CEMMSS Topic Modeling Technical Appendix

# Probabilistic Topic Models

## Introduction to Topic Models

Topic models estimate the distribution of abstract concepts (topics) that make up a collection of documents. Probabilistic topic models obtain this estimate by constructing model of word generation. In this framework, documents are a probabilistic distribution of topics. Topics are a probabilistic distribution of words. To “write” a single word, an author must perform a two-step sampling process. First, the author selects a topic, subject to some probability.[[1]](#footnote-1) Next, the author selects a word from that topic, subject to some probability.[[2]](#footnote-2) The word is recorded and the word and topic are replaced. For a 300-word document, this two-step process of sampling with replacement is repeated 300 times, 500 times for a 500-word document, and so on. The objective of topic modeling is to estimate these probability distributions, noted below.

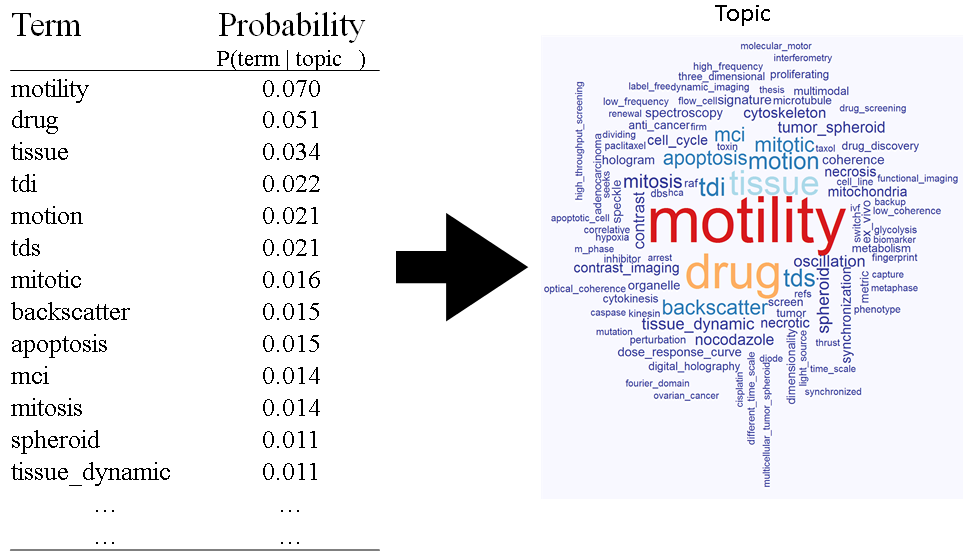


Figure 1: Topics are probability distributions over words

A useful framework is to think of the distribution of topics over documents, , as the proportion of the document that is “about” each topic. The formal definition is that it is the probability that any random word in a document was generated by each topic. The probability distribution of topics over documents is often misinterpreted. Therefore, it is safest to consider this distribution as a percentage. For example, if , then document is 60% about topic .

