Improving Quality of LDA Models

RP#76

Henri Trenquier

Supervisor:
Dr. Carlos Ortiz Martinez
MSc Security and Network Engineering

UNIVERSITY OF AMSTERDAM

July 5, 2018

System and Network Engineering

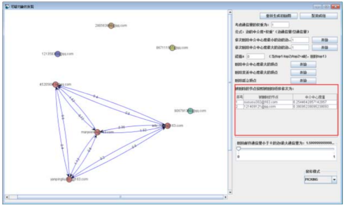
Context

Forensics

- Accelerate forensic investigations
- Large document collections

A Forensic Analysis Solution of the Email Network Based on Email Contents

- L Xie, Y Liu, G Chen (2015)
- Email network analysis



考虑通信量的权重为: 1

公式: 边的中介度+权重*（边通信量/总通信量）

依次删除中介中心度最小的边的边... 1

依次删除中介中心度最大的边的边... 1

阈值a 0 （当(top1-top2)/top2>a时，删除top1）

删除中介中心度最大的顶点

删除紧密中心度最大的顶点

删除孤立顶点

被删除的节点按照被删除顺序依次为:

序号	被删除的节点	中介中心度值
1	susuşu302@163.com	0.2544642857142857
2	121409121@qq.com	0.38095238095238093

删除邮件通信量小于 0 的边(最大通信量为: 1.59999999999...

State of the art

Latent Dirichlet allocation

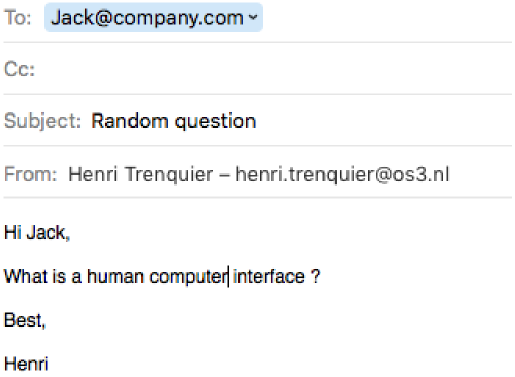
- David Blei, Andrew Ng, and Michael I. Jordan (2003)
- Cited over 23K times
- Machine learning

Statistical model

- Bayesian
- generative & probabilistic
- for a collection of discrete data
- Topic discovery

Example

Document



To: Jack@company.com

Cc:

Subject: Random question

From: Henri Trenquier – henri.trenquier@os3.nl

Hi Jack,

What is a human computer interface ?

Best,

Henri

- Preprocessing
- Bag of word: (‘human’, ‘interface’, ‘computer’)

Example

Corpus

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

Example

Corpus

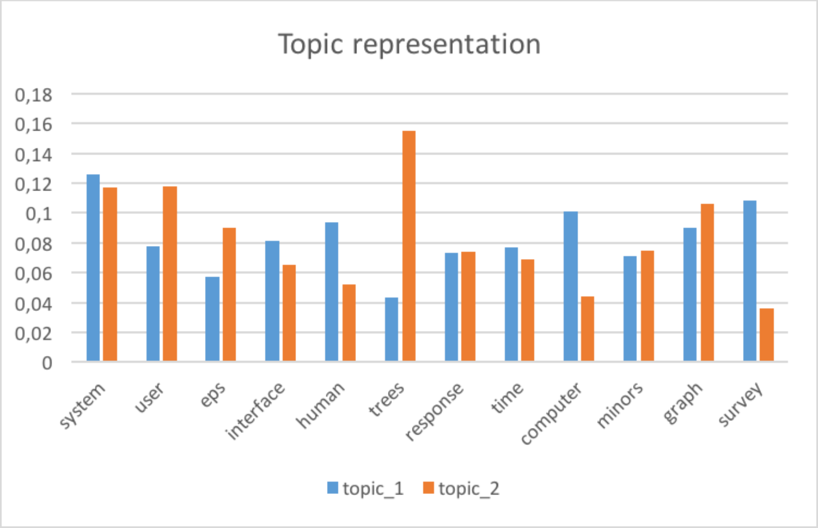
1. ‘human’, ‘interface’, ‘computer’
2. ‘survey’, ‘user’, ‘computer’, ‘system’, ‘response’, ‘time’
3. ‘eps’, ‘user’, ‘interface’, ‘system’
4. ‘system’, ‘human’, ‘system’, ‘eps’
5. ‘user’, ‘response’, ‘time’
6. ‘trees’
7. ‘graph’, ‘trees’
8. ‘graph’, ‘minors’, ‘trees’
9. ‘graph’, ‘minors’, ‘survey’

