TRAC-M-TR-17-008
January 2017

K-means Subject Matter Expert Refined Topic Model Methodology

Topic Model Estimation via K-Means

U.S. Army TRADOC Analysis Center-Monterey
700 Dyer Road, Room 176
Monterey, California 93943-0692

TRAC-M-TR-17-008
January 2017

K-means Subject Matter Expert Refined Topic Model Methodology

Topic Model Estimation via K-Means

Theodore T. Allen, Ph.D.
Zhenhuan Sui
Nathan Parker

This study cost the Department of Defense approximately $149,000 expended by TRAC in Fiscal Years 15-17.
Prepared on 20170103
TRAC Project Code # 060313

U.S. Army TRADOC Analysis Center-Monterey
700 Dyer Road, Room 176
Monterey, California 93943-0692

This page left intentionally blank.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE ***(DD-MM-YYYY)***	**2. REPORT TYPE**	**3. DATES COVERED** ***(From - To)***
	Technical Report	October 2015 – December 2016

4. TITLE AND SUBTITLE K-means Subject Matter Expert Refined Topic Model Methodology Topic Model Estimation via K-means	**5a. CONTRACT NUMBER** W9124N-15-P-0022 **5b. GRANT NUMBER** **5c. PROGRAM ELEMENT NUMBER**
6. AUTHOR(S) Theodore T. Allen, Zhenhuan Sui, Nathan Parker	**5d. PROJECT NUMBER** 060313 **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) The Ohio State University 1960 Kenny Road Columbus, OH 43210 TRADOC Analysis Center - Monterey 700 Dyer Road Monterey, CA 93940	**8. PERFORMING ORGANIZATION REPORT NUMBER** TRAC-M-TR-17-008
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) TRADOC Analysis Center – White Sands Missile Range (Forward)	**10. SPONSOR/MONITOR'S ACRONYM(S)** **11. SPONSOR/MONITOR'S REPORT NUMBER(S)** TRAC-M-TR-17-008

12. DISTRIBUTION / AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
We propose an innovative technique using K-means clustering to estimate the posterior topic distributions in Latent Dirichlet topic models as an alternative to the collapsed Gibbs sampling technique. This research also develops a topic modeling software instantiation of the K-means Subject Matter Expert Refined Topic methodology using the Visual Basic for Applications programming language. This topic modeling software is deployable across the majority of the Department of Defense computing environments and allows analysts to develop topic models using a graphical user interface.

15. SUBJECT TERMS
Text Analysis, Topic Models, K-means, clustering, LDA, Latent Dirichlet Allocation, SMERT, KSMERT, Subject Matter Expert Refined Topic

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Nathan Parker
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** ***(include area code)***
U	U	U	U	45	(831) 656-7580

This page left intentionally blank.

Acknowledgements

We thank the members of the Brigade Modernization Command at Fort Bliss, Texas for their perspectives regarding text analytics in a military specific environment. In addition, we extend a specific thanks to Varun Mohan for his assistance in computational work and code development. In total, the following people contributed through support or oversight other than the authors.

The Ohio State University
Jeremiah Lawson
Varun Mohan
Brian Mulh

U.S. Army Training and Doctrine Command (TRADOC) Analysis Center (TRAC)
Omar Gutierrez
Major Adam Haupt

U.S. Army Brigade Modernization Command
Colonel Roger Lemons

This page intentionally left blank.

Table of Contents

List of Figures

List of Tables

This page intentionally left blank.

Executive Summary

In this project, we develop a text analysis tool, employing a topic model methodology, for distribution to military analysts. Currently, analysts across the Department of Defense often encounter data sets containing vast amounts of unstructured free text documents. The majority of these analysts lack either the technical expertise or system availability to employ code-based native language processing and topic modeling software tools. Additionally, many of the current topic modeling methodologies function as a black box and do not allow the analyst employ their domain expertise during the development of the topic model.

The requirement to deliver a text analysis tool that is easily deployable across the Department of Defense computing environment limits the programming languages available for software development. Since the Microsoft Office Suite, including Excel, is ubiquitous across the DoD computing environment the Visual Basic for Applications (VBA) programming language presents the option as our programming language of choice.

We propose an innovative topic modeling technique to overcome the limitations of the Visual Basic for Applications language, while still delivering a software solution capable of supporting the analysis of large datasets. Specifically, we propose a topic modeling methodology using K-means clustering to estimate the posterior probabilities of the topic distributions within the Latent Dirichlet Allocation family of topic models. This estimation method replaces the collapsed Gibb sampling estimation technique currently in use in most Latent Dirichlet Allocation topic models. The K-means clustering estimation method produces results with similar or better accuracy versus the collapsed Gibbs sampling method while significantly reducing the computational time. We then incorporate K-means Latent Dirichlet Allocation topic models into the Subject Matter Expert Refined Topic methodology to arrive at K-means Subject Matter Expert Refined Topic methodology.

We also develop a software instantiation of the K-means Subject Matter Expert Refined Topic methodology that is available both as a standalone Excel spreadsheet and as an Excel add-in. This software allows analysts without a coding background, or access to other computational programming environments, to build topic models from free text datasets using a familiar Excel based graphic user interface.

This page left intentionally blank.

Chapter 1. Introduction

Our primary objective is the development of a text analysis tool, employing a topic model methodology, for distribution to military analysts. This technical report documents the key elements of the K-means Subject Matter Expert Refined Topic (KSMERT) methodology and related software based tool. Chapter 1 describes the project background, use cases, and applications. Chapter 2 covers the major innovation for this project, which is the development of a topic model estimation method using K-means clustering. Chapter 3 describes the development of a software instantiation of the KSMERT methodology suitable for deployment across the Department of Defense (DoD) computing environment. Finally, Chapter 4 communicates our conclusions and recommendations for future research.

Background

Analysts across the DoD often encounter data sets containing vast amounts of unstructured free text documents. The majority of these analysts lack either the technical expertise or system availability to employ code-based native language processing and topic modeling software tools. Additionally, many of the current topic modeling methodologies function as a black box and do not allow the analyst employ their domain expertise during the development of the topic model. With these challenges in mind, the objective of this project is the development of a text analysis tool, employing a topic model methodology, which is: easily deployable across the majority of DoD computing systems, allows analysts to incorporate their domain knowledge into the topic model development, and does not require analysts to have a specific coding background.

We build on the foundational work by Allen, Xiong, and Afful-Dadzie (2015) that proposes Subject Matter Expert Refined Topic (SMERT) models. This includes a method to incorporate SME domain specific knowledge as an intermediate step in fitting Latent Dirichlet Allocation (LDA) topic models. This type of refinement is often necessary so that the topics align with pre-existing classifications, context specific needs, and have dispersion across the observation space of interest.

Use Cases and Applications

We use the Network Integration Exercise (NIE) events run by the Brigade Modernization Command (BMC) at Fort Bliss, Texas as the primary use case. The BMC, with analytic support from TRADOC Analysis Center – White Sands Missile Range (TRAC-WSMR), is responsible for assessing the viability and maturity of multiple systems during each NIE event. One of the primary data collection methods for these events is direct observation reports written by observer controls and observer analysts that each contain multiple free text data fields. This use case provides direct parallels to the larger problem scope in that analysts of varied technical acumen must analyze a body of unstructured, free text data in an environment the restricts access to other software based text analytic tools. Opportunities to deploy developmental versions of the methodology and software provide extensions to the base use case and further demonstrate the need for a deployable topic modeling tool.

These use cases only capture a small subset of the potential applications for both the underlying methodology and deployable software that this research seeks to produce. Multiple communities across the DoD face the challenge of rapidly analyzing large corpora of free text data. The potential application include human intelligence and signals intelligence reports in the intelligence community, network and administer logs in the cyber community, and free text survey responses in the training and doctrine community.

Chapter 2. A Topic Model Estimation Method Based on K-Means Clustering

Overview

Latent Dirichlet Allocation is a clustering method widely used for creating interpretable topics from text corpora. SMERT models are a generalization of LDA, invented by our researchers, that permits SMEs to edit and improve the topic definitions. Unfortunately, the current methods for fitting LDA models, collapsed Gibbs sampling and variational inference estimation, lack repeatability and are computationally expensive.

To address these limitations, we propose an innovative topic modeling technique that uses K-means clustering to fit LDA models. We then combine this novel LDA technique with the previously developed SMERT model to arrive at our KSMERT model. This method is able to take advantage of users' knowledge in directing data manipulations to achieve much more accurate and meaningful results within a reasonable duration, despite the high computational loads when handling large text data sets. We illustrate the methodology using four small case study examples and conclude that KSMERT offers desirable accuracy and computational cost tradeoffs with wide applicability in military and other contexts.

Literature Review

Text analytics is the process of analyzing unstructured text, extracting relevant information and transforming it into a structured form for further use in analytic process (Packiam and Prakash, 2015). For example, text analytics could aid in systematic summarization of field reports, interviews of relevant leaders, and insights from analysts. Here, we focus on one type of text analytics called topic modeling (e.g., see Blei, Ng, and Jordan, 2003). By dividing the corpus into clusters or topics, these models can clarify what is missing and what is present in the entire corpus. The identification of the latent structure of the corpus also allows topic models to inform information retrieval reminiscent of a "Google" search but for unstructured, unindexed, and untagged data sets.

Blei, Ng, and Jordan (2003) propose the use of mean field variational inference methods to estimate the parameters of LDA based topic models. Teh (2007) criticizes the accuracy of mean field variational inference and argues that unbiased collapsed Gibbs sampling is comparably computationally efficient with improved accuracy. Blei, Ng and Jordan (2003) also argue that topic models are qualitatively more relevant than non-generative clustering models, including K-means. We postulate that they overlook the potential of transforming clustering results to create generative models and propose a K-means based estimation method of the model parameters as a computationally faster and reproducible method for fitting LDA models.

Topic model methods that incorporate expert knowledge are a topic of ongoing research. For example, Zhao, Li, Li, Wang, Ding, and Li (2012) propose two supervised topic models devised by tracking previous history text data as background knowledge using on interaction matrix. Sun (2014) proposes a term frequency-inverse document frequency model, which is a keyword model

reflecting how important a word is in a document. However, these models lack semantic structure, especially for multiple probabilistic distributions over the vocabulary. For this reason, the LDA method Blei, Ng, and Jordan (2003) propose is the more widely accepted method for clustering unsupervised images or text documents.

