University of Amsterdam

MSc System and Network Engineering
Research Project 2

Improving Semantic Quality of Topic Models for Forensics Investigations

by
Henri Trenquier
11758929

July 8, 2018

6 ECTS
July 8, 2018

Supervisors: Dr Rob van Nieuwpoort,
Dr Carlos Martinez Ortiz

Abstract

In this research, we study the quality of a topic modeling technique called *Latent Dirichlet Allocation*. We wonder if it could be reliably used in forensic investigations to cope with large evidence datasets. We make several measurements to understand the impact of two important parameters on the quality of *LDA* models. For that, we consider two existing evaluation metrics to estimate the quality of a model and design our own. The dataset used in the experiments is the Enron e-mail database which is often used in *Natural Language Processing*.

Contents

1 Introduction

E-mails have been the most popular means of communication in companies since the birth of the Internet. When a company is accused of running fraudulent schemes, it can be asked to hand over registered communications. E-mails as the oldest communication channel, will probably represent the largest dataset for forensic investigations.

The main purpose of this project is to know how Natural Language Processing (NLP) can alleviate forensics tasks in analyzing large datasets. Our dataset will be the popular Enron's email database. This dataset is used a lot to train models in NLP. In this paper, we will focus on the Latent Dirichlet Allocation (LDA) model [4]. LDA infers topics out of the frequency of occurrence of words in a set of documents. In other words, LDA can organize a collection of documents according to their prominent topic. This would allow the investigators to focus on the most relevant topics according to criminal activity under investigation. Some considerations should be taken such as assuming the e-mails that contain the incriminating information might not be obvious but could form considerable exhibits to create a pattern of behavior. Thus, this research studies how accurate can the tool be.

We deal, in this research, with the idea of *coherence* of a topic. This should be seen as how precisely a specific list of words define a topic. In topic modeling, the coherence of each topic and the coherence over the set of topics cannot be measured in a classic scale nor has a unit. Indeed the coherence in a topic is a human appreciation. Several coherence measures have been developed such as C_V and C_{UMASS} [13]. We compare them in this study. The aim of the study is to evaluate the quality of the topics given by LDA depending on the number of topics and the number of iteration

In section 2, we describe the State of art of topic modeling and coherence. Section 3 defines the research question. Then, in section 4 we describe and try to give an intuition of LDA. Section 5 describes coherence as it can be found in the literature. The preprocessing phase of the experiment is explained in section 6. We describe and analyze the relevance of these coherence measures for our case and define a novel coherence metric. Finally, in section 7, we try to find an optimal number of topics according to the tested coherence metrics. We also study the impact of the number of iteration on the quality of LDA models in section 8. We discuss the results and point out some limitations of this research in section 9.

2 State of the art

Natural Language Processing (NLP) is a field of computer science and Artificial Intelligence (AI) that aims at analyzing and interpreting natural language. The past 20 years, many algorithms processing texts on webpages have been developped. They are able to recognize sentences or count words for example. However, even today, it is still not possible for a computer to interpret natural language. Most recent AIs can only understand simple tasks thanks to keywords.

Over the years, NLP research naturally split into many branches in a divide-to-conquer strategy. According to E. Cambria and B. White's review article [2], this research joins the endogenous NLP branch. E. Cambria and B. White list 3 main advantages for this method which are the effectiveness, the low man expert requirements and the portability to different domains. The purpose of Latent Dirichlet Allocation (LDA)[4] is to define the most relevant topics dealt with in a set of documents. Thus LDA only uses endogenous knowledge as opposed to external knowledge bases to generate the topics. Quality of topics can only rely on human appreciation. D. Newman et al. developed a coherence measure to automatically evaluate the coherence of topics [11]. In 2015, H. Hong and T.-S. Moh present a method [7] to sort e-mails based on their prominent topic thanks to LDA. However, their evaluation metric is not standard. In 2011, Mimno et al. benchmark several evaluation metrics that can be found in the literature [9]. Two years later, S. Arora et al. [1] introduce the idea of *inter-topic similarity*. Until then, coherence has not been literally defined nor the difference between coherence of a topic and coherence of a topic model. In this paper, the coherence of a LDA model is computed based on the average coherence of all the topics in the LDA model [10].

L. Xie et al. deliver a solution [15] for a forensic analysis but focusing on the network of people. Combined with topic modeling, this solution could be more efficient. Associating people's connections with the prominent topics of their discussions could be effective to find evidence.

3 Research Question

How to improve the quality of LDA models?

In order to answer the main research question, the following sub-questions are defined:

- *What is the optimal number of topics for a LDA model*
- *How does the number of iterations influence the quality of models?*
- *Can we improve the semantic quality evaluation?*

For this research project I will try to answer the previous questions with a precisely defined framework to analyze the effectiveness of LDA. The reproducibility of research is important, this is why the python scripts are available online [6]. We will run the experiments on the Enron e-mail dataset[3], which is frequently used in NLP researches.

4 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a bayesian statistical generative model capable of inferring groups of elements in a dataset based on similarities. In our use-case, the dataset is a collection of e-mails and the similarities that we wish to highlight are the topics dealt with in those messages.

