Topic modelling and social network analysis of online forum posts

1957418

February 20, 2020

1 Introduction

IoT industries benefit from the growth of 5G technology because 5G network can reduce the latency of data transformation between devices and control application. It has raised the interest of business companies and research entities with its remarkable profit and great research potential. How to discover customers' needs and learn the IoT research trend efficiently becomes a problem in IoT industries.

Further most of the topic modelling method research are based on standard document corpora such as 20NewsGroups[1]. There are lack of evidence of their performance on the real-world dataset.

This project aims to solve this problem by modelling topic in the online forum by comparing the performance between modelling algorithm base on a certain evaluation framework. Also, implementing social network analysis to the data, so that identify the poster who are prominent and influential in the forum. Further, the topic model of this project can be accessed from a web interface.

At the end of this project, several topic models over comments should be presented by different modelling algorithms, and performance across modelling algorithms should present(The evaluation method will be described below). Besides, the influencer(who has significant influence in the forum) within posters should be found after social network analysis. Moreover, implement a web interface(based on Django) for presenting performance result and enable user to interact with data and algorithms.

2 Related Work

2.1 Internet of Things

The Internet of Things(IoT) is regarded as the Internet of the future. The essence of IoT is a secure and self-governing data exchange connection between physical devices and real application via the Internet[2]. With the spread of 5G technology helps the establish of the IoT. The IoT can have a more friendly ecosystem because 5G network guarantees the IoT device a shorter latency of a mere one millisecond as well as a faster data transformation speeds[3].

2.2 Topic Modeling

Topic modelling is a method of uncovering hidden structure in a collection of texts(in this project it will be comments). The keys of topic modelling are dimensional reduction and unsupervised learning. Dimensional reduction is to represent text in its topics rather than a text in its feature space. Unsupervised learning method for topic modelling is clustering[4].

2.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation(LDA) is a bayesian topic model method which stands on the facts that a document consists of random multiple latent topics, where a certain distribution over words characterize each topics[5]. It models the topics for all the documents with a shared set of topics, and the model shows the different proportion of topics in each documents[6].

2.3.1 Generative Process

In LDA, each documents is assumed to have a topic distribution θ_i where θ_i is generate from a Dirichlet distribution with hyperparameter α. The process of generating a documents is as follow.

1) Randomly select topic distribution θ_i of document d_i from Dirichlet distribution α.

2) For j-th word in the document d_i:

   i select a topic $z_{i,j}$ from multinominal distribution θ_i.

   ii Randomly select word distribution $\phi_{z_{i,j}}$ from topic $z_{i,j}$ where $\phi_{z_{i,j}}$ is generate from Dirichlet distribution with hyperparameter β.

   iii Randomly select a word $w_{i,j}$ from the multinominal word distribution $\phi_{z_{i,j}}$

2.3.2 Gaussian LDA

Traditional LDA assumed the word type in the document is fixed. Its performance on the documents, which consists out of vocabulary(OOV) is not satisfying. Besides, the topic modelled by traditional LDA has poor semantic coherence. In 2015, Das, Zaheer and Dyer exchanged the categorical distributions of word types in traditional LDA with continuous multivariate Gaussian distribution of space embedding of words(G-LDA) to modelling topics to find the semantic coherence. Besides, they manage the problem of cannot handle out of vocabulary(OOV) by introducing word vector to vectorize words by its semantic similarity[7].

2.4 Latent Semantic Indexing

Latent Semantic Indexing(LSI) or Latent Semantic Analysis(LSA) is a document indexing and information retrieval method. Is has remarkable contribute to automatic indexing, but the theoretical foundation is unsatisfied. Hofmann extends the algorithm based on the likelihood principle to solid the theory foundation called Probabilistic Latent Semantic Analysis(PLSA)[8]

2.5 Neural Network

Miao et al. introduced deep neural networks to topic modelling in order to relax the limit of probabilistic topic model's limit of topic dependencies and exploit conditional information. Their method combined neural networks and traditional probabilistic model(including PLSA and LDA) which makes it be able to trained by backpropagation efficiently.

They introduce three neural structures Gaussian Softmax distribution(GSM), Gaussian Stick Breaking(GSB) and Recurrent Stick Breaking process(RSB). GSM and GSB generate topic distribution from Gaussian process where GSM uses softmax function and GSB applies a stick breaking construction. RSB uses recurrent neural network conditioned on Gaussian draw[9].

The comparison to the three neural structures with traditional probabilistic model(OnlineLDA, NVLDA, probLDA) shows the neural structures achieve the state-of-art performance on topic modelling.

However, their evaluation metrics is the topic coherence on 20NewsGroups dataset which also been used in Gaussian LDA research. Both Gaussian LDA and neural structures claims to be able to enhance the topic coherence of the topic model than the traditional LDA. This research implements both method to identify which method fits the requirement of this dataset better.

3 Research Processes

3.1 Methodology Mind Map

The methodology mind map of this project which shows the workflow of the research processes is attached to the appendices.

