Identifying Subject Matter Experts

Extending Author Topic Modeling

Philip Robinson

Presented to OCIO
NASA - Jet Propulsion Lab

August 23, 2018

Jet Propulsion Laboratory
California Institute of Technology

Introduction - Philip Robinson

Computer Science MSc at Oregon Health and Science University.

Thanks to my mentor Ian Colwell, from the OCIO (1762)

- probabilistic programming
- language processing
- image processing
- audio processing
- stem education
- environmental sciences

Presentation Overview

1. Introduction
2. Problem Description
3. Proposed Approach
4. Investivation
5. Results
6. Conclusion

Problem Description

Our customers, Office of Safety and Mission Success (5x), are interested in identifying experts for resolving anomaly reports in the Problem Reporting System (PRS)

OCIO (17x) has been previously asked for subject matter expert identification systems and document similarity tools, so show particular interest is solutions that provide similarity metrics and can generalize to other teams and corpora.

- A-Team heirarchical frequent item set expert exploration tool
- TechConnect self reported skills host
- Gateway Profiles

Motivating Story

- Domain experts are lost between projects
- Domain experts are often coupled to a single project
- Very few candidates resolve the majority of tickets

- Expert discoverability
- Load balancing employees
- Identification of knowledge gaps

Objective

- Assign & Resolve anomalies quicker
- Support queries for expert discovery
- Find employees with similar domain expertise

Data Provided for Internship

- Problem Reporting System (Anomalies)
  - Problem Failure Report (PFR)
  - Developmental Problem Failure Report (DPFR)
  - Incident Surprise Anomaly (ISA)

- Free Text
  - Title
  - Description
- Experts
  - Responsible Editor
  - Assignee

Approach - Topic Modeling

- Interpret doc as BOW[1]
- Model/Fit topics as mixture of words
- Author & document are projected into topic-space
- Measure distance from author to document