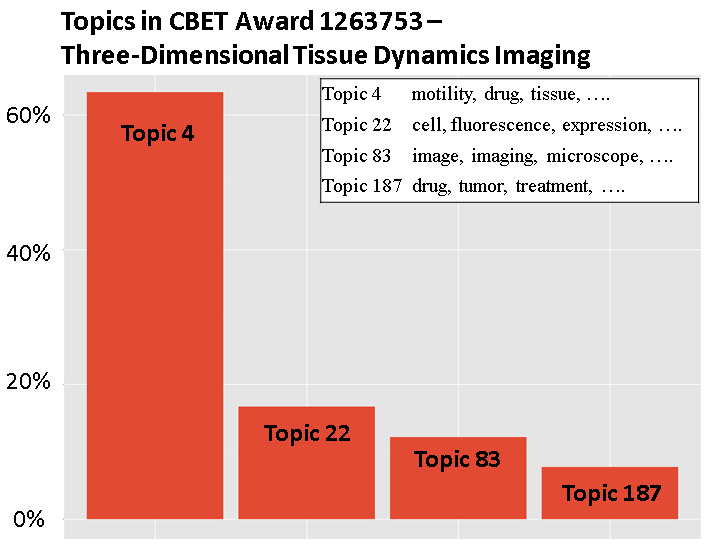


Figure 2: Topics are distributed within documents

## Latent Dirichlet Allocation

STPI chose Latent Dirichlet Allocation (LDA) as the topic model for the CEMMSS analysis. LDA is a Bayesian model; it uses prior probabilities to aid estimation. These probabilities are tuned by the parameters and . These parameters tune the distribution of topics over documents and words over topics, respectively. The number of topics, , is assumed to be fixed before modeling. Choosing is discussed in the following section. A formal mathematical statement of LDA is below.

The objective of LDA is to estimate and for every document and topic. These estimates are combined into matrices and for the whole corpus. The rows of are and the rows of are .

STPI used collapsed Gibbs sampling to estimate the CEMMSS topic model. The collapsed Gibbs sampler is part of the “lda” package for the R programming environment.

# Constructing the Topic Model

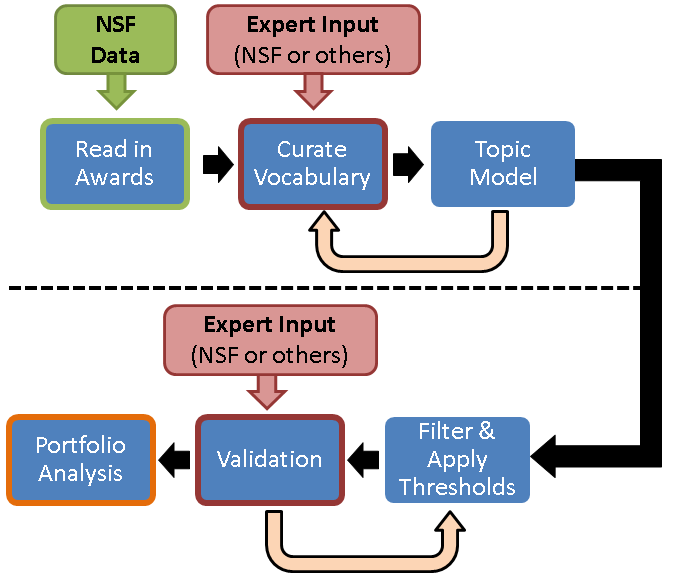


Figure 3: STPI's topic modeling process

## Data Ingestion

STPI used abstracts and project summaries to construct the topic model of 904 CEMMSS projects. The data consisted of 1,201 unique awards. Collaborative awards were collapsed into projects. This reduced the 1,201 awards to 897 projects. Seven projects were funded by more than one CEMMSS program. For example, a project may have received funding from both RI and NRI. STPI “duplicated” these projects to have a representative sample of each program. This raised the number of projects for topic modeling to 904.

STPI used its local copy of NSF’s public-facing awards database for all project data except for project summaries. These data include project titles, abstracts, PIs and Co-PIs, funding amounts and more. None of the 904 projects (and 1,201 awards) are missing data for the above fields.

STPI obtained summaries for 814 of 904 projects. STPI received full grant applications in PDF or XML format from NSF. In some cases, the PDF could not be converted to plain text without corrupting the data. In other cases, the “project summary” field of the XML was blank. As a result, STPI could not use the summaries of 90 projects. STPI concatenated project summaries and project abstracts for modeling. For the 90 projects that did not have a summary, STPI concatenated two copies of the project abstract. This was done to mitigate disparities in document length between project with and projects without summaries.

In many cases text needed additional cleaning to be useful. For example, many project abstracts contained errant HTML code. In other cases, ligatures were missing, leaving words incomplete.[[3]](#footnote-3) Finally, some words were broken by new lines and hyphenated, making them appear as separate words.[[4]](#footnote-4) STPI corrected these issues where possible to preserve the original integrity of the documents.

