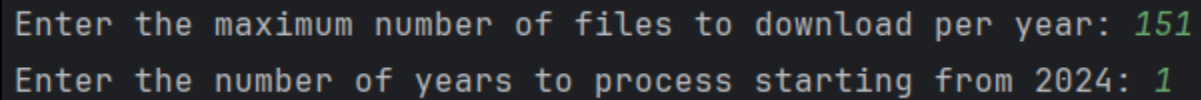
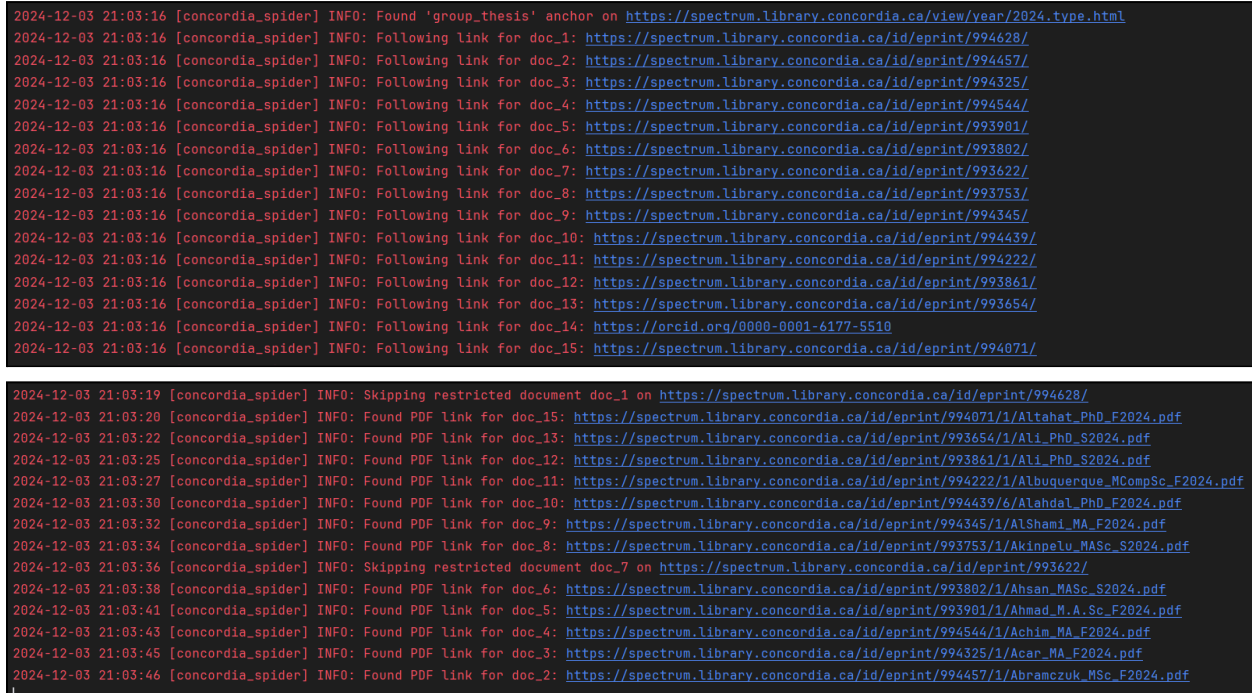


Demo



```
Enter the maximum number of files to download per year: 151
Enter the number of years to process starting from 2024: 1
```

