EVALUATING HIGHLY HETEROGENEOUS DOCUMENT COLLECTIONS

Arun S. Maiya, John P. Thompson, Francisco Loaiza-Lemos, and Robert M. Rolfe

This research effort involved the application of state-of-the-art computational approaches in machine learning and text mining to the discovery of critical information in unstructured collections of text. Unlike search (e.g., Google-like keyword searches), discovery finds information for which one may not have even known to look. The authors, in their article, presented an effective multi-faceted system for exploratory analysis of highly heterogeneous document collections. The IDA Text Analytics (ITA) capability, which incorporates these discovery technologies, is currently being used by the Department of Defense (DoD) to facilitate exploration and understanding of document sets across a number of different domains.

Given a large and diverse collection of unstructured text documents, how does one characterize the subject areas present and use these discovered subject areas to efficiently navigate the collection to locate critical information? Many previous works have investigated such questions within specific domains such as microblog posts or scientific abstracts, but comparatively less attention has been paid to investigating more general and diverse contexts. Unfortunately, in practice, approaches that may work well for domains consisting exclusively of a single document type (e.g., tweets, emails, or scientific abstracts) do not always translate easily or directly to other more heterogeneous and "messy" document collections.

Document collections of interest in our work exhibit a high degree of diversity in content and format. The U.S. government, for instance, is often presented with the challenge of what essentially is exploratory analysis of highly heterogeneous document collections.

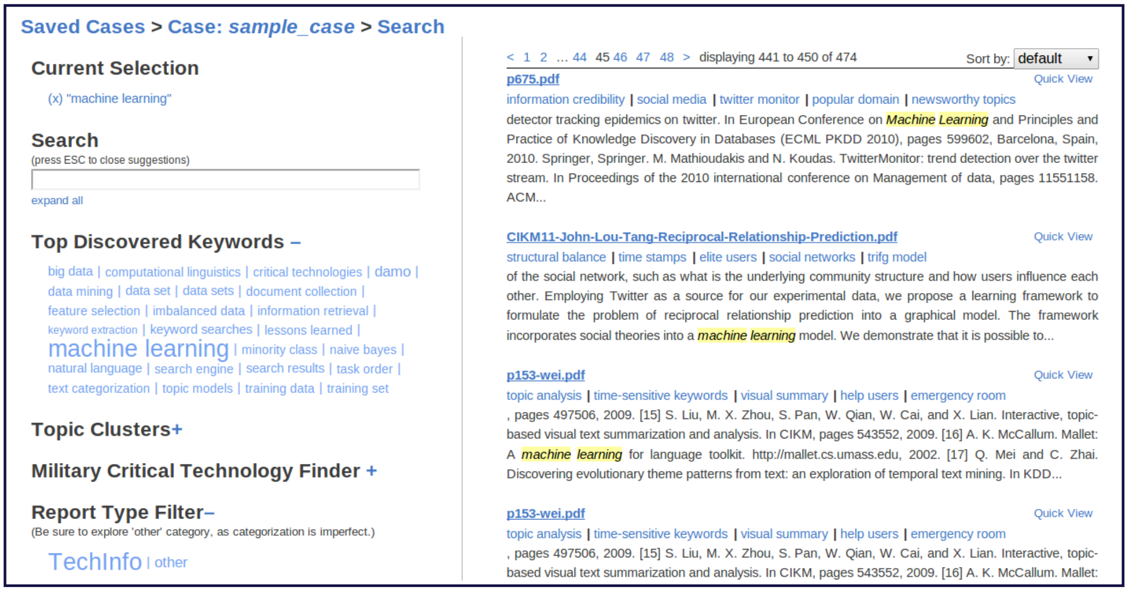
In this work, we present a tag-based system in which tags (i.e., terms or character strings automatically assigned to individual documents) are exploited to efficiently characterize and explore document collections. Document collections of interest in our work exhibit a high degree of diversity in content and format. The U.S. government, for instance, is often presented with the challenge of what essentially is exploratory analysis of highly heterogeneous document collections. Examples include digital investigations, intelligence analysis, and appraisal of electronic records. Our approach to this problem is to mine content from the documents themselves to auto-populate facets: classes of attributes describing objects in an information repository. Such facets can be used to navigate and discover information, as we now describe.