## Constructing the Vocabulary

For a meaningful topic model, STPI constructed a vocabulary of included terms. First, STPI removed common “stop” words from the documents. Stop words are very common words, such as “the”, “and”, “with”, etc. All words were converted to lower case. STPI converted all punctuation and numbers to spaces. STPI included both single words and common ordered-pairs of words. These single words and word pairs are collectively referred to as “terms.” STPI removed some terms based on document frequency. Terms appearing in more than half of the projects, fewer than 5 projects, or appearing no more than twice per-document (on average) were removed. Finally, certain “confounder” terms were removed. Confounder terms include terms like “cps”, “cyber\_physical”, “high\_school”, “research”, etc. These terms, create legitimate but uninteresting topics for the purposes of analyzing the CEMMSS portfolio. STPI focused on discovering scientific topics, rather than topics related to administration or other NSF programs.

STPI discovered and removed confounder terms through an iterative modeling process. STPI constructed a topic model, as described in the next section. The constructed model was used for preliminary analyses. Confounding topics were discovered in these preliminary analyses. One modeling iteration, for example, all CPS documents contained high proportion of a topic whose most probable terms were “cps”, “cyber\_physical”, “physical\_systems”, “cyberphysical” etc. These confounding terms were flagged for removal in the next modeling iteration.

## The Final Topic Model

The final topic model was an “ensemble” topic model. Ensemble models combine the output of many different models into a single aggregate model. Ensemble models have been shown to be more accurate than any single model in many contexts, including topic modeling. Ensemble models’ predictions tend to be more robust on new data. The final model contained 96 topics, aggregated across 50 modeling runs. The entire modeling process is described in the following paragraphs.

### Initial parameter settings

Latent Dirichlet allocation has three parameters. Two of these parameters are Bayesian priors, sometimes called “smoothing” parameters. These parameters— and —shape the distributions of topics within documents and words within topics respectively. The third parameter, , is the number of topics. The settings that STPI chose for these parameters were:

* was considered for a range of values, described in the next paragraph.

The and parameters influence the estimates for and , however the initial impact of and dissipate with more iterations of the Gibbs sampler. However, is considered “fixed”; it does not update during the model fitting process. As a result, STPI considered a range of values for and selected one with the most statistical support in the data. Figure [4], below, plots the log likelihood of the data, assuming the resulting model is correct. STPI chose 125 topics as the setting for in the 50 ensemble runs.

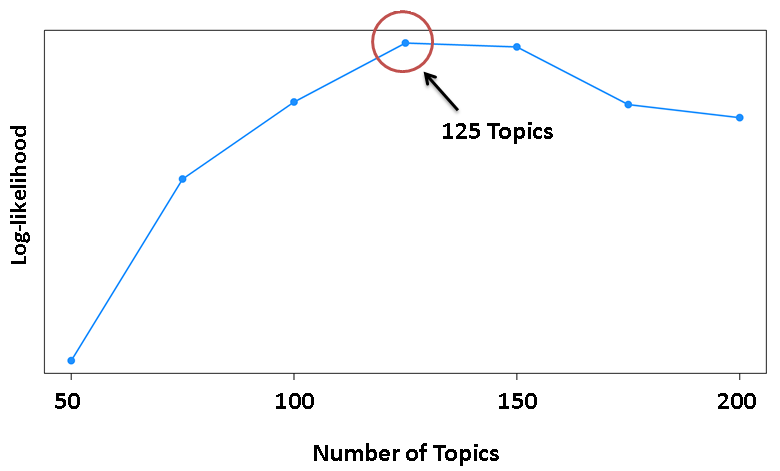


Figure 4: Selecting the number of topics by log likelihohood

### Filtering & Applying Thresholds

STPI combined topics into an ensemble by clustering topics across model runs. Fifty models of 125 topics resulted in 6,250 total topics. Topics were combined using linguistic similarity only, not co-occurrence within documents. STPI measured the cosine similarity between topics’ distributions over terms. STPI then used hierarchical agglomerative clustering with Ward’s decision rule for combining clusters to aggregate topics. Since the log likelihood method indicated 125 topics, STPI partitioned the hierarchical clustering tree at 125 topics. Topic distributions over words () and over documents () were combined according to cluster assignment. The resulting ensemble model, then, contained 125 topics at this preliminary stage.