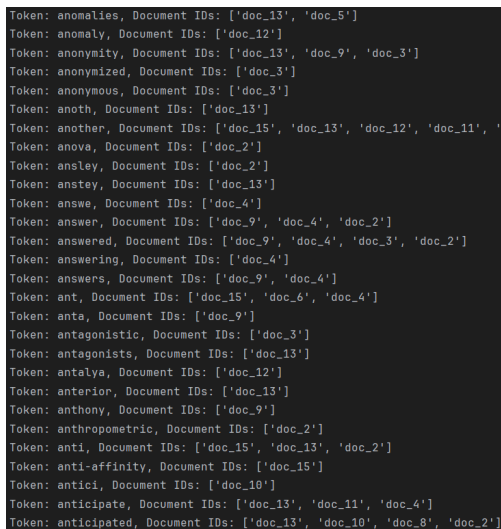
Figure 1: Crawler accepting inputs for the maximum number of pdfs to try and process per year, as well as for the total number of years to go through



```
2024-12-03 21:03:16 [concordia_spider] INFO: Found 'group_thesis' anchor on https://spectrum.library.concordia.ca/view/year/2024.type.html
2024-12-03 21:03:16 [concordia_spider] INFO: Following link for doc_1: https://spectrum.library.concordia.ca/id/eprint/994628/
2024-12-03 21:03:16 [concordia_spider] INFO: Following link for doc_2: https://spectrum.library.concordia.ca/id/eprint/994457/
2024-12-03 21:03:16 [concordia_spider] INFO: Following link for doc_3: https://spectrum.library.concordia.ca/id/eprint/994325/
2024-12-03 21:03:16 [concordia_spider] INFO: Following link for doc_4: https://spectrum.library.concordia.ca/id/eprint/994544/
2024-12-03 21:03:16 [concordia_spider] INFO: Following link for doc_5: https://spectrum.library.concordia.ca/id/eprint/993901/
2024-12-03 21:03:16 [concordia_spider] INFO: Following link for doc_6: https://spectrum.library.concordia.ca/id/eprint/993802/
2024-12-03 21:03:16 [concordia_spider] INFO: Following link for doc_7: https://spectrum.library.concordia.ca/id/eprint/993622/
2024-12-03 21:03:16 [concordia_spider] INFO: Following link for doc_8: https://spectrum.library.concordia.ca/id/eprint/993753/
2024-12-03 21:03:16 [concordia_spider] INFO: Following link for doc_9: https://spectrum.library.concordia.ca/id/eprint/994345/
2024-12-03 21:03:16 [concordia_spider] INFO: Following link for doc_10: https://spectrum.library.concordia.ca/id/eprint/994439/
2024-12-03 21:03:16 [concordia_spider] INFO: Following link for doc_11: https://spectrum.library.concordia.ca/id/eprint/994222/
2024-12-03 21:03:16 [concordia_spider] INFO: Following link for doc_12: https://spectrum.library.concordia.ca/id/eprint/993861/
2024-12-03 21:03:16 [concordia_spider] INFO: Following link for doc_13: https://spectrum.library.concordia.ca/id/eprint/993654/
2024-12-03 21:03:16 [concordia_spider] INFO: Following link for doc_14: https://orcid.org/0000-0001-6177-5510
2024-12-03 21:03:16 [concordia_spider] INFO: Following link for doc_15: https://spectrum.library.concordia.ca/id/eprint/994071/

2024-12-03 21:03:19 [concordia_spider] INFO: Skipping restricted document doc_1 on https://spectrum.library.concordia.ca/id/eprint/994628/
2024-12-03 21:03:20 [concordia_spider] INFO: Found PDF link for doc_15: https://spectrum.library.concordia.ca/id/eprint/994071/1/Altahat_PhD_F2024.pdf
2024-12-03 21:03:22 [concordia_spider] INFO: Found PDF link for doc_13: https://spectrum.library.concordia.ca/id/eprint/993654/1/Alf_PhD_S2024.pdf
2024-12-03 21:03:25 [concordia_spider] INFO: Found PDF link for doc_12: https://spectrum.library.concordia.ca/id/eprint/993861/1/Alf_PhD_S2024.pdf
2024-12-03 21:03:27 [concordia_spider] INFO: Found PDF link for doc_11: https://spectrum.library.concordia.ca/id/eprint/994222/1/Albuquerque_MCompSc_F2024.pdf
2024-12-03 21:03:30 [concordia_spider] INFO: Found PDF link for doc_10: https://spectrum.library.concordia.ca/id/eprint/994439/6/Alahdel_PhD_F2024.pdf
2024-12-03 21:03:32 [concordia_spider] INFO: Found PDF link for doc_9: https://spectrum.library.concordia.ca/id/eprint/994345/1/AlShami_MA_F2024.pdf
2024-12-03 21:03:34 [concordia_spider] INFO: Found PDF link for doc_8: https://spectrum.library.concordia.ca/id/eprint/993753/1/Akinpelu_MASc_S2024.pdf
2024-12-03 21:03:36 [concordia_spider] INFO: Skipping restricted document doc_7 on https://spectrum.library.concordia.ca/id/eprint/993622/
2024-12-03 21:03:38 [concordia_spider] INFO: Found PDF link for doc_6: https://spectrum.library.concordia.ca/id/eprint/993802/1/Ahsan_MASc_S2024.pdf
2024-12-03 21:03:41 [concordia_spider] INFO: Found PDF link for doc_5: https://spectrum.library.concordia.ca/id/eprint/993901/1/Ahmad_M.A.Sc_F2024.pdf
2024-12-03 21:03:43 [concordia_spider] INFO: Found PDF link for doc_4: https://spectrum.library.concordia.ca/id/eprint/994544/1/Achim_MA_F2024.pdf
2024-12-03 21:03:45 [concordia_spider] INFO: Found PDF link for doc_3: https://spectrum.library.concordia.ca/id/eprint/994325/1/Acar_MA_F2024.pdf
2024-12-03 21:03:46 [concordia_spider] INFO: Found PDF link for doc_2: https://spectrum.library.concordia.ca/id/eprint/994457/1/Abramczuk_MSc_F2024.pdf
```