Expected topics

1. Human machine interface
2. Graph theory

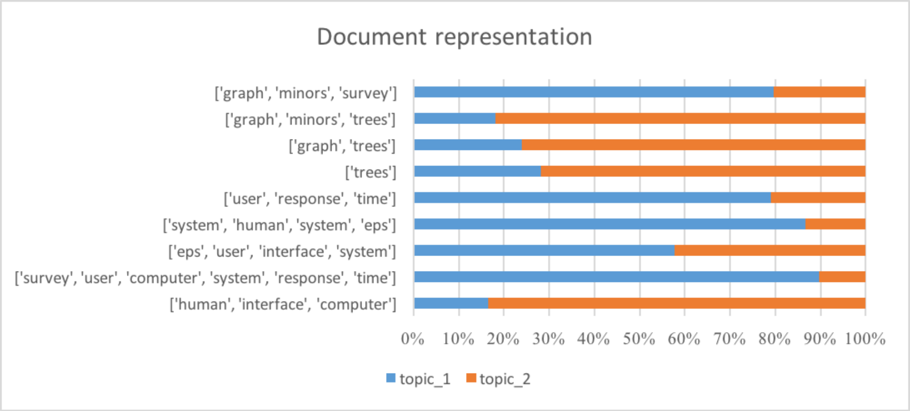
Example

Topic modeling LDA



Example

Topic modeling LDA



Example

Topic modeling LDA

Expected topics:

1. Human machine interface
2. Graph theory

Model	Topics
Good_Model	(‘system’, ‘user’, ‘eps’, ‘human’, ‘interface’)
	(‘graph’, ‘trees’, ‘minors’, ‘survey’, ‘time’)
Bad_Model	(‘computer’, ‘system’, ‘user’, ‘trees’, ‘graph’)
	(‘system’, ‘graph’, ‘trees’, ‘user’, ‘eps’)

Table: Good and Bad models

Example

Topic modeling LDA

Expected topics:

1. Human machine interface
2. Graph theory

Model	Topics
Good_Model	('system', 'user', 'eps', 'human', 'interface')
	('graph', 'trees', 'minors', 'survey', 'time')
Bad_Model	('computer', 'system', 'user', 'trees', 'graph')
	('system', 'graph', 'trees', 'user', 'eps')

Table: Good and Bad models

- More words over all topics
- **More similar** words **within** a topic
- **Less similar** words **across** topics

Context

Enron

- Accounting fraud
- ˜500K e-mails database
- Topic modeling dataset
- quickly target incriminating e-mails

Research Question

How to improve the quality of LDA models?

- *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 semantic quality evaluation?*

Scope: Enron e-mail dataset

System and Network Engineering

State of the art

Coherence

- Evaluation metric for topic modeling

Optimizing Semantic Coherence in Topic Models

- D Mimno et al. (2011)
- 542 citations

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \qquad (1)$$

State of the art

Coherence

- Evaluation metric for topic modeling

Optimizing Semantic Coherence in Topic Models

- D Mimno et al. (2011)
- 542 citations

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \tag{1}$$

Measure evaluated by a survey:

- "good", "intermediate" or "bad"
- no literal definition of coherence
- lack of "inter-topic" evaluation
- C_v and U_{MASS}

System and Network Engineering

State of the art

Coherence

A Practical Algorithm for Topic Modeling with Provable Guarantees

- S Arora et al. (2013)
- 229 citations
- introduces "inter-topic similarity"

Evaluation metric

Topic Coherence

$C_{word2vec}$ coherence measure

- Semantic space
- word2vec model trained on Google News

Evaluation metric

Topic Coherence

$C_{word2vec}$ coherence measure

- Semantic space
- word2vec model trained on Google News

1. *intra_topic_similarity*
2. *inter_topic_similarity*

$$C_{word2vec} = \frac{avg(intra\_topic\_similarity)}{avg(inter\_topic\_similarity)} \quad (2)$$

Evaluation metric

Topic Coherence

$C_{word2vec}$ coherence measure

- Semantic space
- word2vec model trained on Google News

1. *intra_topic_similarity*
2. *inter_topic_similarity*

$$C_{word2vec} = \frac{avg(intra\_topic\_similarity)}{avg(inter\_topic\_similarity)} \tag{2}$$

Model	Topics	$C_{word2vec}$
Good_Model	('system', 'user', 'eps', 'human', 'interface') ('graph', 'trees', 'minors', 'survey', 'time')	0.887
Bad_Model	('computer', 'system', 'user', 'trees', 'graph') ('system', 'graph', 'trees', 'user', 'eps')	0.604

