Analyzing Prevalence of Topics in Hillary Clinton's Emails Over Time, and Sentiment of Hillary Clinton's Emails Across Regime Types:

Shane Nordquist, Evan Cornelius, Adika Witoelar, Jason Stergiou

General Findings:

We find that the topics of Hillary Clinton's emails varied quite substantially from the period of her tenure as Secretary of State from 2008-2011, and 2012-2015. Moreover, the time periods in which these topics are most prevalent generally coincide with major domestic political reforms and foreign policy crises. Furthemore, we find that Hillary Clinton generally speaks more positively about autocracies than any other regime type, and intriguingly, speaks least positively about democracies.

Structural Topic Modeling:

Given that we know little about the contents of the emails prior to beginning this project, it seems sensible to utilize an unsupervised machine learning model to conduct exploratory data analysis. It seems that a structural topic model would be the most sensible choice in this context, given that we expect many of the emails Hillary Clinton sent as Secretary of state to pertain to public policy, and we expect many public policy concerns to be fundamentally intertwined. Foreign policy and policy on energy, per se, are inexorably intertwined due to the dependence of the United States on petroleum which can be produced by other nations with a comparative advantage in oil extraction, the growing demand for sources of green energy, such as solar panels, which are produced utilizing lithium often from developing countries, and the energy policy of our allies, which may involve dependence on the resources exported by the United States. Thus, it does not seem sensible to utilize k-means clustering, as we should expect many emails to pertain to multiple topics. We find that Unigrams are more useful for analysis with pertinence to term frequency across topics, as well as sentiment analysis, and fitting models involving unigrams is substantially less computationally expensive than with Bigrams or Trigrams. We decided on fitting LDA with K = 30 because to us, it provided us with the most

distinctive, substantive, and interpretable topics. We tried fitting models with K = 10, K = 15, and K = 20, but we found that some of them are not sufficiently interpretable.

Topic 2, 8, 15, 23, 26, 27	Domestic Affairs
Topic 4, 6, 7, 9, 13, 14, 17, 19, 20, 22, 24, 28, 29, 30	Foreign Affairs
Topic 1, 3, 5, 10, 11, 12, 16, 18, 21, 25	Unknown

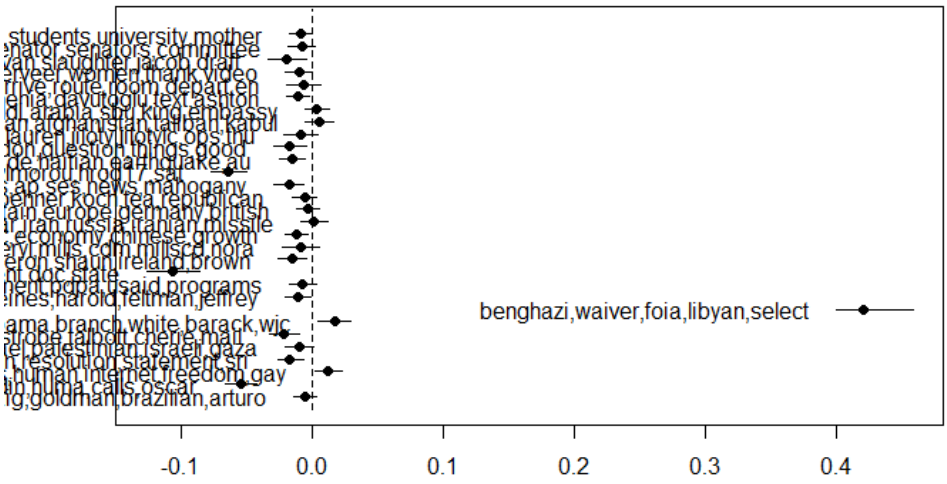
From doing K = 30, we observed that most of the topics in Clinton's emails discussed Foreign Affairs, with a wide range of countries such as Haiti, Israel, Afghanistan, and many others[1]. Many of these topics were exceedingly detailed, such as topic 14, which contained the names of former British Prime Minister Tony Blair, former French President Nicolas Sarkozy, current German chancellor Angela Merkel. This is, intuitively what we would expect, as her role as Secretary of State is exceedingly important in the determination of United States foreign policy.

