Document Clustering:
A Case Study on HRC's Emails

Megan Barnes
University of Chicago / Chicago, IL
meganbarnes@uchicago.edu

Abstract

This paper presents the results of an exploration of Hillary Clinton's private emails (released in 2015). The results indicate that clustering has a useful capacity as an exploratory tool for extracting topics and functions from unlabelled document collections. Refinement of a k-means clustering model through non-random initial seeds ('Buckshot') provides similar results to the baseline model, while recursive k-means clustering ('Bisecting') can improve the model's ability to identify salient topics.

1 Introduction

Supervised learning methods are often considered the preferred approach to NLP as they provide straightforward metrics to evaluate their performance, like accuracy or F-score. They require labeled datasets, however, which can be scarce or expensive to construct. The abundance of unlabelled text data relative to its labelled counterpart provides motivation to explore unsupervised learning methods. The official release of sets of documents (transcripts of government addresses, etc.) that come from a single source but have very little labeling present a useful application of unsupervised clustering methods; insights from them have a potentially high impact, but their form as a sea of text makes them opaque. The aim of this investigation is two-fold: to discern structure from an unlabelled data source through clustering and to compare the performance and peculiarities of different clustering methods.

Section (2) will review related work on the subject of clustering models and document collection exploration. Section (3) will describe the dataset being explored (Hillary Clinton's email server) and the clustering models being compared. (4) will discuss the experiments being performed and (5) will review their results.

2 Related Work

Previous work in document clustering is abundant, and methods in hierarchical and flat clustering have become standard. Willett claims that agglomerative (or hierarchical) clustering techniques can improve the quality of clusters compared to flat methods. The tradeoff between them, however, is that agglomerative methods require the calculation of pairwise similarities between all the documents in a given collection, which can be computationally prohibitive for document collections of nontrivial size. Willett focuses on algorithmic approaches to make agglomerative methods more feasible, but in practice these methods are either still too expensive to implement on large datasets or sacrifice cluster quality.

Flat clustering, on the other hand, is much more efficient, but the cluster quality of the most common method, k-means clustering, is subject to randomly chosen 'seeds' (cluster centroids initialized at the beginning of the algorithm) and is, for this reason, non-deterministic. Research in the area of flat clustering has introduced methods which attempt to improve the k-means algorithm. Cutting et al. propose a method called 'Buckshot' in which agglomerative clustering techniques are applied to a small sample of the documents in a collection and the centroids of those clusters are used as the initial seeds of the k-means algorithm. Steinbach et al. propose another method, bisecting k-means, in which the k-means algorithm is applied to iteratively to the document collection to recursively split the collection into two sub-clusters (this is a hierarchical implementation of a traditionally flat clustering algorithm).

Scatter/Gather, an investigation into using clustering to browse large document collections, contributes an important metric for qualitative evaluation of unsupervised document clustering: the cluster 'digest'. The digest object consists of a collection of documents that is closest (given a similarity measure) to the centroid of a learned cluster and the most frequent words in a document cluster (Cutting et al.). Cluster digests will be used to evaluate the clusters in this investigation and formal description of the digest object will be in section (4).

3 Methods

3.1 Dataset

The dataset is comprised of Hillary Clinton's recently released emails from her private server. Each document is represented as a vector of word frequencies, normalized by document length. The vocabulary being counted has stop words removed and also has terms removed that represent document metadata (each email's unique identifier, labels related to when the document was released, Clinton's case, etc.). Frequencies in the word count vectors are expressed as TF-IDF (term frequency * inverse document frequency).

3.2 Algorithms

3.2.1 K-Means Clustering

This algorithm is the baseline against which more refined methods will be compared. Random 'seed' centroids are chosen from the set of document vectors and the rest of the documents are assigned to the cluster of whichever centroid they are closest to using euclidean distance. Centroids are then updated to be the mean of each cluster. The cluster update/centroid update steps are iterated until convergence (or, for the sake of efficiency, distortion ceases to increase by a certain threshold).

3.2.2 K-Means Clustering with Buckshot

In order to improve the selection of seed centroids, the initial seeds are chosen by taking a random sample of the documents (of size $\sqrt{kn}$ if k is the number of clusters and n is the number of documents) and using agglomerative clustering to learn k clusters in the sample. The centroids of the clusters are used as the initial seeds for the k-means algorithm. Group average linkage clustering is the agglomerative method, which merges clusters based on the average euclidean distance between pairs of documents in two clusters (Cutting et al.).

3.2.3 Bisecting K-means

This algorithm picks the largest cluster and finds 2 sub-clusters using the k-means algorithm. This step is repeated a fixed number of times and the split that produces the clustering with highest overall intracluster similarity (minimal distortion) is chosen. Splitting is repeated until the desired number of clusters is achieved (adapted from Steinbach et al.).

4 Experimental Setup

The dataset consists of 7,945 email documents, sent or received by Hillary Clinton. After removal of stop words, the data contains 110,367 unique types comprising the vocabulary. The word count vectors are normalized by document length and multiplied by inverse document frequency (TF-IDF).

In order to evaluate the structure being identified by the clusters in the data, the browsing technique used for *Scatter/Gather*, cluster digest, will be used. The two components to the digest of a cluster are:

1. The m documents whose similarity (euclidean) to the centroid of a cluster is the largest.

2. The z highest weighted words in a cluster. If $p(\alpha)$ is the transformed word count vector of document α, the weighted word count of the entire cluster is:

   $p(\Gamma) = \sum_{\alpha\epsilon\Gamma}(p(\alpha)/$ distance from centroid)

   the highest weighted words in the cluster are the set of words corresponding to the maximum values in $p(\Gamma)$.