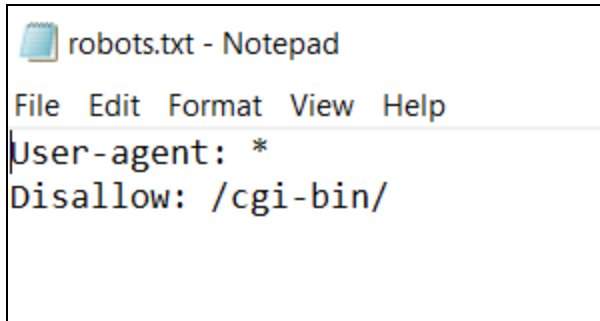
Figure 2: Process of finding x amount of files and their respective urls, as well as skipping any urls that lead to a /cgi redirect



```
Token: anomalies, Document IDs: ['doc_13', 'doc_5']
Token: anomaly, Document IDs: ['doc_12']
Token: anonymity, Document IDs: ['doc_13', 'doc_9', 'doc_3']
Token: anonymized, Document IDs: ['doc_3']
Token: anonymous, Document IDs: ['doc_3']
Token: anoth, Document IDs: ['doc_13']
Token: another, Document IDs: ['doc_15', 'doc_13', 'doc_12', 'doc_11', 'doc_10', 'doc_9', 'doc_8', 'doc_6', 'doc_5', 'doc_4', 'doc_3', 'doc_2']
Token: anova, Document IDs: ['doc_2']
Token: ansley, Document IDs: ['doc_2']
Token: anstey, Document IDs: ['doc_13']
Token: answe, Document IDs: ['doc_4']
Token: answer, Document IDs: ['doc_9', 'doc_4', 'doc_2']
Token: answered, Document IDs: ['doc_9', 'doc_4', 'doc_3', 'doc_2']
Token: answering, Document IDs: ['doc_4']
Token: answers, Document IDs: ['doc_9', 'doc_4']
Token: ant, Document IDs: ['doc_15', 'doc_6', 'doc_4']
Token: anta, Document IDs: ['doc_9']
Token: antagonistic, Document IDs: ['doc_3']
Token: antagonists, Document IDs: ['doc_13']
Token: antalya, Document IDs: ['doc_12']
Token: anterior, Document IDs: ['doc_13']
Token: anthony, Document IDs: ['doc_9']
Token: anthropometric, Document IDs: ['doc_2']
Token: anti, Document IDs: ['doc_15', 'doc_13', 'doc_2']
Token: anti-affinity, Document IDs: ['doc_15']
Token: antici, Document IDs: ['doc_10']
Token: anticipate, Document IDs: ['doc_13', 'doc_11', 'doc_4']
Token: anticipated, Document IDs: ['doc_13', 'doc_10', 'doc_8', 'doc_2']
```

```
Token: zigzag, Document IDs: ['doc_10']
Token: zijl, Document IDs: ['doc_13']
Token: zilles, Document IDs: ['doc_13']
Token: ziman, Document IDs: ['doc_5']
Token: zimmer, Document IDs: ['doc_9']
Token: zimmermann, Document IDs: ['doc_13']
Token: zinterhof, Document IDs: ['doc_15']
Token: zipped, Document IDs: ['doc_9']
Token: zis, Document IDs: ['doc_15', 'doc_11']
Token: zisserman, Document IDs: ['doc_12', 'doc_6']
Token: zlokovic, Document IDs: ['doc_13']
Token: zoe, Document IDs: ['doc_4']
Token: zonation, Document IDs: ['doc_13']
Token: zone, Document IDs: ['doc_13', 'doc_10']
Token: zones, Document IDs: ['doc_10']
Token: zontal, Document IDs: ['doc_10']
Token: zoom, Document IDs: ['doc_9', 'doc_2']
Token: zoomed, Document IDs: ['doc_13']
Token: zooming, Document IDs: ['doc_13']
Token: zou, Document IDs: ['doc_6']
Token: zuend, Document IDs: ['doc_13']
Token: zuo, Document IDs: ['doc_13', 'doc_6']
Token: zur, Document IDs: ['doc_5']
Token: zurich, Document IDs: ['doc_12']
Token: zurlo, Document IDs: ['doc_2']
```