Estimated Effects and Topic Prevalence Over Time:

Utilizing an estimated effects plot, and regular expressions to group the emails by those being sent in 2011 or prior, and 2012 or later, we garnered several worthwhile insights from the data. Generally, it seems that topic prevalence corresponds quite closely to the occurrence of major domestic political reforms, or foreign policy challenges. An example for this notion is the topic corresponding to Haiti, which explicitly includes the term earthquake, and is more prevalent prior to 2012. This is precisely what we would expect, given that a major earthquake struck Haiti in 2010, wherein circa 300,000 were injured, and 250,000 died. Moreover, the

[1] Appendix contains the 30 topics outputted by our final model.

Benghazi incident occurred in 2012, and the associated topic is far more prevalent following 2012. Similarly, Obergefell V. Hodges was decided in 2015, and the topic containing "gay" and "rights" is more prevalent following the case. Moreover, topics we should expect to have consistently occurred throughout her tenure, such as discussions pertaining to allies, or vague logistics are not more prevalent in either period. Thus, the contents of the emails appear to be roughly what we would expect in the communications of the acting secretary of state in each time period.



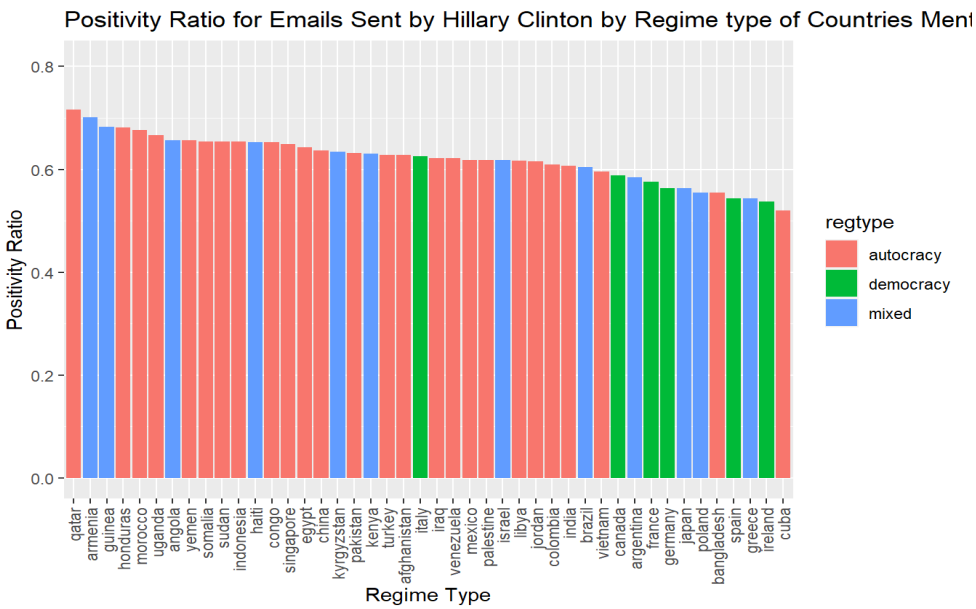
Sentiment Analysis Across Regime Types:

Utilizing the positive words and negative words dictionary provided in lecture, we evaluated positivity ratios as a metric for sentiment both across countries, and across regime types. Regime types were assigned on the basis of the Reporters Without Borders Press Freedom Index. We found 43 countries explicitly referenced in the dataset, with 6 being democracies, 16 being "mixed" regimes, utilized to denote fragile or struggling democracies with less freedoms, and 26 autocracies. We have observed that Clinton seems to have a more positive sentiment in

her

emails regarding autocratic regimes, followed by mixed regimes, then lastly democractic regimes.[2] Clinton favoring autocratic regimes is a surprising finding, which could be due to Clinton attempting to establish stronger diplomatic relations with opposing regime types by representing the interests of the United States. An alternative explanation of this would maybe be that countries classified as Autocratic regimes through the Reporters Without Borders Press Freedom Index might also be developing countries that the United States has an interest in[3]. Moreover, the United States is formally aligned with several states labeled as autocracies via this index, which could potentially explain some degree of positivity in emails pertaining to autocracies. This conjecture is supported by the fact that whilst emails about Cuba, with which the United States has had hostile relations for several decades, are far more negative than emails pertaining to Qatar, a United States ally. This trend could also be accounted for by the relatively small quantity of democracies present in the data, but does provide an intriguing, and somewhat counterintuitive insight into the content of Hillary Clinton's emails.