   (Cutting et al.)

The second form of evaluation, that of cluster quality, will be constituted by a metric called the silhouette coefficient. It takes the form:
$s = (b - a)/max(a, b)$
where a is the mean distance between a sample and all other points in the same cluster and b is the mean distance between a sample and all other points in the next nearest cluster. This value ranges from -1 to 1, with -1 meaning a high occurrence of assigning a point to the wrong cluster, 0 signifying

significant cluster overlap, and 1 meaning maximum cluster quality (clusters are dense and well separated).

The three algorithms listed in (3) will be implemented, with basic k-means clustering as the baseline, and their performance, varying k, will be evaluated for what it illuminates in the email data as well as the quality of their clusterings. Agglomerative clustering, as it applied in the k-means with buckshot method, is implemented using the scikit-learn toolkit for Python.

5 Results and Analysis

5.1 Topic Salience in the Data

	Topics (cluster size)		
K	**K-Means**	**Buckshot**	**Bisecting**
2	Israel-Palestine Conflict (size10)	Middle East (size17)	Unclear (size54)
2	Middle East (size9)	Speeches (size16)	
5	Time/Logistics (size113), Libya/Qaddafi (size6), Libya/Qaddafi (size 28)	Speeches (size16), Korea (size7)	Foreign Affairs Officials (size84), Insurgents (size15), Press Surrounding Clinton (size17)
5	Ambassadors/ Embassies (size150), Middle East (size24)	Speeches (size16), Korea (size7)	
10	Speeches (size16)	Speeches (size16), Korea (size7), Logistics (size 13)	Logistics (size33), Speeches (size12), Foreign Affairs Officials (size6), Press Surrounding Clinton (size8), Ambassadors/ Embassies (size252)

Table 1: Discovered Topics

Table (1) shows the salient topics that could be manually gleaned from the k clusters produced in each of the trials (full highest-weighted word lists from the cluster digests are available in the appendix, but are summarized here for brevity). Almost every trial produced one or more clusters of nontrivial size (greater than five documents) and every trial produced one ’catchall’ cluster, that contained the vast majority of the documents in the collection. Possible topics of catchall clusters and clusters of trivial size are omitted. Although some trials produced less cohesive clusters, many maintained a very clear topic; the cluster appearing to contain documents related to the Israel-Palestine conflict listed ’gaza’, ’israeli’, ’rocket’, ’militants’, ’fired’ as five of the top 6 most common words, for instance. Other trials, however, produced clusters that were less clearly related. The bisecting k-means trial, with k=2, produced a word list that included terms like ’device’, ’climate’, ’research’, and ’letter’ that did not have an obvious connection. Inspection of the documents closest to the centroid did not reveal an obvious connection, either, and because the cluster was fairly large (54 documents), manual inspection of the entire cluster was unrealistic. Although not every cluster served to illuminate the structure of the email dataset, the repetition of certain topics between models was significant. Many trials produced clusters appearing to pertain to issues in the middle east - particularly referring to Libya and Muammar Qaddafi, as well as the Arab/Israeli conflict. Occurrences of the terms ’qaeda’ and ’qaeda-linked’ (in reference to Al-Qaeda) in highest weighted word vectors are common. It is clear from this relevant frequency of MIddle East-themed cluster discovery that Clinton’s emails contain a substantial amount of correspondence regarding situations of global importance in the Middle East (the situation surrounding the Arab Spring and events linked to the Al-Qaeda, specifically).

Discussion of the Middle East in Clinton’s emails is not particularly surprising, given her position as Secretary of State and the region’s relevance to global affairs. In general, the topics discovered in all versions of the model were related to some aspect of Clinton’s career. One cluster seemed to be related to scheduling and logistics, having highest weighted terms like ’arrive’, ’depart’, ’schedule’, ’airport’, and many times of day. This is reflective of the functional purpose that Clinton’s emails served in scheduling. Another cluster contained discussion of speeches, with words like ’teleprompter.’, ’talking’, ’points’, and ’opinion’. This is likely because of Clinton’s many speaking engagements. One of the most opaque clusterings contained names in its highest weighted terms list: ’espinosa’, ’eikenberry’, and ’feingold’, among others. After web search it became clear that this cluster pertained to discussion of ambassadors and embassies. This is not particularly surprising because of Clinton’s job function, but does have relevance in combination with the cluster topic of Libya in that it pertains to the discussion of Clinton’s involvement in the 2012 Benghazi attack: these email clusterings show that Clinton is discussing both the situation in Libya and U.S. ambassadors and embassies over her private email server.

5.2 Model Evaluation

With small k, k-means clustering has a tendency to pull out small clusters that differ from the bulk of the emails in the dataset in some salient way, producing a topical cluster and a ’catchall’ cluster. The ’Catchall’ clusters had highest weighted words that would be common to emails as a document type, including scheduling words like days of the week, and words that might be common for a state official like ’president’ and ’obama.’ Because of the dependency on random initial seeds of k-means, the size of the topical cluster varied, was often trivial (less than five documents), but sometimes produced clusters of several hundred documents. Topics of smaller clusters were generally more clear in cluster digests, which makes k-means for this dataset most useful for identifying small threads about particular topics, rather than large clusters that share more general commonalities. Increasing k in the basic k-means model did not necessarily increase the number of salient topics that could be gleaned from the clustering, as many clusters in the models with larger k value were of trivial size.