In our work, we focus on the method of incorporating expert knowledge proposed in Allen, Xiong, and Afful-Dadzie (2015). They introduce the SMERT model to permit analysts to incorporate their domain specific knowledge to edit the topics while maintaining the LDA topic model structure. Sui, Milam, and Allen (2015) shows that SMERT could estimate the proportion of words in the overall corpus on each topic and incorporate "high-level" inputs from a SME to adjust the topics by confirming or denying the membership of words in the topic definitions. We incorporate our K-means clustering method for fitting LDA models into the SMERT model to arrive at our KSMERT model.

Review of Data Preparation and Topic Models

This section reviews the preparation of text data for clustering and information retrieval and describes the likelihood that defines both SMERT and LDA. Estimating the parameters in the likelihood is the objective in the next section.

Preparing Data for Text Modeling

Feldman and Sanger (2006) propose a generalized view of text mining system architectures composed of four main phases: preprocessing tasks, core mining operations, presentation layer components and browsing functionality, and refinement techniques. The K-means estimation technique we propose in this research follows these same four phases shown in Figure 2.1.

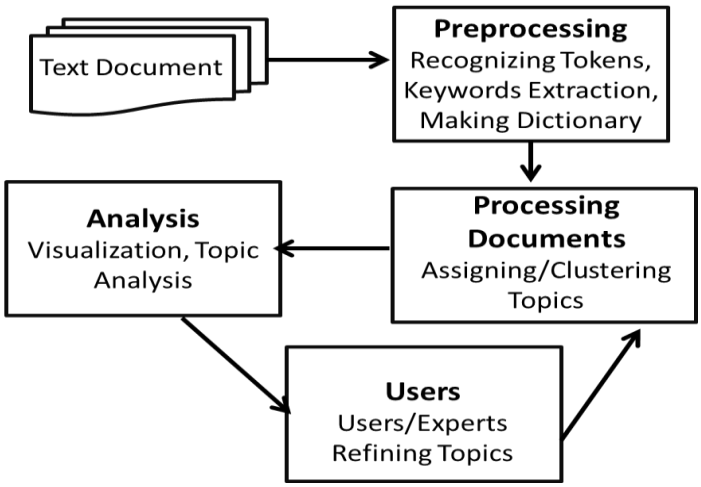


Figure 2.1. Four phase method for text analysis adapted from Feldman and Sanger (2005).

Phases 1. Preprocessing Tasks include all the preparation of raw data for text mining core operations. These preparations include cleaning the data source to convert it into a canonical

format. In our methods, this processing includes recognizing tokens or words based on space, trimming words into word roots, making a dictionary and assigning a corresponding number index for each word. Here, we use the algorithm from Porter (1980).

Phases 2. Core Mining Operations are the core operations in text analytics including pattern discovery, trend analysis, and modeling algorithms. In our method, processing documents is the core part with pattern discovery through word frequency. The main goal of our algorithm is the clustering and grouping of topics and words. Here, we fit the LDA and K-means model forms for mining.

Phases 3. Presentation Layer Components include browsing functionality and visualization tools. In this phase, the primary visualization tool to view different topics and their contents is the Pareto chart.

Phases 4. Refinement Techniques include methods to filter information through pruning, generalizing or suppressing approaches to achieve discovery optimization. Based on Allen, Xiong, and Afful-Dadzie (2015), in this phase our method incorporates users' human domain knowledge and enables analysts to directly supervise model results and refine the results while keeping the model simple.

Topic Model Notation

We use a topic model notation that closely follows notation that Blei, Ng, and Jordan (2003) propose for their LDA models. In our notation, $w_{d,n}$ is the n^{th} word in document $d = 1,\ldots,D$ and $n = 1,\ldots,N_d$. Therefore, "D" is the number of documents and "N_d" is the number of words in the d^{th} document. The number of words created in the dictionary is denoted as *"WC"*. The *WC*-dimensional random vector $\emptyset_t$ represents the probability that randomly selected words are assigned to each pixel in the topic indexed by $t = 1\ldots T$.

The posterior mean of ϕ_t defines the topics. The prior parameters α and β are usually scalars in that all documents and all words are initially treated equally. Generally, low values, or "diffuse priors", are assigned so that only a small amount of shrinkage is applied and adjustments are made on a case-by-case basis (Griffiths and Stuyvers, 2004). The sampled topic assignments ($z_{d,n}$) permit estimation of the topic definitions (β). The most common words in each topic are often the most relevant outcome. The *T*-dimensional random vector θ_d represents the probability that a randomly selected word in document *d* is assigned to each of the *T* topics or clusters.

Subject Matter Expert Refined Topic (SMERT) and Latent Dirichlet Allocation (LDA)

With the parameter definitions above, SMERT joint distribution or likelihood that defines the initial SMERT model is simply the product of the individual conditional densities:

$$
\begin{aligned}
&P(z,w,x,\theta,\phi|N,N_t,\alpha,\beta) \\
&\quad=\left[\prod_{t=1}^{T}P(\phi_t|\beta_t)\right]\left[\prod_{d=1}^{D}P(\theta_d|\alpha_d)\right] \\
&\quad\times\left[\prod_{d=1}^{D}\prod_{j=1}^{N_d}P\left(z_{d,j}\middle|\theta_d\right)P\left(w_{d,j}\middle|\phi_{z_{d,j}}\right)\right] \\
&\quad\times\left[\prod_{t=1}^{T}\prod_{c=1}^{WC}P\left(x_{t,c}\middle|N_{t,c},\phi_{t,c}\right)\right]
\end{aligned}
\tag{2.1}
$$

where w and x are matrices of the data and θ_d and ϕ_t are vector model parameters to be estimated. Specifically, ϕ_t has elements $\phi_{t,c}$. The vectors, α and β_t, contain prior parameters that might be assumed to have all their elements equaling the same values, α_0 and β_0. Also, effective constants include N the vector of document lengths and N_t the matrix of trial counts.

The constituent parts of equation (2.1) are the Dirichlet, categorical, and binomial densities. Collecting all the parts, equation (2.1) becomes:

$$
\begin{aligned}
&P(z,w,x,\theta,\phi|N,N_t,\alpha,\beta)= \\
&=\left[\prod_{t=1}^{T}\frac{\Gamma\left(\sum_{c=1}^{WC}\beta_{t,c}\right)}{\prod_{c=1}^{WC}\Gamma\left(\beta_{t,c}\right)}\prod_{c=1}^{WC}\phi_{t,c}^{\beta_{t,c}-1}\right]\left[\prod_{d=1}^{D}\frac{1}{B(\alpha_d)}\prod_{t=1}^{T}\theta_{d,t}^{\alpha_{d,t}-1}\right] \\
&\times\left[\prod_{d=1}^{D}\prod_{t=1}^{T}\theta_{d,t}^{n_{d,(.)}^{t}}\right]\times\left[\prod_{t=1}^{T}\prod_{c=1}^{WC}\phi_{t,c}^{n_{(.),c}^{t}}\right] \\
&\times\left[\prod_{t=1}^{T}\prod_{c=1}^{WC}\binom{N_{t,c}}{x_{t,c}}\phi_{t,c}^{x_{t,c}}\left(1-\phi_{t,c}\right)^{N_{t,c}-x_{t,c}}\right]
\end{aligned}
\tag{2.2}
$$

where

$$
\begin{aligned}
n_{d,(.)}^{t}&=\textstyle\sum_{j=1}^{N_d}\sum_{c=1}^{WC}I\left(z_{d,j}=t\ \&\ w_{d,j}=c\right), \\
n_{(.),c}^{t}&=\textstyle\sum_{d=1}^{D}\sum_{j=1}^{N_d}I\left(z_{d,j}=t\ \&\ w_{d,j}=c\right),
\end{aligned}
\tag{2.3}
$$

The left two rectangles in the graphical model representation of Figure 2.2 show the conditional relationships between the variables in the LDA model. In the figure, the rectangles indicate the number of elements in each random vector or matrix. For example, the matrices *z* and *w* have N_d elements for each of the *D* documents. LDA has N_d equal to zero for all *d*.

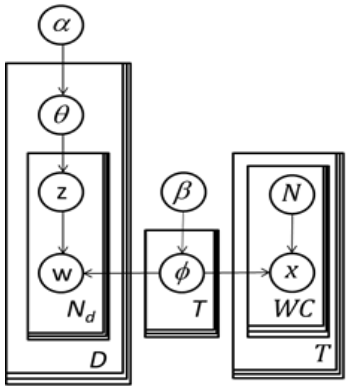


Figure 2.2. Graphical model representation of SMERT.

The posterior mean values for the topic definitions ϕ and topic proportions θ are estimated through a single replicate of the topic assignments after some level of convergence (Blei, Ng, and Jordan, 2003). These posterior means give a conceptual map of the corpus because the words with highest probabilities in the topic definitions offer the most meaningful cluster definitions. The proportions for each topic summarize the corpus and prioritize later visualization parts (Steyvers and Griffiths, 2007).

Techniques for Estimation of the Posterior Probabilities

The main computational challenge for topic modeling in LDA is the approximate estimation for the posterior probabilities. The two previously proposed estimation methods are collapsed Gibbs sampling (Teh, 2007 and Steyvers and Griffiths, 2007) and mean field variational inference (Blei et al., 2003).

Collapsed Gibbs Sampling Methods

Allen, Xiong, and Afful-Dadzie (2015) fit the distribution in SMERT using collapsed Gibbs sampling which is a type of Markov Chain Monte Carlo process. Collapsed Gibbs sampling is an iterative process of modifying the topic assignments and distributions. The topic assignments converge to the samples from the new distribution and are then used for estimations for the topics and proportions.

A major component of collapsed Gibbs sampling is the topic selection $\mathbf{Z}_{(m,n)}$ for n^{th} word in the m^{th} document. Let v be the word and the topic assignments for other words be $\mathbf{Z}_{-(m,n)}$. Let $q^{i}_{j,r}$ br the counts of the number of samples of topic i in document j of the r^{th} word. We use $(.)$ to denote the sum over counts. Then each sampling is calculated as:

$$P(Z_{(m,n)} = k | Z_{-(m,n)}; \alpha, \beta) = \frac{(q^{k,-(m,n)}_{m,(.)} + \alpha)\frac{(q^{k,-(m,n)}_{(.),v} + \beta)}{\sum_{r=1}^{W}(q^{k,-(m,n)}_{(.),r} + \beta)}}{\sum_{t=1}^{T}(q^{t,-(m,n)}_{m,(.)} + \alpha)\frac{(q^{t,-(m,n)}_{(.),v} + \beta)}{\sum_{r=1}^{W}(q^{t,-(m,n)}_{(.),r} + \beta)}} \text{ for } k = 1,\ldots,T \qquad (2.4)$$

$\mathbf{Z}_{(m,n)}$ is therefore from a single multinomial draw and then the iteration moves to the next word.