LDA is based on several assumptions:

- Documents (e-mails) with similar topics will use similar lexical fields
- Topics are represented as a statistical distribution of words such as Figure 2(a)
- Documents are represented as a statistical distribution of topics such as Figure 2(b)

A very simplistic explanation of the algorithm used to build LDA models is given by Algorithm 1

Data: docs (list of doc), doc (list of word), topics (list of t)
Result: LDA Model
initialization;
set all words equally likely to appear in each topic
set all topics equally represented in each documents
for *doc* ∈ *docs* **do**
  for *word* ∈ *doc* **do**
    for *t* ∈ *topics* **do**
      Update the probability of *word* in *t*
    end
  end
end

Algorithm 1: Model training algorithm

The weight of each word is updated depending on:

- the frequency occurrence of **word** in **doc**
- other topics
- prior variables arbitrarily defined by David M. Blei et al. [4]

In the *Gensim* python library [12], the initial state of the LDA implementation is randomly chosen. This choice introduces a non deterministic effect on the results of the LDA model.

4.1 Example

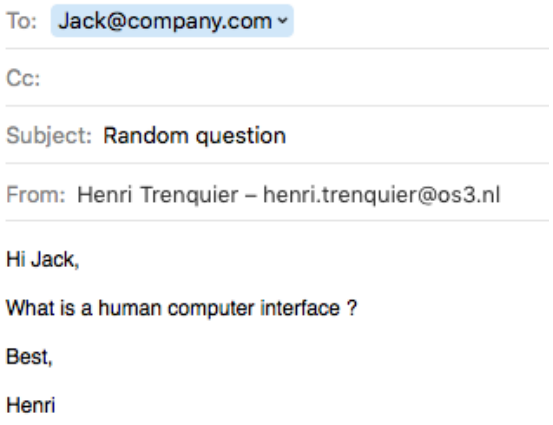


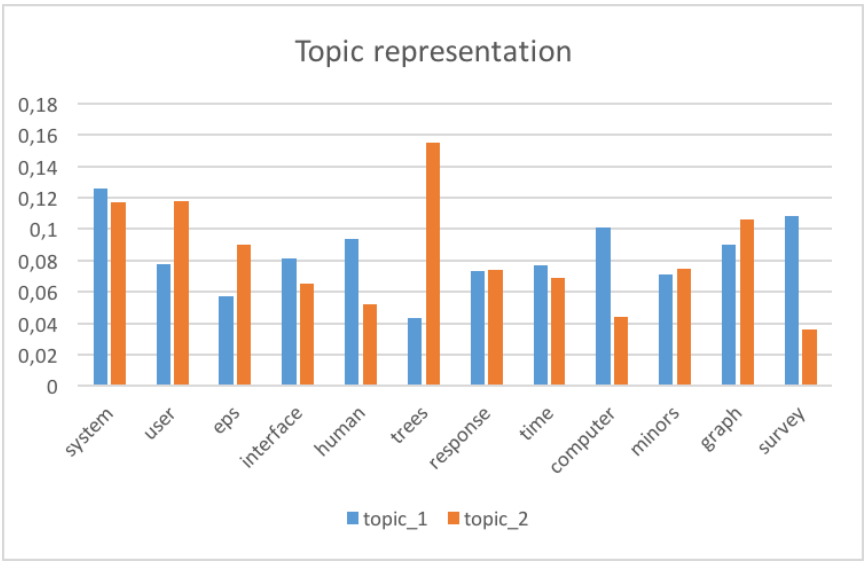
Figure 1: E-mail example

To draw an example, we assume the following e-mail Figure 1. In order to extract the most meaningful words from this e-mail, we need to remove the

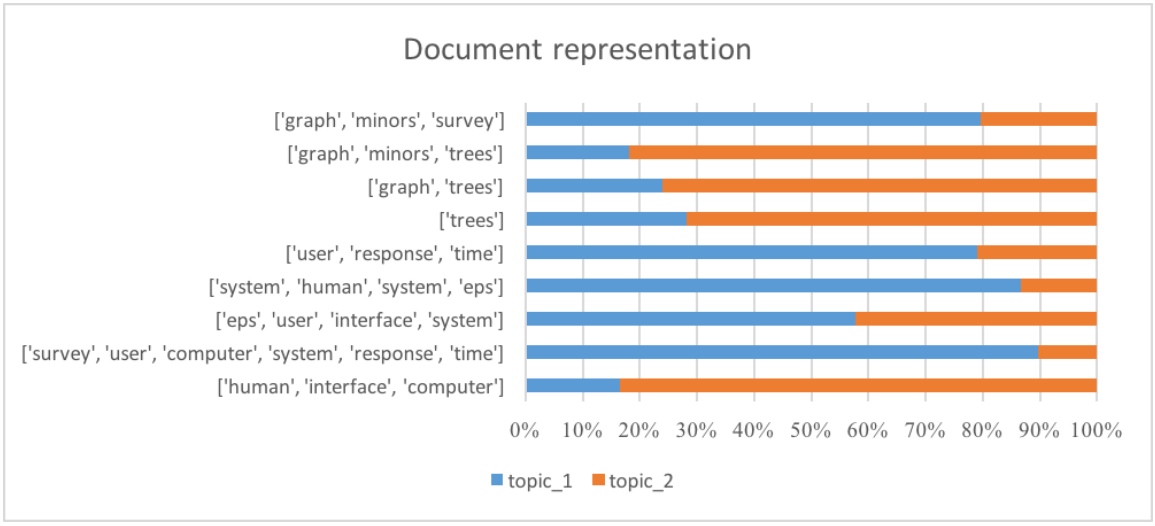
e-mail headers, signature, courtesy parts. This step is called preprocessing and is explained in section 6. In this example, we only want to keep the 3 words: 'human', 'computer' and 'interface'. The words are put in a list that will represent the document. In this list, the order doesn't matter and we keep the multiplicity of the words. Such list is called a *bag of words*, or *bow*. We can now assume the following set of documents, each formatted in a *bow*.