APPLICATION OVERVIEW

IDA Text Analytics (ITA) is a multi-faceted system for exploratory analysis of highly heterogeneous document sets. Although initially intended for facilitating the review of military-related technical reports, the system has been designed in a way that it serves as a general-purpose tool for search and discovery in arbitrary text collections.

Figure 1 shows a screenshot of one of the main interfaces. On the surface, it appears to be a standard search engine interface in which users can type ad hoc search queries and view search results. However, the standard search functionality is enhanced (on the left in Figure 1) with numerous facets. These facets are populated in an automated fashion by intelligently tagging each document in the collection along various dimensions. Most (but not all) of the facets take the form of tag clouds.

Figure 2 shows a sample tag cloud displaying topic-representative keywords discovered using Keyword Extraction for Reports and Articles (KERA), our unsupervised algorithm for key term extraction, which we describe later. This tag cloud facet can be viewed as a "lens" into document collections. The remaining facets can be viewed as controls used to point, zoom, and focus this "lens" to areas of high interest in the corpus. In actuality, each facet can play either the role of a "lens" or a "lens control." For instance, using other facets (not shown in the screenshot but described later), the search results can be filtered by folder location. Subsequently, the tag cloud shown in Figure 2 will dynamically regenerate to display the top discovered keywords



Saved Cases > Case: *sample_case* > Search

Current Selection

(x) "machine learning"

Search

(press ESC to close suggestions)

expand all

Top Discovered Keywords –

big data | computational linguistics | critical technologies | damo | data mining | data set | data sets | document collection | feature selection | imbalanced data | information retrieval | keyword extraction | keyword searches | lessons learned | machine learning | minority class | naive bayes | natural language | search engine | search results | task order | text categorization | topic models | training data | training set

Topic Clusters+

Military Critical Technology Finder +

Report Type Filter–

(Be sure to explore 'other' category, as categorization is imperfect.)

TechInfo | other

< 1 2 ... 44 45 46 47 48 > displaying 441 to 450 of 474 Sort by: default ▾

p675.pdf Quick View

information credibility | social media | twitter monitor | popular domain | newsworthy topics

detector tracking epidemics on twitter. In European Conference on *Machine Learning* and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2010), pages 599602, Barcelona, Spain, 2010. Springer, Springer. M. Mathioudakis and N. Koudas. TwitterMonitor: trend detection over the twitter stream. In Proceedings of the 2010 international conference on Management of data, pages 11551158. ACM...

CIKM11-John-Lou-Tang-Reciprocal-Relationship-Prediction.pdf Quick View

structural balance | time stamps | elite users | social networks | trifg model

of the social network, such as what is the underlying community structure and how users influence each other. Employing Twitter as a source for our experimental data, we propose a learning framework to formulate the problem of reciprocal relationship prediction into a graphical model. The framework incorporates social theories into a *machine learning* model. We demonstrate that it is possible to...

p153-wei.pdf Quick View

topic analysis | time-sensitive keywords | visual summary | help users | emergency room

, pages 497506, 2009. [15] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian. Interactive, topic-based visual text summarization and analysis. In CIKM, pages 543552, 2009. [16] A. K. McCallum. Mallet: A *machine learning* for language toolkit. http://mallet.cs.umass.edu, 2002. [17] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In KDD...