Mean Field Variational Inference

Blei et al. (2003) proposes the variational inference estimation method to approximate an intractable posterior distribution over hidden variables with a much simpler one with free variational parameters. Topics β_k is described by a V-Dirichlet distribution λ_k. Topic Proportion θ_d is described by a K-Dirichlet distribution γ_d. Topic assignment $z_{d,n}$ is described by a K-multinomial distribution $\phi_{d,n}$. The main iteration is then:

Step 1. For each topic *k* and word *v*, $\lambda_{k,v}^{t+1} = \boldsymbol{\eta} + \sum_{d=1}^{D} \sum_{n=1}^{N} \mathbf{1}(\boldsymbol{w_{d,n}} = \boldsymbol{v})\, \boldsymbol{\phi_{n,k}^{t}}$ (2.5)

Step 2. For each document *d:*

(a) Update γ_d: $\gamma_{d,k}^{t+1} = \boldsymbol{\alpha_k} + \sum_{n=1}^{N} \boldsymbol{\phi_{d,n,k}^{t}}$ (2.6)

(b) For each word *n*, update $\phi_{d,n}$ $\boldsymbol{\phi_{d,n,k}^{t+1}} \propto \boldsymbol{exp}\{\boldsymbol{\Psi}(\gamma_{d,k}^{t+1}) + \boldsymbol{\Psi}(\lambda_{k,w_n}^{t+1}) - \boldsymbol{\Psi}(\sum_{v=1}^{V} \lambda_{k,v}^{t+1})$ (2.7)

where $\boldsymbol{\Psi}$ is the digamma function, the first derivative of log $\boldsymbol{\Gamma}$ function. The iterations are repeated until the minimization of Kullback-Leibler function for the variational parameters converges.

K-means based Subject Matter Expert Refined Topic (KSMERT)

In practice, not all of the distribution is relevant to the user nor will they find the expression of all topics as an ordered list of words interpretable. To address this shortcoming of LDA topic models Allen, Xiong, and Afful-Dadzie (2015) developments the SMERT method for probabilistic clustering of texts.

The third rectangle in Figure 2.2 is the additional part necessary for a SMERT model. The two left-hand-side rectangles are identical to LDA with multinomial response data, *w*. The right-hand-side begins with the arrow from *N* to *x*, which introduces binomially distributed response data, $x_{t,c}$ for $t = 1,\ldots,T$ and $c = 1,\ldots,$WC. $x_{t,c}$ represents the number of times in a given topic, *t*, word *c* is selected in $N_{t,c}$ trials. Note that the choice of $N_{t,c}$ in the model is arbitrary. Allen, Xiong, and Afful-Dadzie (2015) refers to the right-hand-side portions in Figure 2 as "hierarchal analysis designed latency experiments" (HANDLEs) because it permits users to interact with the model. "Hierarchical analysis" means the use of a hierarchical Bayesian formulation. "Designed" indicates that the users can incorporate their domain specific knowledge about the result to direct further data manipulations. "Latency experiments" indicates that the model has relatively high leverage on specific latent variables, i.e., ϕ. Figure 2.3 shows the flow chart of SMERT. After the initial LDA run, the user can study the results and then apply domain specific expertise as the high-level data to achieve the second stage result model 2.

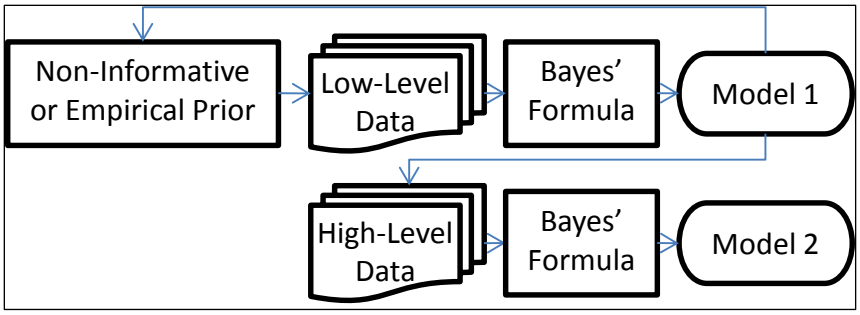


Figure 2.3. Flow chart of SMERT with iterative additions of high-level data.

The primary disadvantages of existing LDA methods is that they require multiple iterative runs making them computationally expensive with noisy and unrepeatable results. Our approach of a topic modeling technique using a K-means estimation method is capable of achieving much faster and repeatable results.

For each data point, the distance to the centroid of the belonging cluster is calculated. The membership function is then computed as the inverse of the distance. If the distance is zero, the membership is set as one. The *T* dimensional membership function vectors $\mathbf{u}_1,\ldots,\mathbf{u}_D$ can be explained as the probabilities that the data point belongs to the associated clusters which are topics, $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_D$. Next the membership functions are scaled

$$\alpha_t = \frac{\sum_d^D u_{d,t}}{\sum_{k=1}^T \sum_d^D u_{d,k}} \text{ for } j = 1,\ldots,T \tag{2.8}$$

as the topic definitions, which show the distribution of words in the topics. The cluster centroids are represented as $\mathbf{z}_1,\ldots,\mathbf{z}_T$, and the centroids are scaled using

$$\phi_{t,c} = \frac{z_{c,t}}{\sum_c^{WC} z_{c,t}} \text{ for } t = 1,\ldots,T \text{ for } c = 1,\ldots,WC. \tag{2.9}$$

as the topic proportions, which show the distribution of topics in all the document lists.

Numerical Studies

This section describes the four test problems and two evaluation metrics used to evaluate the accuracy of the K-means topic modeling methodology. Additionally, the results from the comparisons and related findings follow.

Test Problems

Here, four similar cases provide the ability to compare different estimation methods. Table 2.1 summarizes the computational results for the timing. Appendix C contains an example test problem and the corresponding true model topic distribution. For all test problems in this research we use corpora of 40 documents ($D = 40$) and a dictionary size of 25 words ($WC = 25$) for all case.

Evaluation Metrics

The estimated distribution for topics has no natural ordering so it is hard to compare the results against the assumed ground truths. Therefore, Steyvers and Griffiths (2007) propose the evaluation of each permutation of the cluster labels before selecting the permutation with the closest distance. Define the function $t'(\mathbf{r}, t)$ as the selection of topic *t* in permutation **r**. Use $\phi_{t,c}^{true}$ to denote the true topic definitions and α_t^{true} to denote the true topic proportions for *t* = 1,…,*T and* for *c* = 1,…,W.

The minimum average Kullback-Leibler divergence (KLD) for the topic definitions is:

$$KLD(\phi) = \min_{\mathbf{r}\in S} \frac{1}{T} \sum_{t=1}^{T} \sum_{c=1}^{WC} \phi_{t,c}^{true} ln\left(\frac{\phi_{t,c}^{true}}{\phi_{t'(\mathbf{r},t),c}}\right). \quad (2.10)$$

Further, denote **r*** as the argmax permutation for equation (2.10). Another measure of distance is the average root mean squared (RMS):

$$RMS(\phi) = \frac{1}{T} \sum_{t=1}^{T} \sqrt{\sum_{c=1}^{WC} \left(\phi_{t,c}^{true} - \phi_{t'(\mathbf{r}^*,t),c}\right)^2}. \quad (2.11)$$

The accuracy measures for the topic proportions are thus:

$$KLD(\alpha) = \min_{\mathbf{r}\in S} \sum_{t=1}^{t} \alpha_t^{true} ln\left(\frac{\alpha_t^{true}}{\alpha_{t'(\mathbf{r},t)}}\right) \quad (2.12)$$

and

$$RMS(\alpha) = \sqrt{\sum_{t=1}^{T} \left(\alpha_t^{true} - \alpha_{t'(\mathbf{r}^*,t)}\right)^2}. \quad (2.13)$$

Comparison of Results

The next two sections show how we compare the performance of K-means LDA, as the initial step to KSMERT, to the existing methodologies. The two measures of performance for this comparison are model accuracy and computational time.

Model Accuracy

Figure 2.4 displays the results of K-means LDA, Gibbs Sampling LDA with 10, 100, and 1000 iterations, and variational inference LDA against the true models for all the four cases. Using the RMS metric, K-means LDA could achieve a similar level of distance or even smaller distance to the true model compared with other models. For Gibbs sampling, Monte Carlo simulation introduces uncertainties. A higher number of iterations produces a slightly better RMS than lower numbers, but the trend is highly influenced by the random seed.

Figure 2.5 shows the comparison of the various LDA and associated SMERT/KSMERT implementations for Case 4 only. The accuracy results for all of the SMERT/KSMERT implantations are similar due to the user employing domain knowledge to direct data manipulations.

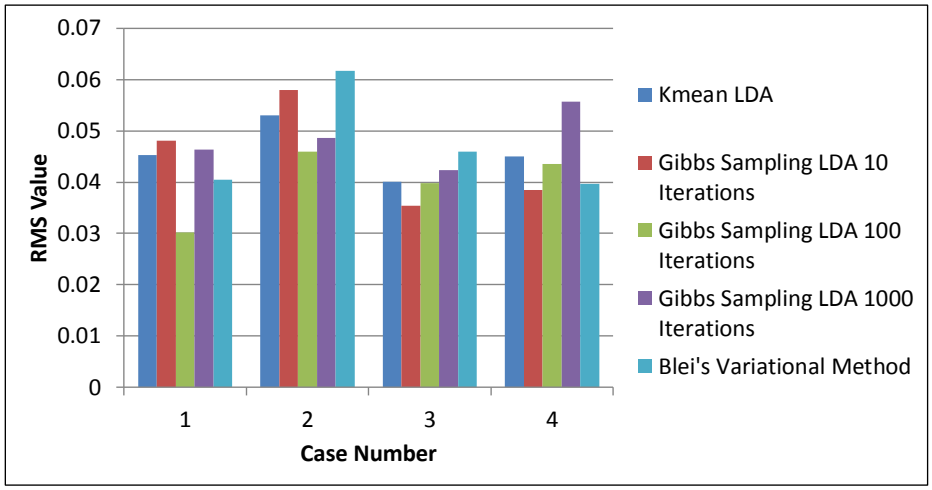


Figure 2.4. RMS comparison for different estimation methods for LDA only.

