Demo for Project 4 of COMP 479 by Silas Kalinowski ID: 40256077

Overview

The program can be started by running the main.py file in a Python 3.10 environment. For the program to work all modules and all input files need to be in the local directory. The required dependencies are named in the requirements.txt. The output files are saved in the local directory and are named in the following manner:

20_most_informative_word_3.txt contains the 20 most informative words with clustering with k=3

20_most_informative_word_6.txt contains the 20 most informative words with clustering with k=6

Files/numofpages_urlwebpage.html for the downloaded web pages.

Filestxt/numofpages_urlwebpage.txt for the text of the web pages.

Doc_name_sentiment_label_k.txt for the name of the documents with sentiment scores and label

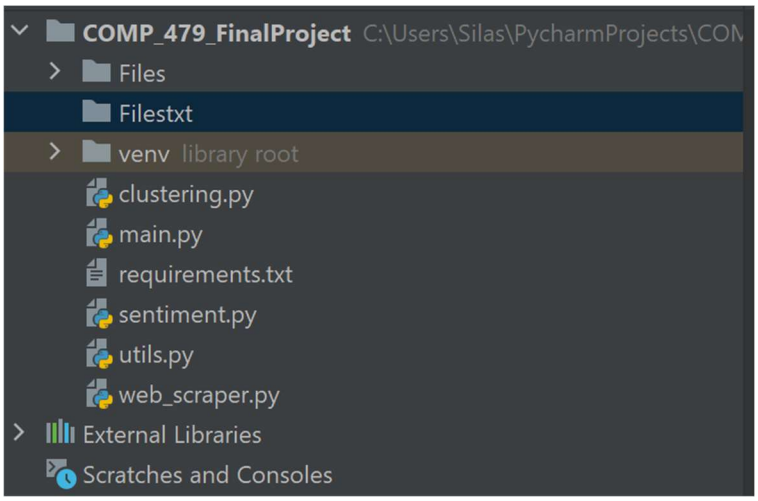
Sentiment_of_cluster_k.txt for the different methods of scoring the sentiment of the clusters.

Running the main file saves the text from the html files in the Filestxt directory and performs clustering with k=3 and k=6. To convert the documents into vectors the tfidf vectorizer is used. With calling the read_txt_and_tfidf function with the input path and False, True the texts are not stemmed but stopwords are removed. Also, the cluster assessment function is ran which plots and prints the different measures for assessment. Lastly, the sentiment analysis is performed, and the results are plotted and printed. The clustering results should be the same as the clustering results I talked about in the report.

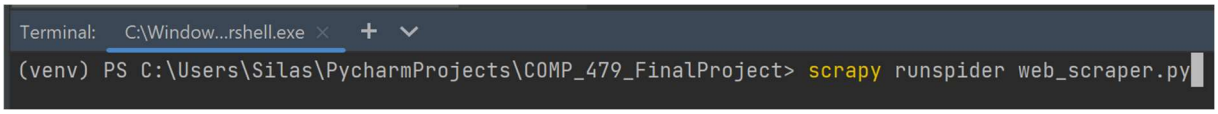
If the user wants to start the crawler and not use the web pages I used for the report, the user needs to open the Terminal and type scrapy runspider web_scraper.py. The output directory can be changed in the parse method in the open statement. By default, the documents are output to the Files directory. I recommend creating a new directory because if not some documents of the old documents in the Files directory might be replaced. Also, the number of pages to downloaded, the download delay, the start url, and the allowed domains can be changed in the attributes of the query.