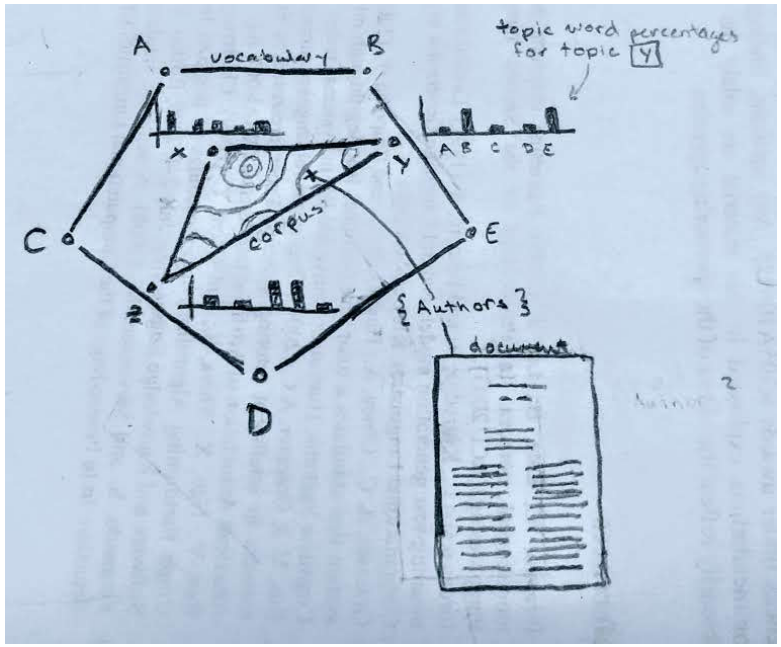


Figure: Latent Dirichlet Allocation

$$T(x) = \texttt{Project } x \texttt{ into topic-space}$$

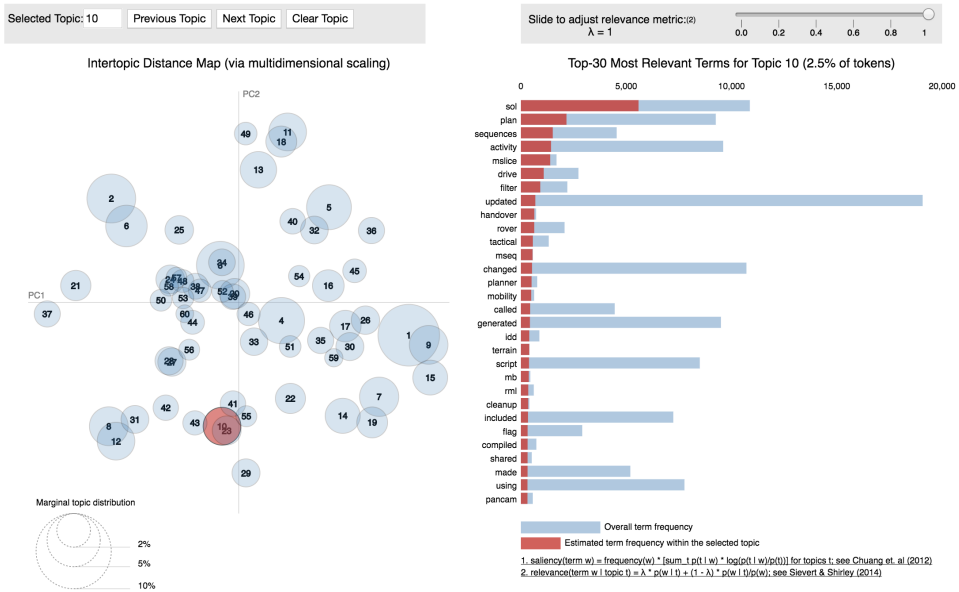
$$R_d = \underset{a \in A}{\text{argsort}} \{ Distance(T(a), T(d)) \}$$

[1] Bag of Words

Latent Dirichlet Allocation (LDA)

LDA is a generative model that describes each document as a mixture of topics, and each topic as a mixture of words.

- Tune hyper-parameters
  - ISA 351 iterations, 40 topics
  - PFR 251 iterations, 25 topics
  - DPFR 351 iterations, 25 topics
- Measures
  - Perplexity (statistic)
  - Coherence (statistic)
  - Visualizations (exploratory)



Author Topic Model

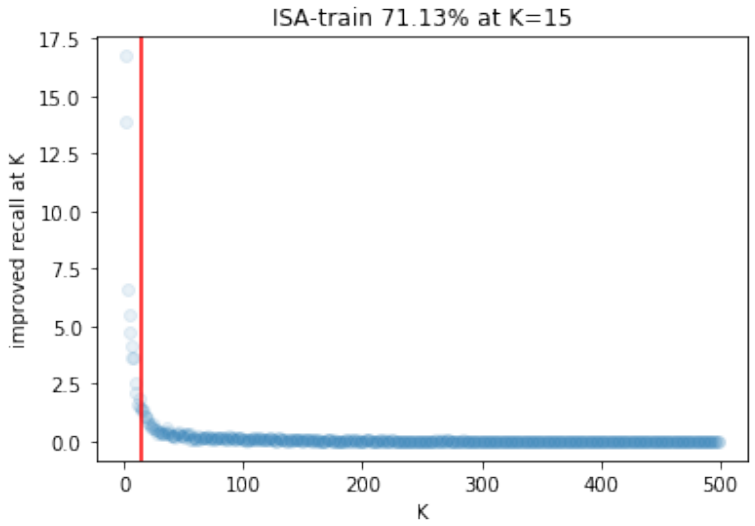
ATM extends LDA to describe authors as a mixture of topics. This allows us to ask questions relating to both documents and authors. Both documents and authors are now mapped to the "Topic Space" described in LDA.

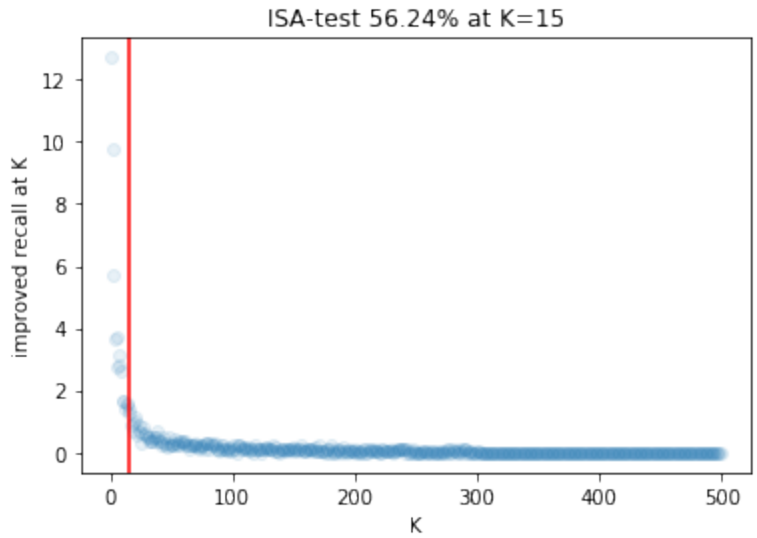
- Down sample to documents with a Responsible Editor / Assignee
- Split into Train and Validation
- Train model

Ranking

Given Authors/Experts in a "Topic Space" and a mapping from document to "Topic Space", we can rank likely authors for a document. Ideally this is done with the probability of an author given a document, but presently we use distribution similarity metrics to rank authors.