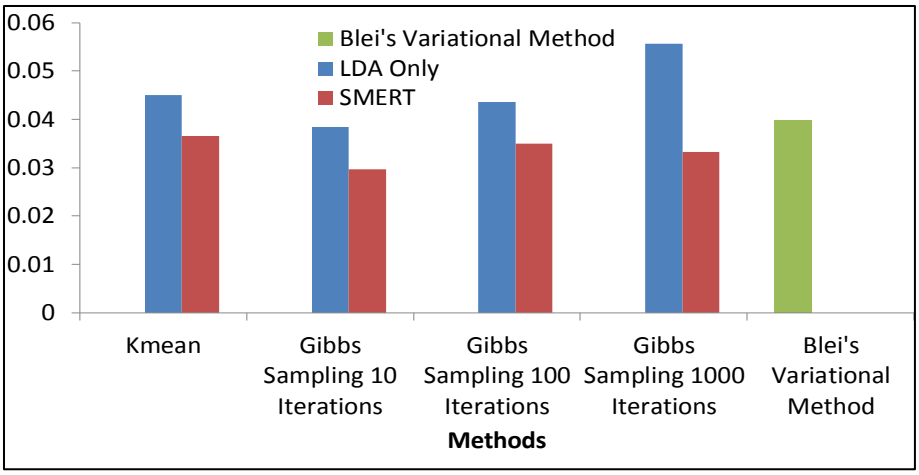


Figure 2.5. RMS comparison for different estimation methods for LDA and SMERT (Case 4).

Computational Time

The running time in minutes for Collapsed Gibbs sampling LDA is roughly predicted using:

$$\text{Run time(collapsed Gibbs)} = \frac{-1899.1766 + 0.2736 * numberDocuments + 17.9817 * numberFields + 38.5079 * TopicNumber.Value + 0.9342 * MaxNumberIteration.Value}{60} \quad (2.14)$$

For K-means-based estimation, the run time grows as:

$$\text{Run time(k-means-based)} = \frac{-391.4993 + 0.1001 * numberDocuments + 11.1628 * numberFields + 9.3168 * TopicNumber.Value + 0.58986 * MaxNumberIteration.Value}{60} \quad (2.15)$$

As the data sets grow in size, the runtime of collapsed Gibbs sampling estimation method grows at a rate more than twice that of the K-means estimation method. The K-means LDA achieves these reductions in runtime while providing a level of accuracy similar to the true models.

Chapter 3. KSMERT Software Development

Overview

In addition to proposing the K-mean estimation method and KSMERT models, we develop a corresponding software instantiation of the methodology that is easily deployable to analysts operating in a DoD computing environment. Since the Microsoft Office Suite, including Excel, is ubiquitous across the DoD computing environment the Visual Basic for Applications (VBA) programming language presents the best overall value for development of the KSMERT software instantiation. We acknowledge that VBA has some drawbacks, including slow runtimes and lack of external libraries, but believe it's ability to support easy deployment to the DoD analytic workforce outweighs these drawbacks.

Developmental Concept

The four lines of effort in the development of the KSMERT software are:

- **Core subroutines** which transform text to numbers and create clusters and word assignments,
- **Human-computer interaction** which include data visualizations and knowledge elicitation from the SMEs,
- **Method and code testing** which are built-in ways to evaluate the core subroutines for verification and validation (the outcome of our quality assurance plan), and
- **Code sharing methods** which include the ability to install the code as an add-in similar (in some respects) to the excel solver.

Figure 3.1 shows how the key subroutines and features of the software development fit into the four areas of effort. The items highlighted with double thick boards are those specifically developed over the course of this project while the other items represent those carried forward from previous effort by the research team.

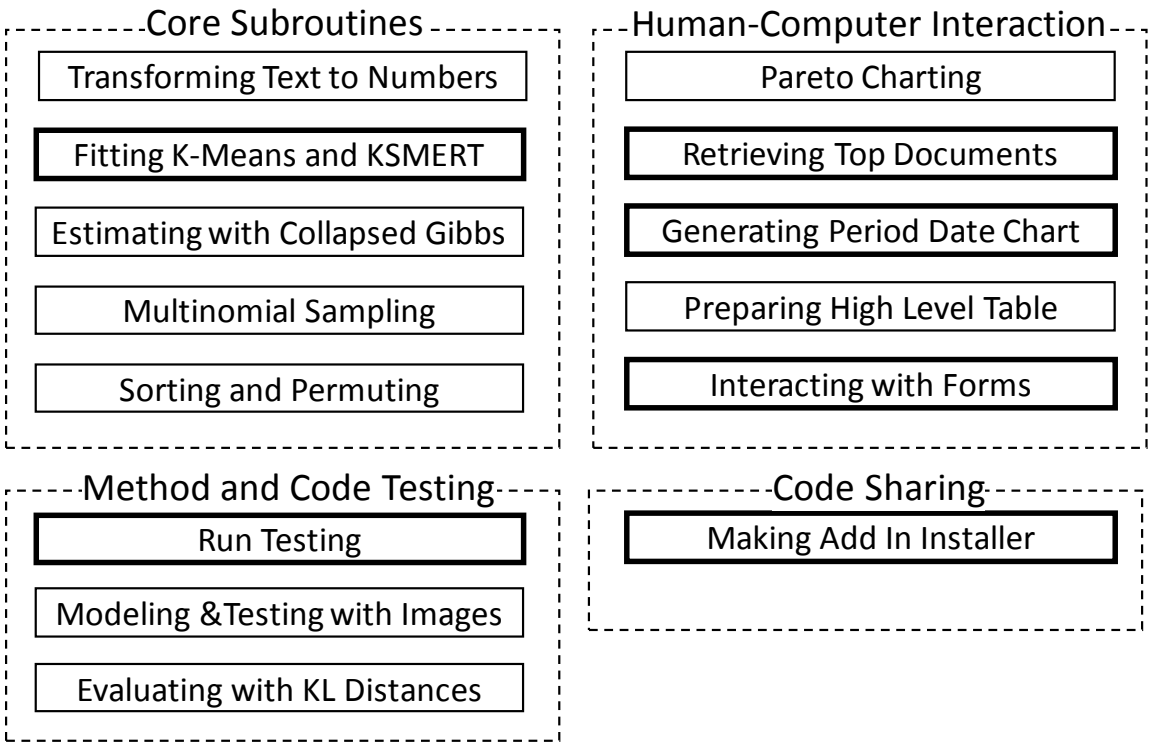


Figure 3.1. Categories and modules in KSMERT with thick borders indicating innovations.

Details Related to the Code Additions

This section contains additional details for each of the modules of the software we develop to provide a deployable software instantiation of the KSMERT topic modeling methodology.

Fitting K-means LDA and KSMERT.

In our initial evaluations, the most significant usability issue was the speed. The initial version of SMERT software, using collapsed Gibbs sampling based LDA models, took too to fit corpora of interest to the user population. Our sponsor's user population requires a tool capable of producing topic models from corpora containing thousands of documents in less than half a minute. In response to this need, we develop the K-means based estimation method described in Chapter 2. The software incorporates this methodology as an option to fit both the initial LDA models and the SMERT models.

Run Testing.

It is critical for quality assurance to test the core subroutines using examples with known inputs and outputs. It is desirable to be able to run these tests at any time, particularly after major code changes, to assure that the code has not regressed. To address this requirement, the code includes built-in testing features. The user can "Unhide" the test cases worksheet and click "Run

Tests". The resulting outputs assure the user that the KSMERT method offers comparable accuracy to collapsed Gibbs sampling and variational inference methods as shown by four test cases. Allen, Xiong, and Afful-Dadzie (2015) further describes development and use the test problems and metrics.

Retrieving Top Documents.

During beta software testing with our BMC user group at Fort Bliss, Texas, they requested an additional application of the topic model. The users have a need fit the topic model, edit the definitions, and then retrieve the documents most relevant to a specific topic on demand. This new application for topic model methodologies led to the development of an information retrieval feature. The user simply specifies the topic number and the desired number of top documents. The feature then retrieves the documents (often simply rows of the database) which have the highest proportion of their words estimated to be on the relevant topic.

Generating Period Date Chart.

The original SMERT software divides the corpus into 10 parts effectively assuming a time-based ordering of the base data. This allows the user to identify changes in the topic with the highest proportion across fixed buckets. Our current KSMERT software instantiation now allows the user to specify time intervals such as day, month or year if the data contains a date/time field. This allows users to identify changes in the top topic across time domains containing different quantities of documents.

Interacting with Forms.

A key feature of the software development for this project period is a graphical user interface including user forms to elicit directions for data fitting and table creation. Additionally, we add icons the Excel ribbon to allow users to run KSMERT in any worksheet without needing to copy the data into a specific sheet. The user interface greatly improves the usability and professionalism of the software.

Including an Add-In Installer.

The user has two options. First, the user can copy the data into the KSMERT workbook and perform analysis. Alternatively, the user can apply KSMERT as an Add-In. After loading the path information for the KSMERT file, the user can then use KSMERT inside any Excel workbook. These two options increase the flexibility of the software and may aid its usefulness in classified environments. Appendix D contains detailed instructions for the installation of the add-in.

This page intentionally left blank.

Chapter 4. Summary and Conclusions

We propose an innovative estimation technique using K-means clustering to fit LDA topic models. We also integrate our K-means clustering technique with the original SMERT model methodology to produce KSMERT models. We demonstrate through test problems that KSMERT can achieve improved repeatability and comparable subjective accuracy. Specifically, we use four cases to test our new model against the true models. The improved efficiency is important for enabling spreadsheet applications or the use of topic modeling techniques on large data sets.

A number of areas for future improvement in fitting topic models remain for future study. Other techniques besides K-means based estimation, such as Fuzzy-C clustering, deserve further research. In addition, additional comparison metrics and test cases might better clarify the accuracy limitations of KSMERT methods. New evaluation metrics could be more objective and interpretable than RMS. Currently, the running time experiments involve only small test corpora from Allen, Hui, and Afful-Dadzie (2015). Larger corpora from the literature may serve as more respective test cases. The development of additional visualization methods beyond Pareto charts may increase the interpretability of results for the analyst and customers.

We believe that KSMERT is a valuable software tool that enables analysts without a coding capability or access to other analytic tools the ability to conduct accurate and reproducible analysis on large free text data sets. Additionally, we believe that the development of the K-means estimation technique for LDA models is a significant advancement in the topic modeling field. At the same time, we see opportunities for further improvements. Below are our top five potential directions for further improvement.

1. **Further improvements in computational speed** are possible by simply coding the core operations in C++. This likely would require the creation two files instead of one, making transportability more difficult, but it may yield 20× or 50× the speed increase.

2. **The comparison of alternative estimation methods deserves expansion** beyond the 4 small, arbitrarily generated cases from Chapter 2. A more thorough comparison should make use of test problems from other literature besides Allen, Xiong, and Afful-Dadzie (2015). Additionally, ground truth topic models could generate corpora with the different estimation methods measured against their ability to reconstitute the ground truth models.