1. 'computer', 'human', 'interface',
2. 'computer', 'response', 'survey', 'system', 'time', 'user'
3. 'eps', 'interface', 'system', 'user'
4. 'eps', 'human', 'system', 'system'
5. 'response', 'time', 'user'
6. 'trees'
7. 'graph', 'trees'
8. 'graph', 'minors', 'trees'
9. 'graph', 'minors', 'survey'

When the LDA model is generated, each topic is represented as a probability distribution over all the words of the corpus as in Figure 2(a). Also, each document can be given a share of all K topics as in Figure 2(b). In this example, we have 2 topics and 9 documents.



(a) Topic representation



(b) Documents representation

Figure 2: Topic and document representation in LDA models

We will use the N most probable words also called *top words* of each topic to evaluate these topics. For this example, N is equal to 5. We arbitrarily choose to use a N equal to 10 to represent the topics in the main experiment.

5 Coherence

In this study, we define *coherence* as the humans' semantic appreciation of a topic represented by its N top words. Coherence is closely related with

human impression and thus is hard to translate in a mathematical formula. Hence, the scoring model performance at predicting human scores is usually evaluated with surveys such as in Newman's work [11].

In order to evaluate our models, we used 2 different coherence measurements from the literature: U_{MASS} and C_V. They are respectively the fastest and the most accurate according to M. Röder et al. [13]

The C_V measure is based on the *Pointwise Mutual Information* score (PMI score). The PMI is defined as in Equation 1

$$PMI(w_i, w_j) = log \frac{P(w_i, w_j) + \epsilon}{P(w_i)P(w_j)} \tag{1}$$

$P(w_i)$ is the probability of finding the word w_i in a random document of the trained model. $P(w_i, w_j)$ is the probability of finding both words w_i and w_j in a random document of the trained model. The probabilities are estimated from the trained model. U_{MASS} estimates the probability as "the number of documents in which the word occurs divided by the total number of documents" [5](CoherenceModel module). According to M. Röder et al. C_V estimates the probability with a sliding window over an external training dataset such as *Wikipedia*. But the implementation of the coherence measures are computed on the LDA model itself in the Gensim library.

The U_{MASS} metric accounts for the ordering of the top words of a topic. The PMI score used for U_{MASS} is slightly different and noted $PMI*$ in Equation 2

$$PMI*(w_i, w_j) = log \frac{P(w_i, w_j) + \epsilon}{P(w_j)} \tag{2}$$

5.1 word2vec coherence

A limitation in the implementation of these measures is that they are computed from the trained topic model itself. Thus, we want to make an estimation that is based on an external model to avoid the bias of our own model. We also want to introduce the semantic value of the top N words. We choose to use a word2vec model [14] trained on Google News.

Word2vec[8] is an open source project developped by Google researchers. It implements a vector representation for words in a semantic space. The model has been trained on a large set of Google News articles.

We will use the similarity feature of the word2vec model to compute the semantic similarity of the *top words* of the topics. We base our coherence measure on 2 criteria:

1. *intra_topic_similarity*: The similarity of words in the same topic should be as high as possible. This means a topic should refer to the same lexical field.

2. *inter_ topic_ similarity*: The similarity of words across topics should be as low as possible. This means that a set of topics is more interesting if they differ from each other.

similarity(v,w) is the similarity between words v and w computed thanks to the word2vec model. It is part of the *word2vec* library [12, 5]. *n_ similarity(lv,lw)* is also function of this library. It computes the similarity between two sets (lv and lw) of words. These functions compute *cosine similarities*. Therefore, we will add 1 to keep the result always positive.

To compute the *intra_ topic_ similarity*, we will compute the average similarity of every possible pair in the top N words of a topic. The mathematical formula is defined in Equation 3 with w_i the i^{th} most frequent word of the topic t. The *intra_ topic_ similarity* is computed per topic.

$$intra\_topic\_similarity(t) = \frac{\sum_{j>i\geq 0}^{N} similarity(w_i, w_j) + 1}{C_2^N} \tag{3}$$

To compute the *inter_ topic_ similarity* between 2 topics, we will use the *n_ similarity* function of the word2vec library which computes the similarity between 2 sets of words.

