# Data Science in Practice - FieldLab 3 - Support implementation Archive law with e-mail box analysis - Individual essay

Atish Kulkarni s2483122
Nick van der Linden s2971976
Maxime Casara s2465124
Thijs van Meurs s1828983

January 7th 2021

## Contents

**1 Introduction** **2**
1.1 Objective of the FieldLab . . . 2
1.2 Introduction of the Team and mode of collaboration, details of competences, division of tasks 2
1.3 Introduction to the client, the relationship with the client, mode of operation . . . 2
1.4 Canvas-considerations . . . 2
1.5 Operational objective and plan . . . 3
1.6 Methodology (including CRISP-DM) . . . 3

**2 Understanding the problem** **4**
2.1 Interviews, literature review and understanding of the problem . . . 4

**3 Considerations on Data** **4**
3.1 Availability . . . 4
3.2 Selection . . . 4
3.3 Data quality . . . 5

**4 Model** **5**
4.1 Description . . . 5
4.2 Design steps . . . 5
4.2.1 Pre-processing . . . 5
4.2.2 Exploratory data analysis with graph theory . . . 6
4.2.3 Topic extraction . . . 7
4.2.4 Deep Neural network model . . . 7
4.3 Results . . . 8
4.4 LDA applied on Enron dataset . . . 8
4.5 Reddit dataset . . . 8
4.5.1 Content . . . 8
4.5.2 LDA applied to Reddit dataset . . . 9
4.6 Limitations . . . 9
4.7 Analysis . . . 10

**5 Evaluation** **10**
5.1 Ethical and regulatory consideration . . . 10
5.2 Data FAIRness . . . 10
5.3 Questions for the FieldLab Owners (clients) . . . 10
5.4 Considerations for deployment . . . 11

**6 Own assessment** **11**
6.1 Group process . . . 11
6.2 Participation . . . 11
6.3 Effectiveness . . . 11
6.4 Lessons learned . . . 11
6.5 Points for improvement . . . 12

# 1 Introduction

*The submission method chosen is hybrid: part 1 to 5 is of common work, while part 6 was redacted individually*

The following report summarizes our work on the FieldLab *Support implementation Archive law with e-mail box analysis* under the supervision of the Data Science in Practice team of Leiden University and in collaboration with the Ministry of Economic affairs and Climate policy and the Ministry of Agriculture, Nature and Food quality of The Netherlands. In this project, we follow the Cross-industry standard process for data mining (CRISP-DM) [3] and carefully covers each steps. Our result is a text pre-processing model coupled with an LDA [4] algorithm that can train on a corpus of documents and accurately assign topics for each documents, as well as a Locality Sensitive Hashing (LSH) [5] method to find out most important or most similar emails.

## 1.1 Objective of the FieldLab

E-mail filtering is a common task in the field of Machine Learning [1]. The main challenge of the project will be to implement such a classification algorithm for our specific purposes. This algorithm will gather similar mails into specific topics and then order them by importance.

## 1.2 Introduction of the Team and mode of collaboration, details of competences, division of tasks

Our team consists of 4 members:

- Atish Kulkarni :- Data science
- Nick vd Linden :- Data science
- Thijs van Meurs :- Data science
- Maxime Casara :- Advanced Data Analysis

Our team originally applied for FieldLab 16 as our preferred project. Our shared interest at that time was the opportunity to develop an image recognition method on natural ecosystems that assesses the effects humans have had on it. Unfortunately, the FieldLab 16 was discontinued, and our very first challenge as a group was to find a new FieldLab to apply for. Thijs rapidly suggested FieldLab 3, and after some discussions, we decided it was the better compromise between shared interest and simplicity of execution. From there on, the distribution of task became sort of natural. Everyone seemed to have fit the position appealing them the most. Thijs took the responsibility of spokesperson of the team, regularly sending mails to the client and preparing the meetings. Nick took care of the deliverables of the group, mainly assignments, by making sure everything was in order and submitted in time. Atish and Maxime focused on implementing the solution for the FieldLab, with Atish working on exploring and pre-processing the data and developing an LSH method, and Maxime evaluating and tuning the LDA model for the task at hand.

## 1.3 Introduction to the client, the relationship with the client, mode of operation

The project is conducted internally by Stephany N. and Tomer .G, currently trainees at the Ministry of Economic Affairs and Climate Policy. They are both are participating in a data analysis trainee course and are our direct contacts. We contribute to their work by proposing a solution for mail classification and showing that the implementation is feasible.

## 1.4 Canvas-considerations

Our very first deliverable is the project Canvas. This is a useful step to define the scope of our project and what actors, resources and cost are at stake.