3. **The development of additional visualizations** can increase the interpretability of the information for both the analyst and the supported decision maker. Plotting the top topic in the time series is useful but methods to visualize all the topics over time that are present in other software and articles could be included. New visualizations, designed to facilitate the focus on specific contrasts relevant to specific issues or technologies, deserve further research and development.

4. **Further improvement to the accuracy of fast surrogate estimation techniques**, like K-means, is likely possible from a careful study of the SMERT and LDA likelihoods to fashion an improved surrogate. Improvements in computational efficiency and reductions

in bias in estimation are likely possible. KSMERT is a promising start but additional research is possible.

5. **Porting KSMERT to a coding language used by the wider analytic community.** The use of VBA as the language the best supports software deployment only makes sense in the restricted environs of the DoD computing abyss. By coding KSMERT, ideally as a deployable package, in a language such as R or Python we significantly expand the power of the methodology. Additionally, this allows wider populations form the analytic and academic communities to use, review, and improve on our research.

Appendix A – References

Allen, T. T., Xiong, H., and Afful-Dadzie, A.. A directed topic model applied to call center improvement. *Applied Stochastic Models in Business and Industry* 2015; 32(1): 57-73.

Allen T.T., Vinson S.M., Raqab A., and Alam Y. Using SMERT to Identify Actionable Topics in Student Feedback. *Integrated Systems Engineering Technical Report 2013*.

Allen TT, Xiong H. Pareto charting using multifield freestyle text data applied to toyota camry user reviews. *Applied Stochastic Models in Business and Industry* 2012; 28(2):152–163.

Allen, T.T., and Bernshteyn, M., Supersaturated designs that maximize the probability of identifying active factors. *Technometrics* 2003; 45:90–97.

Blei, D.M., Probabilistic topic models. *Communications of the ACM* 2012; 55(4):77–84.

Blei, D.M., Lafferty, J.D., A correlated topic model of science. *The Annals of Applied Statistics* 2007:17–35.

Blei, D. and Lafferty, J.D., Correlated topic models. Advances in neural information processing systems, 2006; 18, 147.

Blei, D.M., and McAuliffe, J.D., Supervised topic models. In *Advances in Neural Information Processing Systems*, J Platt, Koller Singer Y, R Roweis(eds). MIT Press: Cambridge, MA, 2008; 121–128.

Blei, D.M., Ng, A.Y., Jordan, M.I.. Latent Dirichlet allocation. *The Journal of Machine Learning Research* 2003; 3:993–1022.

Britt, R., *How you and 'The Rock' Turned His Movie Around*. Retrieved from http://www.marketwatch.com/story/how-hollywood-is-using-social-media-to-tell-if-a-movie-will-be-a-hit-Accessed June 19, 2015.

Feldman, R. and Sanger, J., The Text Mining Handbook –Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, 2006.

Griffiths, T.L., Steyvers, M., Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America* 2004; 101(Suppl 1):5228–5235.

Lee, S.H., Comparison and application of probabilistic clustering methods for system improvement prioritization. *Ph.D. Thesis*, The Ohio State University, 2012.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. *University of California Press*. pp. 281–297.

Packiam, R.M. and Prakash, V.S.J., December. An empirical study on text analytics in big data. In *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* . IEEE. , 2015; 1-4.

Porter, 1980, An algorithm for suffix stripping, *Program*, Vol. 14, no. 3, pp 130-137.

Steyvers, M. and Smyth, P., Chemuduganta C. Combining background knowledge and learned topics. *Topics in Cognitive Science* 2011; 3(1):18–47.

Steyvers, M., Smyth, P., and Rosen-Zvi, M., and Griffiths, T., Probabilistic Author-Topic models for information discovery. *The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2004; 10:306–315.

Steyvers, M. and Griffiths, Probabilistic Topic Models. In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum, 2007.

Sui, Z, Milam, D, and Allen, T. T., A visual monitoring technique based on importance score and twitter feeds, *INFORMS Social Media Analytics Student Paper Competition*, 2015.

Sun, X., Textual document clustering using topic models. InSemantics, Knowledge and Grids (SKG), *2014 10th International Conference on IEEEE*, 2014; 1-4.

Y. W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In NIPS 19, pages 1353–1360. 2007.

Xiong, H., Combining subject expert experimental data with standard data in Bayesian mixturemodeling. *Ph.D. Thesis*, The Ohio StateUniversity, 2011.

Zaman, T. R., Herbrich, R., Van Gael, J., and Stern, D., Predicting information spreading in twitter. In *Workshop on computational social science and the wisdom of crowds, nips* 2010; 104: 17599-601.

Zhao, T., Li, C., Li, M., Wang, S., Ding, Q. and Li, L., Predicting Best Responder in Community Question Answering Using Topic Model Method. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*, IEEE Computer Society, 2012; 1: 457-461.

Zheng, N., Discovering interpretable topics in free-style text: diagnostics, rare topics, and topic. *Ph.D. Thesis*, The Ohio State University, 2008.

Appendix B – Glossary

BMC	Brigade Modernization Command
DoD	Department of Defense
KSMERT	K-mean Subject Matter Expert Refined Topic
LDA	Latent Dirichlet Allocation
NIE	Network Integration Exercise
SME	Subject Matter Expert
SMERT	Subject Matter Expert Refined Topic
TRAC	TRADOC Analysis Center
WSMR	White Sands Missile Range
VBA	Visual Basic for Applications

This page left intentionally blank.

Appendix C – Numerical Examples for KSMERT

This appendix contains data for the case studies including the true model that originally appeared in Allen, Xiong, and Afful-Dadzie (2015).

Table C-1. Synthetic data for the numerical example.

Doc#	Document
1	The operator cut aluminum and dropped it at station1.
2	The inspector drilled plastic and overheated it at station2.
3	The manager milled steel and misaligned it at station3.
4	The engineer saw stone and over torqued on the truck.
5	The supplier welded and misdimensioned the titanium offsite.
6	The inspector drilled plastic and overheated it at station2.
7	It was drilled and overheated.
8	It was drilled and overheated.
9	The engineer and the manager at station3 and on the truck.
10	The welded titanium was misdimensioned.
11	The titanium was welded and misdimensioned offsite.
12	The steel was misdimensioned.
13	The operator cut the steel and plastic.
14	The manager welded it and misdimensioned it.
15	The operator cut and dropped the aluminum at station1.
16	The operator cut and dropped it at station1.
17	The engineer welded and misdimensioned the titanium.
18	It was drilled and overheated.
19	It was drilled and overheated.
20	The manager milled steel and misaligned it at station3.
21	The operator cut and dropped the steel at station1.
22	The engineer and the manager at station3 and offsite.
23	It was drilled and overheated.
24	The engineer saw stone and over torqued on the truck.
25	The stone was drilled and overheated.
26	It was drilled and overheated.
27	It was drilled and overheated.
28	It was drilled and overheated offsite.
29	The supplier welded titanium and misdimensioned it offsite.
30	The operator cut and dropped the titanium at station1.
31	The operator cut and dropped it at station1.
32	It was steel.
33	The steel was drilled and overheated.
34	It was drilled and overheated at station3.
35	The engineer and the manager at station1 and on the truck.
36	The welded titanium was misdimensioned.
37	It was drilled and overheated.
38	It was drilled and overheated.
39	The supplier welded titanium and misdimensioned it offsite.
40	It was drilled and overheated.

Table C-2. True model for the numerical example.

T1	0.4	T2	0.2	T3	0.15	T4	0.125	T5	0.125
Word	Prob	Word	Prob	Word	Prob	Word	Prob	Word	Prob
Oper	0	oper	0	oper	0.23	oper	0	oper	0
Cut	0	cut	0	cut	0.23	cut	0	cut	0
aluminum	0	aluminum	0	aluminum	0.08	aluminum	0	aluminum	0
Drop	0	drop	0	drop	0.23	drop	0	drop	0
station1	0	station1	0	station1	0.23	station1	0	station1	0
inspector	0.1	inspector	0	inspector	0	inspector	0	inspector	0
Drill	0.35	drill	0	drill	0	drill	0	drill	0
plastic	0.1	plastic	0	plastic	0	plastic	0	plastic	0.1
overh	0.35	overh	0	overh	0	overh	0	overh	0
station2	0.1	station2	0	station2	0	station2	0	station2	0
manag	0	manag	0	manag	0	manag	0.25	manag	0.1
mill	0	mill	0	mill	0	mill	0	mill	0.1
steel	0	steel	0	steel	0	steel	0	steel	0.5
misalign	0	misalign	0	misalign	0	misalign	0	misalign	0.1
station3	0	station3	0	station3	0	station3	0.25	station3	0.1
engin	0	engin	0	engin	0	engin	0.25	engin	0
saw	0	saw	0	saw	0	saw	0	saw	0
stone	0	stone	0	stone	0	stone	0	stone	0
overtorqu	0	overtorqu	0	overtorqu	0	overtorqu	0	overtorqu	0
truck	0	truck	0	truck	0	truck	0.25	truck	0
supplier	0	supplier	0.05	supplier	0	supplier	0	supplier	0
weld	0	weld	0.3	weld	0	weld	0	weld	0
misdimens	0	misdimens	0.3	misdimens	0	misdimens	0	misdimens	0
titanium	0	titanium	0.3	titanium	0	titanium	0	titanium	0
offsit	0	offsit	0.05	offsit	0	offsit	0	offsit	0

Appendix D – SMERT Users Guide

Instructions for Installing KSMERT as an Add-In

It is important to note that KSMERT is most easily used as a spreadsheet, loading data into it and obtaining results. However, frequent users might prefer to have KSMERT loaded into their environments so that they can use KSMERT in any sheet as an add-in similar (perhaps) to the excel solver. The following are steps for using KSMERT as an add-in.

Step 1. Download the software "KSMERT_DLL_v….xlsm" (current version) and save it on the desktop or in any designated folder. Rename the file to "SMERT.xlsm" (optional). In Figure D-1, the illustration saves the results to a SMERT folder on the desktop.

