# Unsupervised Text Analysis of Released Clinton Emails

By: Manuel Duran, Sally Rong, Jenny Wu, Xixi Zheng

March 10, 2021

**Major Finding**

In Secretary of State Hillary Clinton's emails, we discover a shift of topics from an initial more concentrated focus on national issues in 2009, to an emphasis on international issues in the later part of her term as Secretary of State in 2010 and after. We also prove that there was a shift in the sentiments of Clinton's emails before and after years 2009 and 2010, from generally more litigious to strongly modal and expressing uncertainty, and we confirmed the significance of this sentiment analysis shift by running t-tests. Our text analysis of Clinton's emails provide insights on Clinton's shifting sentiments that may be attributed to her increased time in her position as well as the increased complexity of foreign affairs that she worked on while she served as Secretary of State.

**Choosing and Processing Data**

To gain a general understanding of the provided "Clinton.csv dataset," we first employed the use of k-clustering and topic models, with varying values of K in the topic models. We found that these strategies didn't appear to produce useful results in terms of frequency nor distinctiveness of words, even after using varying values of K or subsetting the data between unigrams and bigrams due to the presence of stop words in the data. From this initial analysis, we decided to instead process the data ourselves, using the raw emails text dataset "emails.csv."

Furthermore, we noted that among the released emails, a majority were sent either in 2009 or 2010, so we narrowed our focus down to a temporal analysis of the dataset. Years before 2010 and years after and including 2010 were separated into two groups, giving us an almost even split of the entire dataset. We chose to include years other than 2009 and 2010 because their smaller numbers would play a negligible role in the creation of the model, but in the subgroup after 2009, might add to continuous focus on certain topics.

Frequency of Emails Released Sent by Clinton by Year

| | 2008 | 2009 | 2010 | 2011 | 2012 | 2014 |
|---|---|---|---|---|---|---|
| Frequency | 1 | 3495 | 4013 | 52 | 250 | 2 |
| Total | | 3496 | | | | 4317 |

**Structural Topic Modeling**

First, to understand the themes and issues that are contained within the dataset of emails Clinton sent, we chose to implement structural topic modeling. Through this unsupervised learning model, we wanted to specifically know what national and international issues that the emails contain. We know that Clinton sought to increase the influence of the state department especially in U.S. diplomatic power and global womens' rights, and want to confirm our qualitative knowledge with topic modeling analysis. In line with our knowledge, our topic modeling outputs reveals that Clinton's emails do indeed have the themes we were looking for. Top frequency words that relate to diplomacy in different topics include "afghanistan," "benghazi," and "women." We also note that other topics fall under two broad categories: national themes and individuals. Many topics reveal a focus on national themes that relate to the State Department as well as bi-partisan issues. Other topics were centered around the names of individuals, including "sullivan" who was the Director of Policy Planning.

More specifically, our goal was to see what specific topics Clinton was mainly focused on around 2009 and 2010, and to analyze how her focus may have shifted to different topics during this time. The two K values we tried were 10 and 20 and we decided to keep the latter because of the additional specificity of the topic results. Running structural topic modeling analysis on the two subgroups based on year sent with K = 20, we find that Clinton had a

continued interest in womens' rights as it showed up as top words in the two separate groups, but starting from 2010, more focus was shifted from national affairs to international affairs. We can see this trend from the increased frequency of different countries in the top words, such as "haiti," "iran," "israel," "ireland," and "benghazi." Not much can be gauged from the distinctive words (FREX) within the topic models, however, they do support the fact that Clinton was working on different issues related to national and international relations and was in contact with a wide array of individuals.

**Sentiment Analysis**

We chose to implement sentiment analysis to explore how Clinton's tone in her emails may have changed further in her position and with regards to what we observed above with STM: a shift towards more international issues after 2009. In order to evaluate the sentiment contained in the emails we analyzed them temporally with the same split from our topic modeling analysis. Using the Harvard IV-4 Dictionary, we analyzed that Clinton's emails were generally more positive in sentiment regardless of time period as seen in the graphs depicting the positive ratios of each time period (Figure 1, 2).