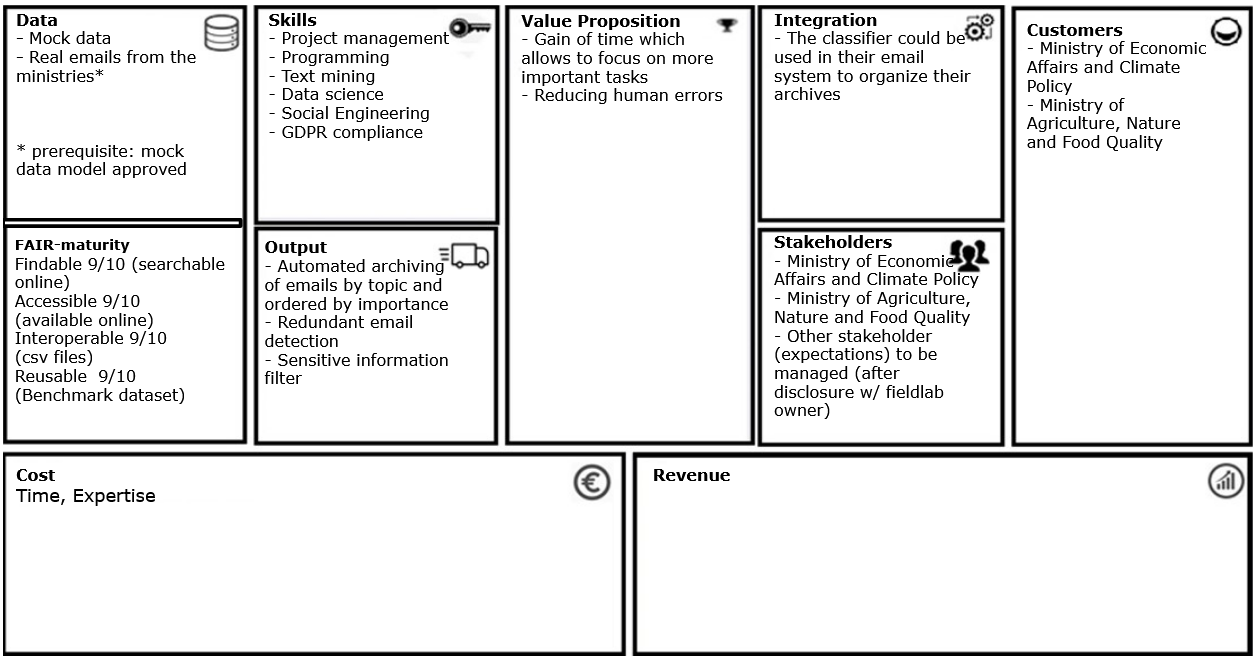


Figure 1: Project Canvas. Formally setting up a project canvas helps us not steer away out of scope from our objectives.

This document is not meant to be substantially updated. We regularly came back to it to check whether our work aligned with the Output and Value Proposition stated in the document.

## 1.5 Operational objective and plan

There are a few main steps we uncovered to carry out the short-term goals (Operational objectives) of our project:

- the first step was to inspect the data. In this phase, we collected the data and explored its properties. By doing so, we got an overview of how the data is structured and its general quality, so that we could determine how to make the best use of it.
- the second step was to pre-process the data so we can work with it. This involves selecting the relevant data points, cleaning the data, and formatting it in a way that is practical to us. The goal of this step was to make sure the data would be ready to work with in our next step. This step is the most important one and directly influenced our results. In most text mining project, this is usually where the challenge lies.
- the third step was to build a modelling plan. This means selecting the right modelling technique for the job, building the model, and building some way to verify the accuracy of the model. We explored different models after pre-processing the data. This gave us more insight on the task at hand and allowed us to choose a model that fit the best.
- the fourth step was to carry out the experiments to evaluate the performance of our final model. To this end, we had to generate our own database on which to perform LDA algorithm, as the original data proved to be unfit for the task.

## 1.6 Methodology (including CRISP-DM)

Throughout the development of the project, we carefully followed the Cross Industry Standard Process for Data Mining (CRISP-DM) as a guiding line. This method allowed us to successively do the following:

- Understand the business, its objectives and success criteria. It required us to formally define the environment of the project such as the background, the actors and the context. It concluded with producing a project plan.
- Understand the data, explore it, describe it and assess its quality.

- Prepare the data and clean it for the purposes of our project. The Enron mail Dataset proved to be very challenging to clean.
- Choose a modelling technique. For the data modelling, we used common Natural Language Processing (NLP) methods, which are very efficient for text mining tasks. For topic classification, one of the state-of-the-art methods is the Latent Dirichlet Allocation (LDA) algorithm. This is the algorithm we studied in our project.
- Evaluate our model. We created a benchmark dataset to assess the performance of our LDA model.
- Deploy and implement the solution. This is the final step of the project. Unfortunately, our client never provided us with real data that would allow us to tune our model for a trial implementation.

# 2 Understanding the problem

## 2.1 Interviews, literature review and understanding of the problem

After our first interview with the stakeholder, we were able to structurally analyse the project definition. The problem at hand was quite clear from the start: divide and segment e-mails based on topic and ranked on (subject)importance. However, since the GDPR is in place, and we would be working with real data from real people, we are not allowed to build our models using their data initially. We will be using mock data (Enron dataset) to test our process and models. If they perform well, and all the caveats are taken care of, we can move on to the next stage. This would mean we can use their data. This would also mean that the cycle begins anew, since we are starting with fresh data. This data will need to be pre-processed and cleaned thoroughly, because real data will contain noise. Another caveat we must keep in mind, is that the mock data is in English, whereas many e-mails in the ministry would be in Dutch. If the models are set up correctly, this would only pose a minor inconvenience at best. We have made clear that the only resources we will need are ourselves and possibly the LIACS servers. The latter depends on both the magnitude of the data, as well as the confidentiality in which we must act: LIACS servers are more secure than home servers.

