Expert Modeling System

Philip Robinson

NASA: Jet Propulsion Laboratory

September 13, 2018

Abstract

In a large community, enlisting potential collaborators and subject matter experts greatly impacts the success of projects. Candidate discovery and expertise ranking against a task or project is necessary to inform recruiting of impactful teams [5]. NASA's Jet Propultion Lab is interested in better tools for expert discovery and matchmaking to tasks in misssion critical, late stage, anomalies. Topic modeling, such as Biterm Topic Model (BTM) [3, 10], Latent Dirichlet Allocation (LDA), and Correlated Topic Model (CTM), have long been used to as discovery tools, usually focused on exploratory analysis, finding topics for text [2]. Likewise, author modeling has been used to measure attribution [6] and contribution [1]. Author-Topic Modeling (ATM) establishes a strategy to map both authors and documents to the same topic-space over a vocabulary [8].

1 Introduction

Building effective teams, especially against specialized projects, is essential to project success. With a greater candidate pool, matchmaking often falls on managers whose scope is socially limited. In order to best support the scale of an institution like JPL/NASA, and domain specific nature of the problems they address, they are interested in strategies to explore and recomend subject matter experts (SME). An effective SME recomender system significantly reduces social coordination overhead of electing contributors to to complex, domain specific, problems. Additionally, tools that allow exploratory and comparitive view of experts have potentially great benifits to workload balancing and identifying company knowlege gaps.

Latent Dirichlet Allocation (LDA) is a topic modeling strategy that empowers exploratory analysis, topic discovery, and dimmentionality reduction. Since we are looking for SMEs, author-modeling is a closer fit. The phrase "author modeling" has also been used in techniques which are more concerned with literal text-content document contribution and attribution [6]; we are not interested in these techniques. Although LDA is best used as an exploratory tool, many derivatives exist that leverage LDA for more powerful or specific applications. One extension to LDA is the author-topic model, which attempts to describe authors in LDA's learned topic space.

2 Objectives

The Jet Propultion Lab uses a ticketing system called the Problem Reporting System to manage work assignment of experts to mission critical late stage anomalies. This ticketing sytem contains Pre-Launch Failure Reports (PFR)s and Incident Surprise Anomalies (ISA)s. Without an expert assignment strategy, candidates are usually elected by the manager of a ticket, which is sucsptable to bias of their prior collaborators, may require understanding of candidate skills beyond their perview, and usually leads to over assigning tickets to few candidates. Better management of human resources by exploratory tools and recomender systems, would significantly reduce the difficulty of building effective teams.

3 Method

Traditionally, in representing a document collection, we envision a

`Vocabulary Size x Document Count`

matrix. This is a very high dimmensional structure. In order to effectively traverse this corpus, we hope to express documents in the form of priciple components, a dimmentionality reduction technique that strictly describes an observed document set. This can be very difficult to interpret and has very little predictive power for unobserved documents.

Latent Dirichlet Allocation is a generative model for fitting topics to a corpus. This is done by electing a document, then a topic for that document, then a word for that topic. The yield of this trained model is a mapping from an input document to a topic-space. This is a softer definition allowing us to observe future documents.

The Author-Topic-Model extends LDA to model authors as a mixture of topics. This is, again, a generative model that elects a document, then an author for that document, then a topic for that author, then a word for that topic. This still learns a topic space from the vocabulary, but expresses both authors and documents in this topic space.

For the PRS, persons who are assigned to tickets are considered experts in that topic, because they are able to or have resolved that ticket. In this implementation the asignee is expressed as an author. This process results in a dimmentionality reduction over our topic space, and an inferred dimmentionality increase for our authors (who were previously just singleton tokens).