Using the Loughran and McDonald Sentiment Word Dictionary that contain tone dictionaries related to business communications, we found that emails before 2010 mainly consist of litigious sentiments while the emails from 2010 on expressed greater strength in a strong modal sentiment, possibly reflecting a readiness to establish herself and the United State's policy stance. Litigious words are "words reflecting a propensity for legal contest" as outlined in Loughran & McDonald 2011[1]. Strong modality are words that denote confidence, as defined in

[1] LOUGHRAN, T. and MCDONALD, B. (2011), When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance, 66: 35-65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Jordan 1999[2], and uncertainty can be defined by words that denote uncertainty in one's tone. Emails before 2010 held a more litigious sentiment, indicating that Clinton may have been still figuring out the ropes of her position and setting the foundations to her future policies (Figure 3). Emails from 2010 onwards show an increase by around 40% in strong modality and uncertainty (Figure 4). This change may be due to Clinton having concluded her "listening tour" around the world and a focus on issues surrounding China and the Middle East — as seen in our topic modeling results — emphasizing the continuing primacy of American power and involvement in the world. This may result in the need for decisive action in the face of having to tread lightly but assertively around foriegn policy. It is also possible that the increase of strong modality may be due to Clinton gaining more confidence in her role.

To determine whether or not the changes in sentiment frequencies are statistically significant, we ran t-tests between the sentiments counts of emails sent before 2010 and those sent from 2010 and onwards. Running a t-test on the 5 sentiments and the positive and negative words we did analysis on, we are able to reject the null hypothesis with 95% confidence, proving that, in line with our previous analysis, Clinton's tone in her emails did indeed become less litigious while becoming more strongly modal and uncertain, and her tone becomes more positive.

[2] Jordan, R.R., 1999, Academic Writing Course ( Longman, London).

**Appendix for Graphs and Tables**

Figure 1:

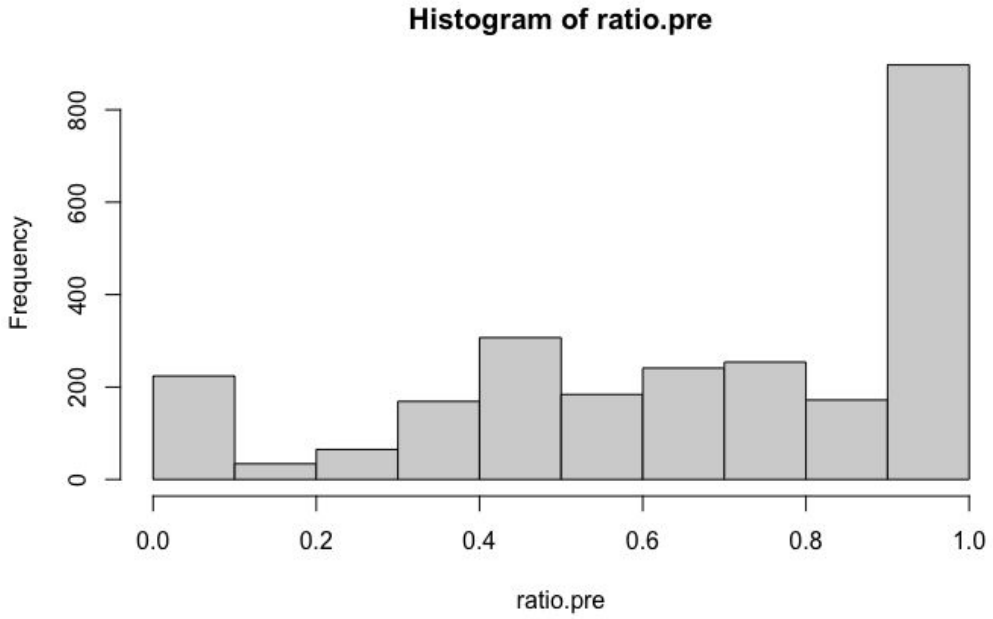


Figure 2:

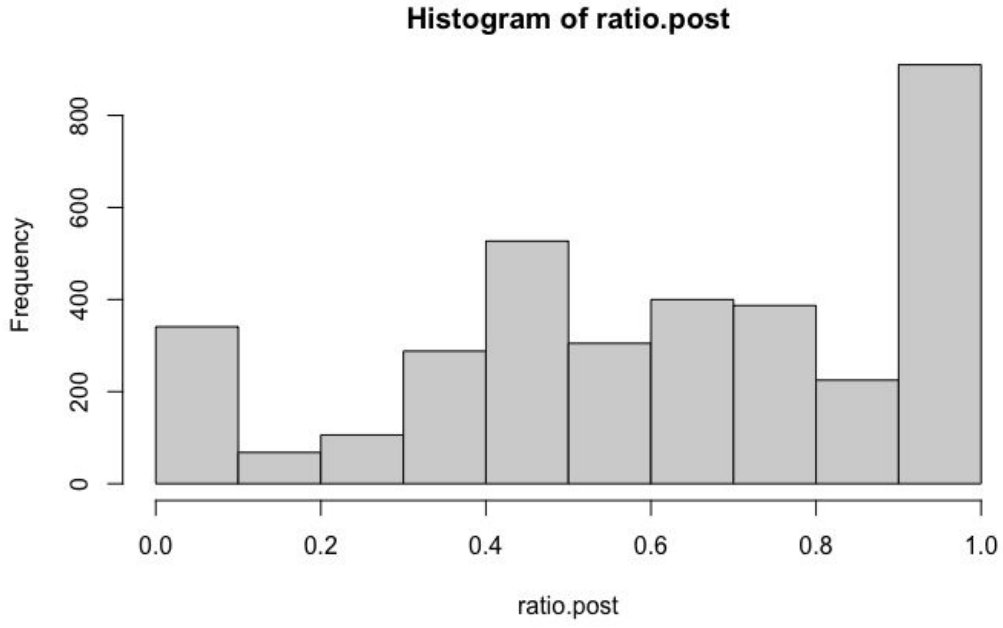

Figure 3:

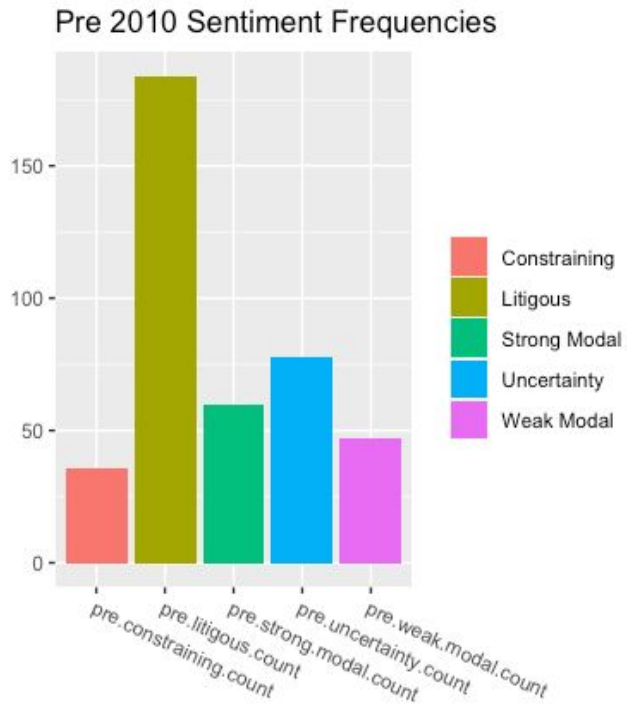


Figure 4:

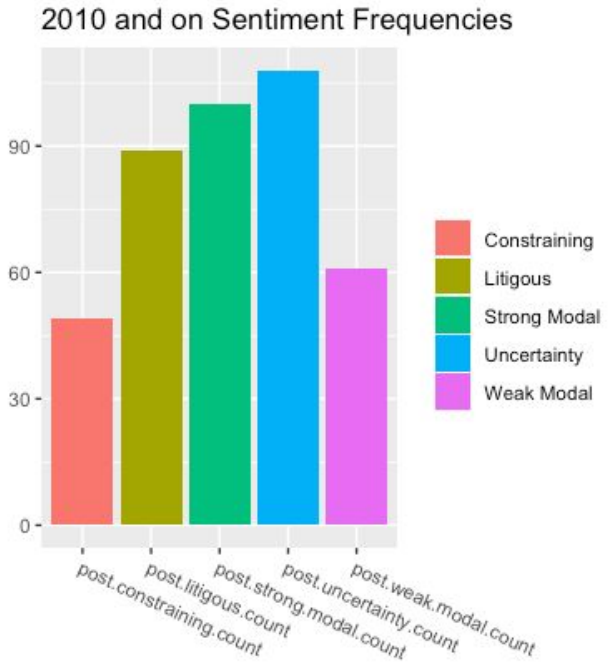

**Table 1: P-Values of Two Sample t-test**

| | **P-Values** |
|---|---|
| **Positive Sentiments** | 5.402e-15 *** |
| **Negative Sentiments** | 4.638e-15 *** |
| **Sentiments of Uncertainty** | 2.969e-09 *** |
| **Sentiments of Constraint** | 7.218e-05 *** |
| **Sentiments of Litigation** | 0.0002448 *** |
| **Sentiments of Strong Modal** | 1.155e-06 *** |
| **Sentiments of Weak Modal** | 1.691e-06 *** |

**Topic Modeling 1: PRE 2010 STM, K = 20**

Topic 1 Top Words:
Highest Prob: state, depart, case, unclassifi, doc, f--, date
FREX: full, hrodclintonemailcom, releas, f--, doc, unclassifi, case
Lift: addresss, attn, autorepli, ballmer, bartz, bing, bren
Score: jiloti, hrodclintonemailcom, lauren, doc, jilotylcstategov, f--, unclassifi
Topic 2 Top Words:
Highest Prob: state, depart, develop, will, date, unclassifi, case
FREX: agricultur, icrw, haiti, haitian, educ, food, donor
Lift: amount-, anti-homosexu, audit, bottleneck, cnpc, coloni, country-level
Score: haiti, agricultur, icrw, global, fellowship, pittman, haitian
Topic 3 Top Words:
Highest Prob: cheryl, mill, subject, sent, millscdstategov, fyi, state
FREX: heather, samuelson, millscdstategov, cheryl, mill, fyi, wendi
Lift: gatestauschercartwright, mckean, feldman, irwin, perla, wendi, —can
Score: cheryl, mill, millscdstategov, nora, cdm, verma, toiv
Topic 4 Top Words:
Highest Prob: state, berlin, depart, date, will, case, f--
FREX: festiv, film, loach, brandenburg, samoa, airlift, amazon
Lift: -british, aam, accent, acquaint, acquiesc, afghanwestern, ajc
Score: festiv, loach, film, berlin, boycott, brandenburg, edinburgh
Topic 5 Top Words:
Highest Prob: depart, offic, secretari, state, room, meet, hous
FREX: managua, spray, franklin, eizenstat, hispan, rout, fill

