Advancing CAPA analysis with NLP: unlocking organizational insights and drawing on past improvements

Evgeny Kiselev[1], Danielle Larese[1], Selen Stromgren[1]

1 FDA/OC/OCS/OSLA

Office of the Chief Scientist

Office of Science and Laboratory Advancement

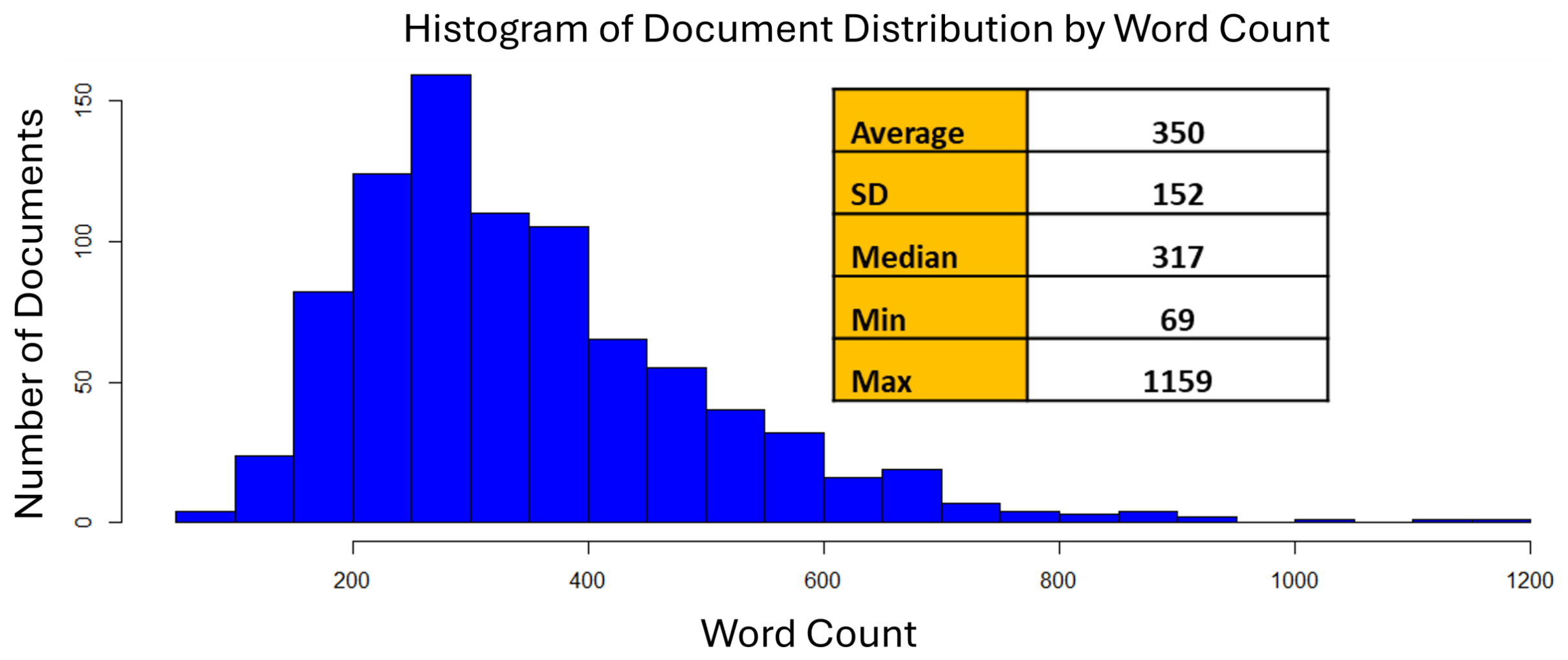
Introduction

The laboratories that are accredited to ISO/IEC 17025:2017 standard are required to "have a procedure that shall be implemented when any aspect of its laboratory activities or results of this work do not conform to its own procedures." The standard also requires that accredited entities shall implement corrective actions when an "evaluation indicates that the nonconforming work could reoccur."

Records detailing nonconforming work contain short text describing the nature of activity, brief evaluation, corrective action steps taken or planned.