Experiment

Pipeline

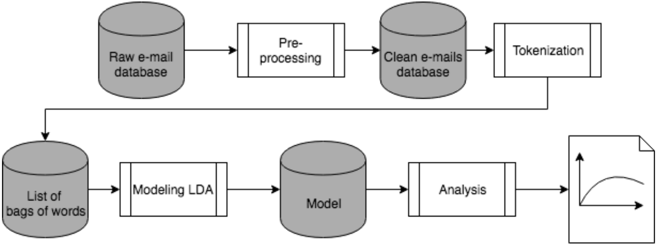


Figure: Similarity measures

- Modeling: I, K
- Coherence analysis: C_v, u_{mass}, $C_{word2vec}$

System and Network Engineering

Results

C_V

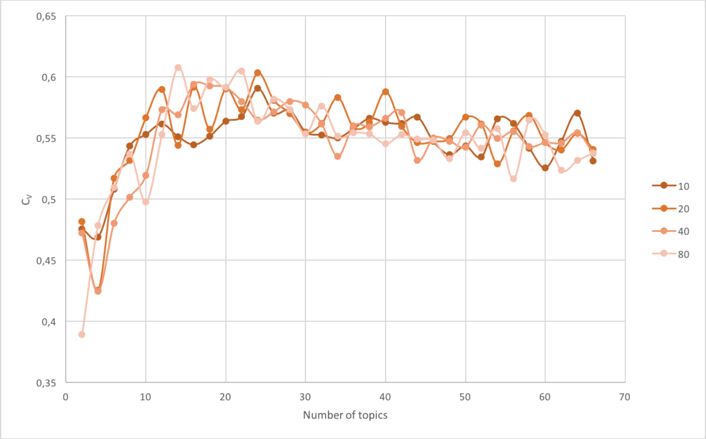


Figure: Influence of the number of topics on the C_V coherence

Results

U_MASS

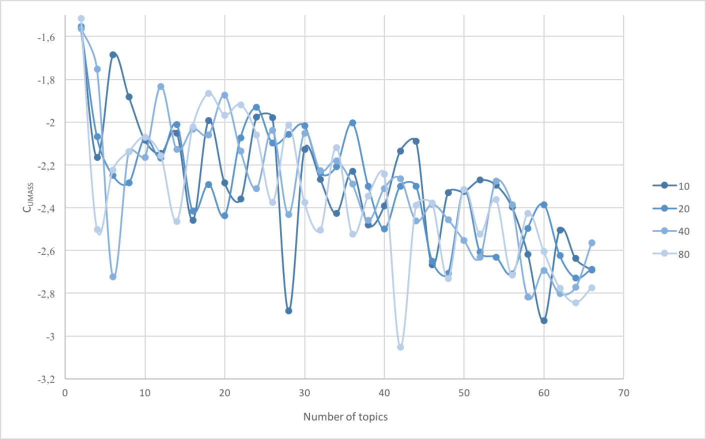


Figure: Influence of the number of topics on the U_{MASS} coherence

Results

$C_{word2vec}$

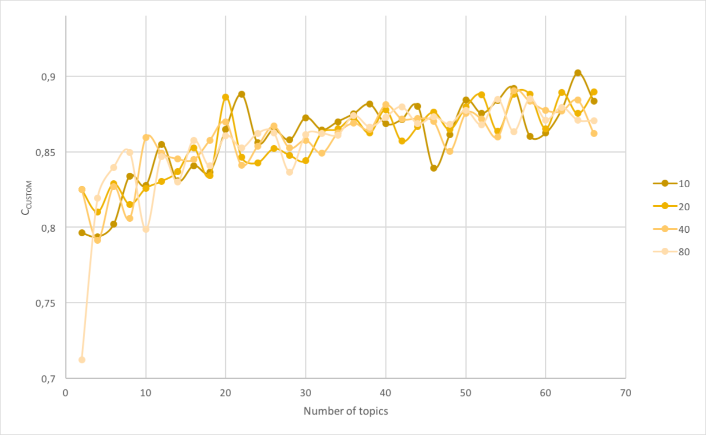


Figure: Influence of the number of topics on the $C_{word2vec}$ coherence

System and Network Engineering

Results

Low & High number of iterations

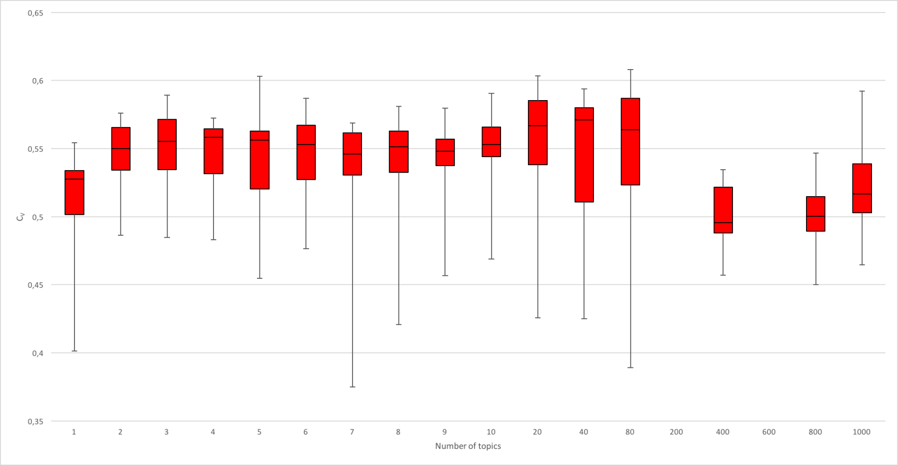


Figure: Influence of the number of iterations on the C_V *coherence*

Results

Low & High number of iterations

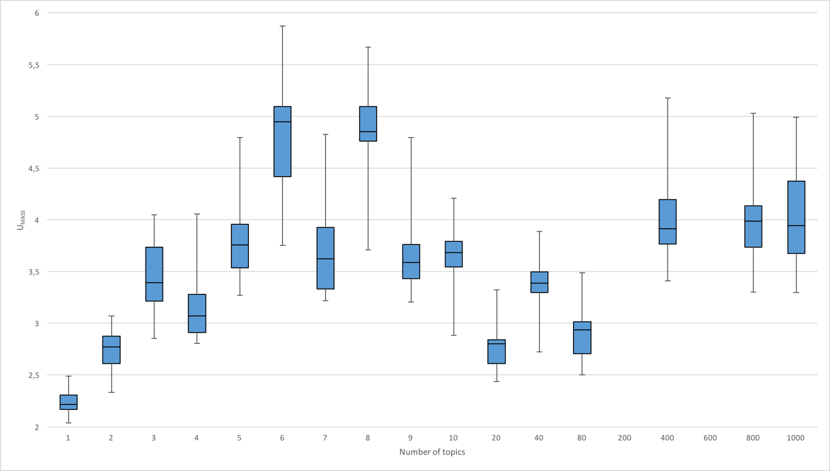


