0	s								
13	8	8	0	7.16381	defensive								
14	7	7	0	0	without								
15	7	7	0	0	using								

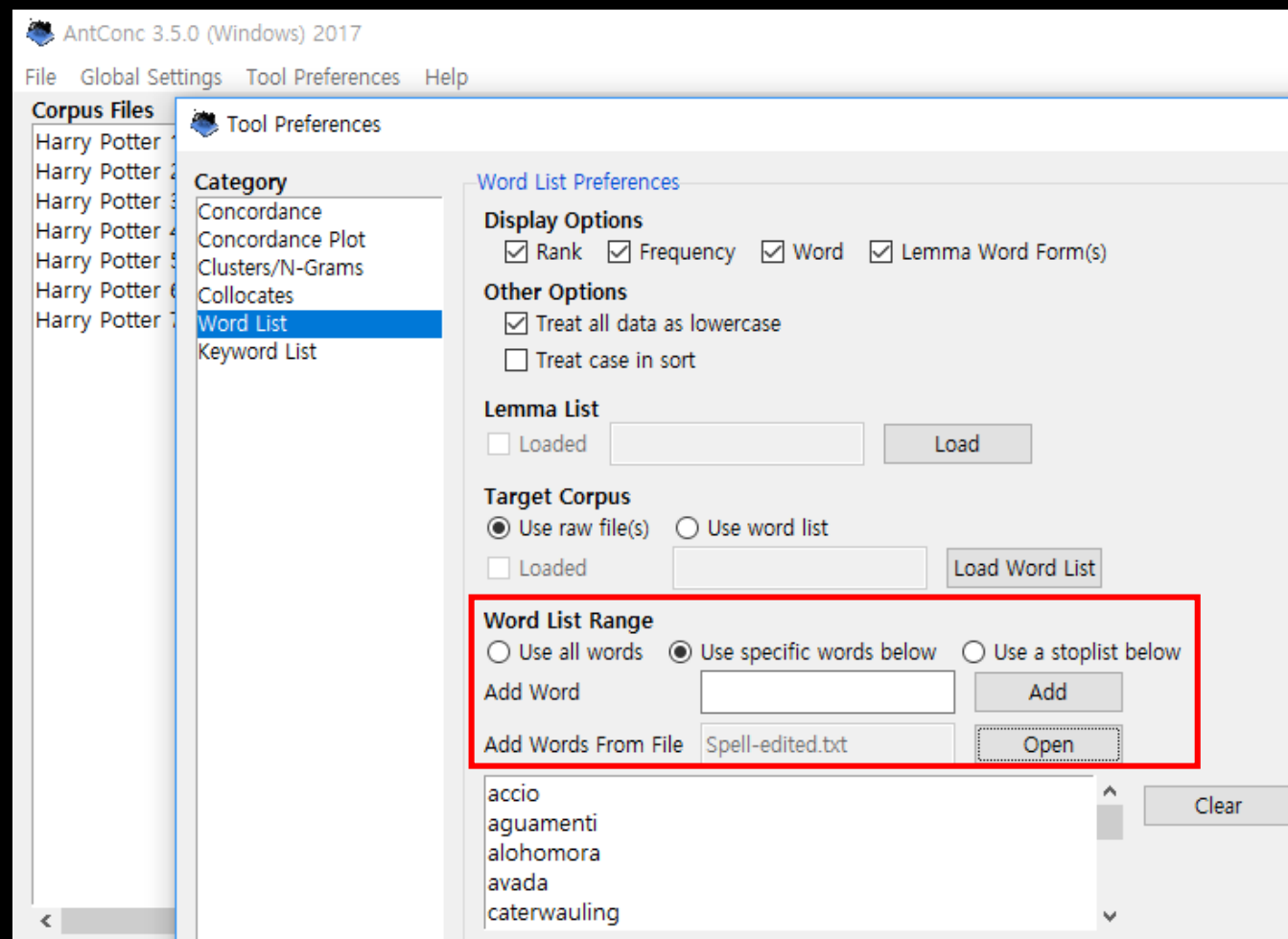
<><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><>&

Ministry of magic,
Minister of magic

Dark magic

Defensive magic

What is the frequent magic spells in Harry Potter?