Although buckshot could potentially improve the k-means model by introducing agglomerative clustering, it did not improve the ability of the model to identify salient topics in the Clinton email dataset. Agglomerative clustering on the sample documents often produced many clusters of trivial size, whose centroids were constituted by the single document in the cluster, effectively making the buckshot initial seeds very similar to truly random seeds. This could be because the documents are not easily separable and most end up in a single cluster, with extremely different documents in clusters by themselves. Performance here was similar to basic k-means.

Bisecting k-means did improve the number of salient topics identified by the model. Because of the persistent nature of the catchall cluster, recursively partitioning the monolith of clustered emails was effective in identifying more salient topics as k increased (as was not the case for basic k-means).

	Silhouette		
K	**K-Means**	**Buckshot**	**Bisecting**
2	0.11	0.49	0.55
2	0.66	0.49	
5	-0.09	0.36	0.37
5	-0.27	0.36	
10	-0.18	0.09	0.37
10	0.25	-0.18	

Table 2: Cluster Quality

Goodness of clustering for the models, as determined by the silhouette coefficient varied widely for each model, but generally decreased as k increased. This is perhaps because as more clusters are introduced for each model, the less separable the clusters are (which effects the silhouette coefficient negatively). It should be noted that silhouette coefficient, or the general quality of the clusters being dense and well-separated, does not necessarily correspond to the informativeness of the clustering. As we see when comparing tables (1) and (2), a trial with silhouette coefficient of 0.25, a fairly good score, can have many clusters of trivial size, which are uninformative for the purpose of identifying topics, but are likely well separated from the rest of the clusters because they may contain documents near the extremes of the dataset.

6 Conclusion and Future Work

K-means clustering has a significant use as an exploratory tool for large, unlabelled document collections. Its use on text documents is relatively efficient and illuminating for the purpose of topic discovery within datasets, yielding clear clusters despite its reliance on random seed documents and arbitrary number of clusters. Specific refinement methods, such as a bisecting k-means model, can potentially aid in its use for topic discovery. Using k-means clustering on Hillary Clinton’s email release revealed that she had significant correspondence on her private server regarding matters of national and global importance, specifically regarding the Middle East, and that her emails did not simply consist of trivial exchanges. Although more detailed inspection of specific documents is needed to determine whether or not document sets like Clinton’s contain classified information, clustering provides a useful heuristic tool for discovering themes within opaque text collections. Possible future work in this area could incorporate different document count normalization in order to prevent the occurrence of ’catchall’ clusters, as well as its application, in conjunction with a user interface, to make document collections and cluster more easily browsable.

References

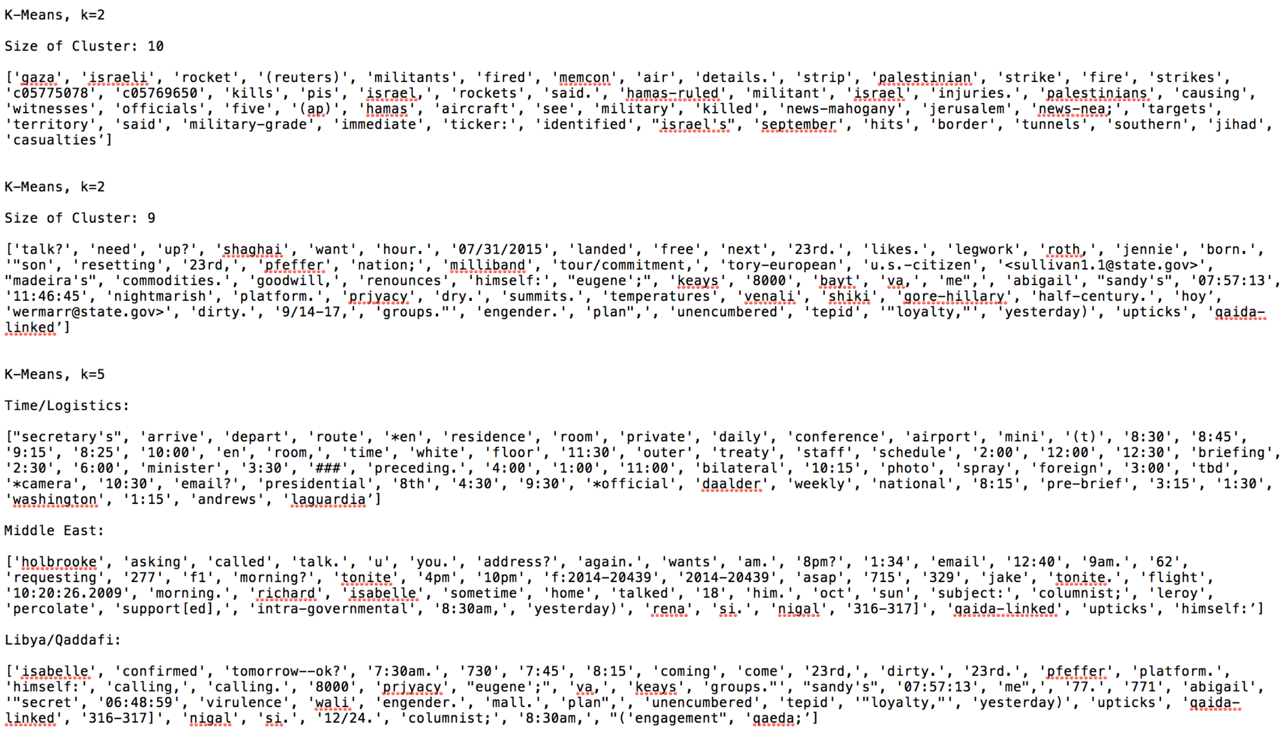
Cutting, D., Karger, D., Pedersen, J. and Tukey, J. W., ?Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections,? SIGIR ?92, 318? 329 (1992).