Figure 3: Example of the inverted index containing all of the tokens, ordered alphabetically, obtained from the valid pdf links



```
robots.txt - Notepad
File Edit Format View Help
User-agent: *
Disallow: /cgi-bin/
```

Figure 4: robots.txt file that restricts redirects to /cgi. An example of this happening when a pdf link redirects to a login page.robots.txt disallows it.

```
Clustering Analysis with k=4 (Faculties)
=====
Number of documents: 120
TF-IDF matrix shape: (120, 108626)
Overall Silhouette Score: 0.565

Cluster Metrics:
-----

Cluster 0:
Size: 47 documents
Average Silhouette Score: 0.686

Cluster 1:
Size: 29 documents
Average Silhouette Score: 0.571

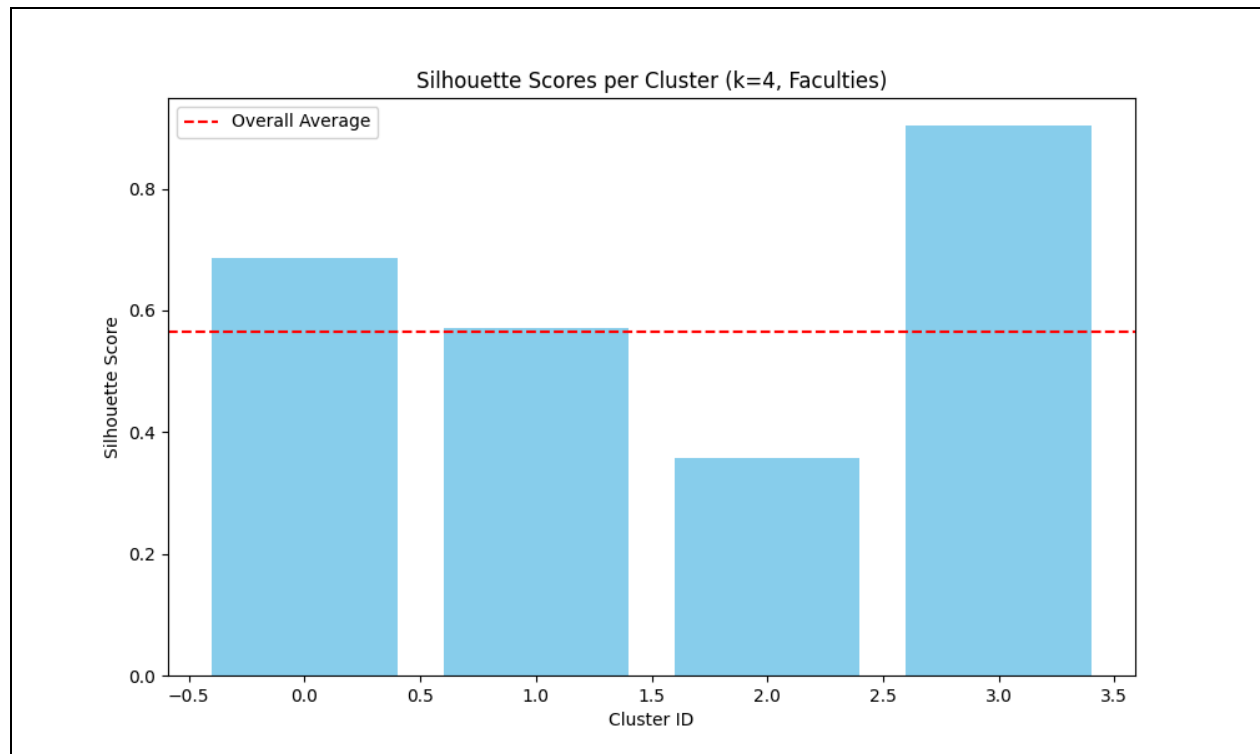
Cluster 2:
Size: 38 documents
Average Silhouette Score: 0.357

Cluster 3:
Size: 6 documents
Average Silhouette Score: 0.904
```