Figure: Influence of the number of iterations on the U_{MASS} *coherence*

Results

Low & High number of iterations

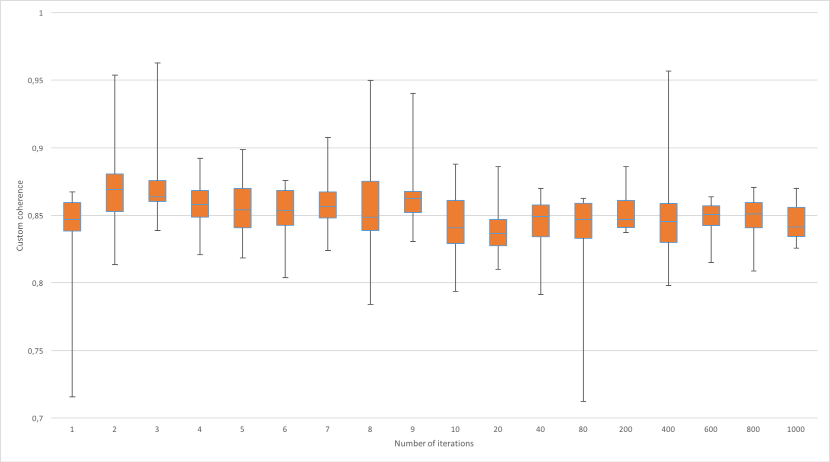


Figure: Influence of the number of iterations on the $C_{word2vec}$ *coherence*

Discussion

- E-mail information density
- Preprocessing phase
- word2vec semantic representation is not perfect sim(['th', 'de', 'er', 'ed', 'ng', 'enron', 'nd', 'es', 'al', 'ing']) $=$ 1.28669572453
- $C_{word2vec}$ coherence still too simplistic

Conclusion

How to improve the quality of LDA models?

- Impression of model coherence
- New semantic coherence
- Results do not reveal an optimum number of topic
- Number of iterations has no visible impact

Future Work

- Better preprocessing: stemming
- Refine $C_{word2vec}$ coherence
  - weight the words of a topic
  - word2vec training dataset
  - compare similar models
- Hierarchical topics

Question ?

Thank you for your attention