4 Fitting

As long as these authors are appropriately represented in the training set, they act as a strong litmus test against this author modeling technique. This is difficult to replicate, due to the small sparse candidate dataset, so we explore rank utility [4, 9] instead.

```
PFR
  :AUTHORS   - 12929
  :ANOMALIES - 10011
Author Types
  :RESPONSIBLE EDITOR
  :ASSIGNEE
```

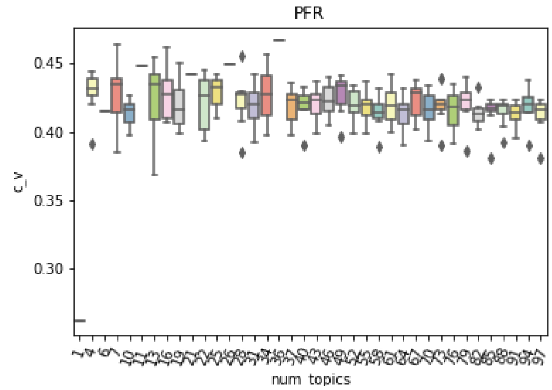
```
DPFR
  :AUTHORS   - 6838
  :ANOMALIES - 5463
Author Types
  :RESPONSIBLE EDITOR
  :ASSIGNEE
```

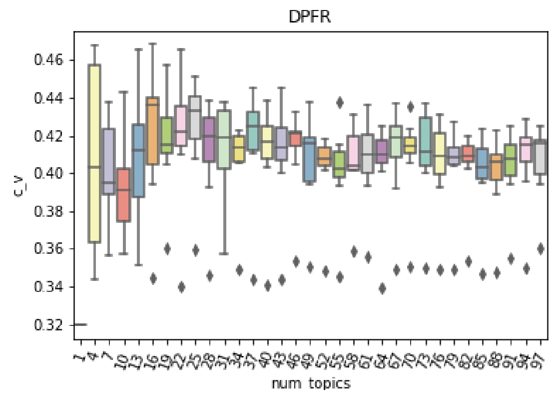
```
ISA
  :AUTHORS   - 8148
  :ANOMALIES - 7932
Author Types
  :RESPONSIBLE EDITOR
  :ASSIGNEE
```

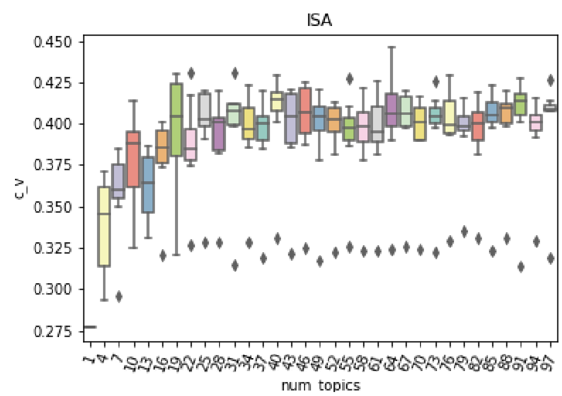
They [8] also state and describe the process for using these single author papers for perplexity.

> Perplexity is the standard measure for estimating the performance of a probabilistic model.

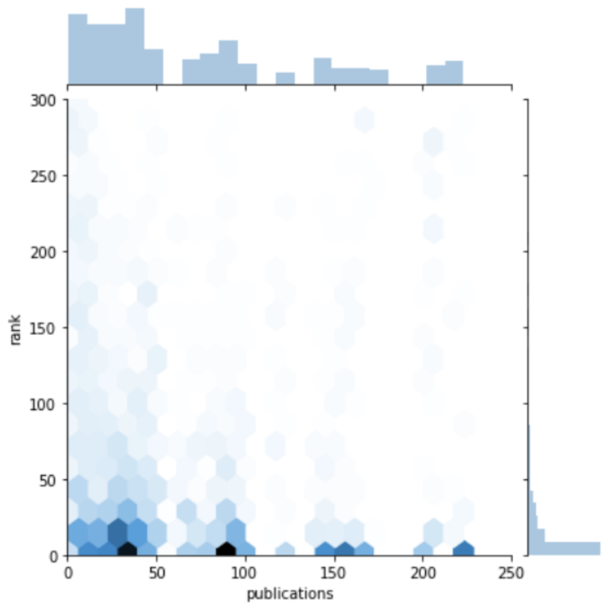
Alternative, exploration using coherence metrics [7], as this has been shown to provide more understandable topics when applied to bag-of-words topic models.