[2] Appendix 2 contains a barplot showing the average positivity ratio by regime type.
[3] Appendix 3 contains choropleth maps showing regime type and positivity ratio by country. A higher value on the first indicates a more autocratic regime, whilst a higher value on the second indicates a higher positivity ratio.



Appendix 1:

Topic 1 Top Words:
Highest Prob: family, school, law, letter, university, year, years
FREX: school, family, university, law, students, director, letter
Lift: palau, quarter, berlusconi, mother, daughter, student, sex
Score: quarter, palau, berlusconi, school, students, mother, university

Topic 2 Top Words:
Highest Prob: senate, bill, house, senator, vote, health, republicans
FREX: senate, bill, senator, republicans, vote, reid, health
Lift: reid, lautenberg, baucus, ohio, corker, senate, senators
Score: reid, senate, republicans, democrats, republican, baucus, vote

Topic 3 Top Words:
Highest Prob: sullivan, jacob, state, message, original, sullivanjj, clintonemail
FREX: sullivan, jacob, sullivanjj, speech, anne, jake, message
Lift: reinesp, jacobi, edits, slaughtera, declassify, sullivanil, derek
Score: reinesp, sullivan, jacob, sullivanjj, muscatine, lissa, slaughter

Topic 4 Top Words:
Highest Prob: source, el, al, state, libyan, government, magariaf
FREX: el, source, libyan, magariaf, al, qaddafi, keib
Lift: juwali, jcp, gnc, ntc, abdel, haftar, los
Score: magariaf, keib, los, libyan, qaddafi, ntc, jalil

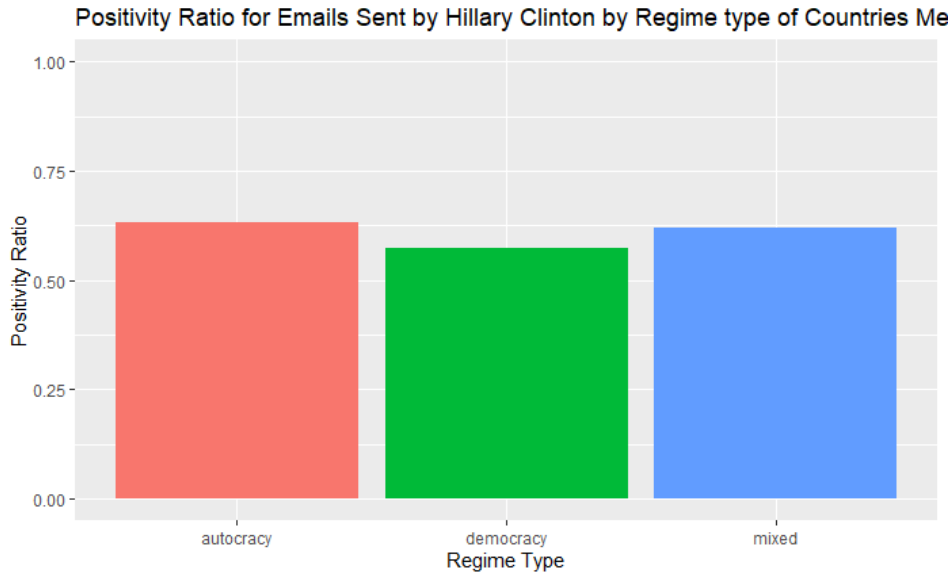
Topic 5 Top Words:
Highest Prob: pm, secretary, office, meeting, state, department, room
FREX: pm, office, secretary, room, meeting, en, arrive
Lift: outer, arrive, route, depart, mini, residence, en