Steinbach, M., Karypis, G., Kumar, V., ?A Comparison ofDocument Clustering Techniques,? University of Minnesota, Technical Report #00-034 (2000).

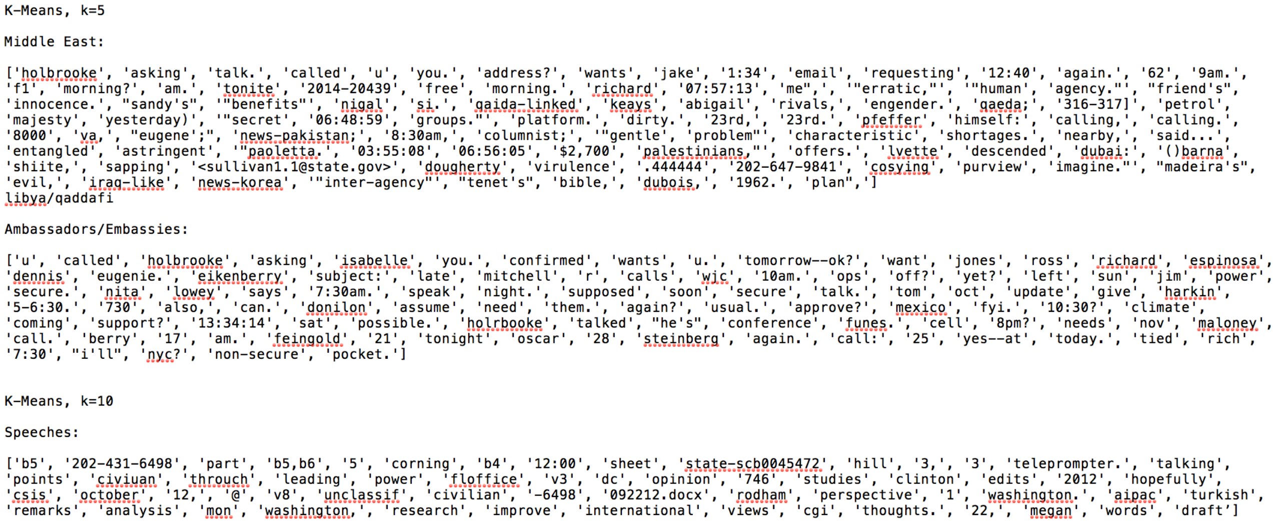
Willett, P., ”Recent Trends in Hierarchic Document Clustering: A Critical Review,” Information Processing and Management: an International Journal, v.24 n.5, p.577-597, (1988).