Lift: acf, acheson, adl, ahern, aman, amsecretari, asi
Score: managua, hispan, rout, mini, fill, spray, room
Topic 6 Top Words:
Highest Prob: sent, subject, sullivan, messag, jacob, origin, sullivanjjstategov
FREX: sheikha, cheri, sullivanjjstategov, hrmycingularblackberrynet, reinespstategov, hongju, harold
Lift: armynavi, asks-, burnswstategov, byer, cheriethi, cornwal, disclaimerinquiriesideacitycom
Score: cheri, sullivan, sullivanjjstategov, jacob, reinespstategov, harold, philipp
Topic 7 Top Words:
Highest Prob: clinton, state, secretari, right, said, human, obama
FREX: marion, mrs, clinton, human, tour, african, cape
Lift: cox, sinek, -gooder, -inclus, -list, aar, abbrevi
Score: clinton, congo, mrs, obama, human, cape, marion
Topic 8 Top Words:
Highest Prob: vote, state, senat, will, date, unclassifi, case
FREX: palau, disapprov, unasur, vote, voter, approv, farrow
Lift: •arturo, •francisco, •lael, •miriam, •thoma, ab-yay, abyei
Score: palau, vote, disapprov, percent, voter, portman, ahmadinejad
Topic 9 Top Words:
Highest Prob: will, case, goldman, famili, richard, verma, sent
FREX: vermarrstategov, rio, youcal, damour, bca, hickey, martha
Lift: administrativa, appel, autoridad, brasil, brito, chamarelli, charsetutf-
Score: goldman, youcal, damour, bca, orna, arola, culver
Topic 10 Top Words:
Highest Prob: berlusconi, korea, north, case, state, report, depart
FREX: berlusconi, dprk, mafia, abdullah, cms, nsoc, pmat
Lift: arcor, avalanch, avvocati, bari, cesar, chengdu, chinese-american
Score: berlusconi, mafia, korea, sbu, korean, opsalert, cms
Topic 11 Top Words:
Highest Prob: branch, wjc, gore, said, clinton, presid, state
FREX: branch, heyman, reno, wjcs, moynihan, pierci, gaf
Lift: -put, aboul-gheit, administration—middl, allergi, apt, asshol, badger
Score: branch, heyman, gore, wjc, moynihan, cvc, lewinski
Topic 12 Top Words:
Highest Prob: huma, abedin, call, sent, subject, messag, origin
FREX: turkey-armenia, abedinhstategov, abedin, huma, davutoglu, call, kaidanow
Lift: difi, falucca, turkeyarmenia, mceldowney, unavail, basescu, bedinhstategov
Score: turkey-armenia, abedinhstategov, abedin, huma, davutoglu, kaidanow, aug
Topic 13 Top Words:

Highest Prob: will, blair, iraq, date, week, former, sid
FREX: chilcot, whitehal, saddam, rickett, mandelson, toni, hoagland
Lift: adion, aea, alton, aluminum, amato, ark, arkadelphia
Score: chilcot, whitehal, meyer, saddam, mandelson, rickett, invas

Topic 14 Top Words:
Highest Prob: hondura, zelaya, state, presid, minist, sent, depart
FREX: zelaya, honduran, hasina, co-sponsor, micheletti, aria, hondura
Lift: aleem, antagon, assad, bdr, bnp, cameroon, dipu
Score: hondura, zelaya, hasina, honduran, co-sponsor, bangladesh, shannon

Topic 15 Top Words:
Highest Prob: sent, subject, messag, origin, schedul, will, lona
FREX: valmoroustategov, lona, valmoro, muscatin, strobe, tue, lissa
Lift: -ham, amdocx, armenia-turkey, barnett, beirn, cafeteria, chay
Score: lona, valmoro, valmoroustategov, muscatin, lissa, strobe, mtg

Topic 16 Top Words:
Highest Prob: state, will, china, said, presid, pakistan, countri
FREX: tec, emiss, taliban, settlement, pakistani, pakistan, israel
Lift: --super, -pass, •merkel, •sarkozi, €bn, accountableand, adumim
Score: missil, taliban, afghan, tori, emiss, cameron, tec

Topic 17 Top Words:
Highest Prob: will, state, depart, case, date, unclassifi, doc
FREX: video, mexico, guy, product, goodwin, angola, idea
Lift: -record, «fob, admonit, agoa, alecross, alrosa, ambitionless
Score: video, cheryl, goodwin, angola, fob, mill, yohann

Topic 18 Top Words:
Highest Prob: state, depart, case, date, secur, unclassifi, guard
FREX: agna, pogo, dyncorp, praisevalid, stateusaidmcc, min, contractor
Lift: armorgroup, gurkha, misconduct, praisevalid, whistleblow, ---info, -cv--rcl
Score: agna, guard, min, contractor, contract, pogo, praisevalid