3.2 Data Resource

The project begins with Data harvest. Reddit forum provides an extensive API which allows the developer to access data on Reddit with JSON format. For the Python program, which will be the language of this project, there is a python package PRAW. PRAW allows the software to access the comment of the Reddit submission, and there is a model names subreddit to get all submission under a specific community (in this project it will be IoT).

3.3 Data Preprocessing

Preprocessing might be the most time-consuming procedure in this project. Comments may contain some misspelt word or irregular word, for example, "happy" may spell as "happyyyyy" in the comment. A spell-check method can be implemented to solve these problems. There may other problems will affect the performance of the modelling which will be found during the processing.

General preprocessing procedure is applied, including tokenization and stemming. Besides, two ways of word embedding will be implemented to the raw data, Bag-Of-Word model and word2vec model. Therefore, the required format of data for different topic model algorithms can be met.

3.4 Topic Modelling

The data then will be used for topic modelling with different algorithms including simple LDA, G-LDA and LSI from gensim package. The result will be plotted using matplotlib as well as on the web. Further,

3.4.1 LDA

Simple multinominal LDA algorithm can be implemented directed from gensim package LdaMulticore function. G-LDA can be implemented based on the Das, Zaheer and Dyer program in their GitHub[10]. They implemented the algorithm based on Java and G-LDA requires all words in documents are embedded to Word2Vec format.

3.4.2 Neutral Network

The Neural Variational Document Model(NVDM) implementation can be access from author's github[11]. Author used tensorflow to implement neutral network, and for this project only NVDM object will be used. Considering original implementation of NVDM is based on 20NewsGroup dataset which is a well preprocessed training corpora, this project requires a high level of preprocessing of the data.

3.5 Social Network Analysis

Each poster and their connection with another poster (comments in others post) will be saved as nodes and edge in the symmetric network. Networkx package will be introduced to doing social network analysis to identify the poster who are prominent and influential in these discussions.

3.6 Evaluation Design

It is not very easy to quantitatively evaluate the performance across algorithms. According to the answer of this stack overflow[12], perplexity and the topic assignment likelihood on randomly selected documents are key metrics to evaluate the algorithm.

In Das, Zaheer and Dyer research[7], they introduced a sampling algorithm Gibbs to a new document to evaluate their algorithm with others. In this project, this method will be used to evaluate the performance of different modelling algorithm.

4 Project Management

4.1 Timeline

The project development will follow the agile development method in order to react to the modification made by analysis and minimize the development risk. The Gantt chart of this project is attached in the appendices.

4.2 Ethical Concern

The dataset is collected from Reddit official API PRAW. Considering Reddit take id number and forum nickname as the index of each data, no real name or data other than comments are used during the process. Therefore there is no ethical concern for this project.

4.3 Risk

The algorithm development and API have clear documentation or solid implementation instruction, so the risk for algorithm implementation can be omitted. The major challenge of this research is that the data is collected from an online forum which may contain massive noise(meaningless word or incomplete sentences). This kind of noise may affect the performance of traditional LDA because traditional LDA uses word type as a feature. It is difficult and time-consuming to fix this noise, which may cause it takes more time to preprocess the data than expected

This risk may be solved by introducing a spellchecker package. It is a python package which can check the spell of words by comparing all permutations (insertions, deletions, replacements, and transpositions) and return a candidates list which shows what the misspelt word likely is[13].

5 Conclusion

This project tends to analysis the topic of IoT in the Reddit forum. It can be done by implementing topic modelling with different algorithm including LDA, Gaussian-LDA and PLSA. Besides, social network analysis can be applied to the dataset to figure out the influencer in the forum. The result social network analysis can be used in topic modelling to add weight on the comment from the influencer.

The current research process is on exploring related work because there is still plenty of research that has not been covered. The algorithm of the topic modelling may be modified while the milestone of each phase of research remains the same.

5.1 Future Work

A wide variety of aspect extraction research based on topic modelling which involved neural network technique. Sentiment analysis also can be implemented to the topic modelling result to identify user's attitude towards a specific topic. Both aspects can be a research direction in the future.

References

[1] Lang, K. (2020). *Home Page for 20 Newsgroups Data Set.* [online] Qwone.com. Available at: http://qwone.com/ jason/20Newsgroups/ [Accessed 20 Feb. 2020].

[2] Chopra, K., Gupta, K. and Lambora, A. (2019). Future Internet: The Internet of Things-A Literature Review. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon).*

[3] Sequeira, N. (2020). *What 5G Means for The Future of Internet of Things - 5G Technology World.* [online] 5G Technology World. Available at: https://www.5gtechnologyworld.com/what-5g-means-for-the-future-of-internet-of-things/ [Accessed 7 Jan. 2020].

[4] NLP-FOR-HACKERS. (2020). *Complete Guide to Topic Modeling - NLP-FOR-HACKERS.* [online] Available at: https://nlpforhackers.io/topic-modeling/ [Accessed 8 Jan. 2020].

[5] M.Blei, D., Y.Ng, A. and I.Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4/5), pp.p993-1022. 30p.