Figure 5: Clustering analysis results with 4 clusters (4 faculties). A total of 120(out of 151) successful pdf links were processed.Total documents and the silhouette score is calculated for each individual cluster as well as the entire clustering to allow for comparison within and to other clusterings.

Cluster 3 Detailed Analysis:			
Documents: ['doc_103', 'doc_111', 'doc_137', 'doc_32', 'doc_53', 'doc_91']			
Top 50 Most Informative Terms:			
Rank	Term	TF-IDF Score	TF-IDF Rank
1	masculins	0.0436	1
2	rche	0.0396	2
3	tissage	0.0374	3
4	fqs	0.0364	4
5	ekwuyasi	0.0348	5
6	minitaire	0.0348	6
7	camion	0.0321	7
8	bacon-villemaire	0.0318	8
9	co-directrice	0.0318	9
10	guytaine	0.0318	10
11	charles-antoine	0.0318	11
12	trauma-informed	0.0318	12
13	timt	0.0318	13
14	recenser	0.0318	14
15	existerait	0.0318	15
16	reconnaissances	0.0318	16
17	chaleureusement	0.0318	17
18	apportant	0.0318	18
19	infiniment	0.0318	19
20	interpersonnel	0.0318	20
21	stresseurs	0.0318	21
22	interpersonnels	0.0318	22
23	intrapersonnels	0.0318	23
24	gardiens	0.0318	24
25	hussey	0.0318	25
26	sensorimoteur	0.0318	26
27	comportementale	0.0318	27
28	post-traumatique	0.0318	28
29	spt	0.0318	29
30	milot	0.0318	30
31	kendall-tackett	0.0318	31
32	insuffisant	0.0318	32
33	rencontraient	0.0318	33
34	comportementaux	0.0318	34
35	relationnelle	0.0318	35
36	parentale	0.0318	36
37	cliniques	0.0318	37
38	soutenant	0.0318	38
39	levenson	0.0318	39
40	soldats	0.0318	40
41	humaniste	0.0318	41
42	nctsn	0.0318	42
43	saine	0.0318	43
44	bienveillante	0.0318	44
45	communicatif	0.0318	45
46	rubrique	0.0318	46
47	enracine	0.0318	47
48	graduellement	0.0318	48
49	multidisciplinaires	0.0318	49
50	communautaires	0.0318	50

Figure 6: Good example of top 50 vocabulary terms ranked by tf-idf for a specific cluster. In this case, Cluster_3 is the 4th cluster of the clustering in **Figure 5**. Contains 6 different documents, with a good silhouette score of 0.904.



Graph 1: Visualization of the silhouette score of each cluster in faculty clustering $k=4$. Cluster_3, as shown in **Figure 6**, indicates the best results.

Clustering Analysis with k=38 (Departments)	
=====	
Number of documents: 120	
TF-IDF matrix shape: (120, 108626)	
Overall Silhouette Score: 0.144	
Cluster Metrics:	

Cluster 0:	Cluster 30:
Size: 3 documents	Size: 4 documents
Average Silhouette Score: 0.301	Average Silhouette Score: -0.053
Cluster 1:	Cluster 31:
Size: 4 documents	Size: 3 documents
Average Silhouette Score: 0.171	Average Silhouette Score: -0.010
Cluster 2:	Cluster 32:
Size: 5 documents	Size: 2 documents
Average Silhouette Score: 0.018	Average Silhouette Score: 0.385
Cluster 3:	Cluster 33:
Size: 2 documents	Size: 2 documents
Average Silhouette Score: 0.047	Average Silhouette Score: 0.507
Cluster 4:	Cluster 34:
Size: 6 documents	Size: 1 documents
Average Silhouette Score: 0.450	Average Silhouette Score: 0.000
Cluster 5:	Cluster 35:
Size: 6 documents	Size: 1 documents
Average Silhouette Score: 0.098	Average Silhouette Score: 0.000
Cluster 6:	Cluster 36:
Size: 8 documents	Size: 2 documents
Average Silhouette Score: 0.066	Average Silhouette Score: 0.144
Cluster 7:	Cluster 37:
Size: 5 documents	Size: 1 documents
Average Silhouette Score: 0.078	Average Silhouette Score: 0.000

