NLP Analysis of Email Interactions to Find Automation Opportunities

José Miguel Hernández Cabrera

Universitat Politècnica de Catalunya
Facultat d'Informàtica de Barcelona

20th October, 2022

Email messaging is one of the most important tools for any kind of industry.

- Up to 65 % of employee interactions are done through social technologies, email messages included.
- About 28 % of their working time responding, reading and writing emails.
- Around 19 % of working hours tracking down important emails
- Up to 10 % categorizing email messages.
- Companies can lose a significant amount of productivity due to a lack of spam management.

Automation offers a range of benefits.

- Automatic email answering system can help companies to save labor force in email customer service.
- Automatic expertise discovery using emails can improve organizational efficiency.
- Automating tasks of derived from emails may reduce time consumed and even improve the life quality.

Objectives

General Objective

- To propose a system based on Natural Language Processing and Unsupervised Machine Learning to look for opportunities of automatization arising from recurrent email patterns found in email texts.

Secondary Objectives

- Compare different clustering algorithms based on density and partition capabilities.
- Evaluate cluster quality using several indices that do not require ground-truth assumptions.
- Explore the feasibility of using unsupervised methods to group together email chains or detect interactions between emails.

Email Mining

Definition

Email mining is a sub-field from data mining that involves techniques focused on email data.

- Spam detection
  - From contents.
  - From senders.
- Email categorization.
- Contact analysis.
  - Contact identification.
  - Contact categorization.
- Email network property analysis
- Email visualization.
- Automatic Email Answering.

Parts of an email

- Header: “From”, “To”, “CC”, “BCC”, “Subject” and “Date”.
- Body: Unstructured Data. May contain graphic elements, URL links, markup tags, and attachments.

Data representation

- **Feature based approach.**
- Social structure base approach.

Enron Dataset

- It possess around 562k messages distributed along 150 users.
- At the personal request of users, folders and emails have been removed, making the data set even more incomplete.
- Common practice: ignore `discussion_threads` and `all_documents` folders.
- Present challenges such as incomplete information, inconsistent chain format, it is not enough to apply the typical workflow for tokenization and data cleansing.

Enron Dataset

Structure of email dataset

1. Header

Message-ID: <7457472.1075845189537.JavaMail.evans@thyme>
Date: Mon, 21 May 2001 06:21:47 -0700 (PDT)
From: elyse.kalmans@enron.com
To: kenneth.lay@enron.com, rosalee.fleming@enron.com
Subject: FW:

3. Header Features

Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Kalmans, Elyse </O=ENRON/OU=NA/CN=RECIPIENTS/CN=EKALMANS>
X-To: Lay, Kenneth </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Klay>, Fleming, Rosalee </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Rflemin>
X-cc:
X-bcc:
X-Folder: \Lay, Kenneth\Lay, Kenneth\Inbox
X-Origin: LAY-K
X-FileName: Lay, Kenneth.pst

4. Computer-generated metadata

2. Body

Per Holly's request, please see below.

Elyse

5. Most recent message

-----Original Message-----
From: "Holly Korman" <holly@layfam.com>@ENRON [mailto:IMCEANOTES-+22Holly+20Korman+22+20+3Cholly+40layfam+2Ecom+3E+40ENRON@ENRON.com]
Sent: Monday, May 14, 2001 3:49 PM
To: Modad, Jessica; Fleming, Rosalee
Cc: Siegel, Misha; Kalmans, Elyse
Subject:

Rosie,
Per Jessica's request I have attached the most updated copies of Mrs. Lay's information. Elyse and Misha, I just thought that you might be interested as well.