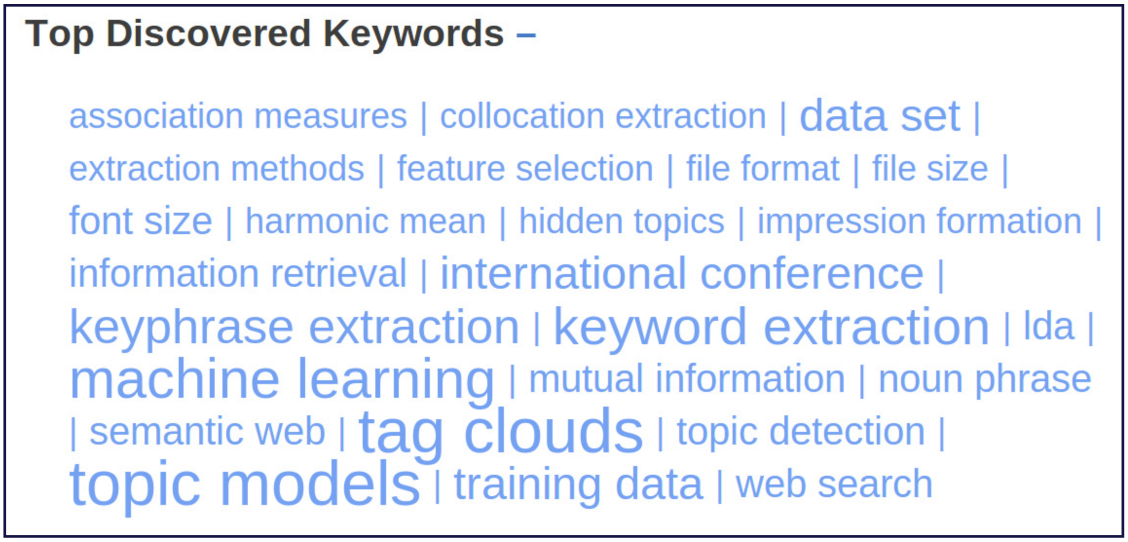
p153-wei.pdf Quick View

topic analysis | time-sensitive keywords | visual summary | help users | emergency room

, pages 497506, 2009. [15] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian. Interactive, topic-based visual text summarization and analysis. In CIKM, pages 543552, 2009. [16] A. K. McCallum. Mallet:

On the left, a rich set of information facets are provided for exploratory analysis. Only a subset (i.e., the Topic Facets) is viewable in this screenshot. The application also provides standard search engine functionality powered by Solr, as shown.

Figure 1. A Screenshot of Our System.



Top Discovered Keywords –

association measures | collocation extraction | data set | extraction methods | feature selection | file format | file size | font size | harmonic mean | hidden topics | impression formation | information retrieval | international conference | keyphrase extraction | keyword extraction | lda | machine learning | mutual information | noun phrase | semantic web | tag clouds | topic detection | topic models | training data | web search

Figure 2. A Tag Cloud generated by the KERA algorithm for 64 documents used as references in this paper.

of only the refined search results (i.e., documents residing in the folder selected). In this way, users can quickly "triage" noisy document collections for information of interest (in some cases, even before opening and reading documents).

In the following sections, we describe the five different categories of facets that we employ: (1) Topic Facets, (2) Mention Facets, (3) Format Facets, (4) Location Facets, (5) Time Facets, and (6) Author Facets.

TOPIC FACETS

Topic Facets are intended to help discover and characterize the subject areas present in a document collection. We employ three different types of *Topic Facets.*

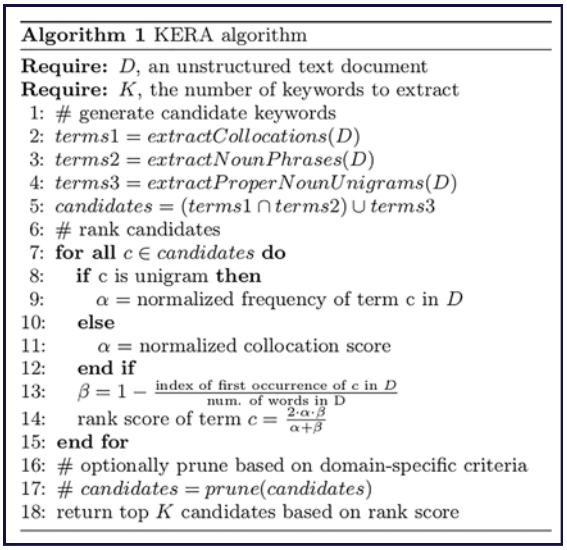
Automated Keyword Extraction

Our first approach to populating a Topic Facet is based on extracting topic-representative terms (i.e., keywords) from documents. (This is shown as *Top Discovered Keywords* in Figures 1 and 2). Here, we present KERA [Keyword Extraction for Reports and Articles], which is an unsupervised algorithm to extract keywords from individual text documents. KERA, at its core, is a descriptive model for keyword assignment based on observations of human-assigned keywords. The KERA algorithm, shown in Figure 3, comprises the following components:

- **Collocation extraction.** We first employ the use of collocation extraction to identify candidate key terms (shown in Line 2 of Figure 3). A collocation is "an expression consisting of two or more words that corresponds to some conventional way of saying things" (Manning and Schütze 1999). Using the log-likelihood ratio test, the collocation score for a bigram (i.e., two-word phrase) of words w^1 and w^2 is

$$2\sum n_{ij} \log \frac{n_{ij}}{m_j}$$

where n_{ij} are the observed



```
Algorithm 1 KERA algorithm
Require: $D$, an unstructured text document
Require: $K$, the number of keywords to extract
 1: # generate candidate keywords
 2: $terms1 = extractCollocations(D)$
 3: $terms2 = extractNounPhrases(D)$
 4: $terms3 = extractProperNounUnigrams(D)$
 5: $candidates = (terms1 \cap terms2) \cup terms3$
 6: # rank candidates
 7: for all $c \in candidates$ do
 8:     if c is unigram then
 9:         $\alpha$ = normalized frequency of term c in $D$
10:     else
11:         $\alpha$ = normalized collocation score
12:     end if
13:     $\beta = 1 - \frac{\text{index of first occurrence of c in } D}{\text{num. of words in D}}$
14:     rank score of term $c = \frac{2 \cdot \alpha \cdot \beta}{\alpha + \beta}$
15: end for
16: # optionally prune based on domain-specific criteria
17: # $candidates = prune(candidates)$
18: return top $K$ candidates based on rank score
```

Figure 3. The KERA Algorithm.

frequencies of the bigram from the contingency table for w^1 and w^2 and m_{ij} are the expected frequencies assuming that the bigram is independent (Dunning 1993; Manning and Schütze 1999).

- **Part-of-speech filtering.** Next, in Line 3 of Figure 3, we filter the set of collocations by removing terms that do not match the pattern (ADJECTIVE)*(NOUN)+, since expressive keywords tend to be noun phrases. Phrases greater than two terms are truncated. To this filtered set, we add extracted unigrams (i.e., one-word phrases) that are proper nouns (in Line 4) since we find such terms can be critical to the topic of documents. The criticality of these terms is especially true of government, scientific, and technical publications since proper nouns often refer to a system, algorithm, program, or initiative being described.

- **Ranking keywords.** Finally, we rank the extracted terms, as shown in Figure 3, and return the top K candidates. Our ranking methodology takes into account the position of terms within a document and the collocation score and term frequency. The final score is taken as the harmonic mean of these metrics. Before returning the final set, one might optionally prune the candidates based on domain-specific criteria. For instance, in our case, the set of proper noun unigrams can be pruned to contain only those unigrams that are upper case since it is those terms that often signify important technical systems and programs.

Topic Modeling and Clustering

A second Topic Facet that we employ is based on the concept of topic clusters. Topic modeling and clustering algorithms segment documents into different groups such that documents in the same group pertain largely to the same topic or theme. Whereas many clustering algorithms produce "hard" clusters or disjoint sets of documents, topic models typically produce "soft" or overlapping clusters. Topic models

[1] Currently, we set K = 5 or K = 10 for KERA.

[2] Other possible variations include discarding candidates when proper noun unigrams also appear as part of extracted bigrams, removing unigrams that do not first appear until later in the document, performing significance testing to filter the set of collocations, and setting α always as normalized frequency.

and clustering strategies may also tag clusters with topic-representative words. Latent Dirichlet Allocation (LDA) is the topic modeling algorithm that we currently employ for topic clustering (Blei, Ng, and Jordan 2003). Documents are assigned to a topic only if the topic proportion assigned by LDA is greater than 0.3, and documents are tagged using the top 10 LDA-derived topic tags. The facet populated by LDA is labeled "Topic Clusters" and appears as a menu showing the list of discovered topics (see Figure 1).

Document Classifier Facets

All of the *Topic Facets* discussed thus far (including topic models) have focused on identifying trends and hotspots within the topic collection. That is, they are not well suited to finding "needles in haystacks." A document pertaining to a lone topic of high interest to a particular user may not be identifiable in the presence of large topic clusters displayed in a tag cloud or other interface. To address this issue, we supplement the facets populated by KERA and LDA with two additional tag cloud facets populated by supervised document classification, which we now describe.

- **Military Critical Technology Finder**. This facet, shown in Figure 1, is populated using a set of binary-supervised machine-learning classifiers. Each binary classifier is trained to identify documents pertaining to a particular critical technology, and each tag in the cloud represents the positive class of a classifier. For any individual document, if no binary classifier categorizes the document as positive, the document is assigned the tag "other," which also appears in the cloud. We use LinearSVM as our main learning algorithm for all classifiers.