[6] Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), p.77.

[7] Das, R., Zaheer, M. and Dyer, C. (2015). Gaussian LDA for Topic Models with Word Embeddings. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).*

[8] Hofmann, T. (2017). Probabilistic Latent Semantic Indexing. *ACM SIGIR Forum*, 51(2), pp.211-218.

[9] Miao, Y., Grefenstette, E. and Blunsom, P. (2020). Discovering discrete latent topics with neural variational inference. *Proceedings of the 34th International Conference on Machine Learning*, 70, pp.2410–2419.

[10] GitHub. (2020). *rajarshd/Gaussian_LDA.* [online] Available at: https://github.com/rajarshd/Gaussian_LDA [Accessed 14 Jan. 2020].

[11] miao, y. (2020). *ysmiao/nvdm.* [online] GitHub. Available at: https://github.com/ysmiao/nvdm [Accessed 20 Feb. 2020].

[12] Stack Overflow. (2020). *LDA topic modeling - Training and testing.* [online] Available at: https://stackoverflow.com/questions/11162402/lda-topic-modeling-training-and-testing [Accessed 14 Jan. 2020].

[13] Barrus, T. (2020). *pyspellchecker.* [online] PyPI. Available at: https://pypi.org/project/pyspellchecker/ [Accessed 14 Feb. 2020].

Appendices

A Methdology Mind Map

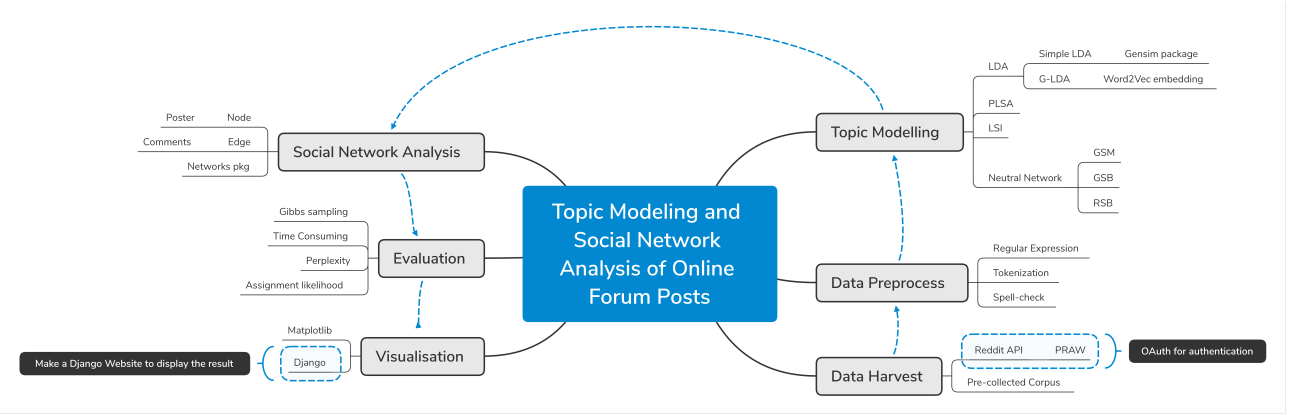

Figure 1: mind mapping of the methodology

B Gantt Chart

The full chart is attached follow.

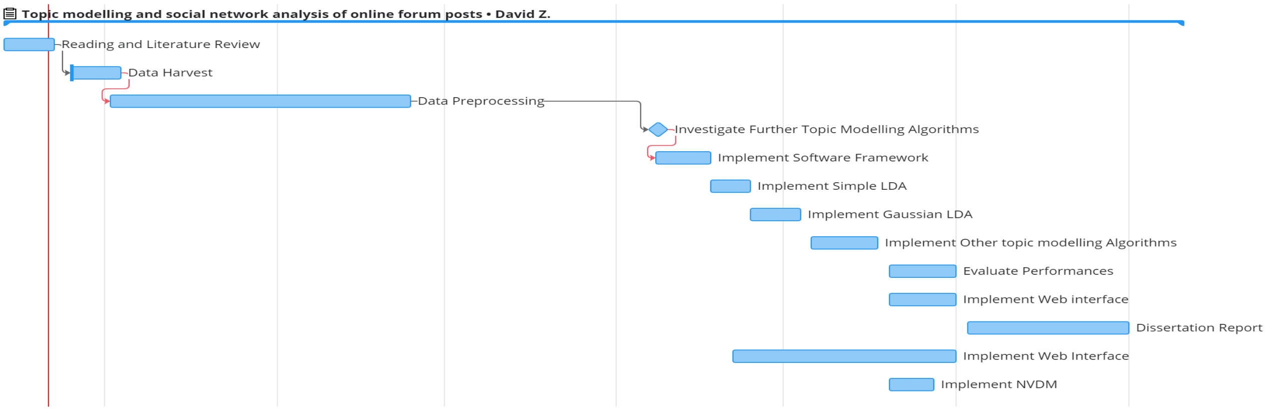

Figure 2: gantt chart