In this project we investigated an approach for text analysis based on descriptive elements.

The library of approximately 900 documents was selected for this work. The documents contain user text which is on average 350 words long, not counting prepositions, articles, conjunctions, and numbers (stopwords).



Average	350
SD	152
Median	317
Min	69
Max	1159

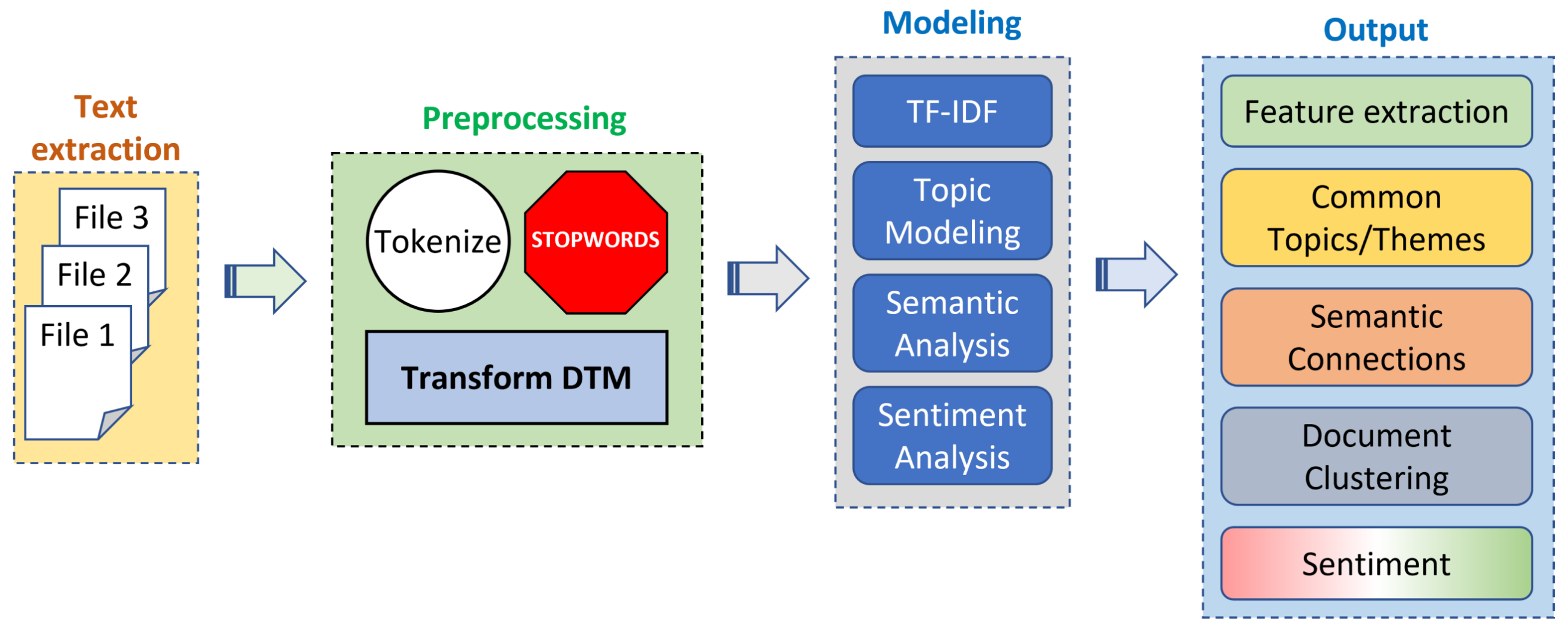
Here we describe an application of Natural Language Processing (NLP) tools for document analysis with the goal of identifying similar documents based on the text within.

NLP tools are used for several reasons. First, the user-input text has spelling errors, typos, and non-standard abbreviations/acronyms, which reduces the effectiveness of simple keyword searches. With the shorter text, there is a lower likelihood that it will contain all the keywords when multiple keywords are searched for, rendering keyword searches ineffective.

The goal of the project is two-fold. First, we would like to assist staff who are completing a new document for non-conforming work by identifying the most relevant existing documents for their reference. Second, we want to provide enhanced tracking and trending possibilities for management by better-categorizing the corpus of documents.