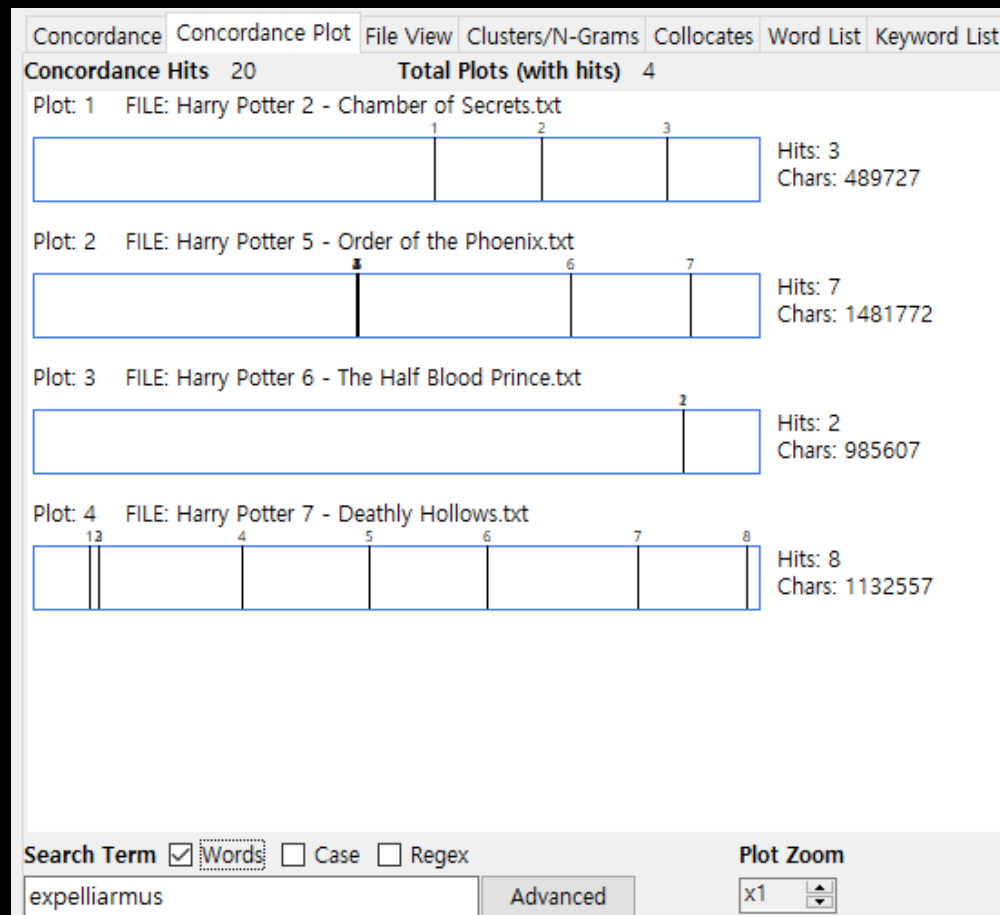


What is the frequent magic spells in Harry Potter?

Concordance	Concordance Plot	File View	Clusters/N
Word Types: 49		Word Tokens: 320	
Rank	Freq	Word	
1	25	accio	
2	20	expelliarmus	
3	18	cheering	
4	17	stupefy	
5	14	avada	Avada kedabra
6	14	lumos	
7	12	protego	
8	11	disillusionment	
9	11	muffliato	
10	10	aguamenti	
11	10	expecto	Expecto patronum
12	10	petrificus	
13	10	crucio	
14	10	reparo	