STPI removed 29 topics whose distributions, , did not have statistical support. STPI measured this support using a metric called probabilistic coherence, described in the “Validation Metrics” section of this appendix. Topics with a probabilistic coherence of less than 0.08 were removed. The final ensemble topic model contained 96 topics.

STPI defined a threshold of 0.05 for definitively assigning topics as “in” or “out” of documents. One issue with LDA is that it assigns positive values for every topic within every document. This is obviously unrealistic. While a corpus may contain topics in total, each document likely contains only a handful of those topics. By assigning a threshold of 0.05, projects in the CEMMSS topic model contained between 1 and 10 topics, with an average of 4. By extension, most topics appeared in 1 to 60 projects. Only 8 topics appeared in 61 or more projects, with topic 99—“robotic systems”—appearing in 220 projects. Figure [5] displays these summary statistics.

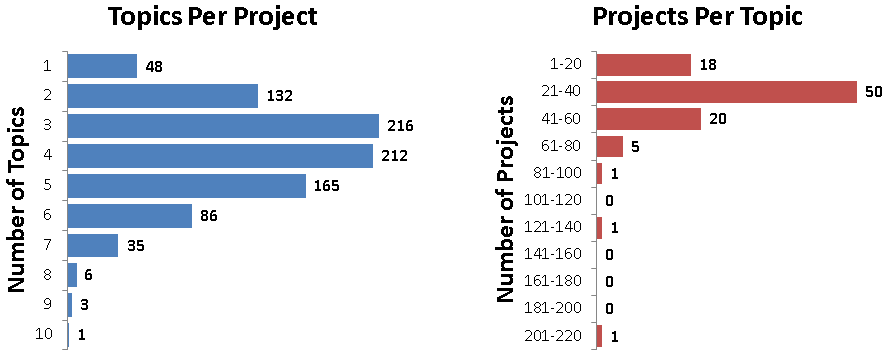


Figure 5: Summary statistics of the final topic model

## Linking Topics to PIs

STPI explored topical links between PIs[[5]](#footnote-5). PIs are implicitly linked to topics through projects. The NSF awards data lists PIs for each award. In aggregating awards to projects, STPI included all PIs on each award. STPI constructed a matrix linking PIs to projects. STPI called this matrix . The rows of represent unique PIs; the columns represent unique projects. The entries of equal 1 if PI­  is an investigator on project , and equal 0 otherwise. Topic proportions for each PI were obtained through simple matrix multiplication: .

## Validation and Other Metrics

STPI used several metrics to validate, aggregate, and otherwise analyze the CEMMSS topic model. This section provides mathematical statements of these metrics and describes their properties. Metrics used for network analysis are not covered in this section. Those metrics are covered in the section titled “Network Analyses with the Topic Model.”

### Probabilistic Coherence

STPI developed a metric to measure statistical support, also known as coherence or cohesion, of the distributions of terms in topics. People tend to view “raw” topic labels as the first several most probable words in . Probabilistic coherence measures the propensity for these top words to appear together within documents. For the top M most probable terms in topic we calculate

where is a term more probable than in that topic. In other words

And is the probability of term being in a randomly selected document in the corpus.

Probabilistic coherence is bound between -1 and 1. Positive values indicate that we are more likely to find word if we search only in documents containing word , rather than choosing documents at random. Values that are close to zero mean that the more probable and less probable terms in a topic are nearly statistically-independent. Not much information is gained by grouping these terms together in such topics. Negative values indicate that we are *less* likely to find word by searching in documents containing word . This topic is likely nothing more than random noise. The 5 topics with highest probabilistic coherence are in the table below.

Table 1: Topics with highest probabilistic coherence

|  |  |  |
| --- | --- | --- |
| **Topic** | **Label** | **Probabilistic Coherence** |
| **t.35** | Supply Chain Management | 0.38 |
| **t.72** | Welding processes | 0.37 |
| **t.73** | Powder sintering | 0.36 |
| **t.36** | Fatigue Life | 0.34 |
| **t.7** | Image Reconstruction / Recognition | 0.34 |

### Topic Prevalence

STPI calculated the prevalence of topics in the corpus overall, as well as in each CEMMSS program. Prevalence is simply the sum of topic shares across a set of documents. We normalize the prevalence score and scale it by 100. Mathematically, the prevalence of a topic,, across a set of documents with topics is

The top 5 most prevalent topics in the whole corpus are listed in Table [2]. The most prevalent topic for each program is listed in Table [3].