# 3 Considerations on Data

## 3.1 Availability

The data discussed in this section is mock data. The structure of the real data may vary in all aspects. It is requested by the problem owner to use the Enron dataset to show our competence before transferring to the real (sensitive) data. Acquiring the dataset was simple. We only needed an email registration at the data source website. Some members of our team downloaded the dataset from an alternative source and found that the layout of the data was stored in a different format, namely a folder of folders. This was hard to work with, because it was too big to upload anywhere and was not stored in a way that was easy to access.

The data consists of approximately 500,000 emails from 150 users, mostly senior managers of Enron. It was initially published by the Federal Energy Regulatory Commission during its investigation, which eventually led to the Enron scandal[1]. It was later collected and prepared by the CALO project, and finally corrected by Melinda Gervasio and her team at SRI International.

## 3.2 Selection

The data is stored in a csv file. Each row is a separate email, and each column is a variable. The variables are column-a, message-id, date, from, to, subject, x-from, x-to, x-cc, x-bcc, x-folder, x-origin, x-filename, content, user, level1, level2 and weight columns for 12 categories, and labelled. Both column-a and message-id are id-numbers for the emails; The difference between the two is not very clear. The next eight columns correspond to usual email headers information. x-folder describes the folder in which the email was stored in the Enron file system. x-origin and user both describe the sender of the email. x-filename is, presumably, the name under which the email was stored locally. Content describes the text of the email. This is the field we are going to use to train our model. The category columns are empty columns, presumably added to aid in classification tasks. 'labelled' describes whether the emails are labelled or not, which often is not the case. A preview of the data is shown in the figure below:

[1] https://www.investopedia.com/updates/enron-scandal-summary/

| | date | from | to | content |
|---|---|---|---|---|
| 0 | 2002-04-17T15:15:53 | frozenset({'chris.germany@enron.com'}) | frozenset({'jackie.adams@cingery.com'}) | Jackie For May 2002, Ponderosa Pines is electi... |
| 1 | 2001-03-30T06:09:00 | frozenset({'chris.foster@enron.com'}) | frozenset({'kim.ward@enron.com'}) | Kim: Read this over and give me a call. ------... |
| 2 | 2000-10-04T07:25:00 | frozenset({'kristian.lande@enron.com'}) | frozenset({'kate.symes@enron.com'}) | The link below is to the West Power Structurin... |
| 3 | 2000-08-09T15:33:00 | frozenset({'john.lavorato@enron.com'}) | frozenset({'jdesouza@royallepage2.com'}) | That's great. Please don't let Len delay. |
| 4 | 2001-09-20T22:41:30 | frozenset({'barry.tycholiz@enron.com'}) | frozenset({'t..lucci@enron.com'}) | Paul, let's try to find some time on Friday to... |

Figure 2: Data description

Only the column 'content' which contains the message in the mail will be useful for training our model. Additional features like the subject of the mail will be considered when trying to improve the model, while other columns will be discarded.

### 3.3 Data quality

The data contains what we expected: a list of mails and their content detailed by several features. Some observations can be made about the content column: first, the content is nicely separated from other features and additional information we are not going to use. This will make the data pre-processing task a bit simpler. However, many mails contain specific information of a 'Forwarded by' chain, which include the email of the recipient, the email of the sender, the email of each of the carbon copies, and the redundant information about subject and date. All this extra information must be removed from the content of the mail, and a first effort to do so proved unsuccessful due to the inconsistent nature of the information in these mails. An eventual solution would be to remove these types of mail from our data, if deemed acceptable after gauging the information loss it would entail. Regarding the subject column, it does hold a few missing values; roughly 3.8 percent of the emails have no subject. Another issue is that the data is all in English, whereas the real data will probably be in Dutch. This is something we will have to consider when working with the data.

## 4 Model

### 4.1 Description

We will be choosing our models based on the data which need to be transformed to form coherent and understandable results. Our data is entirely text based; therefore, we are using models which can operate on textual data. This field is known as Text Mining, and the data is per definition unstructured.

### 4.2 Design steps

#### 4.2.1 Pre-processing

The first step when working with text data is to remove all superficial contents which are not going to help with the classification. For example, after taking a manual look in the data we discovered that every mail id has a tag named "Frozen set" along with it. Such tags do not help our model to learn features hence it was removed. Also, the are some examples in the data where the conversation between employee has short forms used. These mails are small, and the few words present are the only resources in such mail. Hence, we change all such cases with the actual words intended in a normal English language.