Materials and methods

Latent Dirichlet Allocation (LDA) was used to produce thematic topics. Document-specific unique terms were identified with Term Frequency-Inverse Document Frequency (TF-IDF). The results of LDA and TF-IDF were integrated by examining the semantic proximity of topics and unique terms with Latent Semantic Analysis (LSA).



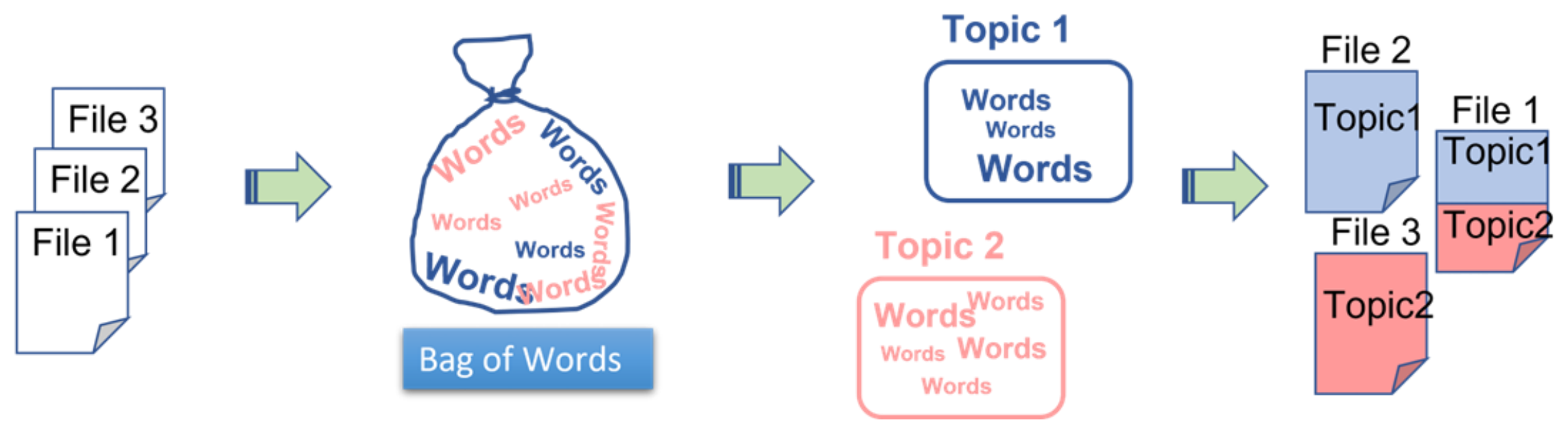
Preprocessing: Each document was preprocessed prior to application of NLP methods. User text for each document was tokenized (split into individual words), stopwords (words that do not contribute to the meaning of the text: articles, prepositions, etc.) were removed. Next, punctuation symbols and numbers were removed, and remaining words were converted to lowercase. Word occurrences per document were counted. A Document-Term Matrix (DTM) was constructed where each column is a word found in this library of documents and where each row is a document. The work frequency per document is specified in each word-document cell of the DTM, with 0 if the word does not occur in a document.

Modeling Methods:

TF-IDF method is applied to DTM. It extracts keywords that characterize each document individually. Each word is assigned a score, TF-IDF, that is maximized for unique (across all docs) and most frequent words in each document. The method considers all documents in the library, identifies words that are most unique to a particular document, making it easy to interpret by the user. TF-IDF identifies keywords that set the document in question apart from the rest of the library.

TF-IDF cannot be easily used to group documents as unique words are prioritized, it is also sensitive to spelling errors and typos across documents.

LDA is an unsupervised modeling technique where each document is considered a "bag of words" where information on grammatical constructs is discarded. Each document is represented as a ranked mix of topics and each topic is a ranked mix of words. Due to the probabilistic nature of LDA, there is variability in topic numbering, also the unsupervised nature of the method makes output validation challenging. Topics might overlap, i.e. have some of the same word/content, or two or more distinct topics may be combined into one.