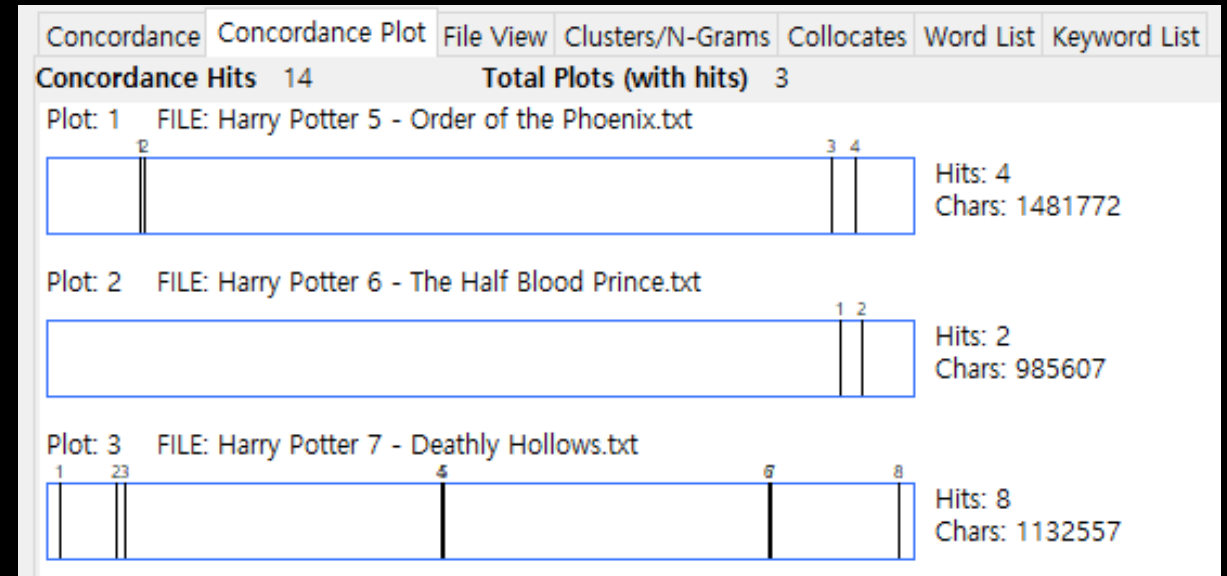
When is the battle scenes?

- Check out the concordance plot of “expelliarmus”