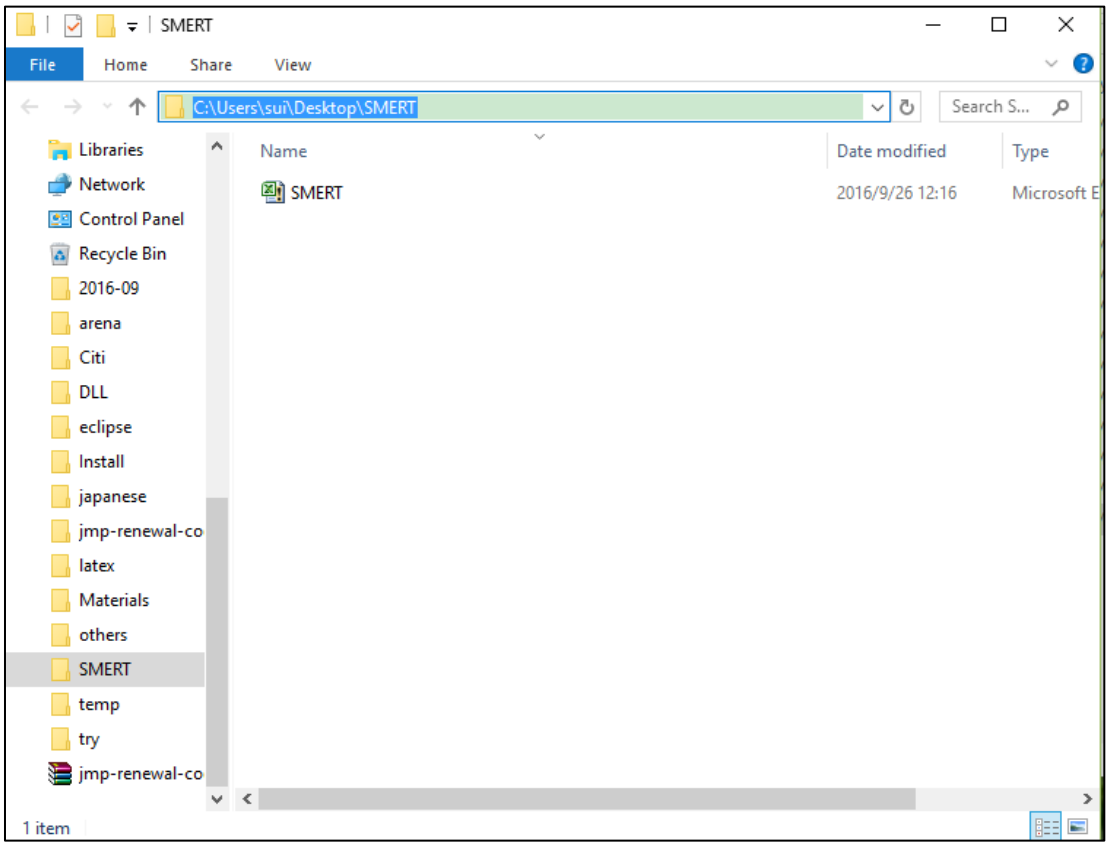


Figure D-1. A directory with SMERT (or KSMERT).

Step 2. Open this file. Click file → Save As. Then Choose "Save as type" as "Excel Add-in". Then the Add-in is automatically be saved to the directory C:\Users\sui\AppData\Roaming\Microsoft\AddIns. This step installs the KSMERT add-in module to excel systems. See Figure D-2 and Figure D-3 to illustrate the steps.

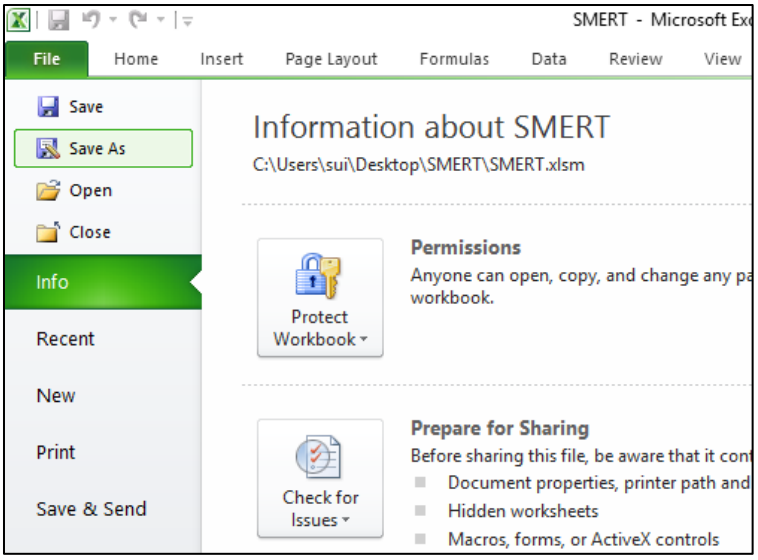


Figure D-2. Depiction of the "Save As" feature in excel.

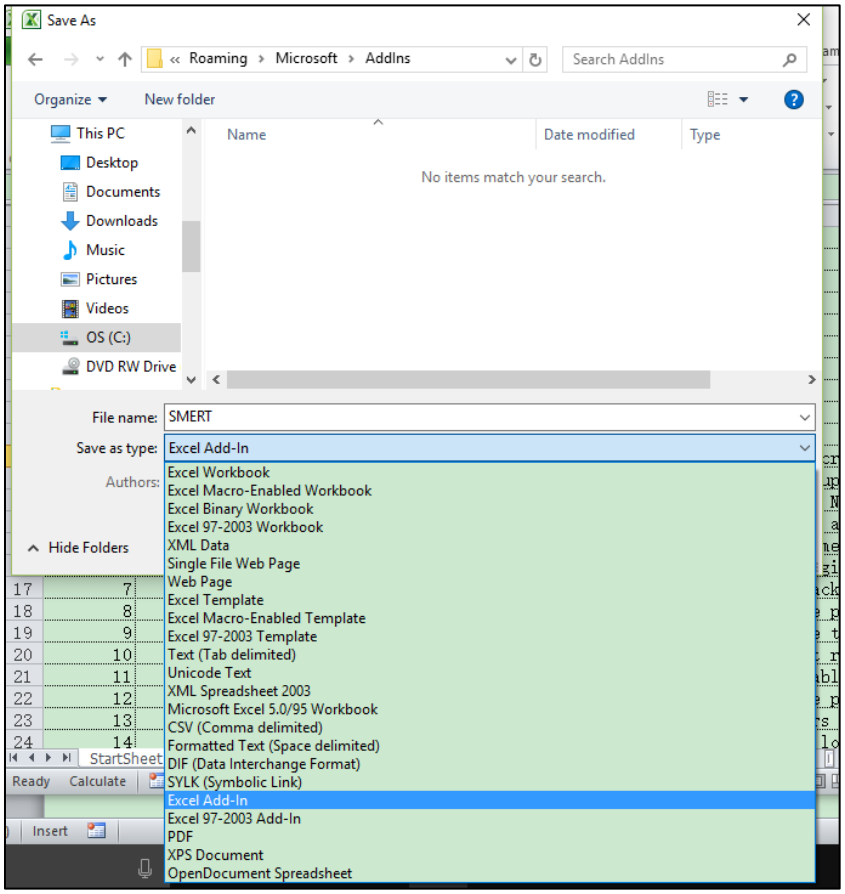


Figure D-3. Illustration of the procedure for generating an add-in.

Step 3. Open a previously saved excel file that you want to analyze. Then go to file → Options. At the very bottom click the button "Go…" The add-in "Smert" should be automatically there. If not, you can click "Browse…" to select. Check the "Smert" add-in. and click "OK". The KSMERT add-in is now available to use. See Figure D-4, Figure D-5, and Figure D-6.

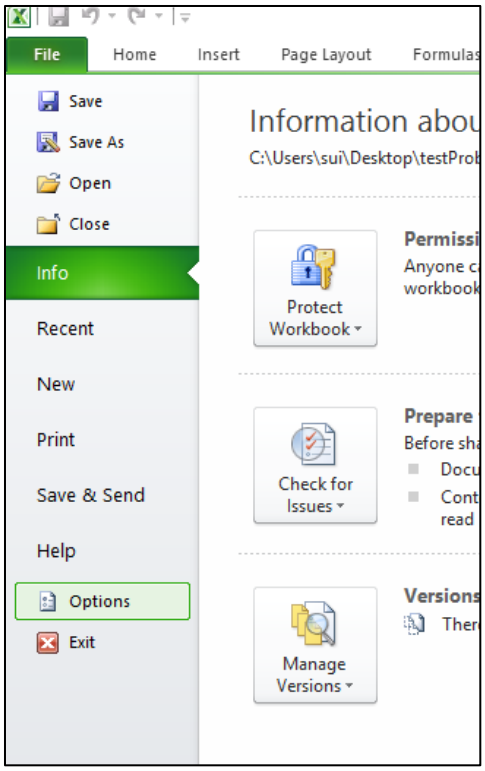


Figure D-4. The selection of options to include an add-in.

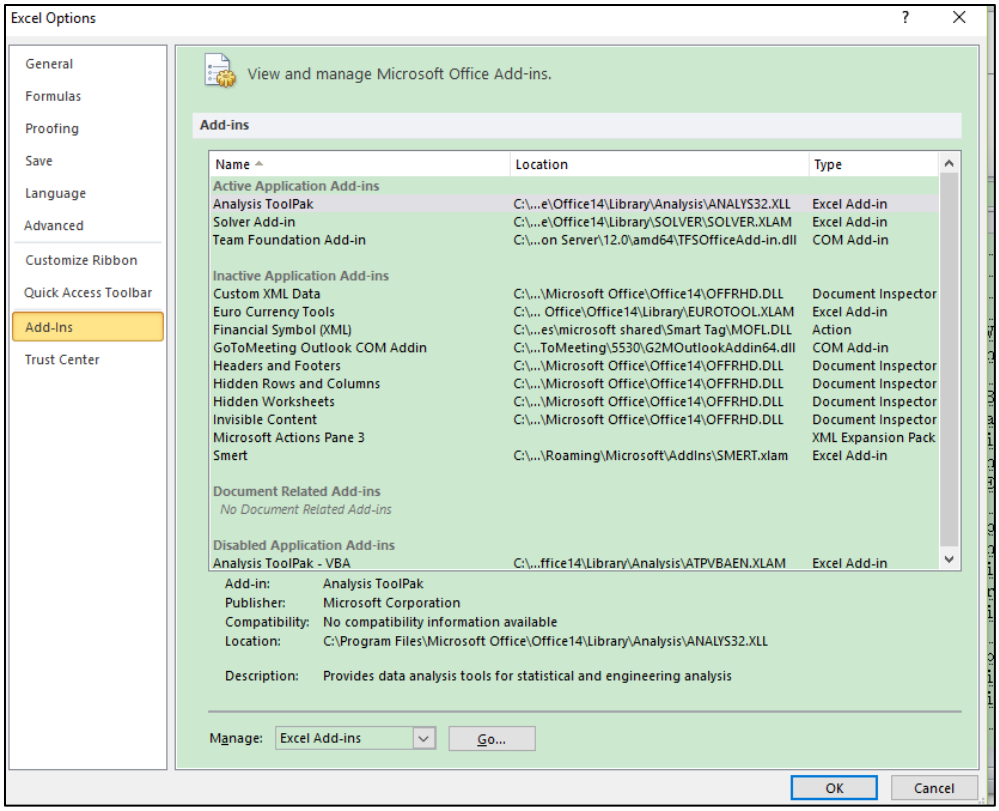


Figure D-5. Illustration showing where KSMERT can be selected before "Go…".

