TOMES Enron Processing

December 2018

Introduction

During the final weeks of the TOMES project, it was decided to utilize the TOMES software tools to process email accounts from the Enron data set.

Processing the Enron data allowed the project team to use the TOMES software[1] with email data created outside any of the project partner institutions.

Additionally, the account files for the Enron dataset are in EML format. This provided a contrast to earlier account processing done by the State of North Carolina in which PST was the source format.

Procedure

A temporary Python script was created to loop through the Enron data and perform the following major tasks for each account:

1) Count the number of messages within the account
2) Convert the EML account data to EAXS
3) Convert the EAXS file to a "tagged" EAXS file
4) Create an Archival Information Package (AIP) consisting of the source EML and derivate EAXS files as well as automatically generated METS files
5) Count the total number of named entities identified within the tagged EAXS file
6) Delete the AIP

File conversion and AIP creation were performed by importing the functionality of the TOMES software tools directly into the Python script. This is possible because each TOMES software module is written in Python and can therefore be imported into other Python scripts.

The last step of deleting the AIP was necessary in order to avoid running out of disk space.

No attempt was made to assess the value of named entities or test for false positives, etc. They were simply counted.

Seven of the 150 accounts within the Enron dataset were completely omitted from the final results as their corresponding AIPs were deleted while testing earlier drafts of the Python script.

Results

Processing the remaining 143 accounts took just under five days, approximately 115 hours. The creation of tagged EAXS files and large METS files containing numerous checksums were easily the most time-intensive tasks.

All processing was performed on an Amazon cloud server with 64 gigabytes of memory. Processing ran as a background process on the server with no manual intervention.

---

[1] For more information on the TOMES software and the workflow it supports, see the TOMES software user guide located at the TOMES project GitHub repository.

477,651 total messages were processed. The total number of entities found was 5,916,455.

The data table below shows the processing results in terms of total messages and entities per account.

A value of "-1" for the total number of entities indicates a processing issue with the account. Three accounts were not processed.

Given the need to delete each AIP upon completion, the error in processing is not currently known for these accounts although an issue has been created in GitHub. Resolving the issues can help improve the TOMES software over time.

Data Table

row_number	account_id	eml_count	entities_found
1	allen_p	3034	22146
2	arnold_j	4898	54105
3	arora_h	654	6801
4	badeer_r	877	10662
5	bailey_s	478	4964
6	bass_e	7823	124101
7	baughman_d	2760	31049
8	blair_l	3415	25096
9	brawner_s	1026	7024
10	buy_r	2429	17194
11	campbell_l	6490	63095
12	carson_m	1400	7673
13	cash_m	2969	51474
14	causholli_m	943	8976
15	corman_s	2025	28528
16	crandell_s	519	7894
17	cuilla_m	1029	21932
18	dasovich_j	28234	824820
19	davis_d	2249	16951
20	dean_c	2429	54631
21	delainey_d	3566	46938
22	derrick_j	1766	23436
23	dickson_s	395	2020
24	donoho_l	1045	7625
25	donohoe_t	1015	10691
26	dorland_c	2127	14572
27	ermis_f	1230	20543
28	farmer_d	13032	110240
29	fischer_m	1498	11455
30	forney_j	729	5453
31	gang_l	590	4623

32	gay_r	1415	13124
33	geaccone_t	1592	14013
34	germany_c	12436	106324
35	gilbertsmith_d	578	4137
36	giron_d	4220	48086
37	griffith_j	2973	22487
38	grigsby_m	2237	24027
39	guzman_m	6054	44354
40	haedicke_m	5246	87006
41	hain_m	3820	82763
42	harris_s	548	3592
43	hayslett_r	2554	20391
44	heard_m	1623	18974
45	hendrickson_s	719	7755
46	hernandez_j	3265	44071
47	hodge_j	1661	14132
48	holst_k	463	7522
49	horton_s	2470	20536
50	hyatt_k	1794	23778
51	hyvl_d	3210	32006
52	jones_t	19950	198015
53	kaminski_v	28465	388005
54	kean_s	25351	598070
55	keavey_p	2177	36879
56	keiser_k	1113	8521
57	king_j	462	4821
58	kitchen_l	5546	91465
59	kuykendall_t	1120	12137
60	lavorato_j	4685	27982
61	lay_k	5937	74814
62	lenhart_m	5920	38569
63	lewis_a	2191	42335
64	linder_e	2805	24594
65	lokay_m	5568	56997
66	lokey_t	1156	10829
67	love_p	5002	36058
68	lucci_p	997	23414
69	maggi_m	1991	20320
70	mann_k	23381	257545
71	martin_t	1112	47300
72	may_l	1600	13161

73	mccarty_d	691	10793
74	mcconnell_m	4542	51021
75	mckay_j	998	6428
76	mclaughlin_e	3353	27985
77	merriss_s	1627	5667
78	meyers_a	1099	2408
79	mims_thurston_p	2038	13208
80	neal_s	3268	31658
81	nemec_g	10655	87018
82	panus_s	437	4437
83	parks_j	2284	23751
84	pereira_s	725	6806
85	perlingiere_d	4778	37177
86	phanis_s	35	242
87	pimenov_v	642	6093
88	platter_p	574	5521
89	presto_k	2204	20601
90	quenet_j	395	1947
91	quigley_d	1568	11086
92	rapp_b	563	9097
93	reitmeyer_j	498	3533
94	richey_c	582	4275
95	ring_a	706	5237
96	ring_r	994	10234
97	rodrique_r	2766	10544
98	rogers_b	8009	62583
99	ruscitti_k	1643	15247
100	sager_e	5200	55507
101	saibi_e	1116	6400
102	salisbury_h	1632	12489
103	sanchez_m	256	1971
104	sanders_r	7329	-1
105	scholtes_d	647	5598
106	schoolcraft_d	1859	10260
107	schwieger_j	738	6127
108	scott_s	8022	97716
109	semperger_c	721	5619
110	shankman_j	3856	47862
111	shapiro_r	6071	115517
112	shively_h	1991	15211
113	skilling_j	4139	46581

114	slinger_r	132	787
115	smith_m	1642	10586
116	solberg_g	1081	2373
117	south_s	248	2635
118	staab_t	621	-1
119	stclair_c	3030	28603
120	steffes_j	3331	47445
121	stepenovitch_j	1227	9938
122	stokley_c	1252	11514
123	storey_g	1027	12936
124	sturm_f	1169	7010
125	swerzbin_m	355	4166
126	symes_k	10827	156633
127	taylor_m	13875	170469
128	tholt_j	1885	18277
129	thomas_p	1293	16285
130	townsend_j	646	9225
131	tycholiz_b	1219	7772
132	watson_k	2950	31150
133	weldon_c	1566	16362
134	whalley_g	1878	24097
135	whalley_l	3335	40932
136	white_s	3272	20919
137	whitt_m	807	8054
138	williams_j	1213	13995
139	williams_w3	3440	16143
140	wolfe_j	1587	12976
141	ybarbo_p	1291	-1
142	zipper_a	1563	12030
143	zufferli_j	557	4070