The mathematical formula is defined in Equation 4, with top_{t_1} the top words of topic t_1 , top_{t_2} the top words of topic t_2.

$$inter\_topic\_similarity(t_1, t_2) = \frac{\sum_{j>i\geq 0}^{N} n\_similarity(top_{t_1}, top_{t_2}) + 1}{C_2^N} \tag{4}$$

In order to have a high coherence when the *intra_ topic_ similarity* is high and *inter_ topic_ similarity* is low, we define the coherence measure between 2 topics t_1 and t_2 as in Equation 5, *avg* being the average function.

$$C_{word2vec} = \frac{avg(intra\_topic\_similarity)}{inter\_topic\_similarity(t_1, t_2)} \tag{5}$$

We globalize this measure for K topics in Equation 6

$$C_{word2vec} = \frac{\sum_{i=0}^{K} intra\_topic\_similarity(t_i)/K}{\sum_{j>i\geq 0}^{K} inter\_topic\_similarity(i, j)/C_2^K} \tag{6}$$

5.2 Example

To show that the coherence is relevant, we will use the example defined in Section 4. We generated 10 models and computed the U_{MASS}, C_V coherences and the $C_{word2vec}$ coherence. From these results we show here the best and worst models according to our coherence.

Model	Topics	U_{MASS}	C_V	$C_{word2vec}$
Good Model	('system', 'user', 'eps', 'human', 'interface') ('graph', 'trees', 'minors', 'survey', 'time')	-14.7	0.384	0.887
Bad Model	('computer', 'system', 'user', 'trees', 'graph') ('system', 'graph', 'trees', 'user', 'eps')	-14.7	0.384	0.604

Table 1: Best and worst C_word2vec coherence models

For this very small corpus, we expect the 2 prominent topics to be about "Human machine interface" and "Graph theory".
GoodModel is more coherent because it separates well the two topics and gathers the right words to define them. Moreover, ten different words are used across the two topics making it easy to distinguish.
However, *BadModel* is less coherent because it mixes the words for the two expected topics: 'user', 'system', 'graph' and 'tree' are used in both topics. It makes harder to distinguish each topic. The coherence measured with C_V and U_{MASS} are not relevant since they are trained on the built LDA model itself, and the training corpus is very small.

6 Preprocessing

The Enron's e-mail database has to be processed first. This is a very important step since all the words that we collect will be part of the bag of words used to train our model. The objective of this phase is to get rid of as much irrelevant data as possible to give more value to meaningful words. Several examples where preprocessing is essential:

- Mail signature:
  The automatic signature is a pattern that is present for each mail from the same person. Depending on how many e-mails this person has sent, the frequency of the words forming the signature will increase whereas it doesn't bring any information on the content of the e-mail. It should be removed.

- Dates, addresses and names:
  Even though they could be relevant in specific situations, it will not enhance our LDA model because the word alone does not help to define a topic. We decided to remove as much as possible.

- HTML encoded e-mails:
  We found in the dataset e-mails encoded in HTML. We don't want the tags to be part of processed data, thus we only extract the text content from it.

We also remove the normal e-mail headers, the web links and e-mail addresses from the processed content. These information are valuable in a forensic investigation. But in a topic model, they are taken away from the context and thus do not help shaping the topics.

We tried to make this pre-processing phase as accurately as possible in a reasonable amount of time. We also made a list of words to complete the gensim's list of stop words.

7 Optimal number of topics

From the same preprocessed data, we generated models of K topics and I iterations. K even and going from 2 to 66 and for $I \in 10, 20, 40, 80$. For each model, we computed the coherence with all 3 different measures. We aimed at finding an optimum number of topic K where the coherence was the highest. The three next diagrams show the coherence versus the number of topics. Every diagram has 4 series which are for the number of iterations. For each coherence measure, we take an example of both well rated model and bad rated model for the reader's appreciation of the different output topics.

Note that the values of the different coherences are relative. We will thus compare the global fluctuations and trends of the measure.