Table 2: Top 5 Most Prevalent Topics

|  |  |  |
| --- | --- | --- |
| **Topic** | **Label** | **Prevalence** |
| **t.99** | Robotic Systems | 8.34 |
| **t.10** | Object Recognition | 4.09 |
| **t.90** | Sensing Systems for Medical Devices | 2.35 |
| **t.80** | Computer Vision for Image Search | 2.34 |
| **t.31** | Sensing Systems for Power Grids | 2.27 |

Table 3: Most Prevalent Topic for Each Program

|  |  |  |  |
| --- | --- | --- | --- |
| **Program** | **Topic** | **Label** | **In-Program Prevalence** |
| **AM** | **t.55** | Manufacturing Systems | 5.01 |
| **RI** | **t.99** | Robotic Systems | 16.70 |
| **CPS** | **t.31** | Sensing Systems for Power Grids | 9.29 |
| **NRI** | **t.99** | Robotic Systems | 35.50 |
| **DMREF** | **t.54** | Materials modeling | 21.70 |

### Jensen-Shannon Divergence

STPI used the square root of the Jensen-Shannon Divergence (JSD) as the distance metric for clustering topics and documents and creating similarity networks of the ensemble topic model. JSD is a popular measure in probability theory. It measures the difference in information between two probability distributions. The square root of JSD is a formal mathematical metric.[[6]](#footnote-6) For two probability distributions, and of length , JSD is stated as follows,

where .

### Cosine Similarity

STPI used cosine similarity as a distance metric for aggregating individual topic models into an ensemble. Cosine similarity is not a formal mathematical metric. JSD was preferred; however cosine similarity is a more computationally-efficient measure. There were about 234 billion comparisons to be made in constructing the ensemble model, making JSD inefficient. Cosine similarity between two vectors, and of length , is defined as follows

### Silhouette Coefficient

STPI used the mean silhouette coefficient to aid selecting the number of clusters when clustering topics and documents. The silhouette coefficient scores how well an individual point fits within its cluster. Silhouette is bound between -1 and 1. Negative values indicate that a point is closer to points in another cluster than to points in its own cluster. Over a range of potential cluster partitions, STPI calculated the average silhouette score for all points. The use of silhouette is discussed in the section titled “Clustering with the Topic Model.”

For a clustering solution with clusters,

Let be the average distance from point to all the points in the same cluster as .

Let be the average distance from point to all points in the next-nearest cluster.

Then is the silhouette coefficient for point and is defined as

# Clustering with the Topic Model

## Clustering Topics

STPI clustered topics in the ensemble model based on word distributions (linguistic similarity) as well as on document co-occurrence. STPI measured the square root of JSD between topic distributions over words. STPI also measure the square root of JSD between topic distributions over documents. These two distance metrics were averaged together before clustering. STPI used agglomerative hierarchical clustering with Ward’s method for cluster merging. The silhouette plot for choosing the number of clusters is in Figure [6].

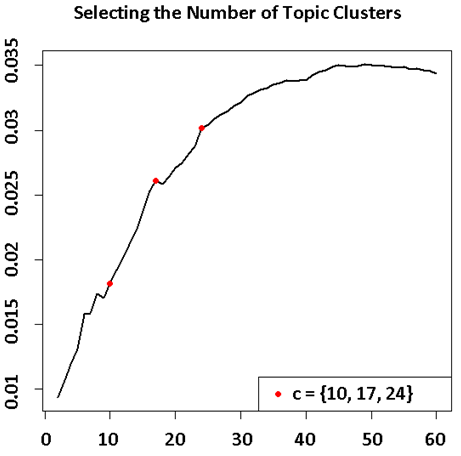


Figure 6: Silhouette Plot of Topic Clusters

## Clustering Documents

STPI clustered documents similarly to how it clustered topics. However, STPI calculated the square root of JSD for documents distributed over topics only. As with topics, STPI used agglomerative hierarchical clustering with Ward’s method. The silhouette plot for clustering documents is in Figure [7].