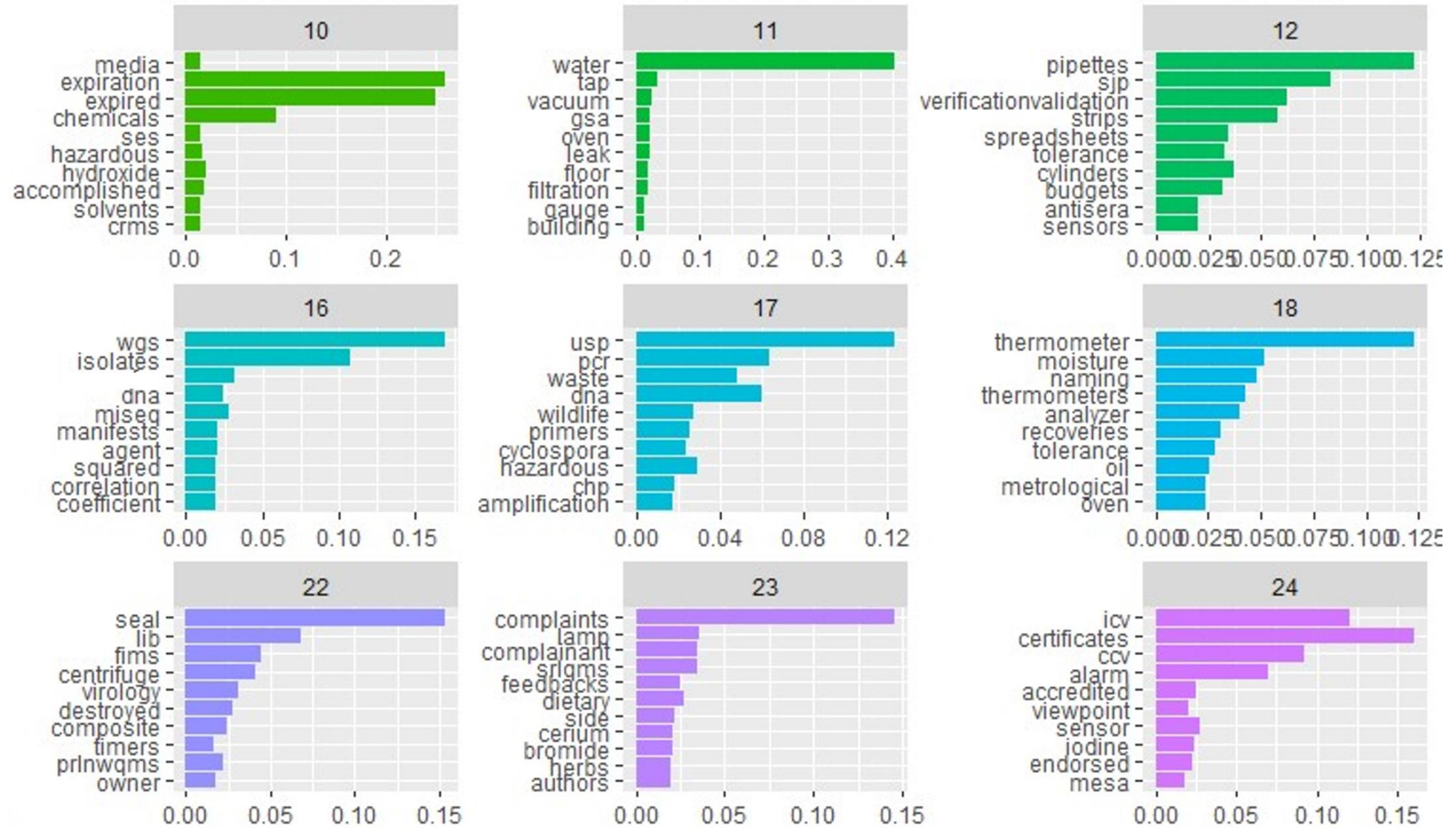


LSA allows for comparison of document or terms based on their "closeness." It addresses issues of having multiple words with the same meaning and words with multiple meanings. It identifies semantic relationships between documents and terms.