For example:
Tbt - throwback times
Tba - to be announced

Along with this all the short forms such as "'ll" are converted to "will". This reduces the overall size of dictionary in terms of unique words. Similarly, techniques such as lemmatization and stemming can be used to put words back to their roots.

For example:
Coming - come
Written - write

Such techniques help to concentrate the dictionary into even fewer words which should help in improving model performance.

Since our task was to determine the topic of emails, there were a lot of columns that did not hold information that was relevant to us. This includes the date, x-cc, x-bcc, x-folder, x-origin and user columns. We also excluded the level1, level2 and weight columns, since for our classifications we will not store the results in the columns, thus they have no use for us.

We also chose to leave out the subject column. This is because we found that the subject title of emails often related to the content well enough that it was not necessarily helpful to include it.

#### 4.2.2 Exploratory data analysis with graph theory

Given the dataset type this data can be also treated as a graph. In such a graph Every email id is a node and link exist if there was an email sent from one node to another. This graph can be treated as directed or undirected. Graph analysis allows to explore data in detail and find important nodes from the network. Other statistical factors like density, in-degree distribution, out-degree distribution, average clustering coefficient can also be found.
When the data was processed in such a way, we can visualize data according to preference. This would allow the author of data to secure the network of emails. Our discoveries of such analysis can be found in figures below. To process the visualization "Gephi" software was used. During this analysis for resolution value 3 about 8 clear communities can be found with modularity score of 0.67. Which means inside a network of emails there are 8 sub networks. For every such network all the details can be extracted in the same way as of main dataset. This exploration helps in getting insights which would not be possible with simple classification. As mentioned in LDA we need to define how many topics we want to classify emails into. As 8 communities are found it is safe to say that we can classify all emails into 8 total topics.

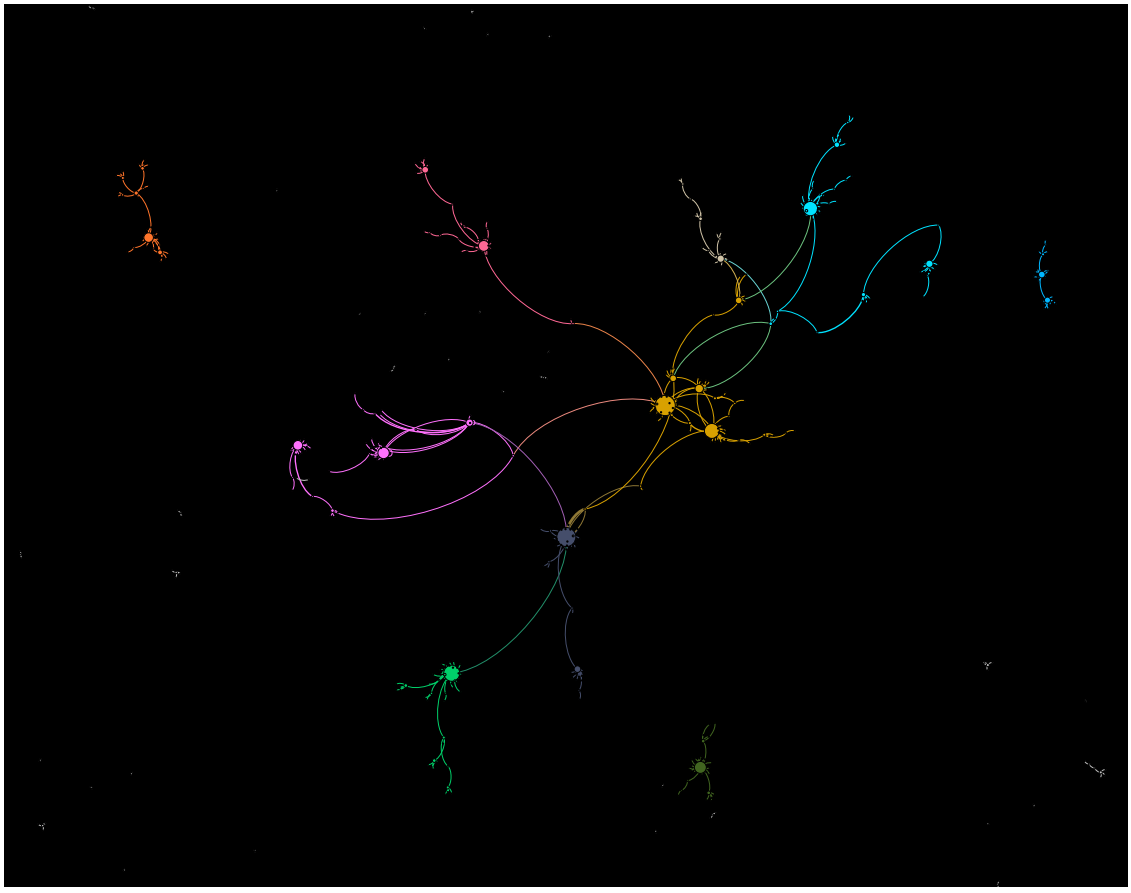

Figure 3: This is a overall visualization for Enron email dataset. In this graph there are 10 clear communities. These communities are obtained using the Louvain Algorithm [2]. Resolution value used is 3 and modularity score obtained is 0.67. Size of nodes was set based on degree of the node to get overall idea of email traffic at a node. In every community there is one large sized node representing the manager or an important position holder for that community.