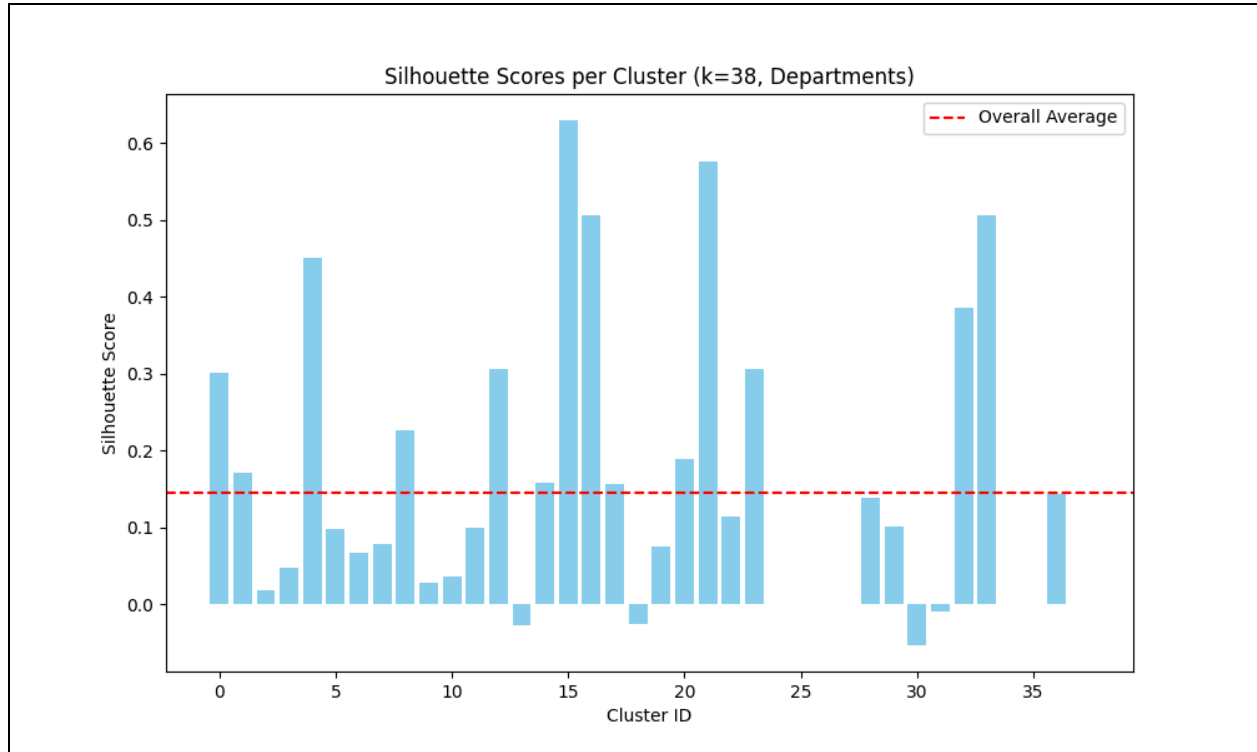
Figure 7: Clustering analysis results with 38 clusters (38 departments). The full clusters list is in its respective .txt file.