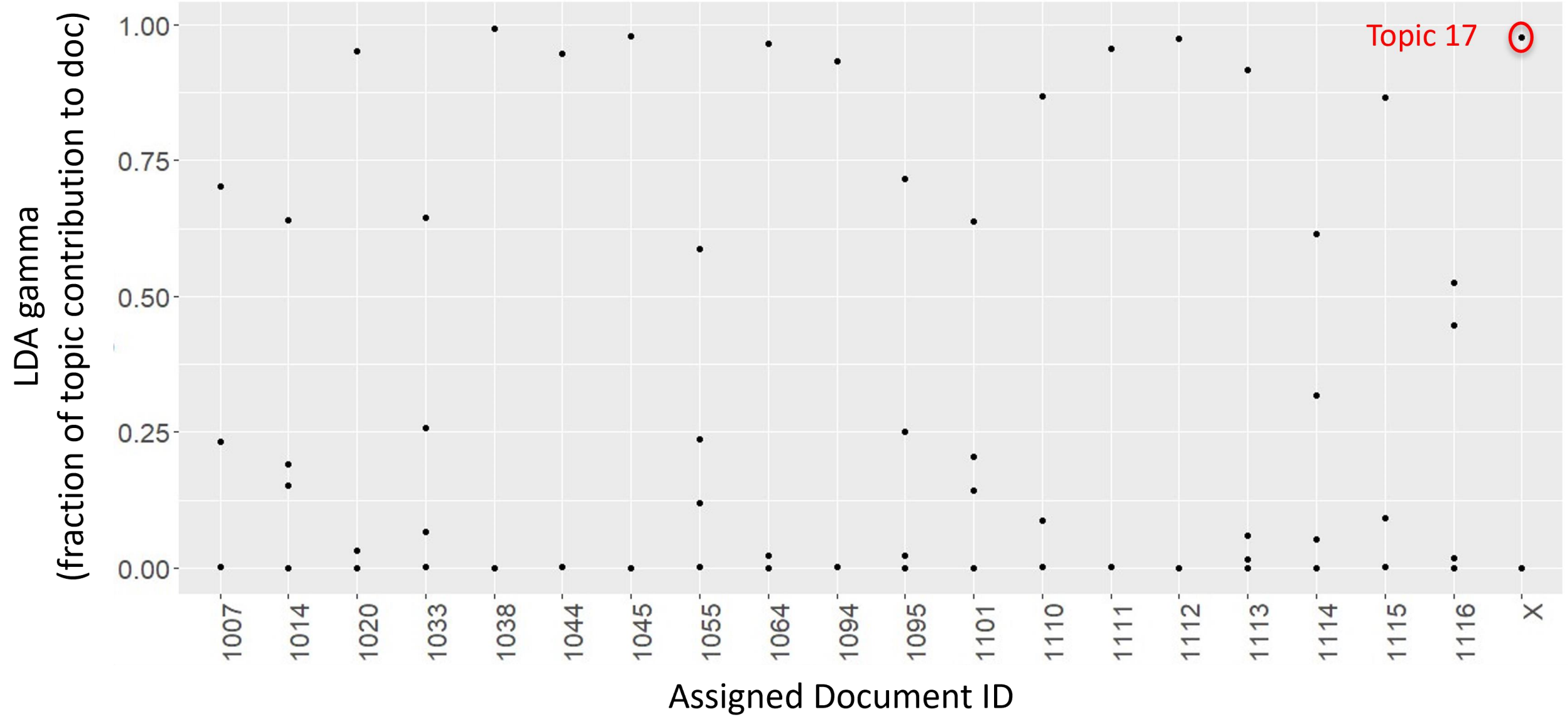
LSA is deterministic and therefore is reproducible by nature, with Single Value Decomposition as its underlying principle. The shortcoming is that it represents documents and words as abstract concepts, making it hard for human users to interpret.