- **Report Type Filter**. Using a very similar methodology to the one described previously, we developed an additional classifier to categorize documents based on report type. That is, documents are categorized into one of four categories: Technical Information (e.g., a research paper), Test Information (e.g., a test plan for a system), Programmatic Information (e.g., details of a program for development of a system), and Other (i.e., everything else).

For more technical details on the development of document classifiers in this domain, one can refer to our earlier work in Maiya, Loaiza-Lemos, and Rolfe (2012).

MENTION FACETS

Users sometimes may be interested in locating documents not by topic but by mentions of particular entities, terms, or expressions of interest (e.g., Internet Protocol (IP) addresses, company names). To address this issue, we employ the use of a *Mention Facet*, which allows users to upload a plain text file containing expressions of interest. These expressions can currently take the form of simple lists of terms, gazetteers (i.e., entity dictionaries), or regular expressions for patterns of interest (e.g., a social security number). The results are displayed as either a tag cloud or menu, where the items are either explicit terms with matches in the document collection or high-level categories described by expressions (e.g., tagging documents containing social security numbers with "PII" [Personally Identifiable Information]). We currently employ such mention facets to navigate and search document sets based on sensitive markings (e.g., "For Official Use Only") and technical entities of particular interest to specific users.

FORMAT, LOCATION, TIME, AND AUTHOR FACETS

Our final set of facets is populated through direct extraction from document metadata. The *Format Facet* (labeled "File Types") is populated by tagging documents based on file type (e.g., .pdf, .doc, .ppt, .txt). The *Location Facet* (labeled "Folders") is populated by tagging each document with the directories in its file path. The *Time Facet* (labeled "Date" in our application) is populated by extracting the *Last-Modified* time from documents. Finally, the *Author Facet* is populated using the *Last-Author* or *Author* name (when available).

The *Location Facet* is displayed as a menu listing the most populous folders, and the Time Facet is displayed as a calendar widget. All other facets are displayed as tag clouds. (Note that none of these facets can be viewed in Figure 1.)

CASE STUDIES

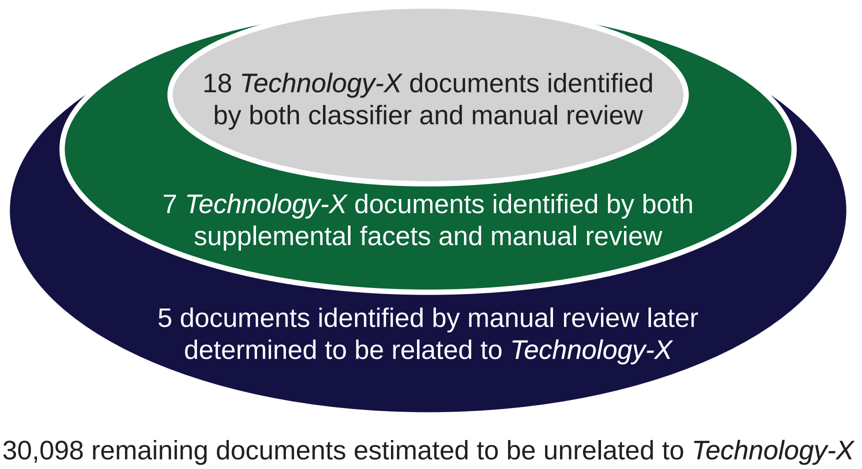
We now briefly describe two case studies that were conducted to more thoroughly evaluate our system. Although our system can be used for many purposes, we focus our evaluation on the current application of interest to our sponsors: locating information pertaining to military critical technologies within heterogeneous document collections. For reasons of sensitivity, we have redacted information, as necessary.

Case Study 1: Search

The search involves the task of finding information pertaining to a particular military critical technology within a document collection. We consider a particular technology of high interest to our sponsors (referred to as *Technology-X*) and assess how well the supervised approaches in our application are able to locate this critical information. A case that contained 30,128 files acquired from workstation hard drives of roughly 11 users was provided to us.

The files spanned numerous file formats including Microsoft Office, HyperText Markup Language (HTML), Portable Document Format (PDF), and plain text. We built machine-learning classifiers and custom-mention searches for *Technology-X*, as described previously. We evaluated these approaches and compared the results to those obtained from a manual review of the case by two analysts using their existing methodology (i.e., ad hoc keyword searches only). Results are shown in Figure 4 as a Venn diagram.