Score: outer, depart, arrive, route, pm, residence, en
Topic 6 Top Words:
Highest Prob: call, huma, text, turkey, abedin, sun, davutoglu
FREX: text, turkey, call, huma, davutoglu, sun, armenia
Lift: scanlon, kaidanow, armenia, amy, tina, davutoglu, esther
Score: kaidanow, davutoglu, armenia, turkey, scanlon, huma, text
Topic 7 Top Words:
Highest Prob: media, embassy, al, report, saudi, sbu, press
FREX: media, embassy, saudi, sbu, al, reported, reports
Lift: hussein, abdullah, kingdom, sbu, ca, gay, saudi
Score: hussein, sbu, saudi, embassy, gay, king, nea
Topic 8 Top Words:
Highest Prob: women, rights, human, people, freedom, melanne, clinton
FREX: women, rights, human, freedom, melanne, internet, africa
Lift: patriarchate, festival, patriarch, ecumenical, congo, women, gender
Score: ecumenical, women, patriarchate, melanne, verveer, orthodox, patriarch
Topic 9 Top Words:
Highest Prob: afghanistan, pakistan, military, afghan, security, department, war
FREX: afghanistan, pakistan, afghan, military, taliban, war, kabul
Lift: taliban, district, guards, afghan, pakistani, kabul, mcchrystal
Score: district, taliban, afghan, pakistan, mcchrystal, afghanistan, kabul
Topic 10 Top Words:
Highest Prob: call, original, message, clintonemail, state, lauren, jiloty
FREX: call, lauren, original, jiloty, clintonemail, thu, message
Lift: pdb, mashabane, hanleymr, anytime, hanley, ops, lauren
Score: mashabane, jiloty, lauren, jilotylc, call, original, clintonemail
Topic 11 Top Words:
Highest Prob: think, one, very, more, know, good, case
FREX: think, very, good, much, one, know, going
Lift: mikulski, things, lot, feel, think, really, chance
Score: mikulski, think, very, much, really, know, don
Topic 12 Top Words:
Highest Prob: huma, abedin, abedinh, message, state, original, clintonemail
FREX: huma, abedin, abedinh, lona, valmoro, valmorou, clintonemail
Lift: tonite, oprah, valmorou, valmoro, oscar, lona, flores
Score: huma, abedin, abedinh, oprah, valmorou, lona, valmoro
Topic 13 Top Words:
Highest Prob: israel, israeli, peace, palestinian, netanyahu, state, arab
FREX: israel, israeli, palestinian, peace, netanyahu, palestinians, arab
Lift: external, palestinians, gaza, israelis, hamas, israel, jerusalem
Score: israel, israeli, external, palestinian, netanyahu, palestinians, jerusalem
Topic 14 Top Words:
Highest Prob: blair, eu, europe, britain, more, germany, british