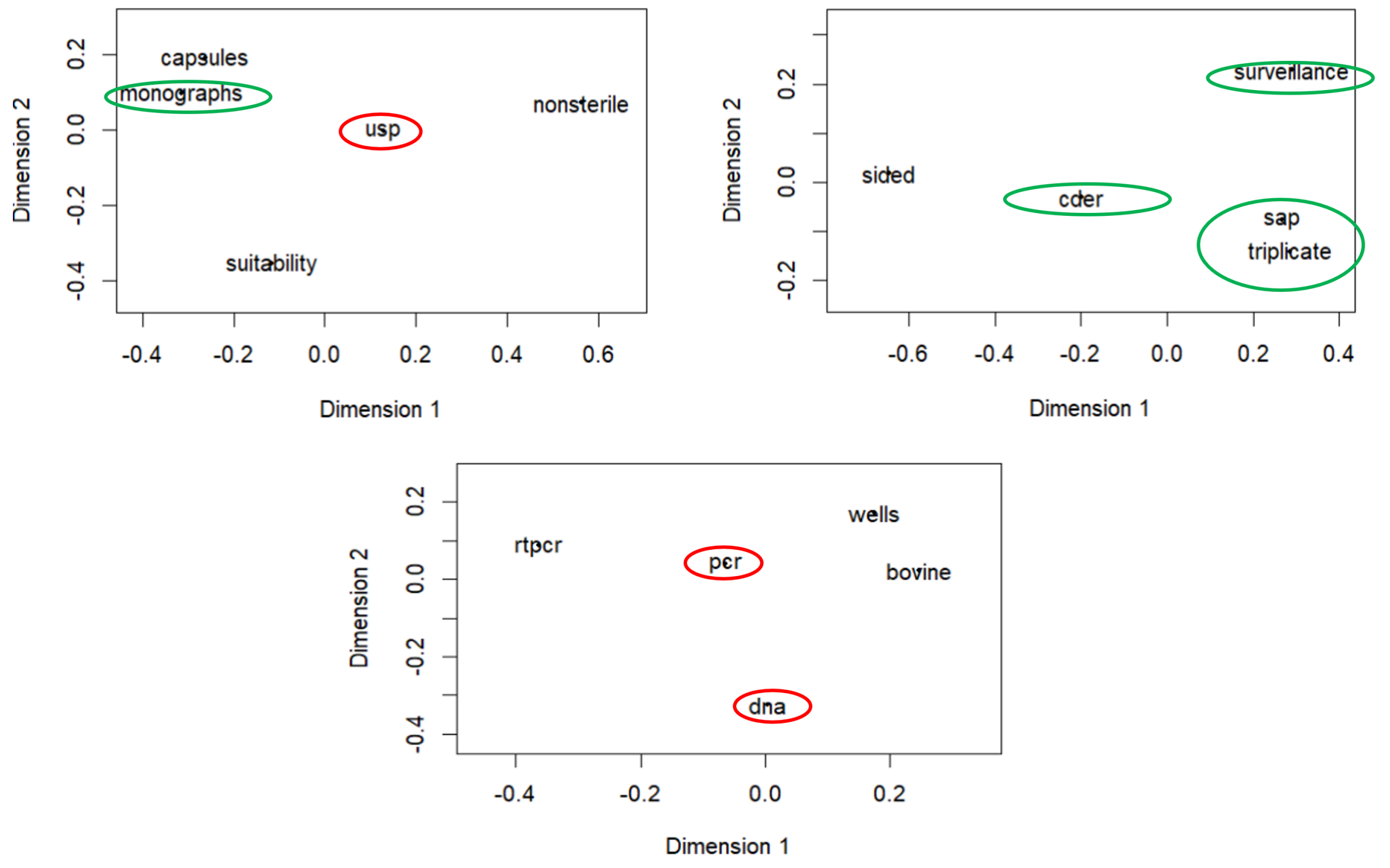
Results and discussion

LDA grouped documents into topics. The number of topics to model with LDA was set to 30. Each topic was characterized with a set of terms specifically associated with that topic.