7.1 C_{UMASS}

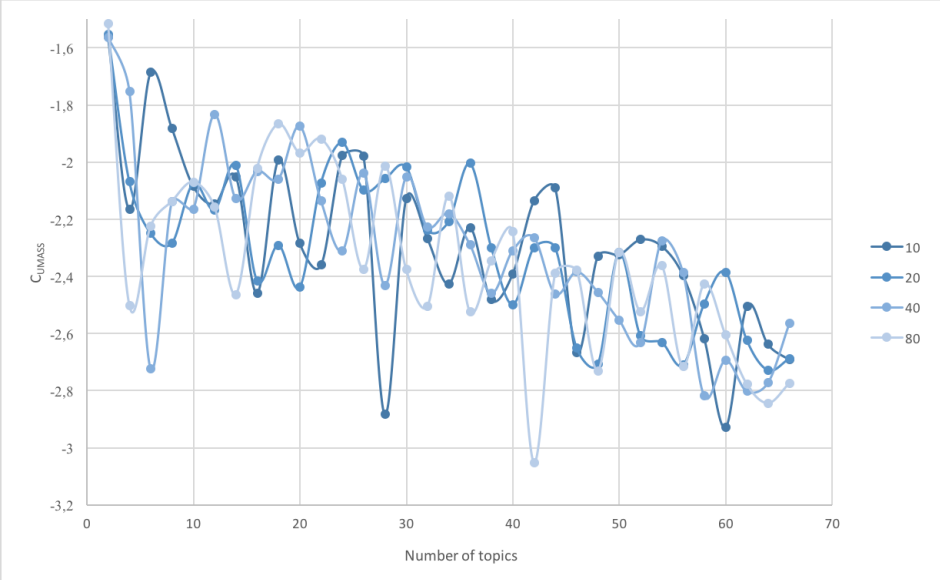


Figure 3: Influence of the number of topics on the U_{MASS} coherence

According to the U_{MASS} coherence measurements, the coherence of the topics globally decreases when K increases. For the analysis, we compare models with K = 6 for 40 iterations which is a local minimum and for 10 iterations which performed better.

7.2 C_V

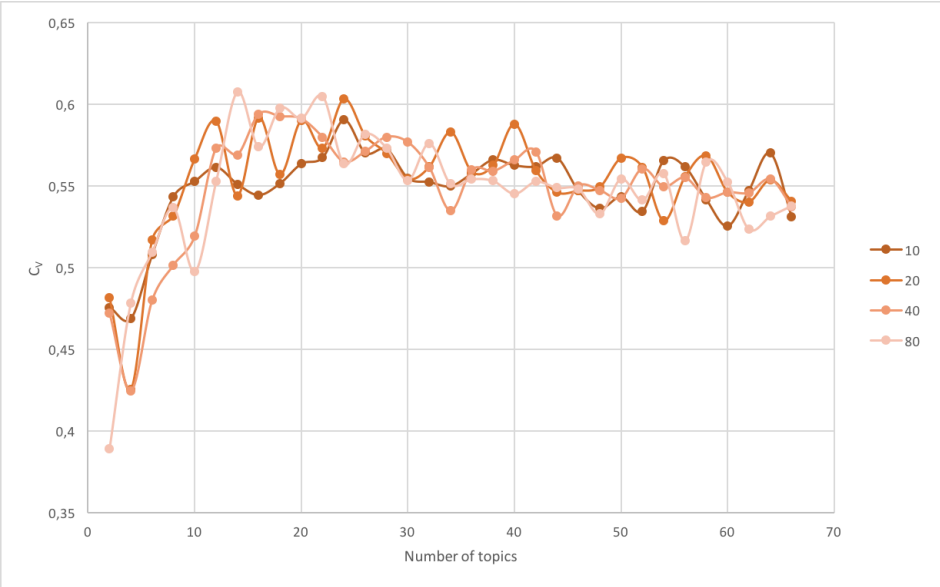


Figure 4: Influence of the number of topics on the C_V coherence

The C_V coherence shows less fluctuations and a global maximum for the trend. A theoretical optimal number of topics would be around 18 for this dataset. Topics for model K4, 40 iterations and K=14, 80 iterations will be shown in the appendix for the reader's appreciation.

7.3 $C_{word2vec}$

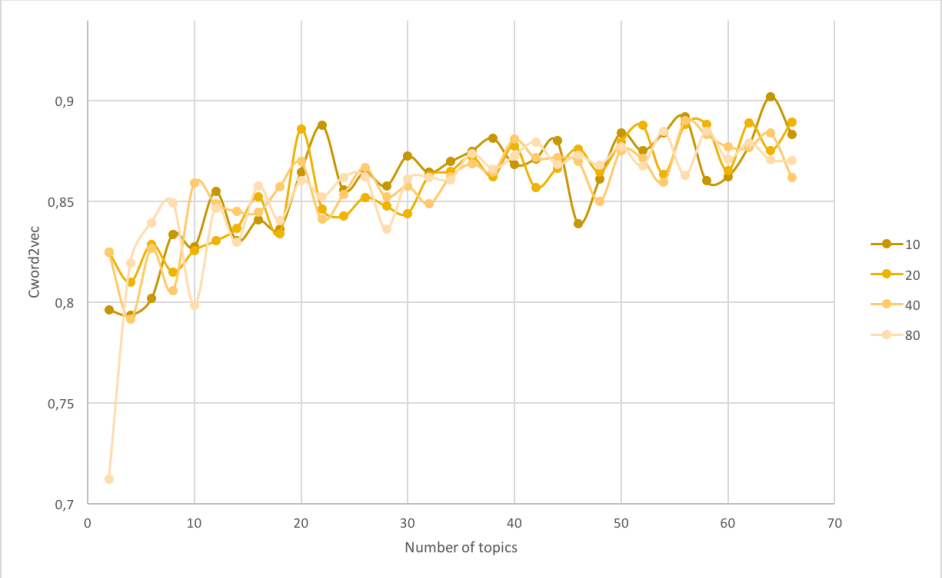


Figure 5: Influence of the number of topics on the $C_{word2vec}$ coherence

The $C_{word2vec}$ coherence is also more stable. It shows a globally increasing coherence for higher numbers of topic. We will compare for the $C_{word2vec}$ coherence the model K=10, 80 iterations which has a bad rating and the K=20, 20 iterations which is better rated.