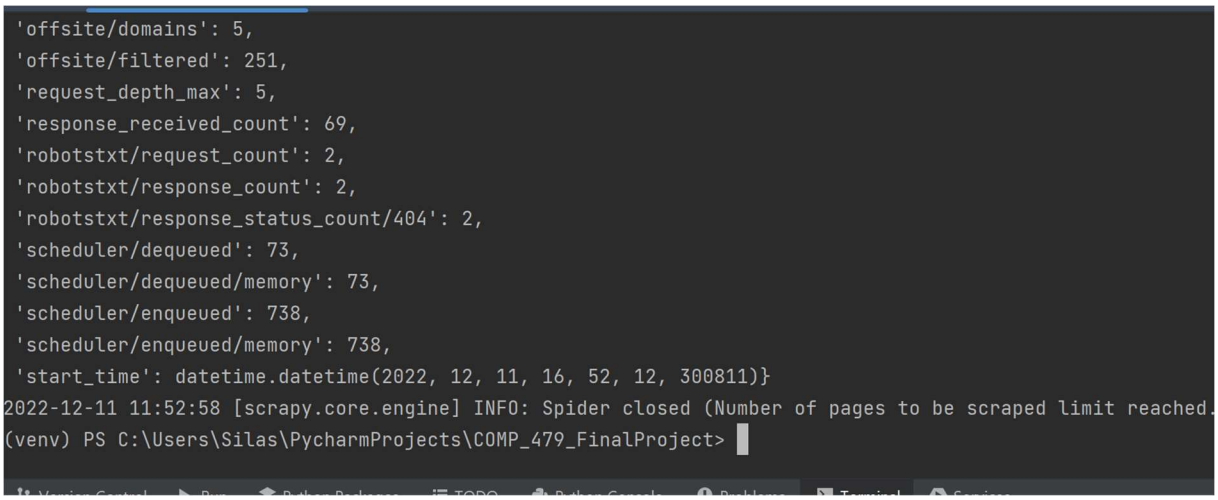
Walkthrough of the program

Project Directory before running the main file



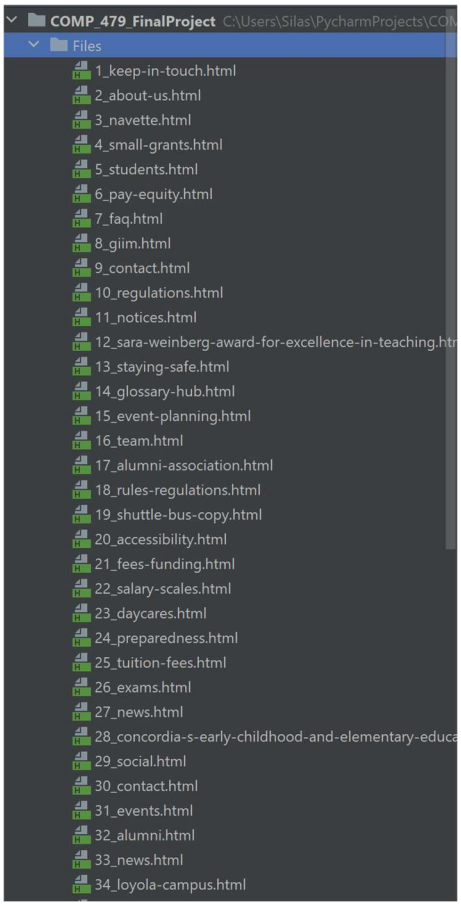
To start the web crawler use the following command in the terminal





After downloading the number of pages that has been set before the spider is closed.

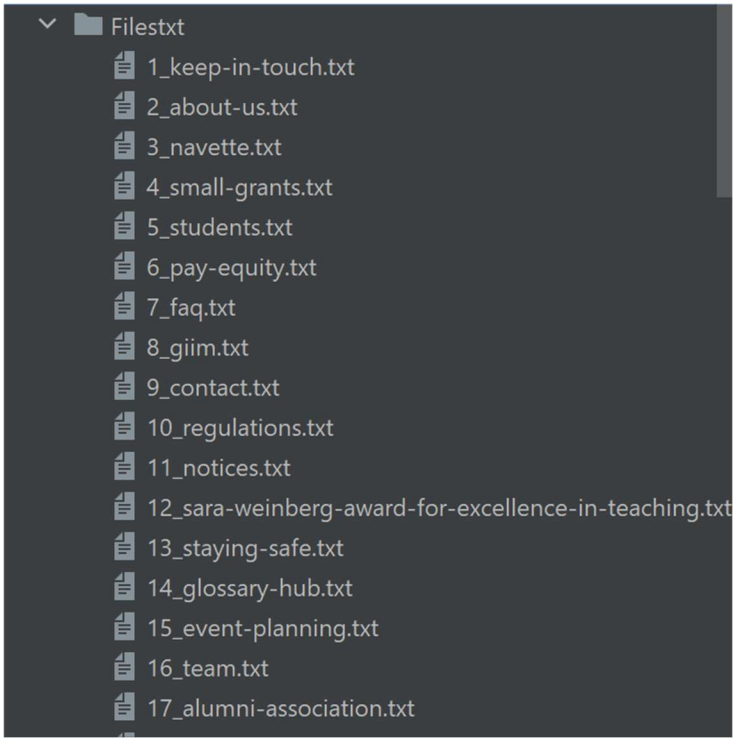
The documents are named in the following manner: number of pages left at the time the page was downloaded_url.html in the Files directory.