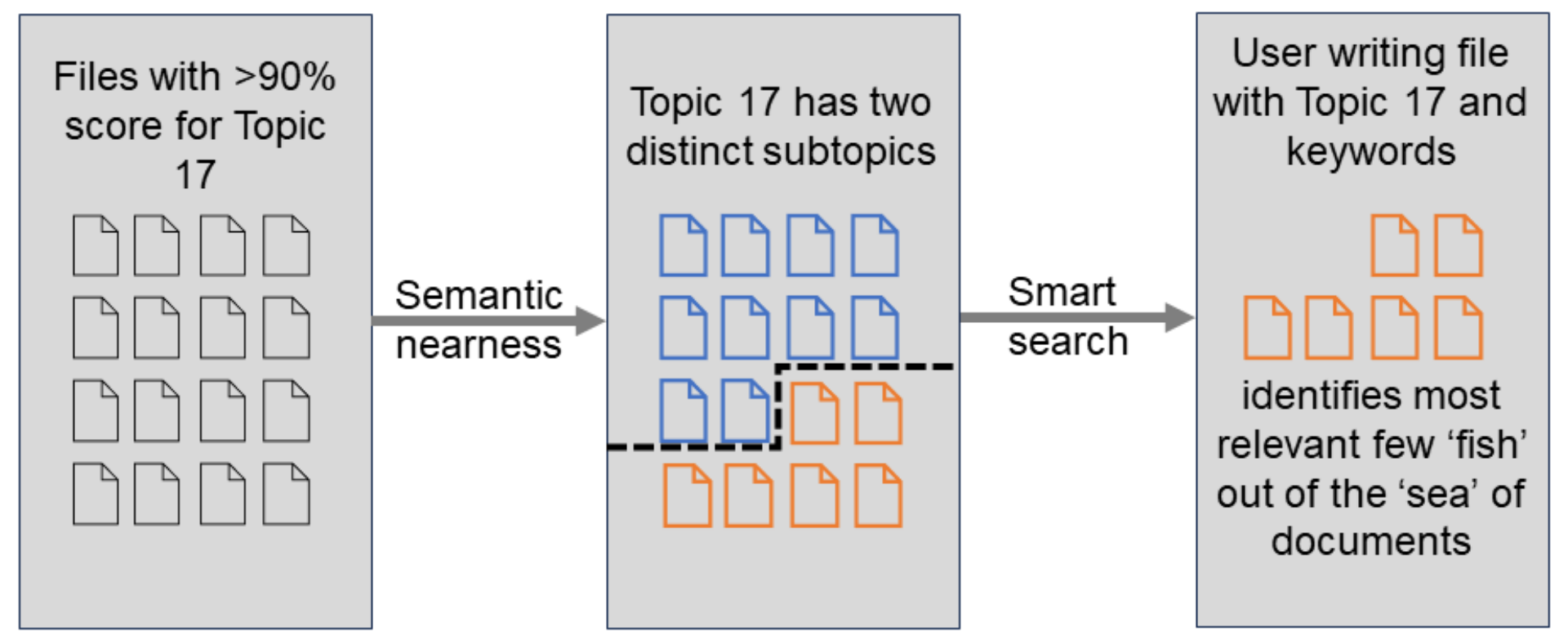
LDA calculates a factor by which each topic contributes to a particular document. While some documents can be characterized by only one topic, other documents' contents are described by multiple topics.





LSA was used to establish connections between top TF-IDF terms and descriptor terms of an LDA topic. For example, the top descripts of topic 17 are USP, PCR, and DNA. LSA shows that 'PCR' and 'DNA' are semantically connected while 'USP' is not, indicating two separate subtopics combined in topic 17.

The semantic nearness information provided by LSA can show that topics determined by LDA are two (or more) unrelated subtopics. Continuing with the example of topic 17, it can be split into its two subtopic components which contain different subsets of files.


Files with >90% score for Topic 17
Semantic nearness
Topic 17 has two distinct subtopics
Smart search
User writing file with Topic 17 and keywords
identifies most relevant few 'fish' out of the 'sea' of documents

Conclusion

Using NLP techniques for text analysis is akin to ingesting a whole book rather than simply judging it by the cover. This methodology offers a novel approach for the analysis of the short texts, which may enable efficient utilization of the past records as knowledge, and facilitates future evaluations of nonconforming work, risk assessment, prioritization, as well as creation and refinement of new documents. Leveraging NLP to group short text documents based on descriptive elements, the organization can gain more nuanced and holistic understanding of the issues, leading to effective solutions and process improvements. Improving corrective actions drives the lab enterprise toward greater efficiency in identifying and addressing public health risks.