A Appendix

Digests Corresponding to Table 1:



```
K-Means, k=2

Size of Cluster: 10

['gaza', 'israeli', 'rocket', '(reuters)', 'militants', 'fired', 'memcon', 'air', 'details.', 'strip', 'palestinian', 'strike', 'fire', 'strikes',
'c05775078', 'c05769650', 'kills', 'pis', 'israel,', 'rockets', 'said.', 'hamas-ruled', 'militant', 'israel', 'injuries.', 'palestinians', 'causing',
'witnesses', 'officials', 'five', '(ap)', 'hamas', 'aircraft', 'see', 'military', 'killed', 'news-mahogany', 'jerusalem', 'news-nea;', 'targets',
'territory', 'said', 'military-grade', 'immediate', 'ticker:', 'identified', "israel's", 'september', 'hits', 'border', 'tunnels', 'southern', 'jihad',
'casualties’]


K-Means, k=2

Size of Cluster: 9

['talk?', 'need', 'up?', 'shaghai', 'want', 'hour.', '07/31/2015', 'landed', 'free', 'next', '23rd.', 'likes.', 'legwork', 'roth,', 'jennie', 'born.',
'"son', 'resetting', '23rd,', 'pfeffer', 'nation;', 'milliband', 'tour/commitment,', 'tory-european', 'u.s.-citizen', '<sullivan1.1@state.gov>',
"madeira's", 'commodities.', 'goodwill,', 'renounces', 'himself:', "eugene';", 'keays', '8000', 'bayt', 'va,', 'me",', 'abigail', "sandy's", '07:57:13',
'11:46:45', 'nightmarish', 'platform.', 'prjvacy', 'dry.', 'summits.', 'temperatures', 'venali', 'shiki', 'gore-hillary', 'half-century.', 'hoy’,
'wermarr@state.gov>', 'dirty.', '9/14-17,', 'groups."', 'engender.', 'plan",', 'unencumbered', 'tepid', '"loyalty,"', 'yesterday)', 'upticks', 'qaida-
linked’]


K-Means, k=5

Time/Logistics:

["secretary's", 'arrive', 'depart', 'route', '*en', 'residence', 'room', 'private', 'daily', 'conference', 'airport', 'mini', '(t)', '8:30', '8:45',
'9:15', '8:25', '10:00', 'en', 'room,', 'time', 'white', 'floor', '11:30', 'outer', 'treaty', 'staff', 'schedule', '2:00', '12:00', '12:30', 'briefing',
'2:30', '6:00', 'minister', '3:30', '###', 'preceding.', '4:00', '1:00', '11:00', 'bilateral', '10:15', 'photo', 'spray', 'foreign', '3:00', 'tbd',
'*camera', '10:30', 'email?', 'presidential', '8th', '4:30', '9:30', '*official', 'daalder', 'weekly', 'national', '8:15', 'pre-brief', '3:15', '1:30',
'washington', '1:15', 'andrews', 'laguardia’]

Middle East:

['holbrooke', 'asking', 'called', 'talk.', 'u', 'you.', 'address?', 'again.', 'wants', 'am.', '8pm?', '1:34', 'email', '12:40', '9am.', '62',
'requesting', '277', 'f1', 'morning?', 'tonite', '4pm', '10pm', 'f:2014-20439', '2014-20439', 'asap', '715', '329', 'jake', 'tonite.', 'flight',
'10:20:26.2009', 'morning.', 'richard', 'isabelle', 'sometime', 'home', 'talked', '18', 'him.', 'oct', 'sun', 'subject:', 'columnist;', 'leroy',
'percolate', 'support[ed],', 'intra-governmental', '8:30am,', 'yesterday)', 'rena', 'si.', 'nigal', '316-317]', 'qaida-linked', 'upticks', 'himself:’]

Libya/Qaddafi:

['isabelle', 'confirmed', 'tomorrow--ok?', '7:30am.', '730', '7:45', '8:15', 'coming', 'come', '23rd,', 'dirty.', '23rd.', 'pfeffer', 'platform.',
'himself:', 'calling,', 'calling.', '8000', 'prjvacy', "eugene';", 'va,', 'keays', 'groups."', "sandy's", '07:57:13', 'me",', '77.', '771', 'abigail',
'"secret', '06:48:59', 'virulence', 'wali', 'engender.', 'mall.', 'plan",', 'unencumbered', 'tepid', '"loyalty,"', 'yesterday)', 'upticks', 'qaida-
linked', '316-317]', 'nigal', 'si.', '12/24.', 'columnist;', '8:30am,', "('engagement", 'qaeda;’]
```



```
K-Means, k=5

Middle East:

['holbrooke', 'asking', 'talk.', 'called', 'u', 'you.', 'address?', 'wants', 'jake', '1:34', 'email', 'requesting', '12:40', 'again.', '62', '9am.',
'f1', 'morning?', 'am.', 'tonite', '2014-20439', 'free', 'morning.', 'richard', '07:57:13', 'me",', '"erratic,"', '"human', 'agency."', "friend's",
'innocence.', "sandy's", '"benefits"', 'nigal', 'si.', 'qaida-linked', 'keays', 'abigail', 'rivals,', 'engender.', 'qaeda;', '316-317]', 'petrol',
'majesty', 'yesterday)', '"secret', '06:48:59', 'groups."', 'platform.', 'dirty.', '23rd,', '23rd.', 'pfeffer', 'himself:', 'calling,', 'calling.',
'8000', 'va,', "eugene';", 'news-pakistan;', '8:30am,', 'columnist;', '"gentle', 'problem"', 'characteristic', 'shortages.', 'nearby,', 'said...',
'entangled', 'astringent', '"paoletta.', '03:55:08', '06:56:05', '$2,700', 'palestinians,"', 'offers.', 'lvette', 'descended', 'dubai:', '()barna',
'shiite,', 'sapping', '<sullivan1.1@state.gov>', 'dougherty', 'virulence', '.444444', '202-647-9841', 'cosying', 'purview', 'imagine."', "madeira's",
'evil,', 'iraq-like', 'news-korea', '"inter-agency"', "tenet's", 'bible,', 'dubois,', '1962.', 'plan",']
libya/qaddafi

Ambassadors/Embassies:

['u', 'called', 'holbrooke', 'asking', 'isabelle', 'you.', 'confirmed', 'wants', 'u.', 'tomorrow--ok?', 'want', 'jones', 'ross', 'richard', 'espinosa',
'dennis', 'eugenie.', 'eikenberry', 'subject:', 'late', 'mitchell', 'r', 'calls', 'wjc', '10am.', 'ops', 'off?', 'yet?', 'left', 'sun', 'jim', 'power',
'secure.', 'nita', 'lowey', 'says', '7:30am.', 'speak', 'night.', 'supposed', 'soon', 'secure', 'talk.', 'tom', 'oct', 'update', 'give', 'harkin',
'5-6:30.', '730', 'also,', 'can.', 'donilon', 'assume', 'need', 'them.', 'again?', 'usual.', 'approve?', 'mexico', 'fyi.', '10:30?', 'climate',
'coming', 'support?', '13:34:14', 'sat', 'possible.', 'holrbooke', 'talked', "he's", 'conference', 'funes.', 'cell', '8pm?', 'needs', 'nov', 'maloney',
'call.', 'berry', '17', 'am.', 'feingold', '21', 'tonight', 'oscar', '28', 'steinberg', 'again.', 'call:', '25', 'yes--at', 'today.', 'tied', 'rich',
'7:30', "i'll", 'nyc?', 'non-secure', 'pocket.']


K-Means, k=10

Speeches:

['b5', '202-431-6498', 'part', 'b5,b6', '5', 'corning', 'b4', '12:00', 'sheet', 'state-scb0045472', 'hill', '3,', '3', 'teleprompter.', 'talking',
'points', 'civiuan', 'throuch', 'leading', 'power', 'floffice', 'v3', 'dc', 'opinion', '746', 'studies', 'clinton', 'edits', '2012', 'hopefully',
'csis', 'october', '12,', '@', 'v8', 'unclassif', 'civilian', '-6498', '092212.docx', 'rodham', 'perspective', '1', 'washington.', 'aipac', 'turkish',
'remarks', 'analysis', 'mon', 'washington,', 'research', 'improve', 'international', 'views', 'cgi', 'thoughts.', '22,', 'megan', 'words', 'draft']
```