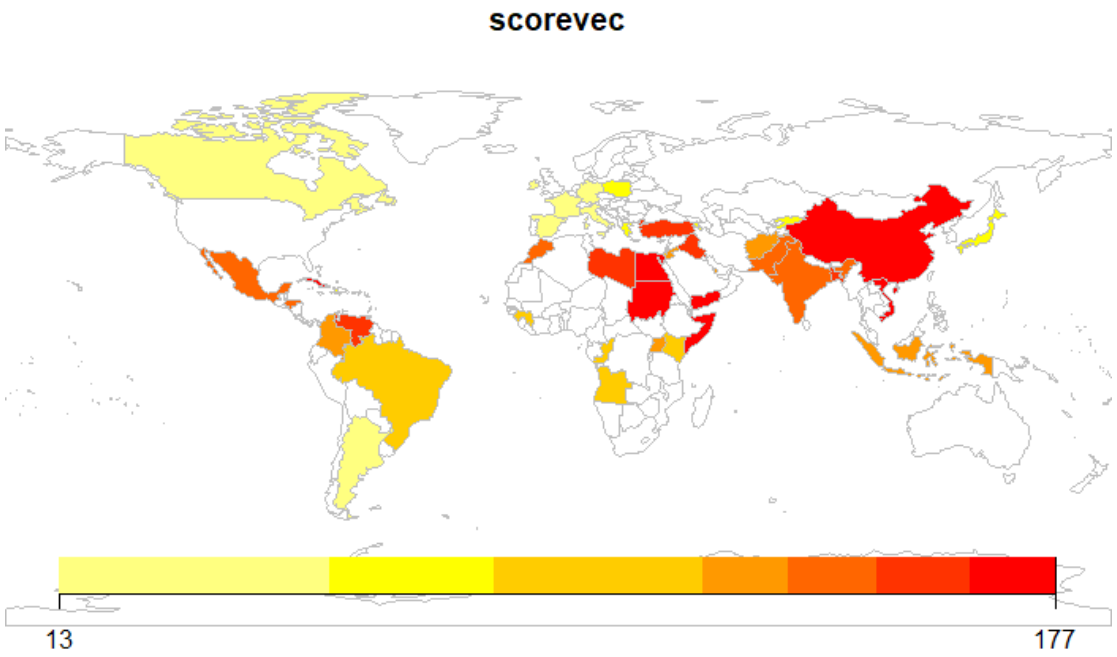
FREX: blair, eu, europe, britain, germany, berlin, british
Lift: berlin, windrush, whitehall, chilcot, saddam, sarkozy, merkel
Score: windrush, chilcot, blair, britain, whitehall, berlin, saddam
Topic 15 Top Words:
Highest Prob: party, case, date, political, doc, department, unclassified
FREX: party, voters, right, political, company, boehner, koch
Lift: koch, heyman, beck, tea, voters, corporate, wing
Score: heyman, koch, boehner, voters, tea, republican, beck
Topic 16 Top Words:
Highest Prob: cheryl, mills, millscd, pm, fyi, state, secretary
FREX: cheryl, mills, millscd, cdm, fyi, friday, wednesday
Lift: laszczych, psa, joanne, katie, ellen, mills, cheryl
Score: cheryl, mills, millscd, psa, cdm, nora, toiv
Topic 17 Top Words:
Highest Prob: iran, united, states, government, nuclear, president, security
FREX: iran, nuclear, united, states, russia, defense, iranian
Lift: pan, colombian, nuclear, wikileaks, cables, iranian, ahmadinejad
Score: pan, iran, nuclear, states, united, iranian, missile
Topic 18 Top Words:
Highest Prob: state, department, date, unclassified, case, doc, release
FREX: state, department, date, unclassified, case, doc, release
Lift: delivered, unclassified, date, doc, case, department, release
Score: doc, unclassified, date, case, delivered, department, state
Topic 19 Top Words:
Highest Prob: deal, uk, sid, brown, party, cameron, gordon
FREX: deal, uk, brown, sid, cameron, dup, party
Lift: unionist, breakthrough, parades, bravo, belfast, uup, sinn
Score: dup, breakthrough, tories, sinn, shaun, labour, clegg
Topic 20 Top Words:
Highest Prob: public, diplomacy, foreign, policy, development, affairs, programs
FREX: diplomacy, public, programs, agency, development, affairs, strategic
Lift: pdpa, audiences, pd, evaluation, expertise, programs, objectives
Score: pdpa, pd, diplomacy, usaid, programs, strategic, development
Topic 21 Top Words:
Highest Prob: march, mar, report, state, jeffrey, message, saturday
FREX: march, mar, jeffrey, report, preines, harold, saturday
Lift: mazen, evergreen, hongju, preines, koh, harold, pir
Score: mazen, preines, feltman, pir, harold, jeffrey, mar
Topic 22 Top Words:
Highest Prob: state, benghazi, foia, information, sensitive, dept, house
FREX: benghazi, foia, information, sensitive, dept, waiver, agreement
Lift: tnc, foia, stevens, dept, waiver, benghazi, select
Score: foia, benghazi, waiver, select, dept, produced, sensitive