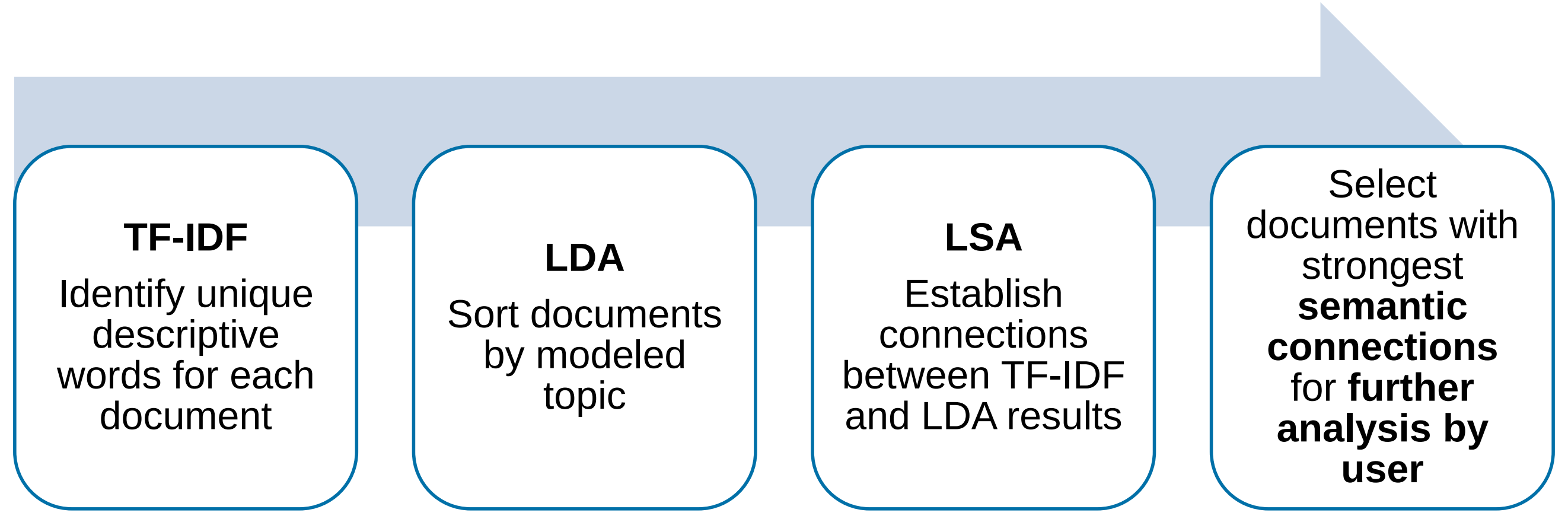
The resulting process can be summarized in few steps:

1. Perform TF-IDF analysis and identify words that characterize individual documents.

2. Model topics over the document space and group the documents based on topics, topics are identified by keywords specific to that topic.

3. Establish semantic connections between LDA and TF-IDF results.

4. Select documents for further analysis.



TF-IDF
Identify unique descriptive words for each document

LDA
Sort documents by modeled topic

LSA
Establish connections between TF-IDF and LDA results

Select documents with strongest **semantic connections** for **further analysis by user**

In order to best assist staff who are completing a new document for non-conforming work by identifying the most relevant existing documents for their reference, we would further develop code and proof-of-concept topic 17 case study into a user-friendly, code-automated app.

We have demonstrated the ability to digest and better-categorize the corpus of documents. This newly-tagged database can be transferred to management to aid in tracking and trending non-conforming work and its resolutions.

Disclaimer: The views expressed are those of the authors and should not be construed to represent the views or policies of the U.S. FDA Any reference to a specific commercial product, manufacturer, or otherwise, is for the information and convenience of the public and does not constitute an endorsement, recommendation or favoring by the U.S. FDA.

FDA | **SYMPOSIUM** | SCIENTIFIC COMPUTING + DIGITAL TRANSFORMATION | 2024