Holly

6. Chain inside message

- LPL & KLL short Bio.doc
- LPL Bio 9 short.doc
- Linda's Associations.doc
- 2001 commitments.xls

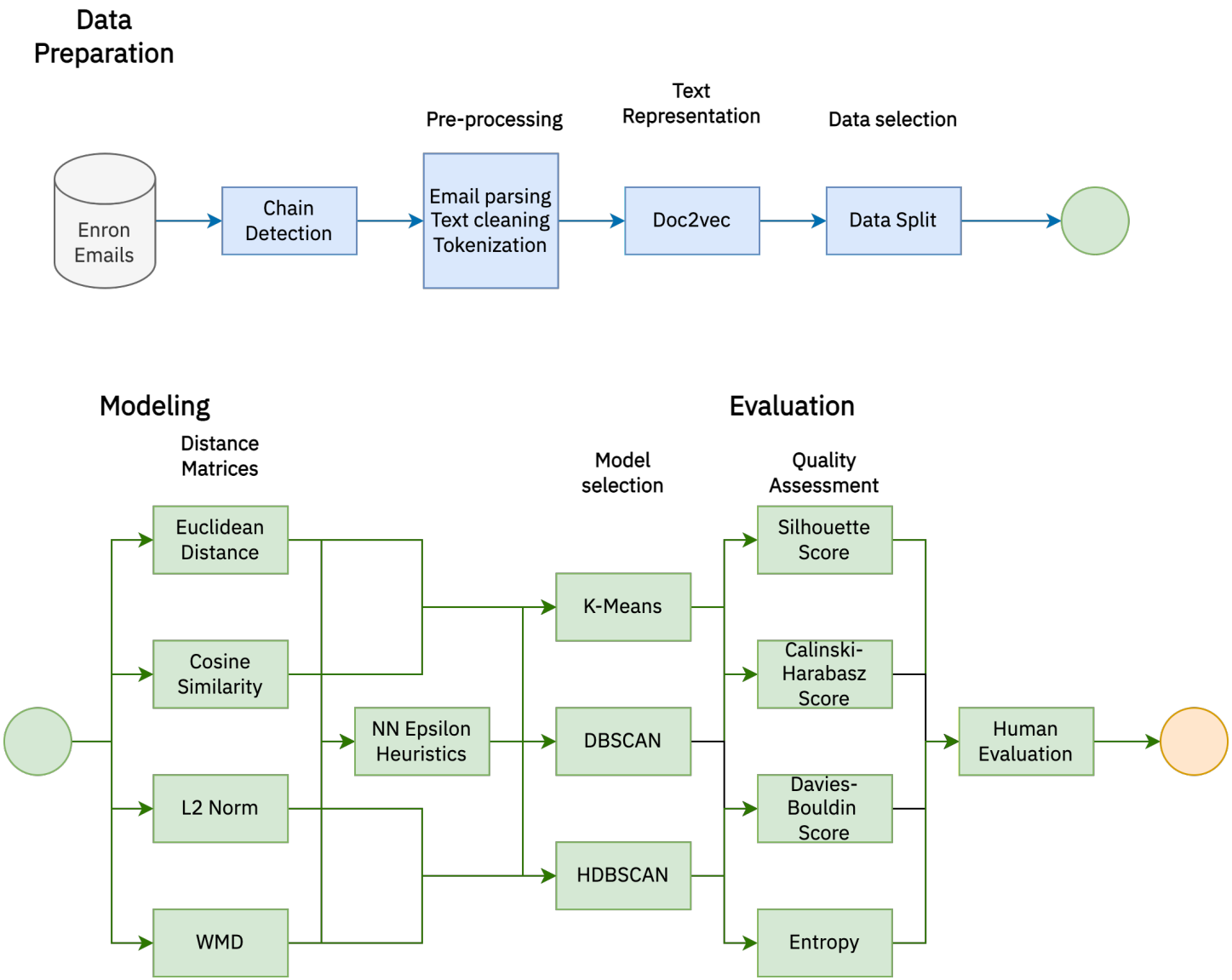
7. Attached documents

Literature Review

Works using Enron Emails.

- Agarwal et al. [1]: Predict the hierarchy of Enron employees. Used Social Network Analysis and NLP.
- Diesner, Frantz, and Carley [3]: Email interactions around organizational crisis. Used Directed Network Graphs and Lehmann's Similarity algorithm.
- Keila and Skillicorn [5]: Word use is correlated to the role within the organization. Used SVD and SDD.
- Kathuria, Mukhopadhyay, and Thakur [4]: Proposed a cohesion score to each cluster to evaluate the intra-cluster quality. Their approach used k-means and hierarchical clustering with TF-IDF vectors.