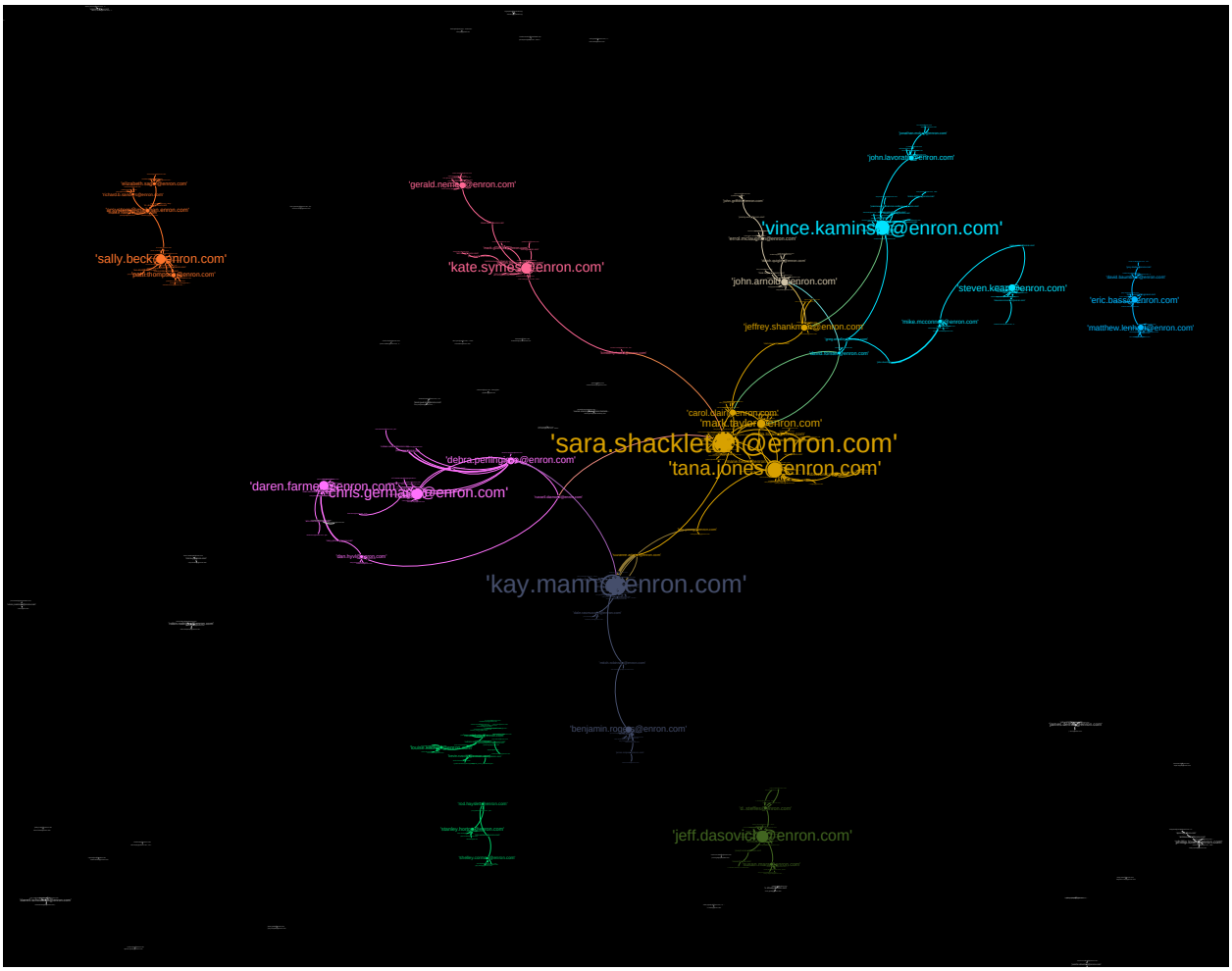


Figure 4: In this image, the same Enron email dataset can be seen along with some filters. When compared to Figure 3, some details become clear. For this image a giant component filter was set along with a range filter for the degree of the nodes. From the filtered results names "Kay Mann", "Tana Jones", "Sara Shackelton" have the highest degree. It is surprising to see that the employee list does not have any record for two of these names but have record for all other employees.

### 4.2.3 Topic extraction

While writing emails as everyone uses different types of formats to summarize their email into subjects. As there is no drop-down option provided on the website, we may get many different phrases in the subject section which ideally should be under one collective subject.

For Example:
Subject – Request for housing allowance
Subject – About the rental benefits

As we can see in the examples above there two different subjects mentioned but they point towards a same parent topic that is finance. To handle such cases Latent Dirichlet Allocation (LDA) can be used. LDA is an unsupervised learning method which extract topics from a given set of documents or sentences. This allows us to assign proper labels for every email. To process every email in LDA it must be converted into vectors first. Every email is represented as vectors. To achieve this NLTK's word2vec model can be used. Number of topics to extract can be set according to how much separation we want between emails. Number of words to assign for every topic can be also customized. This allows us to limit the description about the topic.