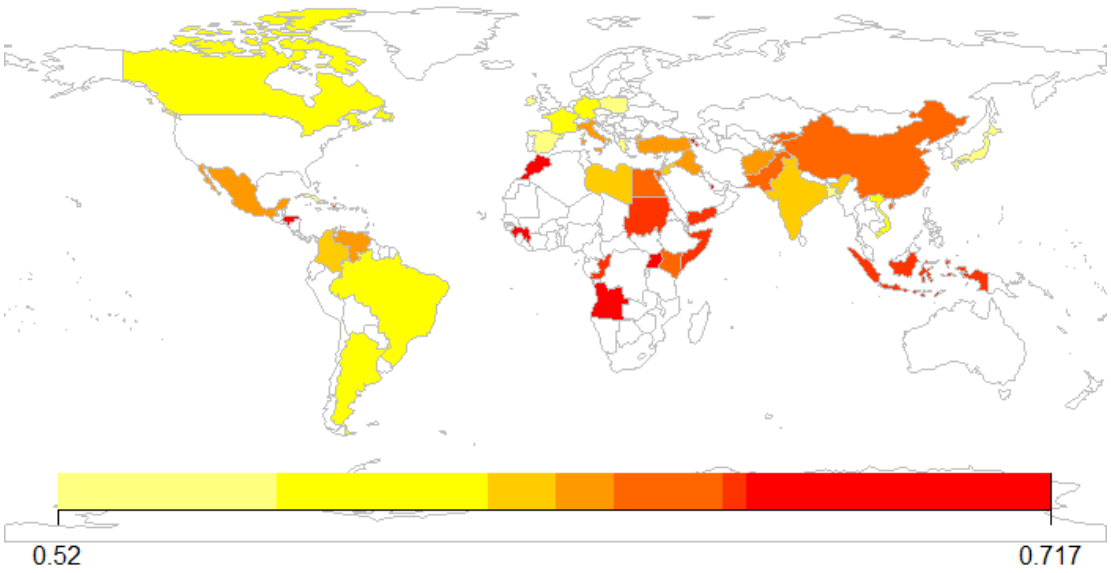
Topic 23 Top Words:
Highest Prob: obama, clinton, president, administration, white, house, policy
FREX: obama, clinton, administration, white, president, policy, bush
Lift: branch, panetta, advisers, emanuel, obama, wjc, barack
Score: branch, obama, clinton, administration, white, bush, romney
Topic 24 Top Words:
Highest Prob: china, global, world, countries, development, economic, health
FREX: china, global, countries, economic, development, food, world
Lift: diseases, agriculture, emissions, chronic, hunger, burden, disease
Score: diseases, china, global, food, economic, chronic, countries
Topic 25 Top Words:
Highest Prob: message, original, please, hillary, mail, clintonemail, clinton
FREX: mail, please, hillary, thank, best, attachments, jan
Lift: cherie, pavilion, kris, balderston, notify, sender, capricia
Score: cherie, kris, strobe, balderston, blair, talbott, expo
Topic 26 Top Words:
Highest Prob: government, percent, more, economy, now, out, prime
FREX: percent, economy, tax, bank, bangladesh, prime, cuts
Lift: grameen, euro, hasina, debt, greece, cuts, bangladesh
Score: grameen, hasina, bangladesh, euro, percent, economy, cuts
Topic 27 Top Words:
Highest Prob: news, state, ses, ap, reuters, department, date
FREX: news, ses, ap, reuters, mahogany, full, north
Lift: gore, reuters, dprk, mahogany, ap, kim, ses
Score: gore, ses, reuters, mahogany, ap, abedinh, news
Topic 28 Top Words:
Highest Prob: haiti, january, haitian, children, de, un, people
FREX: haiti, haitian, january, children, de, relief, un
Lift: et, haitians, bellerive, meghann, haitian, haiti, preval
Score: haiti, et, haitian, bellerive, preval, meghann, au
Topic 29 Top Words:
Highest Prob: secretary, honduras, president, meeting, rights, human, sudan
FREX: honduras, sudan, rights, elections, human, zelaya, resolution
Lift: angola, honduras, uganda, sudan, honduran, cuba, kenya
Score: angola, honduras, zelaya, honduran, sudan, uganda, johnnie
Topic 30 Top Words:
Highest Prob: update, case, david, kelly, craig, unclassified, state
FREX: update, kelly, craig, goldman, david, thomas, brazilian
Lift: venezuela, janice, rio, jacobs, brazilian, kelly, goldman
Score: venezuela, craig, goldman, kelly, brazilian, janice, rio

Appendix 2:


Positivity Ratio for Emails Sent by Hillary Clinton by Regime type of Countries Me
1.00
0.75
0.50
0.25
0.00
Positivity Ratio
autocracy
democracy
mixed
Regime Type

Appendix 3:


scorevec
13
177


newsentvar
0.52
0.717