Pipeline


Data Preparation
Pre-processing
Text Representation
Data selection
Enron Emails
Chain Detection
Email parsing
Text cleaning
Tokenization
Doc2vec
Data Split
Modeling
Evaluation
Distance Matrices
Model selection
Quality Assessment
Euclidean Distance
Cosine Similarity
L2 Norm
WMD
NN Epsilon Heuristics
K-Means
DBSCAN
HDBSCAN
Silhouette Score
Calinski-Harabasz Score
Davies-Bouldin Score
Entropy
Human Evaluation

Data Preparation

Chain Detection

1. For each email allocate in a single array all unique users and sort them alphabetically.
2. Get timestamp and sort emails.
3. Create key by concatenating subject and sorted users. For each key, remove patterns from subject, such as `RE, FWD, FW`. Original subjects must be kept for later steps.
4. To create a chain, allocate key in dictionary with a unique id and values the concatenated key and a boolean array to indicate if email is a reply.
   - If the original subject does not contain `RE:` and there is no email with a prior timestamp than that email, consider it as the initial email from the chain, assigning it a unique ID and assign false to the array value.
   - if the original subject does contain a `RE:`, it is considered as the same chain from the previous timestamped email with the same subject and users. It will be allocated for the same chain id as the previous corresponding email and assign reply to true.
   - if the email contains the same subject and users as others emails but does not contain the `RE:`, it will be considered as a new chain.
5. The final number of emails with the same id is considered as the length of the chain.

Data Preparation

Data pre-processing

- Removed content below `----Original Message----`.
- Cleaned emails with the regular expression

```
'_{4,}.*|\n{3,}|<[^>]*>|-{4,}(.*)(\d{2}:\d{2}:\d{2})\s*(PM|AM)'
```

- Lower case all text.
- Tokenize each email and remove non alphanumeric characters.
- Removed empty tokens.

Data Preparation

Resulting Data set

Worked with 251.068 messages, 119.647 users, 19.993 unique senders.

Distribution of chains length

Chain Length	**Number of chains**
1	203.172
2	12.009
3	2.946
4	1.077
5	463
6	222
7	116
8	67
9	45
$\geq$10	192

Text Representation

Gensim's implementation [6] Doc2vec (Paragraph-vector), parameters according to literature

- 50 and 300 dimensions.
- Window size of 15.
- Minimum count of 1.
- 20 train epochs.
- Alpha of $0,25$.
- Threshold of alpha at 1.0e-5.
- Used PV-DBOW.

Data Selection

Group	**Raw emails**	**% of empty messages**
Chain length $= 1$	203.172	3.4
Chain length $= 2$	24.018	0.7
Chain length $= 3$	8.838	0.8
$10 >$ Chain length ≥ 4	9.708	0.3
Chain length ≥ 10	5.332	0.2