Running the save_text_from_html saves the text from the downloaded documents to the Filestxt directory.

```
save_text_from_html('Files', 'Filestxt')
```

The text of the web pages are named in the same manner as the web pages.




```
matrix, vec, texts, doc_name = read_txt_and_tfidf('Filestxt', False, True)
```

The next command vectorizes the documents. The last two attributes of the function define first if the text is stemmed and second if stop words are removed.

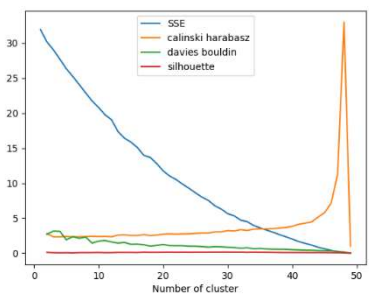
```
labels = kmeans_on_text(matrix, vec, 3, True)
```

The kmeans_on text function performs the clustering and returns the label for other functions. It also prints and saves the top 20 most informative words for every cluster if the last attribute is set to True.

```
cluster_assessment(matrix, 50)
```

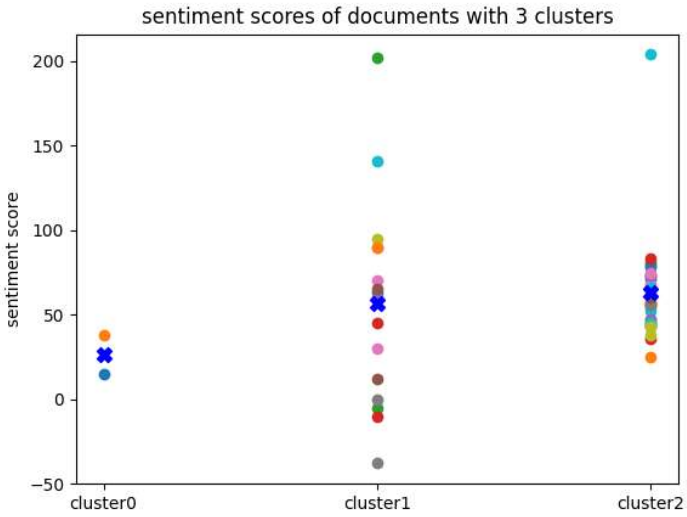
The cluster assessment function prints and plots the different measures to assess the clustering. The following output was returned for my example run:

```
3 clusters silhouette score 0.06720832838874481
3 clusters calinski harabasz score 2.332113941136717
3 clusters davies bouldin score 3.1515640315908846
3 clusters SSE 29.03665849863242
6 clusters silhouette score 0.0392320091892018
6 clusters calinski harabasz score 2.3288478729353463
6 clusters davies bouldin score 2.3544904207045887
6 clusters SSE 25.238944494148253
```



The score function performs the sentiment analysis and plots and prints the result.

```
score(texts, labels, 3, doc_name)
```



A plot with all the sentiment score for each document grouped by cluster. The mean of the sentiment is marked as a blue cross.

	name	scores	label
0	10_regulations.txt	15.0	0
1	26_exams.txt	38.0	0
2	14_glossary-hub.txt	202.0	1
3	16_team.txt	45.0	1
4	17_alumni-association.txt	90.0	1
5	18_rules-regulations.txt	12.0	1
6	22_salary-scales.txt	70.0	1
7	24_preparedness.txt	-38.0	1

A table with the name of the document the sentiment score and the cluster label.

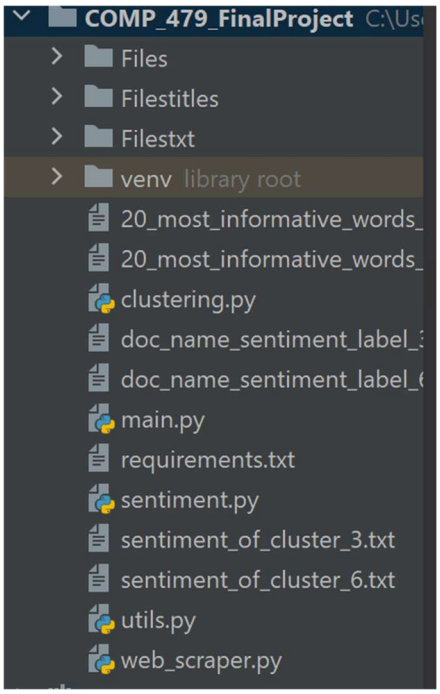
	cluster	min	max	median	mean
0	0	15.0	38.0	26.5	26.5
1	1	-38.0	202.0	61.5	56.9
2	2	25.0	204.0	58.5	63.0