### 4.2.4 Deep Neural network model

LDA method only returns a list of words and their weight for each topic. A neural network model would help to classify these emails based on these given words. We will train this neural network by using the

labels generated with LDA (Latent Dirichlet Allocation). Using a neural network along with LDA allows us to tune and customize the final model which would not be the case if only LDA is used.
The model used for the first trial run is the FCN (Fully Connected Network). As it is the first run, and we are optimizing the hyper parameter values with all the previous processing we are sticking with simple FCN. But as the email classification is a NLP task we would like to explore possibilities with other types of neural network such as LSTM (Long Short-Term Memory) or RNN (Recurrent Neural Network). In many research papers LSTM and RNN have been previously proven to be the best type of neural network to process a language data as it solves the vanishing gradient problem popular with the other types of neural network. In the next stage all the processing before the neural network would be finalized.

## 4.3 Results

## 4.4 LDA applied on Enron dataset

Applying LDA algorithm to the content of the Enron dataset did not bear any cohesive results. This is because the dataset does not contain enough specific topics. In the following figure, we plotted the word clouds for 4 topics generated by LDA.

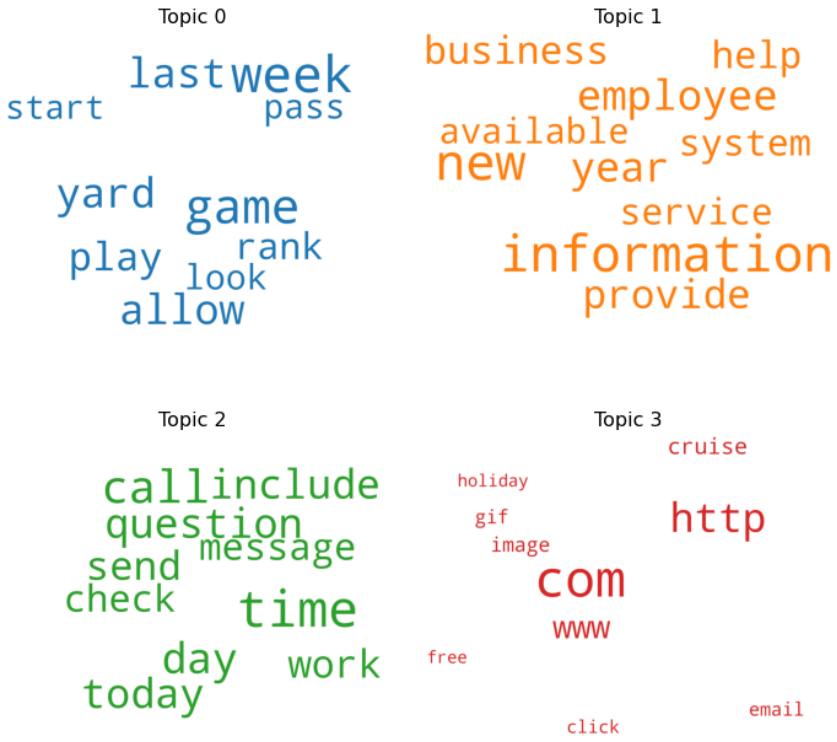


Figure 5: Word clouds of 4 topics generated by LDA on the Enron dataset

As we can see, the words defining each topic do not belong to a same subject. The Topic 3 seemed to have gathered spam mails which usually contain web-links, and which LDA considered a subject. Therefore, the Enron dataset is not fit to evaluate the performance of LDA algorithm. We decided to create a new dataset and select the content ourselves to provide LDA algorithm coherent data to work with.

## 4.5 Reddit dataset

### 4.5.1 Content

Reddit is a community-driven social news website which allow users to create posts in different subreddits. Each subreddit is defined by a topic and gather a varying community size. We chose 3 popular subreddits and collected 1000 post titles from each. The content of our new dataset is shown in the figure below:

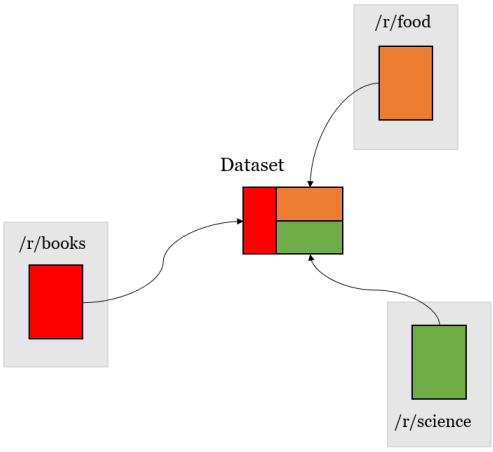


Figure 6: Reddit dataset content. 1000 post titles from the subreddits books, food and science

We now have a dataset with 3 identifiable topics in its content on which we can apply an LDA algorithm. The motivation to use this dataset is that we know there are 3 topics and each item belongs in one. This information will be useful to assess the model performance by calculating how likely items from the same topic are grouped together.

