PA1: WEB CRAWLER AND PROCESSING

CS 4422: Information Retrieval

Joseph Garcia

Table of Contents

Email Statistics Output

Code: Test.py
Result: ttestReport.txt

This section is composed of statistical analysis of the email addresses collected. This table includes the ranking of email addresses, email addresses, and percentage.

Rank	Email Address	Percentage
1	ksugrad@kennesaw.edu	80%
2	diploma@kennesaw.edu	52%
3	registrar@kennesaw.edu	25%
4	WorldLanguagesResourceCollection2@kennesawedu.onmicrosoft.com	22%
5	honors@kennesaw.edu	19%
6	finaid@kennesaw.edu	15%
7	gteixeir@kennesaw.edu	14%
8	radam108@kennesaw.edu	14%
9	asarouji@kennesaw.edu	14%
10	cstaub1@kennesaw.edu	14%

Word Frequency Ranking Before Clean Up

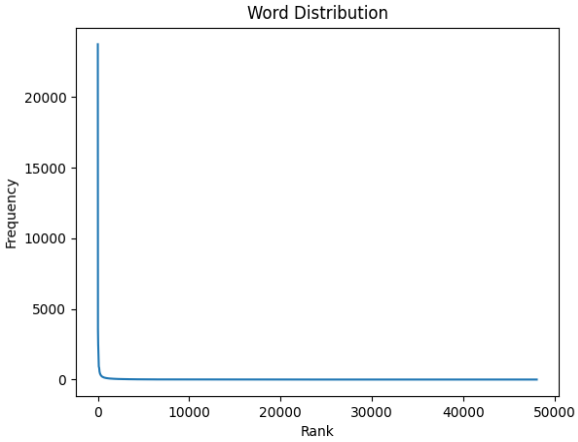
Code: WordFreq.py
Result: WordsbeforeCleanup.txt

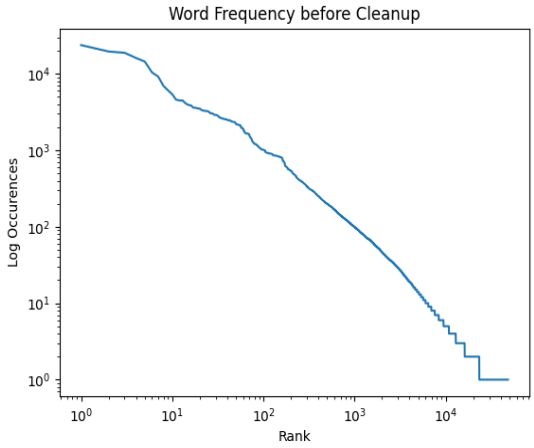
In this section I curated a list of the top 30 words most frequently found in pages before removing stopwords and punctuation. The table below includes the ranking of the word, the word, the frequency, and the percentage used. This section also includes visual representation in the form of plots.

Rank	Word	Frequency	Percentage
1	and	23748	2.94%
2	the	19585	2.42%
3	of	18833	2.33%
4	to	16122	2.00%
5	&	14439	1.79%
6	in	10410	1.29%
7	a	9209	1.14%
8	for	6951	0.86%
9	Students	6057	0.75%
10	Kennesaw	5388	0.67%
11	Campus	4594	0.57%
12	Resources	4483	0.56%
13	KSU	4479	0.55%
14	is	4098	0.51%
15	State	3918	0.49%
16	Alumni	3862	0.48%
17	on	3614	0.45%
18	with	3593	0.44%
19	Business	3530	0.44%
20	The	3509	0.43%
21	or	3350	0.41%
22	Only	3301	0.41%
23	Student	3270	0.40%
24	Faculty	3258	0.40%
25	you	3208	0.40%
26	at	3059	0.38%
27	your	3053	0.38%

28	Staff	2998	0.37%
29	Online	2903	0.36%
30	For	2898	0.36%

Plot of Word Frequency and Distribution (Before Word Cleanup)





Word Frequency Ranking After Clean Up

Code: stopwords_removed.py
Result: WordsAfterCleanup.txt

In this section I curated a list of the top 30 words most frequently found in pages after removing stopwords and punctuation. The table below includes the ranking of the word, the word, the frequency, and the percentage used.

Rank	Term	Frequency	Percentage
1	students	9028	1.55%
2	kennesaw	6529	1.12%
3	campus	5821	1.00%
4	student	4948	0.85%
5	ksu	4907	0.84%
6	resources	4849	0.83%
7	information	4285	0.74%
8	state	4177	0.72%
9	business	4140	0.71%
10	alumni	4067	0.70%
11	faculty	4051	0.70%

12	university	3855	0.66%
13	online	3371	0.58%
14	community	3325	0.57%
15	education	3273	0.56%
16	staff	3252	0.56%
17	programs	3158	0.54%
18	marietta	3139	0.54%
19	program	2987	0.51%
20	college	2858	0.49%
21	research	2850	0.49%
22	2025	2823	0.49%
23	current	2770	0.48%
24	engineering	2742	0.47%
25	family	2655	0.46%
26	friends	2593	0.45%
27	parents	2557	0.44%
28	skip	2472	0.43%
29	financial	2396	0.41%
30	department	2373	0.41%