Topic 19 Top Words:
Highest Prob: state, depart, case, date, doc, women, f--
FREX: fistula, congoles, lyn, lusi, nujood, kivu, sri
Lift: accredit, alshehri, autobiographi, awarde, backgroundpdf, beatric, co-written
Score: fistula, shanghai, melann, heal, verveer, goma, lyn

Topic 20 Top Words:
Highest Prob: obama, said, depart, state, case, presid, say
FREX: cia, interrog, gay, panetta, tortur, black, parlak
Lift: interrog, palin, -commun, -state, adob, amil, anew
Score: obama, panetta, cia, gay, interrog, tortur, bush

**Topic Modeling 2:From 2010 and Onwards STM, K = 20**

Topic 1 Top Words:

Highest Prob: koch, state, case, depart, parti, date, polit

FREX: skousen, jabar, libertarian, birch, geller, formaldehyd, islamophob

Lift: fink, -air, -fbi, —held, abcnewscom, abramoff, accompli

Score: koch, skousen, beck, tea, udal, libertarian, formaldehyd

Topic 2 Top Words:

Highest Prob: obama, clinton, presid, state, say, said, secretari

FREX: romney, rice, yeah, clinton, obama, luce, jone

Lift: abbott, abey, adoptionsi, alphabet, apai, audibl, baffl

Score: obama, romney, yeah, teresa, emanuel, rahm, rice

Topic 3 Top Words:

Highest Prob: sent, huma, call, abedin, subject, origin, messag

FREX: valmoroustategov, valmoro, huma, lona, abedin, pir, sheet

Lift: -april, -mubarak, abedinhstategovi, agenciesi, ah-m, al-missn, astana

Score: abedin, huma, abedinhstategov, lona, valmoro, valmoroustategov, hdrclintonemailcom

Topic 4 Top Words:

Highest Prob: secretari, offic, depart, meet, state, room, arriv

FREX: mini, colombian, urib, monterrey, bajnai, sejm, bogota

Lift: °sob, al-qud, basin, colombian, dowdca, dubos, joneskstategov

Score: mini, bajnai, colombia, pan, sejm, pani, outer

Topic 5 Top Words:

Highest Prob: haiti, state, will, case, depart, unclassifi, date

FREX: beller, mdf, estado, haitian, unido, earthquak, reconstruct

Lift: alissa, anotht, chareyl, chargé, colvill, confab, croix

Score: haiti, haitian, beller, earthquak, donor, mdf, reconstruct

Topic 6 Top Words:

Highest Prob: public, depart, state, unclassifi, media, offic, date

FREX: afpdpa, pao, pelleti, cpao, pd-das, africa, kitti

Lift: -afpdpa, -host, actionschang, agra, aio, airbas, alexey

Score: afpdpa, pao, pelleti, cpao, mchale, pd-das, mellott

Topic 7 Top Words:

Highest Prob: messag, sent, subject, origin, lauren, jiloti, email

FREX: mashaban, windrush, pdb, delet, saba, aclb, obl

Lift: giffin, privilegedconfidenti, aclb, anjana, betsyebel, dep, ghori

Score: jiloti, lauren, mashaban, windrush, pdb, hanley, aclb

Topic 8 Top Words:

Highest Prob: state, depart, date, unclassifi, case, doc, f--

FREX: releas, verma, full, wednesday, f--, doc, unclassifi

Lift: availab, brainard, cvsg, kohhhstategov, madeira, statut, abedinh©stategoy
Score: verma, doc, f--, unclassifi, depart, date, state
Topic 9 Top Words:
Highest Prob: state, american, depart, case, world, countri, f--
FREX: religion, extremist, religi, faith, arabia, expeditionari, muslim
Lift: --scene, -holds-bar, admonit, air-condit, al-hikma, al-wahhab, anti-feminist
Score: saudi, muslim, islam, religion, extremist, arabia, religi
Topic 10 Top Words:
Highest Prob: sullivan, jacob, sent, subject, messag, will, origin
FREX: vikram, sullivan, feltman, palau, jacob, lissa, jacobi
Lift: attitutudes-favor, cui, daniewstategov, gms, godec, hardcopi, helga
Score: sullivan, jacob, lissa, sullivanjjstategov, muscatin, anne-mari, feltman
Topic 11 Top Words:
Highest Prob: state, iran, afghanistan, will, govern, nuclear, said
FREX: mcchrystal, taliban, nuclear, karzai, wikileak, turki, afghanistan
Lift: —even, —washington, abkhaz, abkhazia, aborigin, afghan-l, afghan-pakistan
Score: mcchrystal, taliban, iran, nuclear, karzai, afghanistan, wikileak
Topic 12 Top Words:
Highest Prob: cheryl, mill, sent, subject, millscdstategov, thank, state
FREX: cdm, mill, cheryl, patrick, roberta, millscdstategov, craig
Lift: group-haiti, nyu, -letter-word, -rd, accomod, adjoint, aguerr
Score: cheryl, mill, millscdstategov, haiti, cdm, roberta, arturo
Topic 13 Top Words:
Highest Prob: israel, isra, palestinian, state, peac, arab, netanyahu
FREX: grameen, netanyahus, zionism, zionist, hage, idf, orthodox
Lift: knesset, ramat, -prime, —may, —strike, accost, actions—direct
Score: israel, palestinian, isra, jewish, netanyahu, jerusalem, settlement
Topic 14 Top Words:
Highest Prob: women, will, tori, case, parti, elect, date
FREX: tori, labour, clegg, lds, fisa, stiglitz, cent
Lift: -point, -told--, able-bodi, achieva, adject, alastair, anji
Score: tori, clegg, labour, lds, women, cameron, lib
Topic 15 Top Words:
Highest Prob: work, help, said, state, case, date, one
FREX: mcewen, thoma, correa, juarez, savag, patient, hospit
Lift: afg, ager, alreay, anothermemb, aorta, barclay, bedsid
Score: mcewen, jamal, thoma, thomass, deuel, juarez, savag
Topic 16 Top Words:
Highest Prob: senat, republican, democrat, vote, hous, bill, said
FREX: boehner, phrma, tauzin, baucus, senat, ohio, republican

Lift: tauzin, -hous, ahip, akpd, alabama-bas, altria, ander
Score: boehner, republican, phrma, tauzin, voter, democrat, baucus

Topic 17 Top Words:
Highest Prob: state, depart, unclassifi, case, date, doc, f--
FREX: internet, lgbt, wynn, smit, googl, flood, innov
Lift: abbasi, agroforest, alamgir, arbab, armload, assistancecontribut, baber
Score: bhutto, internet, smit, mchale, qddr, judith, pakistan

Topic 18 Top Words:
Highest Prob: will, deal, sid, parti, ireland, dup, said
FREX: unionist, fein, sinn, belfast, uup, bravo, mcguin
Lift: behalpa, belfast, brava, devolut, donaldson, julien, petroleo
Score: dup, unionist, sinn, fein, shaun, uup, power-shar

Topic 19 Top Words:
Highest Prob: benghazi, state, sensit, inform, select, dept, agreement
FREX: benghazi, redact, foia, libyan, magariaf, qaddafi, el-keib
Lift: —lifg, abdelhamid, abdulbari, ajdabiya, al-senoussi, allagui, belgasim
Score: benghazi, foia, redact, waiver, comm, magariaf, libyan

Topic 20 Top Words:
Highest Prob: said, sent, subject, state, depart, unclassifi, reuter
FREX: news-mahogani, ses-oshift-iii, krista, simmon, ses-oo, reuter, deyo
Lift: opsnewstickerstategov, osd, -admir, —canada, ablaz, americacould, american-islam
Score: reuter, news-mahogani, abedinhstategov, ses-, ses-oshift-ii, abedin, sbu