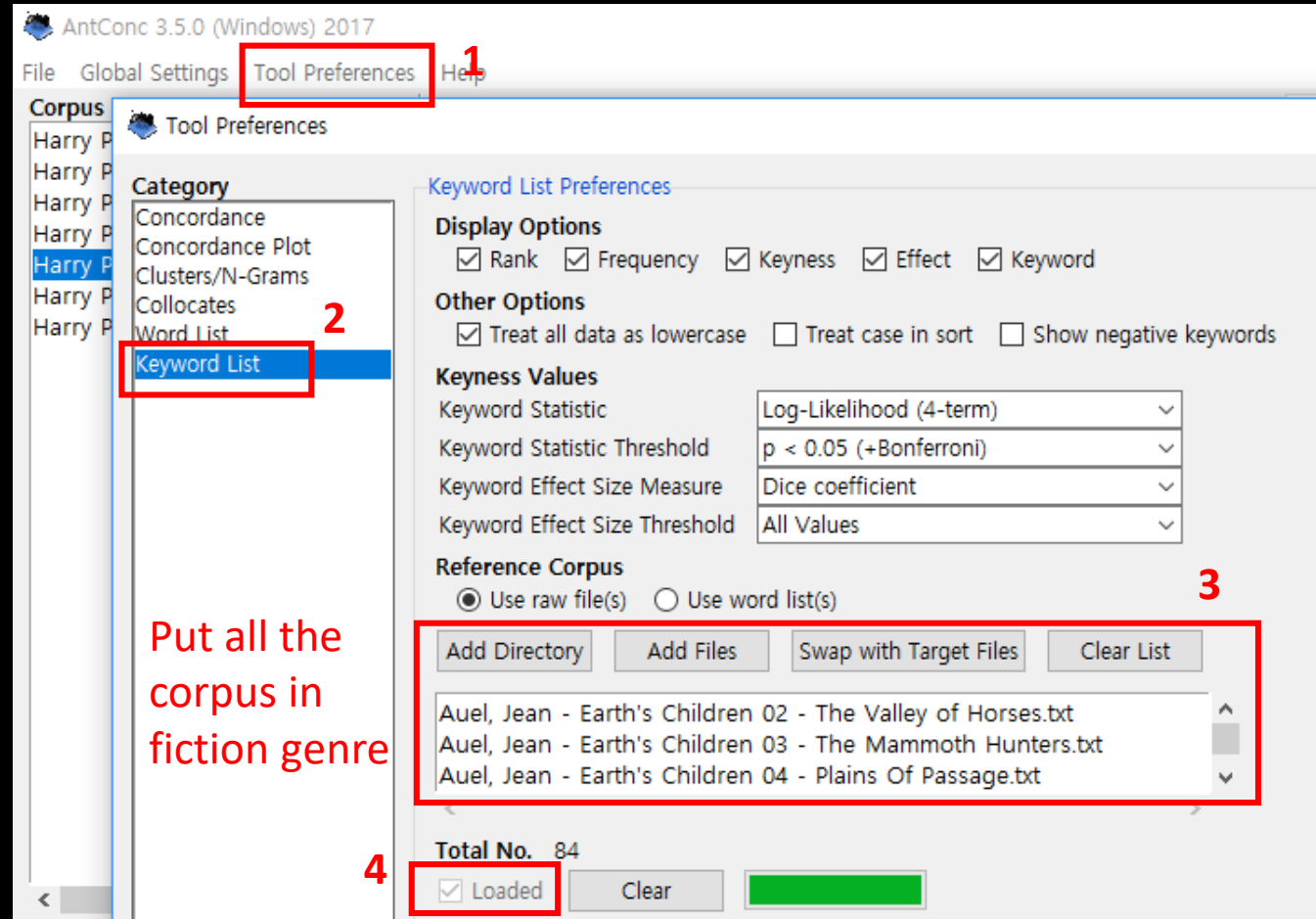


When did the characters killed?

- We may trace the spell “avada kedavra”



Keywords in Harry Potter?



- This process allows to compare the keyness of words in the target novel with general corpus (fiction).

Keywords

Concordance	Concordance Plot	File View	Clusters/N-Grams	Collocates	Word List	Keyword List
Keyword Types: 1247		Keyword Tokens: 133387		Search Hits: 0		
Rank	Freq	Keyness	Effect	Keyword		
1	12963	+ 37597.75	0.0743	harry		
2	4475	+ 12866.56	0.0272	ron		
3	3803	+ 10947.34	0.0232	hermione		
4	2574	+ 7676.94	0.0159	dumbledore		
5	1384	+ 3887.98	0.0086	hagrid		
6	1322	+ 3828.78	0.0082	snape		
7	1329	+ 3626.65	0.0083	professor		
8	1259	+ 3521.63	0.0078	wand		
9	1082	+ 3047.56	0.0067	weasley		
10	999	+ 2957.04	0.0062	malfoy		
11	950	+ 2834.66	0.0059	voldemort		
12	838	+ 2337.58	0.0052	potter		
13	705	+ 2254.34	0.0044	ginny		
14	766	+ 2144.31	0.0048	sirius		
15	631	+ 2090.76	0.004	umbridge		
16	637	+ 1807.77	0.004	fred		
17	584	+ 1722.92	0.0037	neville		
18	594	+ 1669.07	0.0037	hogwarts		
19	546	+ 1546.39	0.0034	george		
20	526	+ 1522.08	0.0033	mcgonagall		

Filter out Proper Nouns

Concordance	Concordance Plot	File View	Clusters/N-Grams	Collocates	Word List	Keyword List
Keyword Types: 1416		Keyword Tokens: 110866		Search Hits: 0		
Rank	Freq	Keyness	Effect	Keyword		
1	1329	+ 3967.97	0.0095	professor		
2	1259	+ 3846.22	0.0091	wand		
3	449	+ 1396.84	0.0033	ministry		
4	476	+ 1270.46	0.0034	school		
5	521	+ 1266.57	0.0038	yeah		
6	319	+ 986.57	0.0023	yeh		
7	292	+ 908.23	0.0021	eaters		
8	259	+ 830.1	0.0019	muggle		
9	319	+ 774.54	0.0023	office		
10	367	+ 763.02	0.0027	uncle		
11	257	+ 652.03	0.0019	students		
12	220	+ 635.01	0.0016	madam		
13	221	+ 587.57	0.0016	potion		
14	1020	+ 570.73	0.0072	door		
15	211	+ 541.26	0.0015	aunt		
16	1053	+ 525.89	0.0074	room		
17	194	+ 520.59	0.0014	dad		
18	213	+ 507.13	0.0015	er		
19	191	+ 470.71	0.0014	owl		
20	146	+ 468.7	0.0011	eater		