8 Influence of iterations on coherence

In the previous experiments, the 4 different series of iterations 10,20,40 and 80 iterations presented similar trends and average coherence. In order to show the influence of this parameter, we generated models with very low number of iteration (1 to 9) and very high 200, 400, 600, 800 and 1000 iterations. For each iteration, we computed 15 measures for $K \in [2, 30]$, K even. Then we plot the box plot for each coherence.

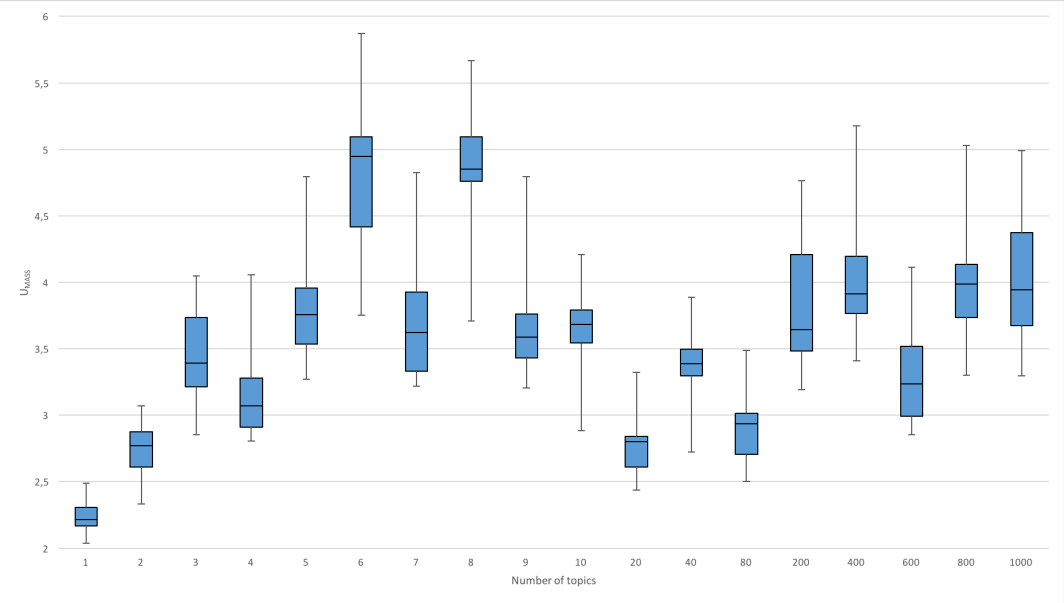


Figure 6: Influence of the number of topics on the U_{MASS} coherence

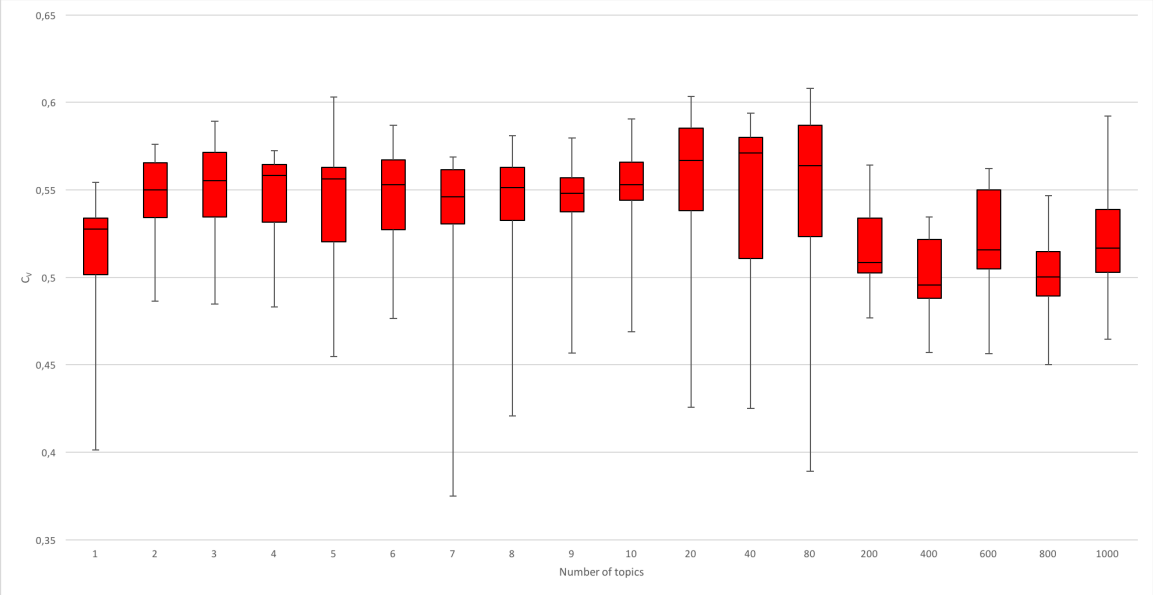


Figure 7: Influence of the number of topics on the C_V coherence

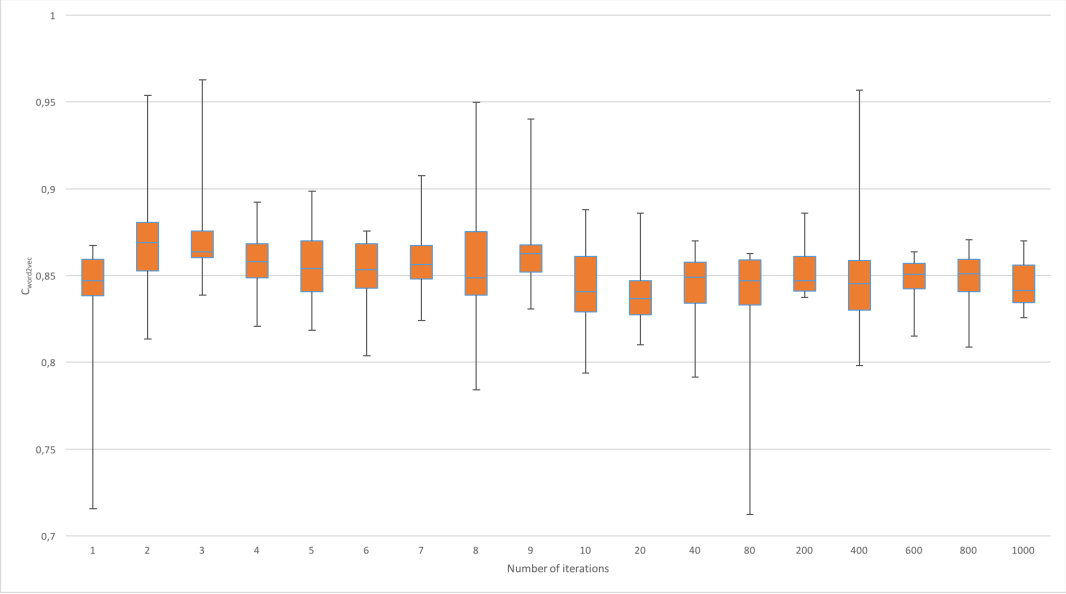


Figure 8: Influence of the number of topics on the $C_{word2vec}$ coherence

Figure 6 shows the absolute values for the U_{MASS} coherence. Since U_{MASS} gives negative values, the graph should be interpreted as "higher means less coherent". The other result show rather flat evolution for a growing number of iteration.

9 Discussion

9.1 Preprocessing

Preprocessing is essential to the quality of LDA models. Reducing the number of words that give poor information out of the context, allows the appreciation of topics to be easier. In order to perform the preprocessing better, we could use two solutions, stemming and part-of-speech tagging. Stemming consists of replacing a word with its stem. e.g. "happiness" becomes "happy". Thus, the semantic value of words has a better statistical impact on the model. Part-of-speech tagging allows to mark up the words according to their part of speech. We could then remove from any sentence the pronouns, the prepositions, articles, conjunctions and interjections. It could also detect what parts of e-mails are not actual sentences and just signature or courtesy parts.