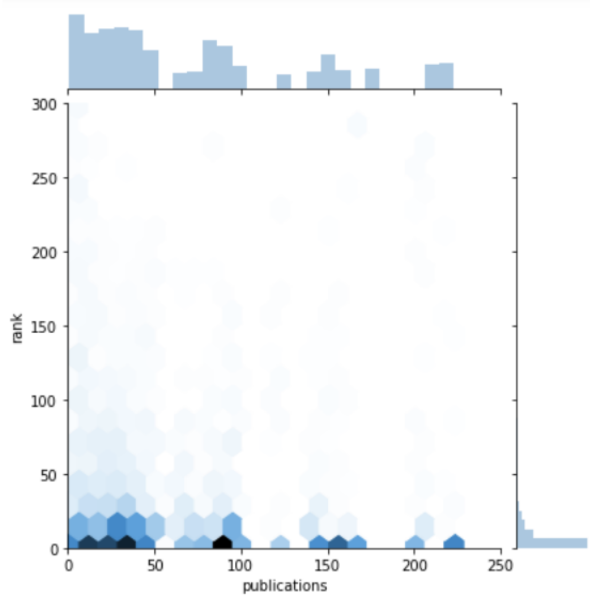




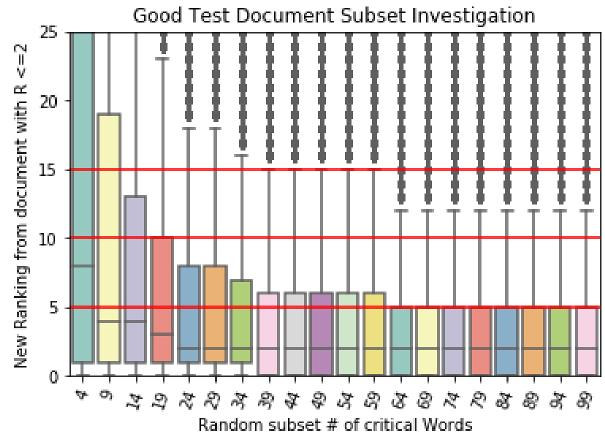


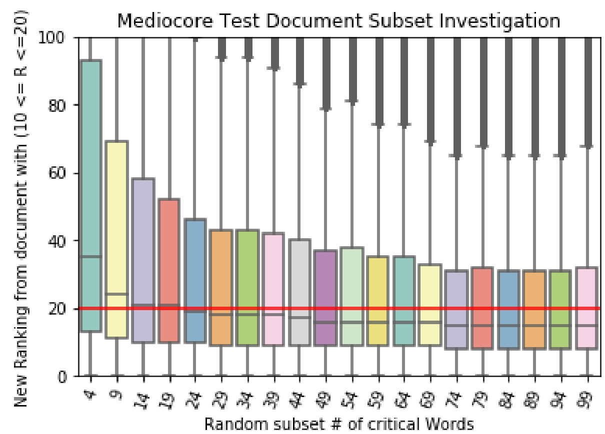
It is worth asking if number of publications impacts model performance, in unexpected ways. From the plots below it is shown that the predictive power does not significantly deviate with publication count.