- Elect a similarity metric (`Hellinger Distance`)
- Rank likely authors on Train and Test document sets
- Inspect/Present results





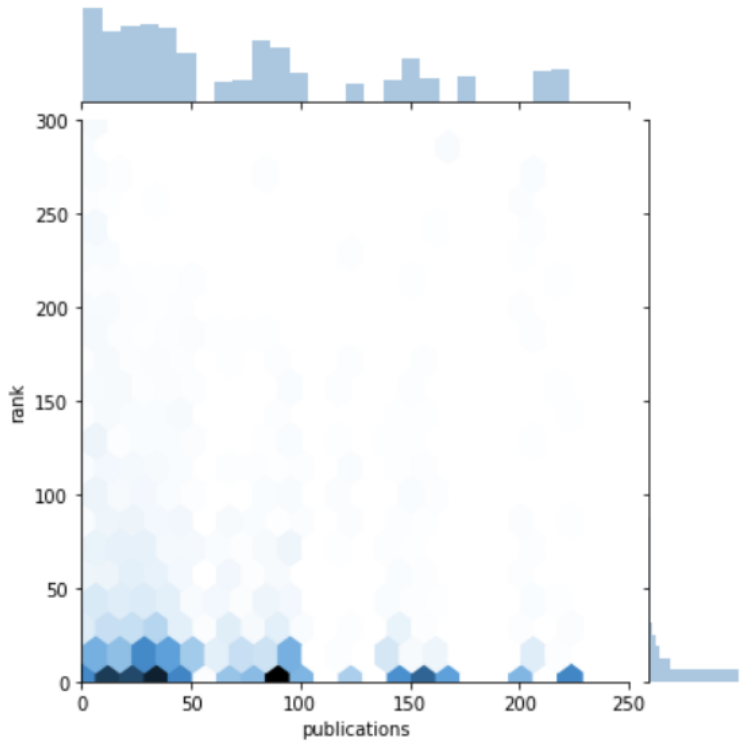
K is cutoff for suggested candidates

Does publication count effect recall?

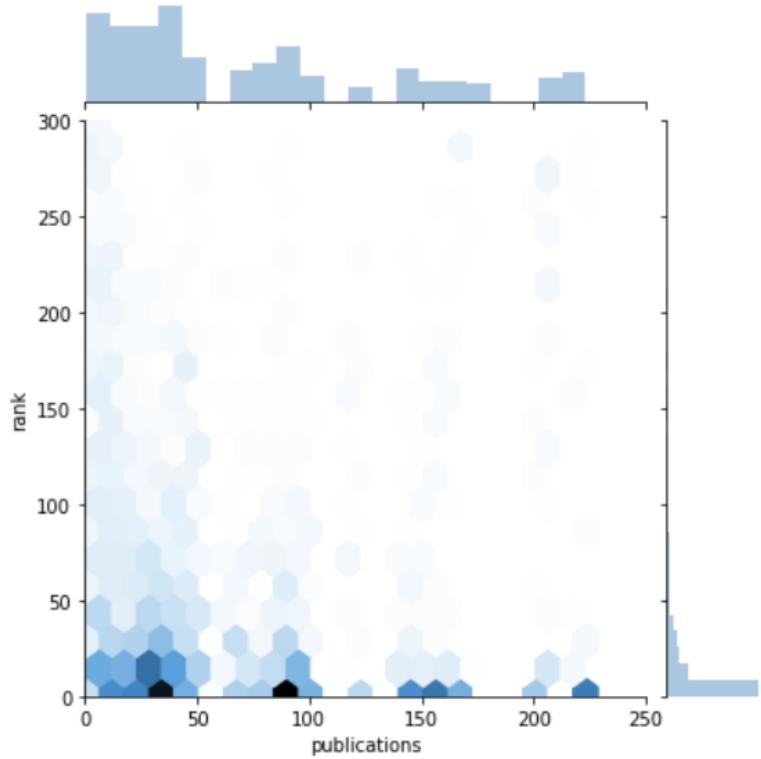
Nope; and thats good

This plot shows consistent identification distribution as a function of publication count from the train set.