9.2 Technical limitations

We performed the experiments on a Dell PowerEdge R230 server (Intel Xeon processor, 16GB RAM). The python script was written within time constraints. The script does not optimize the space management nor the model generation which could be parallelized. Moreover, with the *gensim ldamodel* library on this setup, the maximum number of topics is 66.

9.3 Word2vec semantic coherence

The Word2vec model is not perfect and allows similarity measures on surprising topics. For example, the similarity of the topic defined by ('th', 'de', 'er', 'ed', 'ng', 'enron', 'nd', 'es', 'al', 'ing') is of 1.28. This value is given for words that do not exist in the English language and the result is high. The reason why this topic has a high similarity is not obvious. It could be due to the fact that the words are similar in their meaninglessness. Or because they are also acronyms.

10 Conclusion

The objective of this research was to give an impression on model coherence for LDA models. We defined a simple and intuitive coherence measure based on a different NLP model. The results show that the 3 coherences behave differently when the number of topic increases. According to C_V coherence, a number of 18 topics should give the most coherent topics. This research shows that the number of iteration hos a very low impact on the quality of the models. In this research, we could observe that the *Gensim*[12] LDA implementation is not deterministic. A good forensic implementation for

this model would be to generate several models with the same parameters and keep the most coherent according to a chosen coherence metric.