Buckshot, k=2

Middle East

```
['faxed', 'remarks', 'breakfast', 'prayer', '15', 'minutes', 'house.', 'house?', 'them?', 'it?', 'fax.', 'deliver', 'min', 'coming', 'keays',
'calling.', '8000', "eugene';", 'va,', '07:57:13', 'abigail', "sandy's", 'himself:', 'me",', '"human', 'agency."', '(d-ma)', 'erica,', '22:18:07',
'nigal', 'calling,', 'petrol', 'pfeffer', 'rivals,', 'engender.', 'tipped', 'plan",', 'unencumbered', 'tepid', '"loyalty,"', 'yesterday)', 'upticks',
"('engagement", '316-317]', 'qaida-linked', 'majesty', '9/14-17,', '"secret', '06:48:59', 'groups."', 'platform.', 'dirty.', '23rd,’]
```

Speeches:

```
['b5', '202-431-6498', 'part', 'b5,b6', '5', 'corning', 'b4', '12:00', 'sheet', 'state-scb0045472', 'hill', '3,', '3', 'teleprompter.', 'talking',
'points', 'civiuan', 'throuch', 'leading', 'power', 'flofffice', 'v3', 'dc', 'opinion', '746', 'studies', 'clinton', 'edits', '2012', 'hopefully',
'csis', 'october', '12,', '@', 'v8', 'unclassif', 'civilian', '-6498', '092212.docx', 'rodham', 'perspective', '1', 'washington.', 'aipac', 'turkish',
'remarks', 'analysis', 'mon', 'washington,', 'research', 'improve', 'international', 'views', 'cgi', 'thoughts.', '22,', 'megan', 'words', 'draft’]
```

Buckshot, k=5

Speeches:

```
['b5', '202-431-6498', 'part', 'b5,b6', '5', 'corning', 'b4', '12:00', 'sheet', 'state-scb0045472', 'hill', '3,', '3', 'teleprompter.', 'talking',
'points', 'civiuan', 'throuch', 'leading', 'power', 'flofffice', 'v3', 'dc', 'opinion', '746', 'studies', 'clinton', 'edits', '2012', 'hopefully',
'csis', 'october', '12,', '@', 'v8', 'unclassif', 'civilian', '-6498', '092212.docx', 'rodham', 'perspective', '1', 'washington.', 'aipac', 'turkish',
'remarks', 'analysis', 'mon', 'washington,', 'research', 'improve', 'international', 'views', 'cgi', 'thoughts.', '22,', 'megan', 'words', 'draft’]
```

Korea:

```
['duk-soo', 'korea', '12?', 'campbell', 'kurt', 'fta', 'first.', 'republic', 'shuttle', 'ambassador', 'han', 'u.', '11', 'tomorrow', 'go', 'wants',
'ms', 'returned', '###', 'va,', 'calling,', 'calling.', '8000', "eugene';", 'me",', 'keays', '07:57:13', '8:30am,', 'notch', "sandy's", 'himself:',
'dougherty', 'abigail', 'dirty.', 'pfeffer', '23rd.', 'cdu', 'mall.', 'upticks', 'plan",', '"loyalty,"', 'tepid', 'unencumbered', '12/24.',
'columnist;', '316-317]', "('engagement", 'qaeda;', '(d-ma)', 'majesty', '9/14-17,', '"secret', '06:48:59', 'groups."', 'platform.’]
```

Buckshot, k=5

Speeches:

```
['b5', '202-431-6498', 'part', 'b5,b6', '5', 'corning', 'b4', '12:00', 'sheet', 'state-scb0045472', 'hill', '3,', '3', 'teleprompter.', 'talking',
'points', 'civiuan', 'throuch', 'leading', 'power', 'flofffice', 'v3', 'dc', 'opinion', '746', 'studies', 'clinton', 'edits', '2012', 'hopefully',
'csis', 'october', '12,', '@', 'v8', 'unclassif', 'civilian', '-6498', '092212.docx', 'rodham', 'perspective', '1', 'washington.', 'aipac', 'turkish',
'remarks', 'analysis', 'mon', 'washington,', 'research', 'improve', 'international', 'views', 'cgi', 'thoughts.', '22,', 'megan', 'words', 'draft’]
```

Korea:

```
['duk-soo', 'korea', '12?', 'campbell', 'kurt', 'fta', 'first.', 'republic', 'shuttle', 'ambassador', 'han', 'u.', '11', 'tomorrow', 'go', 'wants',
'ms', 'returned', '###', 'va,', 'calling,', 'calling.', '8000', "eugene';", 'me",', 'keays', '07:57:13', '8:30am,', 'notch', "sandy's", 'himself:',
'dougherty', 'abigail', 'dirty.', 'pfeffer', '23rd.', 'cdu', 'mall.', 'upticks', 'plan",', '"loyalty,"', 'tepid', 'unencumbered', '12/24.',
'columnist;', '316-317]', "('engagement", 'qaeda;', '(d-ma)', 'majesty', '9/14-17,', '"secret', '06:48:59', 'groups."', 'platform.’]
```