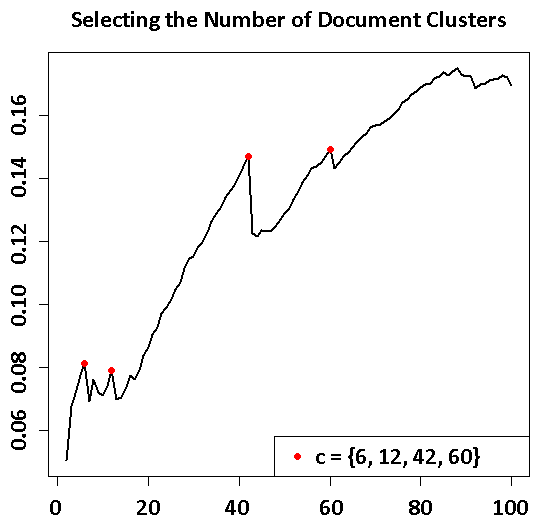


Figure 7: Silhouette Plot for Document Clustering

# Network Analyses with the Topic Model

## Topic Similarity Networks

STPI constructed a topic similarity network to visualize and analyze the topic model. A similarity network is similar to a social network. Nodes are entities (in this case topics). Unlike in social network analysis, edges between nodes indicate latent similarity, rather than an explicit connection. STPI used as the measure of similarity between topics and . This similarity includes linguistic similarity and document co-occurrence. This similarity measure is non-zero for every pair of topics. This would create a graph to “busy” to read. Therefore, STPI deleted edges between topics where . The remaining edges were proportionally weighted.

STPI calculated several network metrics on the topic similarity network. These metrics were betweenness centrality, eigenvector centrality, and closeness centrality. Betweenness centrality measures the centrality of each node for information transfer in the network. Nodes with high betweenness centrality are “hubs” for information transfer. As an example, consider a network of airports in the United States. Atlanta International Airport would have high betweenness centrality as many flights on the east coast must go through Atlanta, regardless of their final destination.

Eigenvector centrality measures the total connectedness of a node in a network. This includes 1st degree connections all the way out to th degree connections. Google’s *PageRank* is a variation of eigenvector centrality.

Closeness centrality is another measure of information spread in a network. Closeness centrality is the reciprocal of the total distance from a node to every other node in the network. Nodes with a high closeness centrality are those nodes most interconnected to the network.

## Document Similarity Networks

STPI constructed a document similarity network as well. The process was identical to creating the topic similarity network with one exception. The rule for edge deletion in the document network was . The document similarity network was used for visualization only. STPI did not calculate network metrics for the document similarity network.

# Labeling Topics

Interpreting the output of topic models is challenging. STPI used three machine-learned topic labels in combination with human judgment to assign topic labels. The three machine-learned labeling techniques are described in the proceeding sections. STPI also examined the three labels and titles of projects with high proportion for each topic in assigning the final label. The three machine-learned labels and the final label for three topics are below in Table[4].