Train



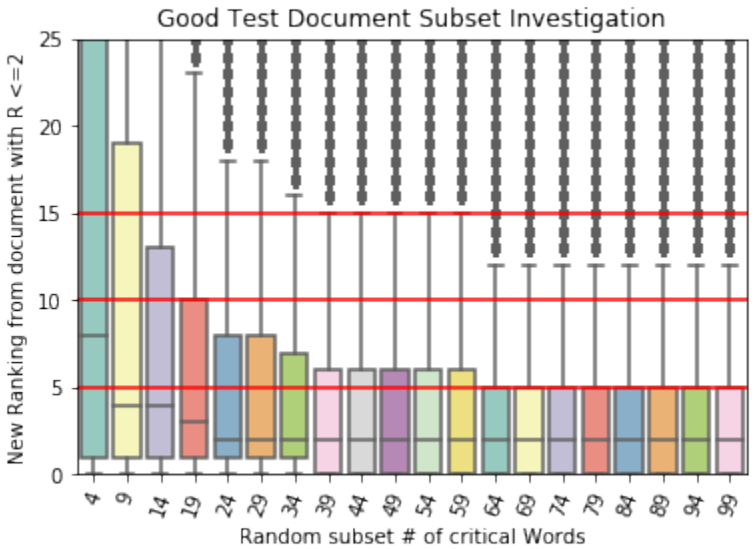
Test



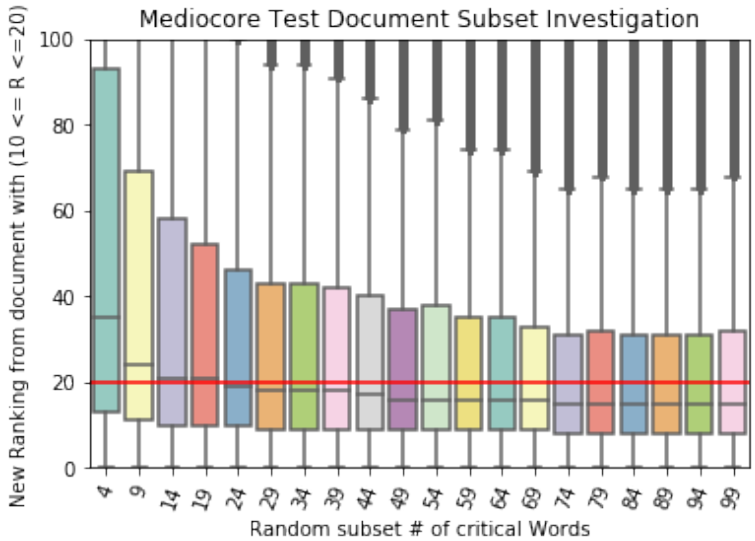
How does word count effect recall?

Best results at 30 words

We are interested in understanding 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.



Expert found in top 2
24 critical words



expert found in 10-20 range
29 critical words

Results

It is possible to get interesting results at a document length of 4 words, however it is hard to know why these results are interesting. This is an example of directly searching for experts.

`'gimbal drive motor friction'`

	Name	Title	Organization
0	Amanda Donner	Mission Assurance Manager	5150
1	John Trager	NaN	337C
2	Mathew Keuneke	Product Delivery Manager	397A
3	Jessica Bowles-Martinez	Systems Engineer	313G
4	NaN	NaN	NaN

`'rtg temperature drive curiosity capacity'`

	Name	Title	Organization
0	John Rakiewicz	NaN	NaN
1	Angela Dorsey	Technologist	335S
2	Otfrid Liepack	Deputy System Manager	394G
3	Mohammad Shahabuddin	Flight Software Engineer	348D
4	Megan Lin	Delivery Manager	397S

Documents queries generate better results. These examples are currently unable to demo.

Next Steps

- Extend tool to provide visualizations supporting for ranking
- Build UI for testing and authoring documents
- Contact high ranked persons to verify expertise
- Filter based on job title
- Open Sourcing

Conclusion

- This looks like it works
- Not computationally or socially prohibitive
- Motivating story to lift limitations on data access
- Further investigation is needed to insure best expert fields
- User interface is required for integration into workflows

Internship Notes

I had never implemented a recommender system, nor used topic modeling in a project prior to this task. I will be leaving JPL with a holistic understanding of topic modeling, and a much better understanding of recommender systems.

This work has encouraged me to look at continued employment in applications of machine learning for information retrieval more realistically. I'm also interested in keeping in touch over future employment opportunities with JPL.

Having multiple interns work with the same data helps. We were able to work through data difficulties together, and share results.

Thanks

I received a lot of input in for this project from my team. Their insight, especially, bridged the gap from academic to practical evaluation of models, and I sincerely appreciate their contributions.

- Ian Colwell
- Valentinos Constantinou
- Jerry Chen
- Leslie Callum
- Bruce Waggoner
- Harald Schone
- Chris Mattmann

et. al.