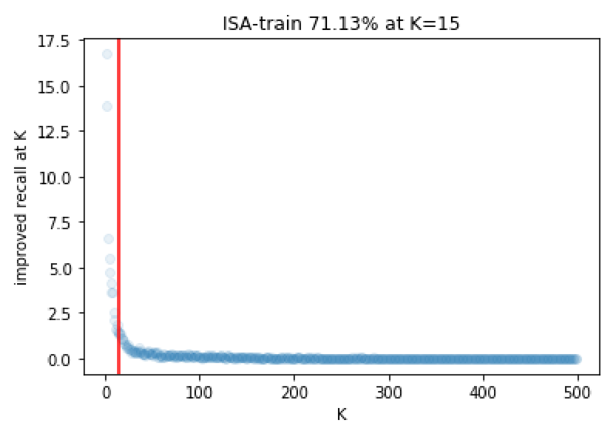


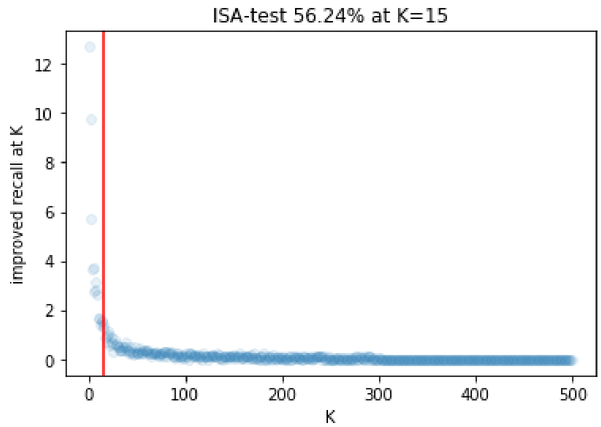
It is also interesting to know how much text is required to inform our model prediction. For these plots, we randomly subset texts for known ticket-expert pairs and observe the expert's new ranking. It is shown that at 30 words, our model is considered consistant with best fit. This is significantly less than our average document size.





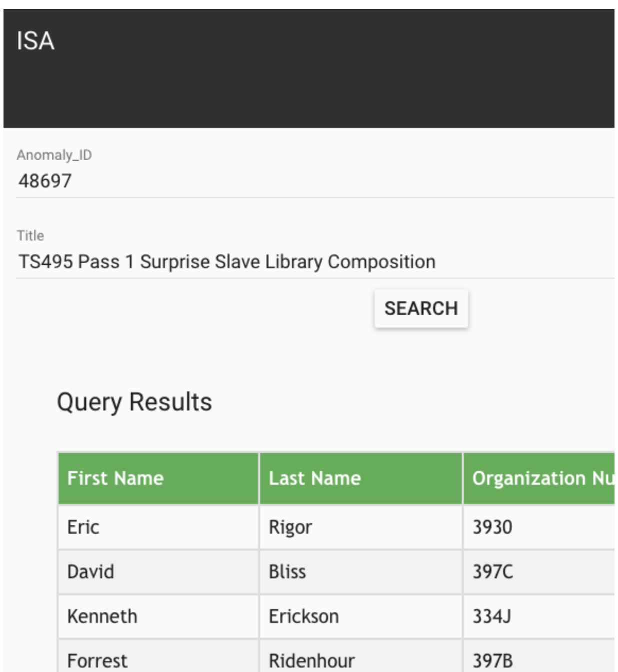
Finally, its important to look at predictive power. These plots show the recall from the top 15 elements for `ISA` documents over `train` and `test`.





5 Results

The yield of this project is a site to allow queries against a learned model. This will be used to beta test and gather more data from customers using the 5x PRS.



The application goals are met sufficient to move this project forward, and incorperate more data. It will also be important that we investigate the stability of models [11], in order to incorporate later documents, while minimizing effect to the user experience. Finally, we need to better understand our false positives. This will require contacting persons who rank high but are not attributed to those tickets.

6 Conclusion

References

[1] Khaled Aldebei, Xiangjian He, Wenjing Jia, and Jie Yang. Unsupervised multi-author document decomposition based on hidden markov model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

[2] Rubayyi Alghamdi and Khalid Alfalqi. A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications*, 6(1), 2015.

[3] Weizheng Chen, Jinpeng Wang, Yan Zhang, Hongfei Yan, and Xiaoming Li. User based aggregation for biterm topic model. In *ACL*, 2015.

[4] Asela Gunawardana and Guy Shani. A survey of accuracy evaluation metrics of recommendation tasks. *J. Mach. Learn. Res.*, 10:2935–2962, December 2009.

[5] Shawn Minto and Gail C. Murphy. Recommending emergent teams. In *Proceedings of the Fourth International Workshop on Mining Software Repositories*, MSR '07, pages 5–, Washington, DC, USA, 2007. IEEE Computer Society.

[6] Andi Rexha, Mark Kröll, Hermann Ziak, and Roman Kern. Authorship identification of documents with high content similarity. *Scientometrics*, 115(1):223–237, Apr 2018.

[7] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408, New York, NY, USA, 2015. ACM.

[8] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press.

[9] Thiago Silveira, Min Zhang, Xiao Lin, Yiqun Liu, and Shaoping Ma. How good your recommender system is? a survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, Dec 2017.

[10] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1445–1456, New York, NY, USA, 2013. ACM.

[11] Yi Yang, Shimei Pan, Jie Lu, Mercan Topkara, and Yangqiu Song. The stability and usability of statistical topic models. *ACM Trans. Interact. Intell. Syst.*, 6(2):14:1–14:23, July 2016.