Table 4: Assigning Topic Labels

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Topic** | **Final Label** | **Most Probable Terms** | **Naïve Label** | **KERA Label** |
| **t.68** | *Biomedical devices* | plant | implant | bone | mechanical | manufacturing | biological | memory\_alloy | medical\_device | natural bone | biological properties |
| **t.22** | *High Strength Steel* | alloy | strength | steel | deformation | plastic | magnesium | strength\_steel | magnesium\_alloy | strength steels | high strength |
| **t.52** | *Machining Processes* | machining | cutting | manufacturing | machine | micro | machining\_process | machining\_process | machine\_tool | physics-based modeling | high speed |
| **t.21** | *Robotics: Dexterous Manipulation* | hand | grasp | grasping | interface | haptic | deformable | deformable\_object | machine\_interface | dexterous manipulation | robotic hands |
| **t.67** | *Nanomaterial growth* | wire | nanowire | nanostructure | growth | nanoscale | multiscale | growth\_process | quality\_control | growth process | new generation |
| **t.86** | *Robotic Bipedal Walking* | locomotion | terrain | actuator | humanoid | walking | muscle | rough\_terrain | humanoid\_robot | bipedal walking | irregular terrain |
| **t.125** | *Semiconductors and Ceramics* | surface | organic | oxide | eager | node | anode | motion\_planning | computer\_vision | metric representations | stereo vision |
| **t.35** | *Supply Chain Management* | chain | supply\_chain | stochastic | inventor | inventory | policy | supply\_chain | based\_model | supply chains | supply chain |
| **t.73** | *Powder sintering* | powder | ceramic | sintering | meter | parameter | metal | fuel\_cell | material\_system | developed concepts | solid oxide |
| **t.75** | *Decision Support Systems for Supply Chain Management* | decision | consume | consumer | policy | service | operation | lead\_time | raw\_material | end-of-life products | inventory management |

## Most Probable Terms

An ordered list of the top 3 to 5 terms is the most common label for topic models. For a topic, , this label is the terms with highest probability in .

## Naïve N-Gram Labeling

N-grams are easier for humans to interpret than single words. STPI extracted all bigrams and trigrams from each document. For each topic, , STPI collected the set of documents, , where . Next STPI constructed two probabilities: the probability of each n-gram appearing in and the probability of each n-gram appearing in the total corpus. The following metric was calculated for every topic, :

The naïve n-gram label is the top 2 n-grams as ordered by .

## KERA Topic Labeling

Arun Maiya and Robert Rolfe, researchers at IDA’s Information Technology and Systems Division, developed a keyword extraction method. This method is named KERA for keyword extraction of reports and articles. This method extracts keywords for a set of documents by finding adjective-noun pairs and ranking them according to various criteria. Some of these criteria include the keywords’ prevalence across documents and where the words appear within the document.

# Bibliography

Blei, D. M. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research, 3*, 993-1022.

Chang, J. (2011). *lda: Collapsed Gibbs sampling methods for topic models.* Retrieved from http://cran.r-project.org/web/packages/lda/

Maiya, A. S.-L. (2013). Exploratory Analysis of Highly Heterogeneous Document Collections. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.*

Nguyen, V.-A. B.-G. (2014). Sometimes Average is Best: The Importance of Averaging for Prediction using MCMC Inference in Topic Modeling. *Empirical Methods In Natural Language Processing.*

1. The intuition: In a document about zoology, a topic about cats is far more probable than a topic about airplane engines, for example. [↑](#footnote-ref-1)
2. The intuition: The words “fur” and “claw” are more probable in a cats topic than words like “turbine” and “jet”. [↑](#footnote-ref-2)
3. Certain characters have special meaning for machine readability. For example “\f” indicates a page break; “\n” indicates a new line. Generally, these characters are not displayed when we read a document; we simply see a new page or new line in the text. The error arose for words such as “fluid” where the machine interpreter processed the ligature “fl” as “\f” in converting the PDF to text. Therefore, “fluid” appeared as “uid” in our plain text documents. This affected both PDFs received from NSF that STPI converted to text and the XMLs received from NSF where NSF converted the PDFs to plain text itself. The problem seems to depend on which program(s) originally compiled the PDF documents. This problem did not affect all projects, only ones whose project summary PDF files had a particular format. [↑](#footnote-ref-3)
4. For example, “regressors” may be converted to “re-gressors” with a new line between “re-” and “gressors.” The text conversion process resulted in “re gressors” appearing as two separate words. [↑](#footnote-ref-4)
5. For the purposes of this analysis, STPI uses “PIs” to refer to both principle investigators and co-principle investigators. Unique PIs were identified by name and email address. This method is likely not robust for large data sets. However, the CEMMSS portfolio is small enough that STPI could manually verify PI identities. [↑](#footnote-ref-5)
6. A formal metric is a distance function in topology. For this document, it is sufficient to say that being a metric gives the square root of JSD nice properties for analysis. [↑](#footnote-ref-6)