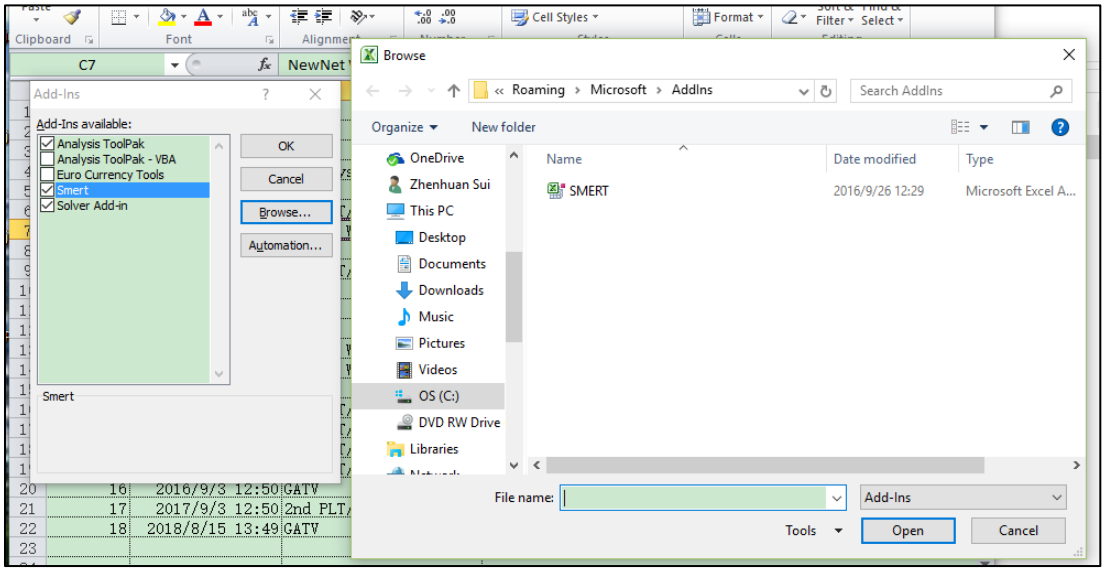
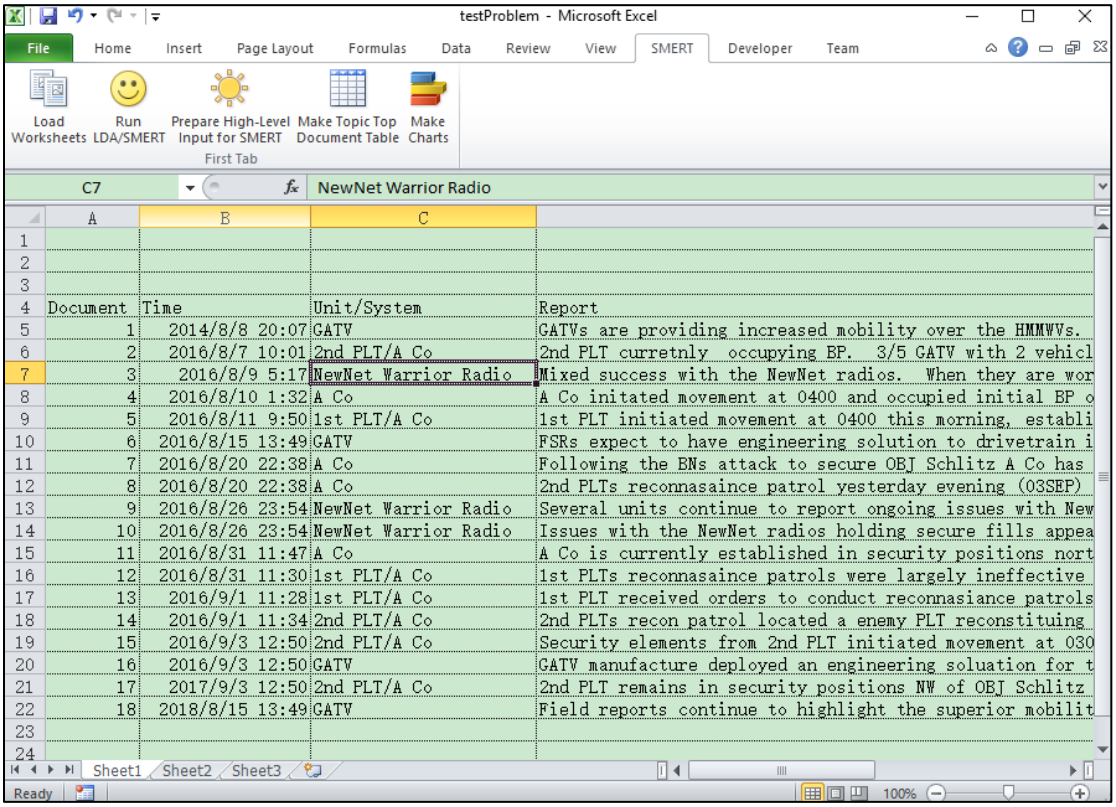


Figure D-6. The last selection for the SMERT (or KSMERT) add-in.

Step 4. The KSMERT button set should now be available on your tool bar. To carry out your analysis, click "Load Worksheets" to load the SMERT worksheets. For the Directory Path,.

"C:\Users\sui\Desktop\SMERT\". **Be sure to include \" at the very end of the directory name**. Then for File Name, input the file *Step 1*. name which, in was set in *Step 1* as "SMERT.xlsm". Then click the "Load Worksheets" button.



	A	B	C	
1				
2				
3				
4	Document	Time	Unit/System	Report
5	1	2014/8/8 20:07	GATV	GATVs are providing increased mobility over the HMMWVs.
6	2	2016/8/7 10:01	2nd PLT/A Co	2nd PLT curretnly occupying BP. 3/5 GATV with 2 vehicl
7	3	2016/8/9 5:17	NewNet Warrior Radio	Mixed success with the NewNet radios. When they are wor
8	4	2016/8/10 1:32	A Co	A Co initated movement at 0400 and occupied initial BP o
9	5	2016/8/11 9:50	1st PLT/A Co	1st PLT initiated movement at 0400 this morning, establi
10	6	2016/8/15 13:49	GATV	FSRs expect to have engineering solution to drivetrain i
11	7	2016/8/20 22:38	A Co	Following the BNs attack to secure OBJ Schlitz A Co has
12	8	2016/8/20 22:38	A Co	2nd PLTs reconnasaince patrol yesterday evening (03SEP)
13	9	2016/8/26 23:54	NewNet Warrior Radio	Several units continue to report ongoing issues with New
14	10	2016/8/26 23:54	NewNet Warrior Radio	Issues with the NewNet radios holding secure fills appea
15	11	2016/8/31 11:47	A Co	A Co is currently established in security positions nort
16	12	2016/8/31 11:30	1st PLT/A Co	1st PLTs reconnasaince patrols were largely ineffective
17	13	2016/9/1 11:28	1st PLT/A Co	1st PLT received orders to conduct reconnasiance patrols
18	14	2016/9/1 11:34	2nd PLT/A Co	2nd PLTs recon patrol located a enemy PLT reconstituing
19	15	2016/9/3 12:50	2nd PLT/A Co	Security elements from 2nd PLT initiated movement at 030
20	16	2016/9/3 12:50	GATV	GATV manufacture deployed an engineering soluation for t
21	17	2017/9/3 12:50	2nd PLT/A Co	2nd PLT remains in security positions NW of OBJ Schlitz
22	18	2018/8/15 13:49	GATV	Field reports continue to highlight the superior mobilit
23				
24				

Figure D-7. The workbook showing where the worksheets can be loaded.

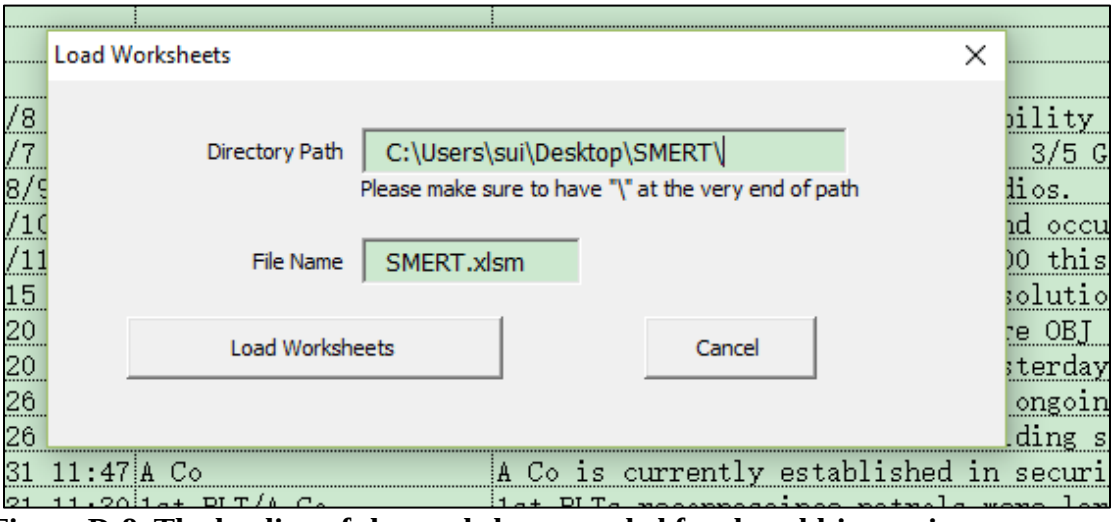


Figure D-8. The loading of the worksheets needed for the add-in version to operate.

After these steps, the program is ready and analysis is possible using KSMERT in different workbooks as indicated in Figure D-9.