A table with the different possible sentiment scoring techniques for the clusters.

```
Top 5 documents with highest sentiment
12_sara-weinberg-award-for-excellence-in-teaching.txt 204.0
14_glossary-hub.txt 202.0
31_events.txt 141.0
28_concordia-s-early-childhood-and-elementary-education-program-celebrates-50-years-of-innovation.txt 95.0
17_alumni-association.txt 90.0
Top 5 documents with lowest sentiment
24_preparedness.txt -38.0
3_navette.txt -10.0
39_security.txt -5.0
8_giim.txt 0.0
18_rules-regulations.txt 12.0
```

Lastly the top 5 documents with the lowest/highest sentiment are printed. These prints are also saved as a txt with the following names: Doc_name_sentiment_label_k.txt for the name of the documents with sentiment scores and label, Sentiment_of_cluster_k.txt for the different methods of scoring the sentiment of the clusters.

The directory after running the program.



All functions with the prints and plots are returned for k=3 and k=6 when running the main.

Good and bad examples of clustering

As mentioned in the Report I tried clustering with just the title of each document. The 20 most influential words were helpful if you want to name the clusters. But overall, the clustering with K=6 was a bad example for clustering: cluster 0, cluster 3, cluster 4 and cluster 5 only contain one document. And the other clusters don't seem to be very meaningful. (there are only 47 documents because some of the documents don't have a title in the hmtl):

	name	scores	label
0	experiential-learning.txt	0.0	0
1	commuting-parking-facilities.txt	0.0	1
2	health-wellness.txt	0.0	1
3	housing-food.txt	0.0	1
4	it-support-software.txt	2.0	1
5	rights-responsibilities.txt	0.0	1
6	safety-security.txt	1.0	1
7	security.txt	0.0	1
8	services.txt	0.0	1
9	sports-fitness-recreation.txt	1.0	1
10	stories.txt	0.0	1
11	tuition.txt	0.0	1
12	about.txt	0.0	2
13	academic-dates.txt	0.0	2
14	academic.txt	0.0	2
15	apply.txt	0.0	2
16	behavioural-integrity.txt	2.0	2
17	birks.txt	0.0	2
18	calendar.txt	0.0	2
19	cancelled-classes.txt	-1.0	2
20	career-planning.txt	0.0	2
21	course-availability.txt	0.0	2
22	current.txt	0.0	2
23	events.txt	0.0	2
24	experts.txt	0.0	2
25	filmshoots.txt	0.0	2
26	funding-research-opportunities.txt	2.0	2
27	industrial-engineering-phd.txt	0.0	2
28	it.txt	0.0	2
29	jmsb.txt	0.0	2
30	life.txt	0.0	2
31	mental-health.txt	0.0	2
32	openings.txt	2.0	2
33	playlist.txt	0.0	2
34	registration.txt	0.0	2
35	reussir-en-francais.txt	0.0	2
36	sgs.txt	0.0	2
37	students.txt	0.0	2
38	team.txt	0.0	2
39	territorial-acknowledgement.txt	0.0	2
40	training.txt	0.0	2
41	udemy.txt	0.0	2
42	units.txt	0.0	2
43	warning-about-covid-19-related-cybersecurity-t...	-5.0	2
44	concordias-next-generation-cities-institute-we...	2.0	3
45	open-a-ticket.txt	2.0	4
46	ginacody.txt	0.0	5

The clustering with stemming and the removal of stop words is an example for good clustering and also the clusterings in the report are both positive examples. The k=3 cluster has some negative aspects: the cluster 1 contains most of the documents but except for that it gives good insight. The k=6 clustering is even better. Every cluster contains documents that talk about related topics and the clusters are mostly small and separate the documents well. The documents names with labels are shown below.