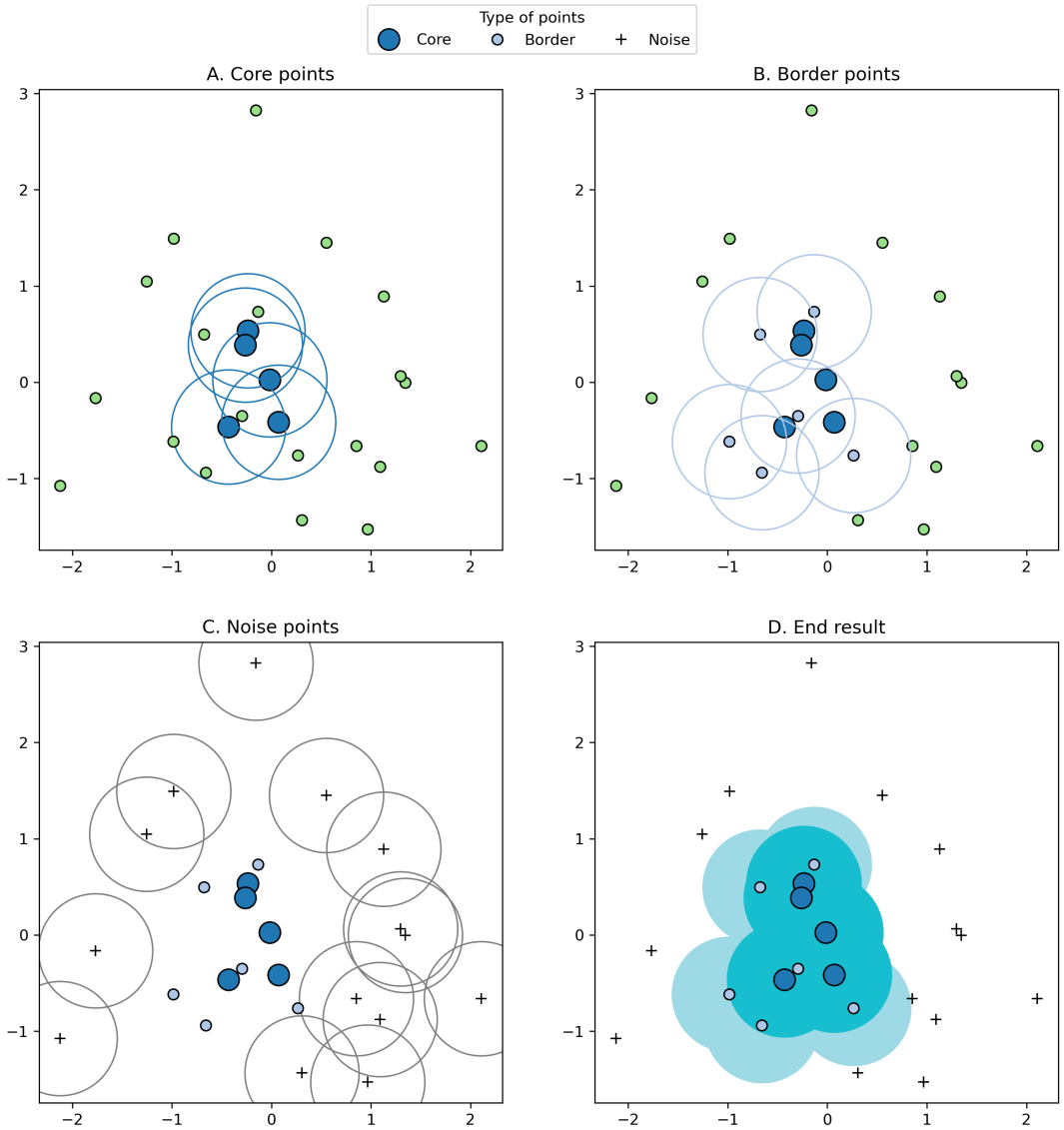
K-Means algorithm is a partition based clustering method. It tries to find k clusters by calculating the distance between each data point and a centroid to assign it to a cluster.

- \+ Cheap to compute: $O(kni)$.
- \+ Easy to implement.
- – Know k *a priori*.
- – Sensible to noise and outliers.

K-means++ : looks to maximize the distance among the initial prototypes. Default initialization in libraries such as `Scikit-learn` and `RAPIDS.AI`.