### 4.5.2 LDA applied to Reddit dataset


Figure 7: Word clouds of 3 topics generated by LDA on the Reddit dataset. Each topic describes a specific theme. Topic 0 gathered the posts from the books subreddit, Topic 1 from the science subreddit and Topic 2 from the food subreddit

From the word clouds above, we see that the 3 original topics have been successfully recovered by the LDA algorithm. Each topic describes a precise theme with a set of words that are coherent with each other. And because we know the content of the dataset and in which theme each post title belongs; we can evaluate an experimental accuracy of the LDA algorithm. In figure 8, we plotted the distribution of post titles over the 3 generated topics for each subreddit. Over the 3 topics, the LDA algorithm achieved an average accuracy of **94.4%**. This verifies our hypothesis that the accuracy of a tuned LDA model will entirely depend on the quality of the data.

## 4.6 Limitations

It is debatable whether post titles from Reddit are comparable to mails. First, a mail is usually longer than a post title. However, post titles do not use as many stop words. Therefore, after pre-processing the data, during which stop words are removed, mails shrink in size whereas post titles remain the same. Then, content in mails can belong to more than one topic. It is a problem if it is true for most mails. If not, and if topics in the corpus can be accurately separated, then the LDA algorithm returns for each mail its score on each topic. This means the LDA algorithm can recognize more than one topic in each mail. In our previous results, each mail would be assigned to the topic it scored highest in. However, this was only a useful way to visualize results and evaluate the accuracy of the model.

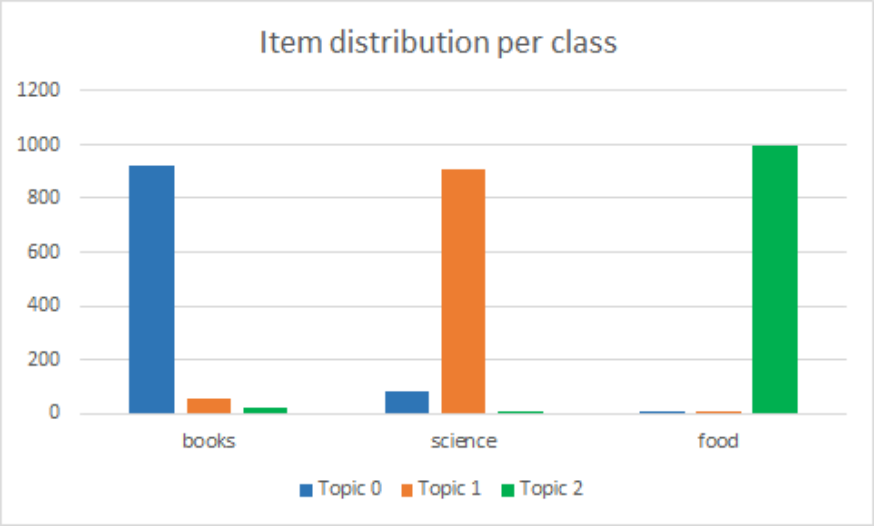


Figure 8: Distribution of post titles for each subreddit

### 4.7 Analysis

We demonstrated the performance of the LDA algorithm on a dataset that contains specific and identifiable topics. This was not the case for the Enron dataset, which we verified as not fit for the task of topic classification. Because of this, we created our own dataset which contains 3000 post titles from the website Reddit. We discussed the limitations of using post titles and whether they are comparable to mails and concluded that their differences were negated after applying our pre-processing method. Furthermore, it is reasonable to believe that the LDA algorithm will behave the same for mails if the quality of the data is like our benchmark dataset.

## 5 Evaluation

### 5.1 Ethical and regulatory consideration

The real data we could be working with might be sensitive for the general public: some mails may contain private information or personal details. The stakeholders wanted to be assured that data which unveils personal information will be dismissed or anonymized. We therefore created a filter to pre-process the data before analysis. The distinction between personal and non-personal information is documented and has been discussed with the stakeholders.

### 5.2 Data FAIRness

We attributed a FAIRness score of 8/10 for our data. We can detail the information from our Project Canvas figure 1:

- Findable 8/10: The dataset is registered in a searchable resource and described with rich metadata
- Accessible 8/10: The dataset is accessible online. It was published by the Federal Energy Regulatory Commission during its investigation on Enron, collected and prepared by the CALO project and finally corrected by Melinda Gervasio and her team at SRI International.
- Interoperable 8/10: The data is stored is csv files, which are common data storage format.
- Reusable 8/10: The Enron dataset has been used in many text mining projects, for example as a support for fraud investigation [6] and as a benchmark for a spam filter algorithm [7].

### 5.3 Questions for the FieldLab Owners (clients)

As stated earlier in the report, the purpose of our work on mock data was to show the feasibility of building a classifier which assign topics to mails and rank them by importance. We also discussed about privacy related issues, and how our model is able to anonymize the data. We believe to have met our objectives regarding the proof of concept. We did not receive any feed back or hint of inclination on implementing the model on real data yet.