Buckshot, k=10

Speeches:

```
['b5', '202-431-6498', 'part', 'b5,b6', '5', 'corning', 'b4', '12:00', 'sheet', 'state-scb0045472', 'hill', '3,', '3', 'teleprompter.', 'talking',
'points', 'civiuan', 'throuch', 'leading', 'power', 'flofffice', 'v3', 'dc', 'opinion', '746', 'studies', 'clinton', 'edits', '2012', 'hopefully',
'csis', 'october', '12,', '@', 'v8', 'unclassif', 'civilian', '-6498', '092212.docx', 'rodham', 'perspective', '1', 'washington.', 'aipac', 'turkish',
'remarks', 'analysis', 'mon', 'washington,', 'research', 'improve', 'international', 'views', 'cgi', 'thoughts.', '22,', 'megan', 'words', 'draft’]
```

Korea:

```
['duk-soo', 'korea', '12?', 'campbell', 'kurt', 'fta', 'first.', 'republic', 'shuttle', 'ambassador', 'han', 'u.', '11', 'tomorrow', 'go', 'wants',
'ms', 'returned', '###', 'va,', 'calling,', 'calling.', '8000', "eugene';", 'me",', 'keays', '07:57:13', '8:30am,', 'notch', "sandy's", 'himself:',
'dougherty', 'abigail', 'dirty.', 'pfeffer', '23rd.', 'cdu', 'mall.', 'upticks', 'plan",', '"loyalty,"', 'tepid', 'unencumbered', '12/24.',
'columnist;', '316-317]', "('engagement", 'qaeda;', '(d-ma)', 'majesty', '9/14-17,', '"secret', '06:48:59', 'groups."', 'platform.’]
```

Logistics:

```
['autoreply:', 'assistance,', '?', 'periodically.', 'immediate', '202-647-9572.', 'checking', 'please', 'need', 'currently', 'e-mail', 'travel',
'official', 'references', '202-647-9573.', 'hariri', '30.', '202-647-9572..', 'november', 'you!', 'thank', '9th.', 'assume', '(laszczychj@state.gov)',
'5548.', 'nr', 'folder', 'yellow', '10:47', 'directly', 'listing', 'ready', 'npr', 'second', 'contact', 'office,', 'tuesday,', 'outo', 'b6', '26th.',
'laszczych', '5th.', "mills'", '202-647-5548.', 'radm', '6th', '3/31)', '(3/22', 'email', 'called', 'sept', '647', 'pc', '4,', 'joanne', 'stone',
'office.’]
```

Bisecting, k = 2

Unclear:

```
['b6', 'talk?', 'details.txt', 'delivered:', 'recipient.', 'attachments:', 'delivered', 'isabelle', 'part', '\\', 'need', 'off?', 'confirmed', '730', 'unscr/
dprk', 'u?', 'climate', 'img00068-20101031-1132jpg', 'summarized', 'device', 'tomorrow--ok?', 'in/', 'memoryhomeuserpicturesimg00015-20100128-0251jpg',
'7:30am.', 'up?', 'shaghai', 'fyi', 'greece?', '11,', 'tripoli', 'references', 'italy?', 'affectionately,', '7:45', '1mg00083-20101031-1645jpg',
'img00083-20101031-1645jpg', '1mg00072-20101031-1141jpg', 'img00072-20101031-1141jpg', 'autoreply:', 'hour.', 'saturday,', 'gina',
'memoryhomeuserpicturesimg00016-20100128-0252:jpg', 'img00016-20100128-0252.jpg', 'april', 'personnel', '8:15', 'research', 'calling', 'letter', 'denis.’]
```

Bisecting, k = 5

Foreign Affairs Officials:

```
['holbrooke', 'b5', 'asking', 'called', 'talk.', 'u', 'you.', 'eugenie.', 'richard', 'daalder', 'wants', 'address?', 'again.', 'arrive?', 'eta?', '202-431-6498',
'ivo', 'email?', '5-6:30.', 'part', 'jake', 'again?', 'u?', 'venezuela', 'holrbooke', '13:34:14', 'am.', '8pm?', 'night.', '1:34', 'email', 'tied', 'b5,b6',
'12:40', 'non-secure.', '277', 'csis', '9am.', '62', 'sun', 'u.', 'back?', '291', 'mitchell', 'requesting', 'greece?', 'f1', 'setting', '1030pm', 'availabe',
'v-2014-20439', 'corning', 'italy?', 'late', 'b4', '167', 'first,', 'morning?', "i'll", '12:00', '11:42:34', 'berry', ‘B’, 'says', '4pm', 'oct', 'them.', 'soon’]
```

Insurgents:

```
['power', 'civilian', '2', '5', '6', '4', '7', 'magazine', '1', '3', 'leading', '9', 'b', '(version', 'affairs', 'b5', '10', '8', 'foreign', 'huma,{', '11',
'12', 'morning)', 'f:2014-20439', 'morning1', 'night)', '###', '21', 'climbing', 'tiiroug', 'team', '(last', 'americas', '.,', '\\', 'hi', 'summit', 'el', '1,',
'20', ',', '13', 'j', '23rd,', 'nearby,', 'said...', 'entangled', 'upticks', '"paoletta.', '"loyalty,"', '06:48:59', 'groups."', 'platform.', 'dirty.', '23rd.',
'shortages.', 'pfeffer', 'himself:', 'calling,', 'calling.', '"secret', 'prjyacy', 'va,', 'innocence.', "friend's", 'agency."', '8000', 'me",', "eugene';”]
```