Clustering Methods

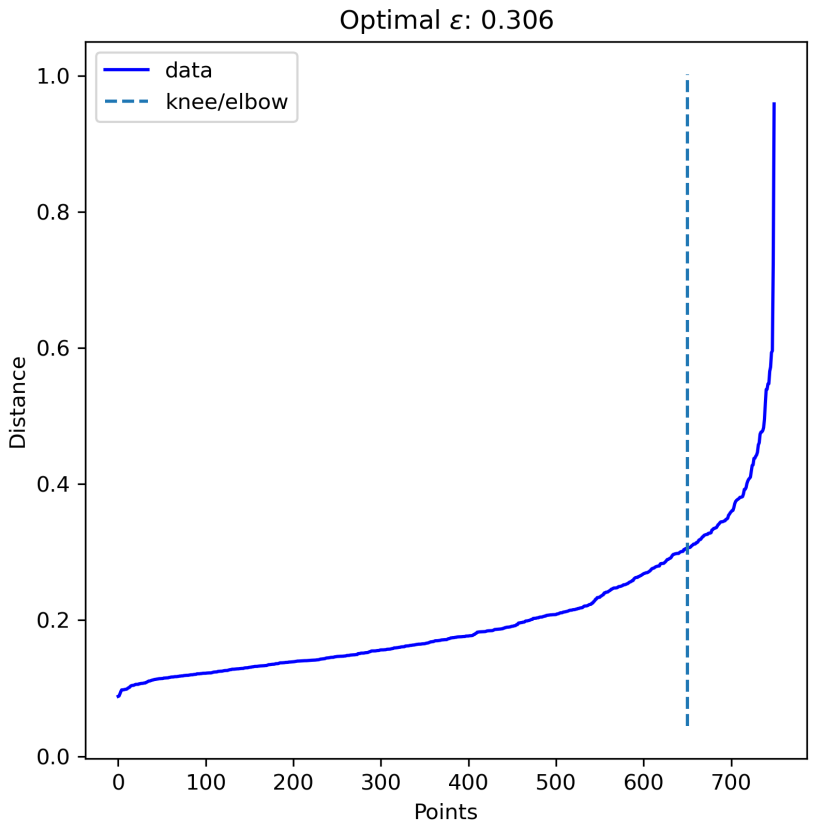
DBSCAN (Density-Based Spatial Clustering of Applications with Noise)



Clustering Methods

DBSCAN Eps Heuristic

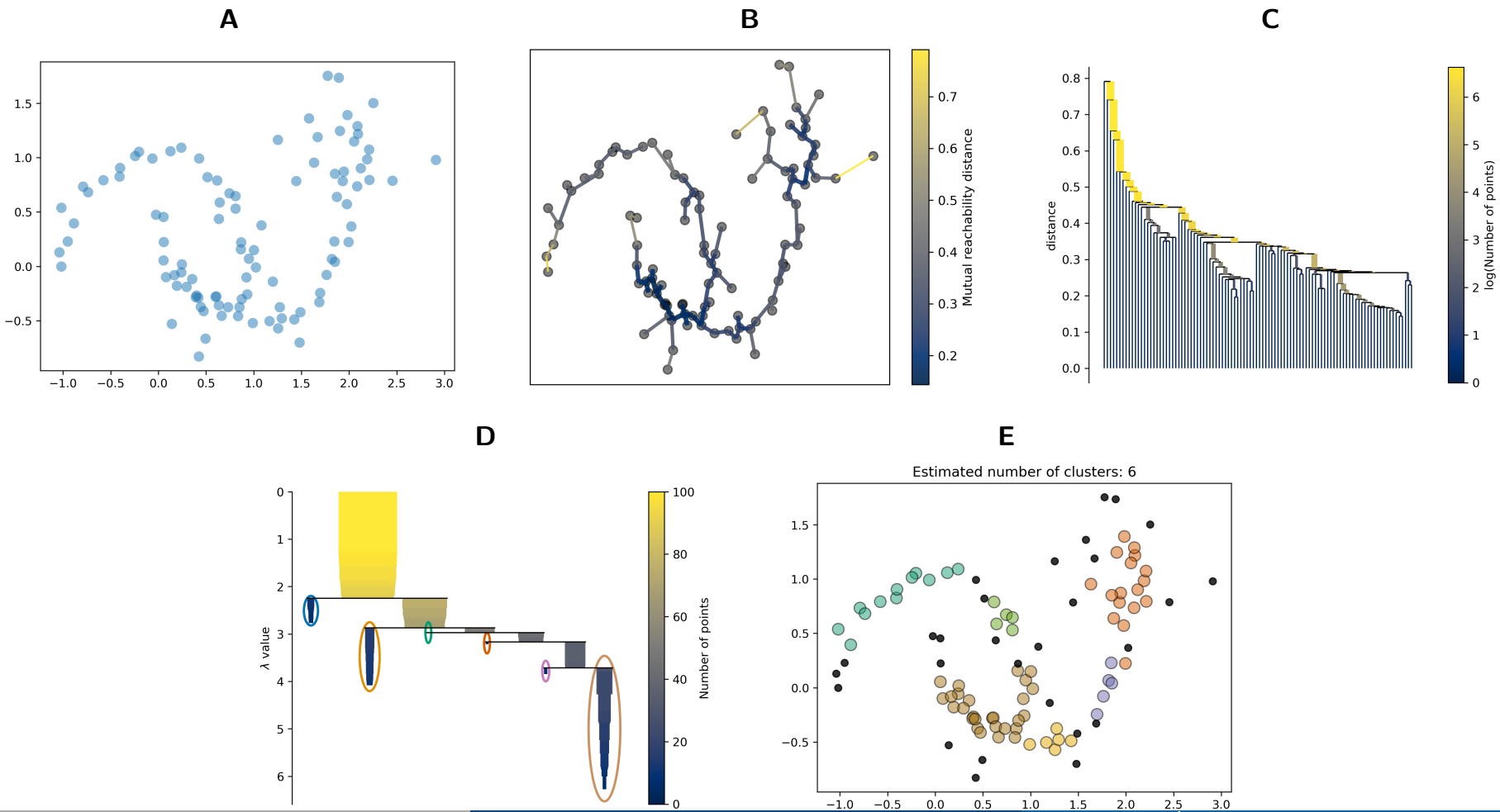
Either use small values of Eps or use the heuristic: obtain Nearest Neighbors, then sorting them in ascending order and then finding the value in which the change rate is minimum.