K=3

	name	scores	label
0	1_comment-sollicitier.txt	-11.0	0
1	21_emplois.txt	-7.0	0
2	2_recrutement-etudiants.txt	-10.0	0
3	31_traductologie-ma.txt	55.0	0
4	3_accessibilite.txt	-7.0	0
5	4_confidentialite.txt	-13.0	0
6	10_associate-dean-faculty-relations-and-inclus...	71.0	1
7	11_rationale.txt	95.0	1
8	12_methodology.txt	69.0	1
9	13_strategic-plan-committee.txt	51.0	1
10	14_milestones.txt	61.0	1
11	15_deadlines.txt	44.0	1
12	16_apply-now.txt	42.0	1
13	17_facilities.txt	150.0	1
14	18_broken-link.txt	43.0	1
15	19_accessibility.txt	45.0	1
16	20_open-a-ticket.txt	48.0	1
17	23_rotten-meat-could-be-easier-to-detect-thank...	40.0	1
18	25_year-2-of-concordias-sustainability-action-...	141.0	1
19	26_concordia-s-early-childhood-and-elementary-...	95.0	1
20	27_contact.txt	57.0	1
21	28_life-in-montreal.txt	109.0	1
22	29_jobs.txt	86.0	1
23	30_entrepreneurship-graduate-certificate-cours...	52.0	1
24	32_cotutelle.txt	105.0	1
25	33_contact.txt	81.0	1
26	34_feedback-forms.txt	47.0	1
27	35_it.txt	56.0	1
28	36_openings.txt	68.0	1
29	37_security.txt	-5.0	1
30	40_birks.txt	38.0	1
31	41_stories.txt	45.0	1
32	42_media-relations.txt	25.0	1
33	43_artsci.txt	80.0	1
34	44_ginacody.txt	64.0	1
35	45_finearts.txt	83.0	1
36	46_jmsb.txt	59.0	1
37	47_sgs.txt	70.0	1
38	48_units.txt	60.0	1
39	49_current.txt	55.0	1
40	50_ginacody.txt	64.0	1
41	6_volunteer.txt	106.0	1
42	8_concordia-pays-tribute-to-the-life-of-sara-w...	105.0	1
43	9_sara-weinberg-award-for-excellence-in-teachi...	204.0	1
44	22_counselling.txt	78.0	2
45	24_events.txt	41.0	2
46	38_mental-health.txt	57.0	2
47	39_events.txt	212.0	2
48	5_services.txt	173.0	2
49	7_counsellors.txt	142.0	2

k=6

	name	scores	label
0	10_associate-dean-faculty-relations-and-inclus...	71.0	0
1	11_rationale.txt	95.0	0
2	12_methodology.txt	69.0	0
3	13_strategic-plan-committee.txt	51.0	0
4	14_milestones.txt	61.0	0
5	27_contact.txt	57.0	0
6	28_life-in-montreal.txt	109.0	0
7	29_jobs.txt	86.0	0
8	45_finearts.txt	83.0	0
9	24_events.txt	41.0	1
10	39_events.txt	212.0	1
11	31_traductologie-ma.txt	55.0	2
12	15_deadlines.txt	44.0	3
13	16_apply-now.txt	42.0	3
14	17_facilities.txt	150.0	3
15	18_broken-link.txt	43.0	3
16	19_accessibility.txt	45.0	3
17	20_open-a-ticket.txt	48.0	3
18	23_rotten-meat-could-be-easier-to-detect-thank...	40.0	3
19	25_year-2-of-concordias-sustainability-action-...	141.0	3
20	26_concordia-s-early-childhood-and-elementary-...	95.0	3
21	30_entrepreneurship-graduate-certificate-cours...	52.0	3
22	32_cotutelle.txt	105.0	3
23	33_contact.txt	81.0	3
24	34_feedback-forms.txt	47.0	3
25	35_it.txt	56.0	3
26	36_openings.txt	68.0	3
27	37_security.txt	-5.0	3
28	40_birks.txt	38.0	3
29	41_stories.txt	45.0	3
30	42_media-relations.txt	25.0	3

31	43_artsci.txt	80.0	3
32	44_ginacody.txt	64.0	3
33	46_jmsb.txt	59.0	3
34	47_sgs.txt	70.0	3
35	48_units.txt	60.0	3
36	49_current.txt	55.0	3
37	50_ginacody.txt	64.0	3
38	6_volunteer.txt	106.0	3
39	8_concordia-pays-tribute-to-the-life-of-sara-w...	105.0	3
40	9_sara-weinberg-award-for-excellence-in-teachi...	204.0	3
41	22_counselling.txt	78.0	4
42	38_mental-health.txt	57.0	4
43	5_services.txt	173.0	4
44	7_counsellors.txt	142.0	4
45	1_comment-sollicitier.txt	-11.0	5
46	21_emplois.txt	-7.0	5
47	2_recrutement-etudiants.txt	-10.0	5
48	3_accessibilite.txt	-7.0	5
49	4_confidentialite.txt	-13.0	5

I certify that this submission is my original work and meets the Faculty's Expectations of Originality.

Signature ID: 40256077 Date 2022-12-12