chen_2019_experimental_explorations_on_short_text_topic_mining_between_lda_and_nmf_based_schemes

Year

2019

Author(s)

Yong Chen and Hui Zhang and Rui Liu and Zhiwen Ye and Jianying Lin

Title

Experimental explorations on short text topic mining between LDA and NMF based Schemes

Venue

Knowledge-Based Systems

Topic labeling

Manual

Focus

Secondary

Type of contribution

Novel

Underlying technique

Manual labeling

Topic labeling parameters

\

Label generation

Two experts who are well familiar with the above short texts are invited to label the generated topics.
Specifically, two tasks (“Topic Labeling” and “Word Labeling”) are conducted by the two judges.

For the first task “Topic Labeling”, the experts are asked to label “coherent” if more than half of the topN terms in a topic are semantically related; otherwise “not coherent”.
[Additionally, at this stage the concept represented by the topic is identified]

Then for the second task “Word Labeling”, the topics that are labeled as “coherent” by both judges are used for word labeling.
Particularly, each topical word is labeled as “correct” if it is coherently related to the concept represented by the topic (identified in the “Topic Labeling”); otherwise “incorrect”.

After labeling, we can compute the number of the coherent topics for different methods and the average Precision for the first topN topical words.
Finally, the overall performances are averaged by the evaluation measures from two experts.

Motivation

“Here we also evaluate the quality of learned topics based on human judgements” (topic coherence)

Topic modeling

LDA, NMF, NMF-based method (with external knowledge)

Topic modeling parameters

Nr. of topics: {20, 40, 60, 80 and 100}
α: 1
β: 0.1

NMF

Iterative times: 100

Nr. of topics

20

Label

Concept associated with the topic.
Coherent / not coherent depending on whether more than half of the topN terms in a topic are semantically related.
Correct / Incorrect depending on whether each topical word is coherently related to the concept represented by the topic.

Label selection

\

Label quality evaluation

No quality evaluation on the label itself, but:
"After labeling, we can compute the number of the coherent topics for different methods and the average Precision for the first topN topical words. Finally, the overall performances are averaged by the evaluation measures from two experts."

Table 2: Average coherent topics evaluated by experts given topic number $K = 20$ for different short text datasets between NMF and LDA methods.

Datasets \ Methods	NMF	LDA
Snippet	**18**	**17**
XinlangNews	**20**	**19**
News	**19**	**18**
StackOverFlow	**19**	**19**
TMNtitles	**19**	**18**

Additionally, the Cohen's Kappa is used to measure agreement between experts

Assessors

\

Domain

Domain (paper): Short text topic modeling
Domain (corpus): Miscellaneous (Google), News, Online Q&A

Problem statement

Learning topics from short texts has become a critical and fundamental task for understanding the widely-spread streaming social messages.
In this context, the basic LDA and NMF are compared with different experimental settings on several public short text datasets;

In the second part, we propose a novel model called "Knowledge-guided Non-negative Matrix Factorization for Better Short Text Topic Mining" (abbreviated as KGNMF), which leverages external knowledge as a semantic regulator with low-rank formalizations, yielding up a time-efficient algorithm.

Corpus

Origin: Google
Nr. of documents: 12.284
Details: Snippet dataset drawn from Google

Origin: Sina website
Nr. of documents: 8966
Details:

- XinlangNews dataset with news crawled from 8/23/2016 to 9/7/2016
- Composed by news titles from 6 domains, "entertainment", "Finance", "Sci Tech", "Society", "Sports" and "Military".

Origin: Various newspaper websites
Nr. of documents: 32.592 (or 29.487 when considering only the title fields)
Details:

- Title and description fields of english news extracted from RSS feeds of popular newspaper websites (nyt.com, usatoday.com, reuters.com)
- Categories are: "Sport", "Business", "U.S.", "Health", "Sci Tech", "World" and "Entertainment".

Origin: StackOverFlow

Nr. of documents: 19.783

Details: Short text dataset collected from an online question-and- answer site (StackOverflow.com) for programmers.

Table 1: Statistics of datasets. (#Docs: the total number of documents; #Words: the average number of words per document; #Vocabulary: the total number of distinctive terms in the dataset; #Label: the number of ground-truth labels or categories.)

Datasets	#Docs	#Words	#Vocabulary	#Label
Snippet	12,284	14.34	4,045	8
News	32,592	11.82	7,771	7
StackOverFlow	19,783	7.25	2,437	20
XinlangNews	8,966	4.38	2,631	6
TMNtitles	29,487	3.46	4,015	7

Document

1. Each snippet is on each line and each one consists of a list of words/terms plus a class label at the end in the data files. Overall, there are 8 categories marked “Business”, “Computers”, “Culture-Arts-Entertainment”, “Education- Science”, “Engineering”, “Health”, “Politics-Society”, and “Sports”.
2. News title
3. News title and description
4. StackOverflow thread

Pre-processing

- Tokenisation
- Filtering the short texts that has less than two terms
- Discard the terms, of which the IDF (Inverse Document Frequency) is less than 5

```
@article{chen_2019_experimental_explorations_on_short_text_topic_mining_between
_lda_and_nmf_based_schemes,
  abstract = {Learning topics from short texts has become a critical and
fundamental task for understanding the widely-spread streaming social messages,
e.g., tweets, snippets and questions/answers. Up to date, there are two
distinctive topic learning schemes: generative probabilistic graphical models
and geometrically linear algebra approaches, with LDA and NMF being the
representative works, respectively. Since these two methods both could uncover
```

```
the latent topics hidden in the unstructured short texts, some interesting
doubts are coming to our minds that which one is better and why? Are there any
other more effective extensions? In order to explore valuable insights between
LDA and NMF based learning schemes, we comprehensively conduct a series of
experiments into two parts. Specifically, the basic LDA and NMF are compared
with different experimental settings on several public short text datasets in
the first part which would exhibit that NMF tends to perform better than LDA;
in the second part, we propose a novel model called ``Knowledge-guided Non-
negative Matrix Factorization for Better Short Text Topic Mining'' (abbreviated
as KGNMF), which leverages external knowledge as a semantic regulator with low-
rank formalizations, yielding up a time-efficient algorithm. Extensive
experiments are conducted on three representative corpora with currently
typical short text topic models to demonstrate the effectiveness of our
proposed KGNMF. Overall, learning with NMF-based schemes is another effective
manner in short text topic mining in addition to the popular LDA-based
paradigms.},
  author = {Yong Chen and Hui Zhang and Rui Liu and Zhiwen Ye and Jianying
Lin},
  date-added = {2023-03-11 11:37:41 +0100},
  date-modified = {2023-03-11 11:37:41 +0100},
  doi = {https://doi.org/10.1016/j.knosys.2018.08.011},
  issn = {0950-7051},
  journal = {Knowledge-Based Systems},
  keywords = {Short text mining, Topic modeling, Latent dirichlet allocation
(LDA), Non-negative matrix factorization (NMF), Knowledge-based learning},
  pages = {1-13},
  title = {Experimental explorations on short text topic mining between LDA and
NMF based Schemes},
  url = {https://www.sciencedirect.com/science/article/pii/S0950705118304076},
  volume = {163},
  year = {2019}}
```

#Thesis/Papers/Initial