We observed significant time savings in this case study. The two analysts took roughly 7 hours (or 14 person-hours) to locate *Technology-X* documents. By contrast, the classifier identified 18 of the 25 files in mere seconds. The remaining files (i.e., all the seven false negatives) were located in less than 30 minutes using the *Mention Search, Report Type Filter*, and *Top Folders* facets in our application. We attribute most false


18 *Technology-X* documents identified by both classifier and manual review
7 *Technology-X* documents identified by both supplemental facets and manual review
5 documents identified by manual review later determined to be related to *Technology-X*
30,098 remaining documents estimated to be unrelated to *Technology-X*

The top (innermost) oval shows documents identified by classifier. The middle oval shows documents identified by other facets. The bottom (outermost) oval shows documents identified via a manual review by two analysts.

Figure 4. Venn Diagram of Search Results.

negatives committed by the classifier to the fact that the positive examples available to us at the time were limited. Given this finding and the breadth and depth of military critical technology information, unsupervised topic discovery is of high importance to this domain, which we evaluate next.

Case Study 2: Discovery

Discovery involves browsing document collections and allows users to locate information for which they did not even know to look. A framework to facilitate discovery can also clearly facilitate a search for something specific. Due to logistical and policy-related issues, we were not able to evaluate discovery on the case described in the previous case study, Case Study 1: Search. Instead, we were provided a new case to evaluate, which contained 39,515 files. Unlike the previous case study, we did not have any approximation of ground truth since the case had not been formally reviewed. Here, we assess the knowledge discovered and summarize lessons learned from execution of our application on this case.

Table 1 shows the two topics pertaining to military critical technologies discovered by our application (referred to as *Technology-Y* and *Technology-Z*). During the search,

Table 1. Critical Topics Found and Effective Usage Patterns

Topic	Documents	Facets Employed
Technology-Y	232	Method A: Topic Clusters (LDA) → KERA Method B: Report Type Filter → KERA
Technology-Z	89	Method A: Topic Clusters (LDA) → KERA Method B: Report Type Filter → KERA Method C: Top Folders → KERA

89 documents were found that pertained to *Technology-Z* and 232 documents were found that pertained to *Technology-Y* (including duplicate files). Through a subsequent exhaustive manual review of the case, we estimate that no additional information on military critical technologies of interest was present on this case. Using our facet-based system, most documents for these two critical topics were identified in less than an hour (and in many cases only minutes). By contrast, domain experts informed us that cases of this size typically require hours or days of "analyst" analysis to produce similar results, which is consistent with our experience during the manual review. Also shown in Table 1 are facet combinations revealed to be effective on this task. For additional results and technical details for this study, one may refer to the full report of this work (Maiya et al. 2013).

Dr. Maiya is a Research Staff Member in IDA's Information Technology and Systems Division. He holds a Doctor of Philosophy in computer science from the University of Illinois at Chicago.

Mr. Thompson is a Research Associate in IDA's Information Technology and Systems Division. He holds Master of Science in physics from the University of Texas at Dallas.

Dr. Loaiza-Lemos is a Research Staff Member in IDA's Information Technology and Systems Division. He holds a Doctor of Philosophy in chemistry from Princeton University and a Juris Doctor from George Mason University School of Law.

Dr. Rolfe is a Research Staff Member in IDA's Information Technology and Systems Division. He holds a Doctor of Philosophy in physics from the University of California, Los Angeles.

The full article was published in *Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, August 2013.

Exploratory Analysis of Highly Heterogeneous Document Collections

http://dl.acm.org/citation.cfm?id=2487575.2488195

REFERENCES

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." Journal of Machine Learning Research 3 (4–5) (March): 993–1022.

Dunning, Ted. 1993. "Accurate Methods for the Statistics of Surprise and Coincidence." Computational Linguistics 19 (1) (March): 61–74.

Maiya, Arun S., John P. Thompson, Francisco L. Lemos, and Robert M. Rolfe. 2013. "Exploratory Analysis of Highly Heterogeneous Document Collections." In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, 1375–1383. New York, NY: ACM. http://dl.acm.org/citation.cfm?id=2488195.

Maiya, Arun S., Francisco Loaiza-Lemos, and Robert M. Rolfe. 2012. "Supervised Learning in the Wild: Text Classification for Critical Technologies." In Proceedings of the Military Communications Conference, 2012 – MILCOM 2012, 1–6. Curran Associates, October. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6415660.

Manning, Christopher D., and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press.