### 5.4 Considerations for deployment

As for most machine learning algorithms, the main challenge will be to build a training set on which our model could be tuned. Although we showed the performance on a personalized dataset, we already practiced with the enron dataset. Hence the mail format is not a problem and can be handled by our pre-processing aglorithm. All in all, taking advantage of our model would only require to have available mails and to define the different topics we wish to assign to them.

## 6 Own assessment

In this section, I (Maxime Casara s2465124) will present my own assessment of the project. I will first discuss about the group dynamics and how the distribution of tasks felt very natural, then I will evaluate my contribution to the project, before analysing the effectiveness of our methods and conclude with the lessons we learned and the points for future improvement. All in all, I'm convinced our group was effective in producing a detailed deliverable.

### 6.1 Group process

*Paraphrasing 1.2 which fit this section*

Atish, Nick, Thijs and I originally applied for FieldLab 16 as our preferred project. Our shared interest at that time was the opportunity to develop an image recognition method on natural ecosystems that assesses the effects humans have had on it. Unfortunately the FieldLab 16 was discontinued, and our very first challenge as a group was to find a new FieldLab to apply for. Thijs rapidly suggested FieldLab 3, and after some discussions, we decided it was the better compromise between shared interest and simplicity of execution. From there on, the distribution of task became sort of natural. Everyone seemed to have fit the position appealing them the most. Thijs took the responsibility of spokesperson of the team, regularly sending mails to the client and preparing the meetings. Nick took care of the deliverables of the group, mainly assignments, by making sure everything was in order and submitted in time. Atish and I focused on implementing the solution for the FieldLab, with Atish working on exploring and pre-processing the data and developing an LSH method, and I evaluating and tuning the LDA model for the task at hand.

### 6.2 Participation

My main contribution to the project are the LDA model and the generated reddit dataset to evaluate its performance. Having already attended the Text Mining class before my teammates, I was the most at ease with implementing the LDA model. It was still quite a challenge: the decision to make a new dataset was on my initiative after repeatedly failing to get valuable results with the Enron dataset. I realized late that implementing an effective topic modelling algorithm to the Enron dataset would require a substantial number of extra pre-processing steps. As it is originally, the dataset contains a large number of spam mails. And unless removing the majority of them, the LDA algorithm would train on flawed inputs and produce bad results. I took the decision to show the performance of the LDA algorithm on a benchmark dataset rather than working further with the Enron dataset. It carried out exceptionally good results. We hoped the demonstration would give justification to our demand to work with real data, which our stakeholder could provide. Unfortunately we could never get hold of it.

### 6.3 Effectiveness

It's hard to say whether our methods were effective or not, for lack of relevant comparison. I am personally content with our work, and believe that everyone contributed to the project to a satisfying degree. There was no tension, and we all got along together very well. Our meetings could last up to 2 hours and were always productive. After every meeting we would divide the short term tasks and decide for a date for the next meeting. By having weekly meetings we made sure no one would be left out. However it is possible to owe the lack of hassle to the continuous guidelines provided by the DiP team. With guidelines, everyone was up to date with the assignments and tasks. That way, there was no place for hard misevents that would disrupt our productivity.

### 6.4 Lessons learned

I really enjoyed the Data Science in Practice class. It came to me during the final presentation that throughout the class, we were experiencing both meaning of 'practice'. The first is training, as in carrying

out a project with the newly learned CRISP-DM methodology, and attending classes from a broad range of different speakers, which helped us develop our Data Science culture. The second one is application, as in cooperating with real stakeholders on a real world data science problem, in a formal environment and with real challenges at stake. It is unfortunate that we couldn't present the project in person. I believe it would have added a layer of valuable experience to the class. The online presentation was still carried out very professionally.

### 6.5 Points for improvement

In regards to the stages of group development, I noticed our group did well in the storming stage, wherein the group starts to work with each other and compatibility problems can arise. We might however have lacked leadership. Us being all students with relatively equal experience, no one really took a stance of the leader of the project. While it prevented status imbalance and tension, I believe a leader in our group could have catalyzed the project to a faster pace.

## References


[1] Application of Machine Learning, https://www.javatpoint.com/applications-of-machine-learning

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, andE. Lefebvre. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008.

[3] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C.; Wirth, R. (2000): CRISP-DM 1.0. Step-by-step data mining guide

[4] David M. Blei, Andrew Y. Ng, Michael I. Jordan (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research 3 (2003).

[5] Aristides G., Piotr I., Rajeev M. (1999) Similarity Search in High Dimensions via Hashing. Proceedings of the 25th VLDB Conference, Edinburgh, Scotland, 1999.

[6] Machine Learning with Python on the Enron Dataset

https://williamkoehrsen.medium.com/machine-learning-with-python-on-the-enron-dataset-8d71015be26d

[7] Empirical Analysis on Email Classification Using the Enron Dataset

https://towardsdatascience.com/empirical-analysis-on-email-classification-using-the-enron-dataset-19054d558697