11 Future Work

NLP is still a very active field in AI research. Even though it has been around for a long time,Scientific community should gather around the same evaluation metrics to be able to compare their results. The semantic coherence introduced in this paper could show more interesting results if it was refined. One could for example take account for the probability of the top words of a topic to weight the similarity measures. Indeed, the similarity of the 2 most probable words of a topic seems to be more important the similarity of the least probable words of the same topic. The impact of the N parameter for the number of top words on the quality evaluation could be studied in later research. A good solution for analyzing large document collection with LDA can also consist of creating hierarchical topics. From a small number of topics, take the documents that relate the most to the interesting topic. And from this new dataset, train a new LDA model to find the "sub-topics" of this dataset. Hierarchical topics is a valuable solution since the Enron dataset[3] is large. Later research could study the coherence of the sub-topics of hierarchical models. Preprocessing is a important step for quality of NLP models. Future work that focus on this essential step could be very useful and enhance the quality of the next researches.

References

[1] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pages 280–288, 2013.

[2] E. Cambria and B. White. Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57, 2014.

[3] W. Cohen. Enron email dataset. http://www.cs.cmu.edu/ enron/.

[4] David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent dirichlet allocation. *"Journal of Machine Learning Research 3"*, 2003.

[5] GENSIM. `https://radimrehurek.com/gensim/models/word2vec.html`, 2010. Source code.

[6] Henri Trenquier. `https://github.com/htrenquier/topic-modeling.git`, 2018. Research Github repository.

[7] H. Hong and T.-S. Moh. Effective topic modeling for email. In *High Performance Computing & Simulation (HPCS), 2015 International Conference on*, pages 342–349. IEEE, 2015.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[9] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272. Association for Computational Linguistics, 2011.

[10] D. Newman, E. V. Bonilla, and W. Buntine. Improving topic coherence with regularized topic models. In *Advances in neural information processing systems*, pages 496–504, 2011.

[11] D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin. Evaluating topic models for digital libraries. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 215–224. ACM, 2010.

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

[13] M. Röder, A. Both, and A. Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM, 2015.

[14] Tomas Mikolov. `https://code.google.com/archive/p/word2vec/`, 2013. Google Code Archive.

[15] L. Xie, Y. Liu, and G. Chen. A forensic analysis solution of the email network based on email contents. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2015 12th International Conference on*, pages 1613–1619. IEEE, 2015.

A Model examples

A.1 C_{UMASS}

K	it	Top 10 words
6	10	enron new business services management group scheduled london risk thru power energy california said state electricity market gas utilities prices know thanks attached let need call meeting enron agreement information gas click day new available information price time date deal get one week good like game time hour schedule back enron company said million inc stock enron's financial year new

Table 2: Well rated LDA model

K	it	Top 10 words
6	40	power energy said california gas state market electricity price prices know thanks let need attached meeting call agreement get like data error database updated dbcaps operation zmzm alias game week get one new day way time like good click great information enron message contact time email e-mail use intended scheduled enron company said new business million energy trading enron's inc

Table 3: Badly rated LDA model

A.2 C_V

K	it	Top 10 words
14	80	agreement attached credit schedule enron hour date final letter energy data scheduled thru sat database outages london error time fri know thanks let need call get meeting like time deal get one like good time going go day people see market power transmission customers order energy court rate commission also enron company said million new stock inc companies enron's financial zmzm de secured surrounding detailed chairperson central description founder detected gas price day prices market natural trading per daily month message intended information e-mail recipient use confidential email received delete game updated week texas season play team fantasy yards last enron business management group new team services risk energy houston power said energy state california electricity davis utilities prices utility information click email new access web online site service internet travel way san city houston new area hotel day available

Table 4: Best rated LDA model

K	it	Top 10 words
4	40	enron new information business group management services time meeting company thanks know attached gas need agreement let information enron call power said energy enron california company state market gas new get time one thru day week sat scheduled way good

Table 5: Badly rated LDA model

A.3 $C_{word2vec}$

K	it	Top 10 words
20	20	enron company said million stock enron's inc financial companies new
		game updated week play season team texas fantasy yards last
		court commission order also issues issue bill committee whether case
		know thanks let call meeting need get time like week
		th de er ed ng enron nd es al ing
		zmzm folder synchronizing concur rpb offline phillip item oneok vud
		houston texas enron street th phone fax center smith employees
		power energy said california state electricity utilities davis utility gas
		gas deal capacity day pipeline natural daily contract area power
		information request report time date access following system questions contact
		agreement attached credit enron draft comments letter contract transaction shall
		message intended e-mail recipient information use confidential email sender privileged
		click information online new email service internet web services free
		enron business group management risk new services energy team development
		get one like think good going people time go i'm
		thru scheduled sat outages london fri outage sun contact time
		data error database schedule hour dbcaps operation alias final variances
		government new project bush president oil india national dabhol international
		travel way hotel san city new day fares per rates
		market price prices markets year power demand high per generation

Table 6: Well rated LDA model

K	it	Top 10 words
10	80	said power energy enron california state electricity utilities prices company gas energy power market services new capacity company service natural get good week know like going time one i'm think message information intended email e-mail thru scheduled contact recipient use company trading stock inc financial new million market year credit thanks know attached need call let meeting questions agreement time one new people world help th home life yahoo like enron business group new management risk information houston team trading data error database schedule hour dbcaps operation zmzm alias variances day click deal price per travel new available deals way

Table 7: Badly rated LDA model