Clustering Methods

HDBSCAN

A. Transform space based on density. **B**. Minimum Spanning Tree. **C**. Robust Single Linkage Hierarchy clustered tree. **D**. Condensed Tree to find clusters' stability. **E**. Final dataset with most stable clusters.



Distance Matrices, Clustering Methods, Cluster Evaluation

For all methods was used `RAPIDS.AI` Python library, which accelerates computations via GPU. Except for

- Word Mover's Distance. Done through Python implementation `dist_matrix` by Baird and Sparks [2], which in turn uses Numba.
- `RAPIDS.AI`'s HDBSCAN only allows Euclidean Distance. To enable pre-computed distances, was used CPU's only version of `HDBSCAN`.

Chains of Length 2

General Results

- For all scores, the best number of clusters was 2.
- DBSCAN is the best method for SL, CH and DB. Entropy is best with K-Means.
- HDBSCAN came second most of the times.
- Euclidean Distance provided the best results across scores and methods, except for the WMD in DB score.
- All best results were obtained with doc2vec of 300 dimensions.
- Most clustered emails were short responses.
- It was not possible to detect a clear interaction pattern.

Score	Method	Label	N. Emails	%
SL	DBSCAN	-1	23.522	98,66
	(Euclidean)	0	314	1,32
		1	5	0,02
CH	DBSCAN	-1	23.522	98,66
	(Euclidean)	0	314	1,32
		1	5	0,02
DB	DBSCAN	-1	23.525	98,67
	(WMD)	0	314	1,32
		1	2	0,01
Entropy	K-Means	1	14.341	60,15
	(Euclidean)	0	9.500	39,85

Chains of Length 3

General Results

- For all cases, the best number of clusters was 2.
- DBSCAN was the best method across all scores.
- More variety of distances: WMD in SL and DB, Euclidean with CH and Cosine with Entropy.
- Best results were obtained with 50 dimension-doc2vec.
- We found no clear pattern of interaction.
- DBSCAN was able to allocate foreign language messages in a single cluster.

Score	Method	Label	N. Emails	%
SL	DBSCAN	0	8.766	99,97
	(WMD)	1	3	0,03
CH	DBSCAN	-1	8.639	98,52
	(Euclidean)	0	126	1,44
		1	4	0,05
DB	DBSCAN	0	8.766	99,97
	(WMD)	1	3	0,03
Entropy	DBSCAN	0	7.905	90,15
	(Cosine)	-1	863	9,84
		1	1	0,01

Chains of Length between 4 and 9

General Results

- Best number of clusters is two.
- DBSCAN is the best method in all scores.
- Euclidean Distance best in SL, CH and DB. Cosine similarity in Entropy.
- Doc2vec with 50 dimensions obtained best results.
- Neither in this group we were able find a clear pattern of interaction.