Cluster 30 Detailed Analysis:				

Documents: ['doc_124', 'doc_125', 'doc_18', 'doc_67']				
Top 50 Most Informative Terms:				
Rank	Term	TF-IDF Score	TF-IDF	Rank

1	originator	0.0596		1
2	fjs	0.0533		2
3	lms	0.0491		3
4	indexed	0.0457		4
5	dielectric	0.0414		5
6	jingyu	0.0414		6
7	yumin	0.0414		7
8	free-energy	0.0414		8
9	time-independent		0.0414	9
10	electromechanical		0.0414	10
11	viscoelast	0.0414		11
12	viscoelasticity	0.0414		12
13	yawu	0.0414		13
14	jundong	0.0414		14
15	qun	0.0414		15
16	yaowei	0.0414		16
17	hesis	0.0414		17
18	dielec	0.0414		18
19	actua	0.0414		19
20	preparat	0.0414		20
21	ipmc	0.0414		21
22	voigt	0.0414		22
23	posse	0.0414		23
24	ielectric	0.0414		24
25	degs	0.0414		25
26	ectro	0.0414		26
27	vhb	0.0414		27
28	polyacrylate	0.0414		28
29	polydimethylsiloxane		0.0414	29
30	elastomers	0.0414		30
31	lce	0.0414		31
32	biocompatibility		0.0414	32
33	deform	0.0414		33
34	rding	0.0414		34
35	to-electrical	0.0414		35
36	silicones	0.0414		36
37	polyurethanes	0.0414		37
38	readi	0.0414		38
39	thickened	0.0414		39
40	intricated	0.0414		40
41	non-adhesive	0.0414		41
42	sandwiched	0.0414		42
43	incompressible	0.0414		43
44	lume	0.0414		44
45	bionic	0.0414		45
46	overcom	0.0414		46
47	flapping	0.0414		47
48	seabed	0.0414		48
49	flutter	0.0414		49
50	jellyfish	0.0414		50

Figure 8: Bad example of top 50 vocabulary terms ranked by tf-idf for a specific cluster. In this case, Cluster_30 is the 31st cluster of the clustering in **Figure 7**. Contains 4 different documents, with a bad silhouette score of -0.053.



Graph 2: Visualization of the silhouette score of each cluster in faculty clustering $k=38$. Cluster_30, as shown in **Figure 8**, indicates the worst results. There are several clusters with scores of 0.00.

Clustering Analysis with $k=3$ (Faculties)

```
=====
Number of documents: 120
TF-IDF matrix shape: (120, 108626)
Overall Silhouette Score: 0.671
```

Clustering Analysis with $k=6$ (Departments)

```
=====
Number of documents: 120
TF-IDF matrix shape: (120, 108626)
Overall Silhouette Score: 0.400
```

Figure 9: Clustering analysis with $k=3$ and $k=6$. All textual results and visualization of the silhouette scores can be found in the current directory.

1. Cluster 15: **The Concordia Institute for Information and Systems Engineering**
 - Size: 2 documents
 - Average Silhouette Score: 0.629
2. Cluster 21: **Civil Engineering**
 - Size: 2 documents
 - Average Silhouette Score: 0.575
3. Cluster 33: **Electrical and Computer Engineering**
 - Size: 2 documents
 - Average Silhouette Score: 0.507
4. Cluster 16: **Health, Kinesiology, and Applied Physiology**
 - Size: 2 documents
 - Average Silhouette Score: 0.506
5. Cluster 04: **French**
 - Size: 6 documents
 - Average Silhouette Score: 0.450
6. Cluster 32: **Marketing**
 - Size: 2 documents
 - Average Silhouette Score: 0.385
7. Cluster 23: **Biology**
 - Size: 3 documents
 - Average Silhouette Score: 0.307
8. Cluster 12: **Creative Arts Therapies**
 - Size: 3 documents
 - Average Silhouette Score: 0.305
9. Cluster 00: **Doctor of Philosophy (Business Administration)**
 - Size: 3 documents
 - Average Silhouette Score: 0.301
10. Cluster 08: **3 different types of Engineering**
 - Size: 3 documents
 - Average Silhouette Score: 0.227

Figure 10: 'Name' (departments) of top 10 clusters in k=38 clustering.

1. Cluster 03: **French**
 - Size: 6 documents
 - Average Silhouette Score: 0.904
2. Cluster 00: **Gina Cody School of Engineering**
 - Size: 47 documents
 - Average Silhouette Score: 0.686
3. Cluster 01: **Faculty of Arts and Science**
 - Size: 29 documents
 - Average Silhouette Score: 0.571
4. Cluster 02: **John Molson/Fine Arts**
 - Size: 38 documents
 - Average Silhouette Score: 0.357

Figure 11: 'Name' (faculty) of top 10 clusters in k=4 clustering.