Princeton Corpus of Political Emails documentation

Corpus version. Release v1.0
You may want to reference this version number when communicating about the corpus as future releases will contain more emails.

Author contact information. Our email addresses are listed here.

Academic publication citation. Please use the citation provided here in your LaTeX project or use the following citation information in your MS Word document:

> "Manipulative tactics are the norm in political emails: Evidence from 100K emails from the 2020 U.S. election cycle". Arunesh Mathur, Angelina Wang, Carsten Schwemmer, Maia Hamin, Brandon M. Stewart, and Arvind Narayanan. 2020. Working paper. https://electionemails2020.org/assets/manipulative-political-emails-working-paper.pdf

Journalism citation. Please cite us using the journalistic norms of your community.

Other citation styles. Please link back to https://electionemails2020.org.

Description. The corpus provided within contains all the emails in the Princeton Corpus of Political Emails starting from December 03, 2019 until election day, November 03, 2020. Each row in this corpus is an email with columns providing additional information such as the sender.

Who are these emails from? The corpus contains emails from over 3,000 political campaigns and organizations in the 2020 election cycle in the U.S. The corpus aims to be comprehensive and includes coverage of emails from the candidates in prominent federal and state races as well as political organizations such as Political Action Committees (PACs) and political parties active in the 2020 cycle. Note that while we attempted to cover the entire population of candidates and organizations, some entities may not have sent us any emails or our automated process may have failed to discover the submission form. A detailed explanation is available in the working draft of our study.

Table 1: Breakdown of the entities in our corpus. Senders differ greatly in how often they send emails: the median federal candidate sent an email every two weeks whereas the 75th percentile candidate sent 5 emails every two weeks. PACs and organizations are more active; the median Hybrid PAC sent an email every week and the 75th percentile Leadership/Single-issue PAC sent 5 emails every week.

Category	Total	Website present	Sent at least one email	Emails per week		
				Q_1	Q_2	Q_3
Federal candidates	4,195	2,552	1,128	0.04	0.43	2.43
State candidates	9,028	5,536	1,540	0.04	0.12	0.56
Leadership/Single-issue PAC	690	262	99	0.26	1.57	4.92
Super PAC	1,728	875	140	0.02	0.14	1.44
Hybrid PAC	164	141	51	0.09	0.91	4.13
Other 527 orgs	68	68	25	0.08	0.45	2.42
Other orgs	209	206	108	0.14	1.06	3.30

How were the emails collected? We automated the process of signing up to receive emails from the websites of the political campaigns and organizations. For each entity's website, if the bot discovered an email sign-up form, it filled it in with the information of a fictional recipient. The bot generated a different, unique email address for each form and monitored all the addresses for incoming email. On receiving an email, the bot opened it exactly once, did not click on any links in the emails, took a screenshot of the email, and relayed the jar of cookies for each recipient, if one existed. Since our list of campaigns and organizations grew over time, we executed the website discovery and automated subscription steps in seven waves starting in December 2019. The released data has been modified to remove the links and obscure the email address we used for signups.

What is in this release? Once you submit a request using the form on the website, we will send you a package containing a CSV file with the emails. We recommend reading the file using any data analysis tool or package like pandas (Python) or R. The package will also include starter R code to read the file. Since the CSV contains several thousand rows, we recommend not using spreadsheet-like software as it might lead to memory issues or unpredictable behavior.

CSV Columns:

- uid_email: A unique id for each email.
- uid_inbox: A unique id for the email's inbox.
- name: The name of the entity (candidate or organization) we signed up for. Note that as we document in our accompanying research paper, signing up for a particular entity results in email leaks to other entities who may also send emails.

- source: Indicates the entity's type: "ballotpedia-campaign" if the email was from a candidate running for office and "orgs" if the email was from an organization.
- office_sought: For candidates, this describes the office pertaining to the election. E.g., "President of the United States". For organizations, this is not applicable and thus a missing value.
- party_affiliation: For candidates, this describes the candidate's political party. E.g., "Democratic Party" and "Republican Party". For organizations, this is not applicable and thus a missing value.
- incumbent: For candidates, this indicates whether the candidate was an incumbent in the election (i.e., "yes" or "no"). For organizations, this is not applicable and thus a missing value.
- office_level: For candidates, this describes what level the office was sought at (i.e., "Federal" or "State"). For organizations, this is not applicable and thus a missing value.
- district_type: For candidates, this describes the office's district. For organizations, this is not applicable and thus a missing value.
- state: For candidates running for a state office, this describes the state. For organizations, this is not applicable and thus a missing value.
- type: For organizations, this describes the type of organization. For candidates, this is not applicable and thus a missing value.
- subtype: For organizations, this describes the subtype of the organization. We do not have this field for all organizations. For candidates, this is not applicable and thus a missing value.
- final_website: The entity's website on which we signed up on to receive emails.
- crawl_date: The date we signed up to receive emails from the entity.
- from_name: The sender's name displayed in the "from" field of each email.
- from_address: The sender's email address.
- date: The date we received the email in ET.
- day: The day we received the email (Mon/Tue etc) in ET.
- hour: The hour we received the email (00-23) in ET.
- subject: The subject of the email.
- body_text: The email body in text format. If an email contained both HTML and plain text parts, we extracted text from the HTML part only. We used the "html2text" library to extract text from the HTML part. We included image alt text but ignored text from tables. We masked all links in the email body ([[URL_REDACTED]]) and only included the link text if it wasn't a link itself. We also masked all occurences of the email id we used to sign up to receive emails ([[EMAIL_REDACTED]]). Finally, we also removed all blank lines from the text.

What about the rest of the data? We will release data collected after November 3, 2020 in future releases. We also have the screenshots and HTML taken when the email was opened but have not released them as it would be substantially harder to effectively mask the signup details and links. If you would like access to these additional materials, please reach out and let us know what your project is about.

What can I do with this corpus? You are free to use this corpus for your research and journalistic purposes such as presenting summary statistics, analyzing content, or showing example emails. You cannot use the corpus for commercial purposes. You must cite the source appropriately and agree that you will not republish the underlying data. If you need to provide replication data for an academic article, you can direct replicators to request a particular version of the corpus. We will not change the data without also changing the version number, so any code you provide to potential replicators should work just as well with data provided from us. The full terms and conditions for access are specified on the data access form here.