Score	Method	Label	N. Emails	%
SL	DBSCAN	-1	9.535	98,54
	(Euclidean)	0	139	1,44
		1	2	0,02
CH	DBSCAN	-1	9.529	98,48
	(Euclidean)	0	139	1,44
		1	8	0,08
DB	DBSCAN	-1	9.535	98,54
	(Euclidean)	0	139	1,44
		1	2	0,02
Entropy	DBSCAN	0	9.397	97,12
	(Cosine)	-1	278	2,87
		1	1	0,01

Chains of Length 10 or greater

General Results

- This subset contains the largest chains, going up to more than 700 emails in a chain.
- 2 clusters are best according to all scores.
- DBSCAN is best for SL, CH, DB, K-Means for Entropy.
- Euclidean Distance obtains best results across all scores.
- Best results with 300 dimension-doc2vec.
- None of the clustered emails contained a discernible interaction pattern.

Score	Method	Label	N. Emails	%
SL	DBSCAN	-1	5.229	98,22
	(Euclidean)	0	93	1,75
		1	2	0,04
CH	DBSCAN	-1	5.230	98,23
	(Euclidean)	0	92	1,73
		1	2	0,04
DB	DBSCAN	-1	5.230	98,23
	(Euclidean)	0	92	1,73
		1	2	0,04
Entropy	K-Means	1	5.323	99,98
	(Euclidean)	0	1	0,02

Discussion I

- DBSCAN dominated in almost every metric,
- Clustering methods with Euclidean distances performed better according to the selected scores and
- The preferred size of clusters in all datasets was $k = 2$.
- Aside from Entropy, Cosine Similarity performed worse than any other metric.
- WMD gave mixed results. Sometimes the best, but most of the time mediocre.
- It is not clear when to choose the size of dimensions doc2vec dimensions.
- DBSCAN's Epsilon Heuristic performed poorly.

Discussion II

- L2-Norm came close to the Euclidean distance in terms of quality (specially in the case of HDBSCAN), but never outperformed the Euclidean distance.
- Despite WMD's mixed performance, during the visual inspection we realised that this metric does detect a humanly-expected relationship of emails, and is even able to detect foreign languages.
- With respect to the visual inspection, the majority of resulting clusters are related to short answers to either formal emails or to personal conversations.
- Working with Enron Emails presented several challenges difficult to solve.

Conclusions

- Our results suggested that the proposed system is not capable of detecting a clear process from the analyzed emails.
- Since most of the clustered emails seemed to be short answers to another incoming message, our system seems to be able to work with Automatic Email Answering with corresponding adjustments.
- Future work with Social Network base approach.
- Should be a Pre-processing Golden Standard for Enron Emails.

References I

[1] Apoorv Agarwal et al. "A comprehensive gold standard for the enron organizational hierarchy". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2012, pp. 161–165.

[2] Sterling Baird and Taylor Sparks. *Distance Matrices*. 2021. URL: https://github.com/sparks-baird/dist-matrix.

[3] Jana Diesner, Terrill L Frantz, and Kathleen M Carley. "Communication networks from the Enron email corpus "It's always about the people. Enron is no different"". In: *Computational & Mathematical Organization Theory* 11.3 (2005), pp. 201–228.

[4] Abhishek Kathuria, Devarshi Mukhopadhyay, and Narina Thakur. "Evaluating cohesion score with email clustering". In: *Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019)*. Springer. 2020, pp. 107–119.

[5] Parambir S Keila and David B Skillicorn. "Structure in the Enron email dataset". In: *Computational & Mathematical Organization Theory* 11.3 (2005), pp. 183–199.

[6] Radim Řehůřek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, May 2010, pp. 45–50.

The End

Thank you

Q&A