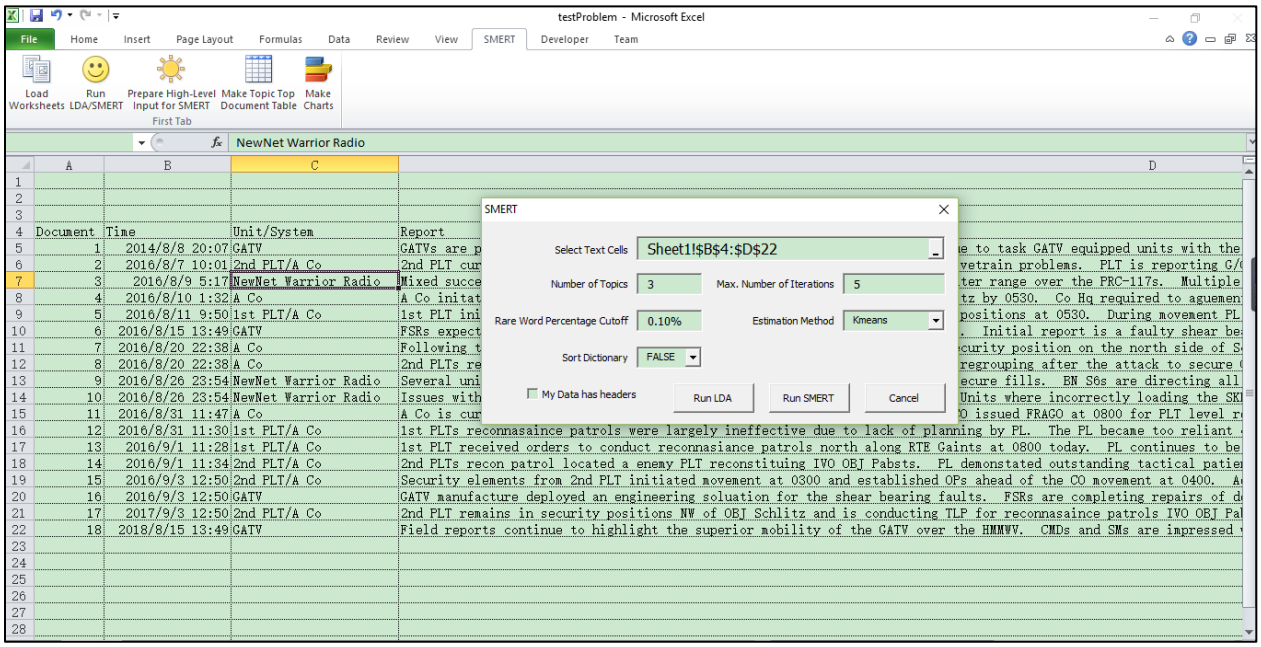


Figure D-9. The final add-in version in operation.

Instructions for Using SMERT and KSMERT

After opening an Excel workbook with KSMERT in it or loading the add-ins, there is a SMERT item in the ribbon. Select Run LDA/SMERT from the SMERT ribbon or, alternatively, click on the Run SMERT button in the Start Sheet. The dialog appears as pictured in Figure D-10. Use the cursor to click on "Select Text" and enter the range of cells with the text. If the data has a header row, click on "My data has headers" in the dialog. Then select the number of "topics" or clusters.

Often, using 10 topics is a reasonable starting point. For preliminary results, keep the maximum number of iterations to be 5. For defensible results select 30 and convergence will often occur automatically before 30 is reached. "Kmeans" is the innovative estimation method in the next chapter. It is much faster than collapsed Gibbs sampling. Click on "Run LDA" since SMERT is not available until after LDA is run.

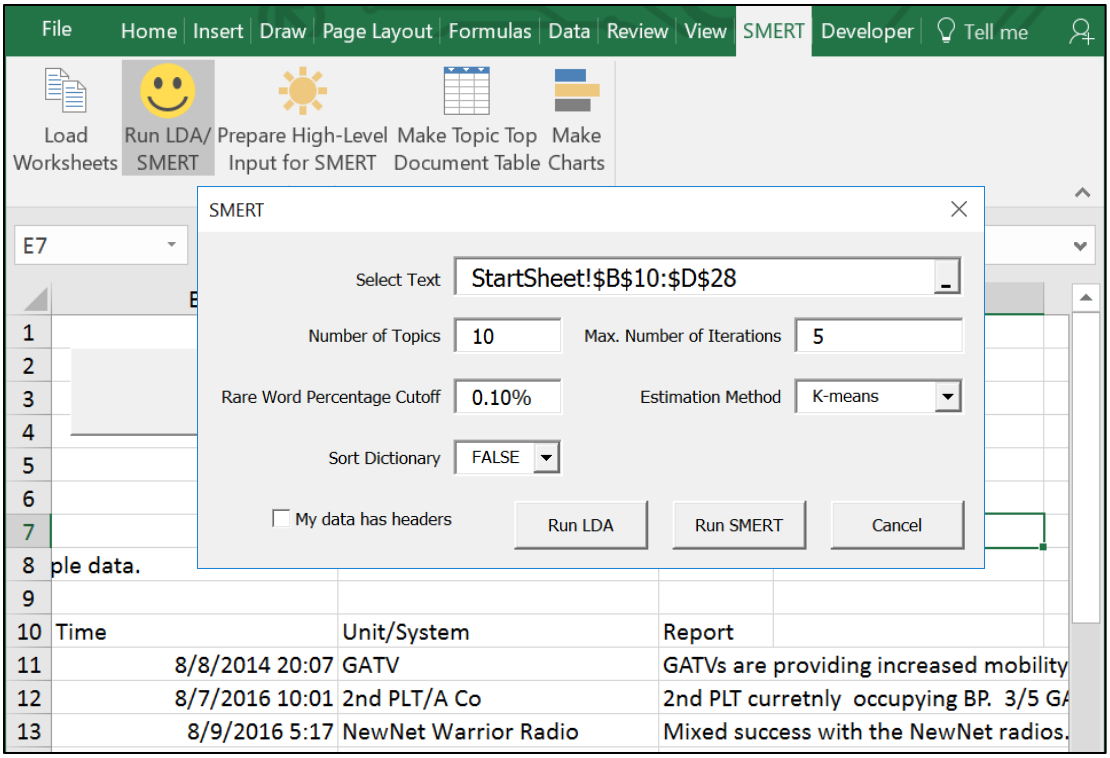
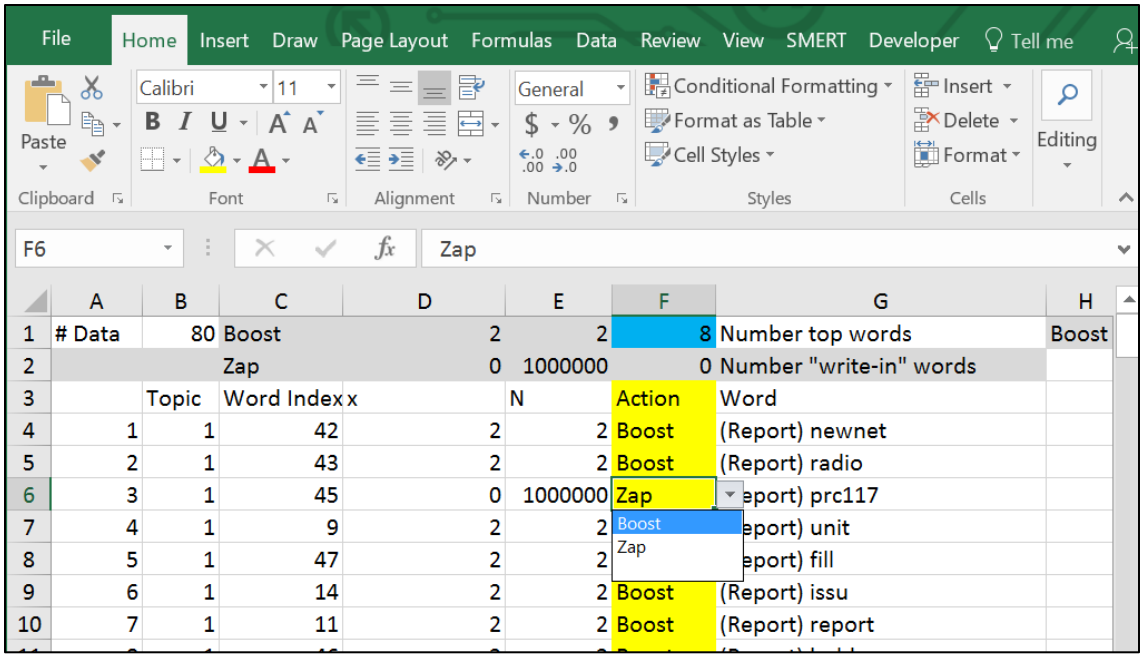


Figure D-10. The basic LDA or SMERT dialog in which data are entered.

After LDA is run, the clusters are represented by the words with the highest posterior mean probability associated with the given topic as a list (Figure D-11). The user can then "boost" to affirm or "zap" to remove any of the top words. This is shown in the figure. In this case, expert judgement suggests that prc117 does not relate to topic one which is about radios and report. After "editing" the topic definitions, rerun by either selecting "Run LDA SMERT" or by clicking "Run SMERT" as before. Click Yes… and Yes…. These relate to a check that the LDA was run or else SMERT cannot be applied.



	A	B	C	D	E	F	G	H
1	# Data	80	Boost	2	2	8	Number top words	Boost
2			Zap	0	1000000	0	Number "write-in" words	
3		Topic	Word Index	x	N	Action	Word	
4	1	1	42	2	2	Boost	(Report) newnet	
5	2	1	43	2	2	Boost	(Report) radio	
6	3	1	45	0	1000000	Zap	eport) prc117	
7	4	1	9	2	2		eport) unit	
8	5	1	47	2	2		eport) fill	
9	6	1	14	2	2	Boost	(Report) issu	
10	7	1	11	2	2	Boost	(Report) report	

Figure D-11. Spreadsheet with boosts and zaps to edit the topic definitions.

After running both the LDA and SMERT steps, then the Pareto and time series visualizations are available. The time series are plots of the top topic definitions by posterior mean probability (approximately) for periods of selected durations (days, months, years,….). Also, either clicking on "Make Top Document Table" or selecting the "topicTopDocumentTable" worksheet permits retrieval of the top documents by posterior probability. The results are similar to the application of a standard search engine as documents on topics are ordered and provided as shown in Figure D-12. At this point the clusters are defined with editing so that they are relevant. Also, the top documents on any topic can be retrieved for inspection as illustrated in Figure D-12.

1	Topic Index	3	Make Top Document Table	
2	Number Top Documents	8		
3				
4	Document	Time	Unit/System	Report
5	2	2016/8/7 1(	2nd PLT/A Co	2nd PLT curretnly
6	8	2016/8/20 2	A Co	2nd PLTs
7	7	2016/8/20 2	A Co	Following the BNs
8	5	2016/8/11 9	1st PLT/A Co	1st PLT initiated
9	4	2016/8/10 1	A Co	A Co initated
10	9	2016/8/26 2	NewNet Warrior Radio	Several units continue
11	3	2016/8/9 5:	NewNet Warrior Radio	Mixed success with
12	6	2016/8/15 1	GATV	FSRs expect to have
13				

Figure D-12. The document retrieval table for a case study example.