Press Surrounding Clinton:

```
['b6', 'part', '\\', 'off?', 'climate', 'fyi', 'in/', 'letter', 'tripoli', 'references', 'affectionately,', '1mg00072-20101031-1141jpg',
'img00072-20101031-1141jpg', 'autoreply:', 'gina', 'personnel', 'born', 'attachments:', '\xe2\x80\xa2thank', '12:12', 'please', 'pverveer', 'women.', 'mrs.',
'17,', 'behalf', 'supporting', 'afghan', '.', 'you,', 'happy', 'december', 'thank', 'friday,', 'hillary', 'clinton', 'help', 'let', 'virulence', '8:30am,',
'columnist;', 'dougherty', '(d-ma)', '.444444', 'agency."', '202-647-9841', 'news-pakistan;', 'himself:', '"secret', 'si.', "eugene';", 'groups."', 'yesterday)',
'nigal', '8000', '12/24.', 'qaida-linked', 'wali', '9/14-17,', 'keays', 'majesty', 'mall.', "('engagement", 'va,', 'qaeda;', 'cosying', '1962.', 'purview’]
Bisecting k = 10

Logistics (33):

['holbrooke', 'asking', 'called', 'talk.', 'u', 'you.', 'again.', 'address?', 'wants', 'am.', '13:34:14', '1:34', '12:40', 'email', '62', 'richard', '9am.',
'277', 'requesting', 'f1', 'back?', '291', 'v-2014-20439', 'morning?', '4pm', 'tonite', 'f:2014-20439', '10pm', 'oct', '2014-20439', 'asap', '04', 'supposed',
'715', '329', 'jake', 'sun', 'tonite.', 'flight', 'morning.', '10:20:26.2009', 'night.', 'others', 'isabelle', 'sometime', 'home', 'late', 'last', 'subject:',
'talked', '18', 'him.', 'engender.', '316-317]', '.444444', 'virulence’]

Speeches (12):
['b5', '202-431-6498', 'b5,b6', 'corning', 'b4', '5', '3,', 'part', 'throuch', 'civiuan', 'leading', 'power', 'hill', 'talking', '746', 'points', '3', 'clinton',
'v3', 'civilian', 'dc', 'aipac', '1', '092212.docx', '-6498', 'unclassif', 'studies', 'rodham', 'washington.', '22,', 'words', 'remarks', '12,', 'march',
'washington,', '2012', 'cgi', 'thoughts.', 'october', 'megan', 'draft', 'hillary', '2o243', 'ed', 'afternoon.', 'international', 'unclass', 'fied',
'092212.doex', 'egi', '202431', 'departm', '202431-6498', 'read', 'ent', 'development’]

Foreign Affairs Officials (6):

['email?', 'daalder', 'ivo', 'kelly', 'craig', 'landler', '70', 'info', 'c', 'story', '8000', 'himself:', 'calling,', 'calling.', 'pfeffer', '23rd.', 'prjyacy',
'keays', "eugene';", 'va,', 'abigail', "sandy's", '07:57:13', 'me",', '"human', 'agency."', '316-317]', 'notch', 'si.', '23rd,', 'platform.', 'dirty.',
'columnist;', "('engagement", '(d-ma)', 'rivals,', 'engender.', 'tipped', 'plan",', 'unencumbered', 'tepid', '"loyalty,"', 'yesterday)', 'upticks', 'cdu',
'mall.', 'nigal', 'qaeda;', 'qaida-linked', 'majesty', '9/14-17,', '"secret', '06:48:59', 'groups."', '8:30am,', 'news-pakistan;', 'dougherty', 'virulence',
'nearby,', 'said...', 'entangled', 'astringent', '"paoletta.’]

Press Surrounding Clinton(8):

['letter', 'yet?', 'left', 'b6', 'chair.', 'ban', 'hakim', 'latest', 'go.', 'draft', 'fine', 'received', 'bill', 'chris', 'hill', 'email', 'offers.', "sandy's",
'calling,', 'calling.', '8000', "eugene';", 'va,', 'keays', 'abigail', '07:57:13', 'pfeffer', 'me",', '"human', 'agency."', "friend's", 'innocence.', '316-317]',
'lockdown,', 'si.', 'himself:', '23rd,', '23rd.', "('engagement", 'tepid', 'yesterday)', '(d-ma)', 'rivals,', 'unencumbered', 'plan",', 'tipped', 'engender.',
'nigal', 'columnist;', 'prjyacy', 'majesty', '9/14-17,', '"secret’]

Ambassadors/Embassies (252):
['u', 'asking', 'talk?', 'ops', 'want', 'isabelle', 'dennis', 'steinberg', 'you.', 'confirmed', 'wants', 'secure', 'eikenberry', 'ross', 'need', 'subject:',
'up?', 'jim', 'holbrooke', 'mitchell', 'calls', 'talk.', 'called', 'r', 'coming', 'u.', 'can.', 'espinosa', 'oscar', 'connect', 'today.', 'call.', 'non-secure',
'feingold', 'sat', '10am.', 'tomorrow--ok?', 'jones', 'aug', "he's", 'free', 'sun', 'pdb', 'secure.', 'soon', 'reminder', '7:30.', 'phone', 'today', 'eugenie.',
'speak', 'call?', 'fm', '8', 'ok?', 'jake', 'sheet', 'to?', 'says’]
```