3-22-2018

Strategic Sourcing Via Category Management: Helping Air Force Installation Contracting Agency Eat One Piece of the Elephant

Theodore S. Holliger

STRATEGIC SOURCING VIA CATEGORY MANAGEMENT: HELPING AIR FORCE INSTALLATION CONTRACTING AGENCY EAT ONE PIECE OF THE ELEPHANT

THESIS

Theodore S. Holliger, Master Sergeant, USAF

AFIT-ENS-MS-18-M-128

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-MS-18-M-128

STRATEGIC SOURCING VIA CATEGORY MANAGEMENT: HELPING AIR FORCE INSTALLATION CONTRACTING AGENCY EAT ONE PIECE OF THE ELEPHANT

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Logistics and Supply Chain Management

Theodore S. Holliger, MS

Master Sergeant, USAF

March 2018

AFIT-ENS-MS-18-M-128

STRATEGIC SOURCING VIA CATEGORY MANAGEMENT: HELPING AIR FORCE INSTALLATION CONTRACTING AGENCY EAT ONE PIECE OF THE ELEPHANT

Theodore S. Holliger, MS

Master Sergeant, USAF

Committee Membership:

Dr. Bradley C. Boehmke
Chair

Dr. Jeffrey A. Ogden
Member

Colonel Matthew A. Douglas, PhD
Member

Dr. Edward D. White
Member

Abstract

The United States Air Force can dramatically reduce resource consumption through strategic sourcing initiatives that leverage sensibly-bound pockets of spend via category management. However, category creation is a particularly daunting task due to the sheer magnitude of purchasing data in large organizations. Text mining is one way to identify categories. Specifically, term frequency analysis, term frequency-inverse document frequency analysis, and topic modeling can identify category membership, unique characteristics of categories, and thematic natures of the categories. This thesis developed an empirical, generalizable, reproducible methodology to analyze historical contract text descriptions to uncover the data's hidden structure. A sample case was transformed into a practical hierarchy, which was internally and externally validated. As a foundational methodology, the impact of token selection, domain expertise, and unique contracting language were identified as considerations for future research.

Acknowledgments

I would like to express my sincere appreciation to my faculty advisors for their guidance throughout the course of this thesis. The encouragement and frankness were most assuredly appreciated. I would also like to thank the ENS faculty for their patience and willingness to explain concepts that were difficult (for me) to digest.

Theodore S. Holliger

To my wife and children whose unwavering support, unconditional love, and implicit acceptance of military service were paramount to the completion of this thesis. Even though I was home, I wasn't always there. You've sacrificed so much for our Air Force. Thank you.

Table of Contents

Page

List of Figures

Page

List of Tables

Page

STRATEGIC SOURCING VIA CATEGORY MANAGEMENT: HELPING AIR FORCE INSTALLATION CONTRACTING AGENCY EAT ONE PIECE OF THE ELEPHANT

I. Introduction

The United States federal government (USG) has always been expected to judiciously allocate or otherwise manage taxpayer funds. Through a memorandum to government agencies, then Deputy Director of the Office of Management and Budget (OMB) highlighted this expectation; "Maximizing value for taxpayers is a top priority for OMB, and I look forward to working with the acquisition community on this important initiative" (U.S. Office of Management and Budget, 2005)

Since the formal directive to implement strategic sourcing (U.S. Office of Management and Budget, 2005) practices, the Federal Government has struggled to comply (GAO, 2012). Subsequently, the United States Air Force (USAF) has struggled to do the same largely because of sequestration (Montgomery, 2015; Muir et al. 2014). However, more recent efforts (i.e. the activation of Air Force Installation Contracting Agency and Air Force Installation and Mission Support Center in 2013 and 2015, respectively) have indicated that the USAF has "turned the corner" and current USAF leadership has recognized the need to maximize the taxpayer dollar. In her first interview as Secretary of the United States Air Force (SECAF), Dr. Heather Wilson summarized her top two priorities for the USAF with the following statements:

> "The highest priority for me is to do those things that only the secretary can do, and that's to try to secure the resources, to fight for the budget, to do all of those things that are 'gotta dos" and "It's not just one big program – it's fighters … and tankers … and bombers … and space assets … and the nuclear deterrent – it's across the board" (Gibson, 2017).

> "[There's] a lot of acquisition going on in the Air Force. We've got to get that right – we've got to value every dollar that's spent, because somebody earned that dollar" (Gibson, 2017).

In a constrained fiscal environment, every allocation is a tradeoff between what was purchased and what opportunities were forgone as a result of the purchase. In order for the USAF to execute its primary mission, leaders must weigh the tradeoffs between installation support acquisitions and organizational needs. Every dollar allocated to installation support is a dollar that *could* have been spent on another organizational need. The federal government spent approximately $50.7 billion on Information Technology (IT)-related products and services during fiscal year (FY) 2015 (Category Management Guidance Document Version 1.0, 2015). During the same time frame, the USAF obligated approximately $21 billion towards IT-related contracts. Moreover, even a small improvement in IT-related acquisitions could "free" substantial resources for fighters, tankers, bombers, space assets, and nuclear deterrence.

Air Force Installation Contracting Agency

The Air Force Installation Contracting Agency (AFICA) is headquartered at Wright-Patterson Air Force Base, OH. AFICA is responsible for managing and executing above-Wing-level operational acquisition solutions across eight Major Commands (MAJCOMS), and provide contracting authority to installation-level operational contracting units, enterprise-wide ("AFICA Flight Plan," n.d. [accessed July

21, 2017]). To do this, AFICA focuses on four mission areas; MAJCOM support, mission execution, **enterprise sourcing**, and expeditionary operations.

Within the mission focus area of enterprise sourcing, AFICA has identified a goal to reduce costs and improve mission effectiveness through the application of strategic sourcing concepts and practices ("AFICA Flight Plan," n.d. [accessed July 21, 2017]). To achieve this goal, AFICA leadership identified four focus areas:

- Structured, data-driven processes to deliver cost efficient and mission efficient solutions;
- Collaborate with AFICA partners to develop innovative solutions;
- Focus on rate (better price), process (**eliminate waste and redundancies**), and demand (**reduce consumption** and cost drivers) savings;
- Conduct informed spend analysis to **leverage buying power, improve efficiencies and manage consumption** ("AFICA Flight Plan," n.d. [accessed July 21, 2017])

Decision makers at AFICA need to know what themes historical contract descriptions contain so the contracts may be grouped into sub-categories.

Research Question, Purpose, and Scope

The fundamental research question is: *How can AFICA group a historical list of IT-related contracts into sub-categories?* The answer to this question will provide AFICA with a methodology to classify their sourcing activities at a granular level, which will reduce costs and improve mission effectiveness.

The purpose of this research is to aid AFICA in their endeavor to apply strategic sourcing practices to USAF contracting operations that will:

1. facilitate Federally mandated strategic sourcing efforts
2. enable the USAF to leverage its buying power
3. identify opportunities to consolidate redundant contracts
4. maximize value of the American taxpayer's dollar

There are ten large (Level-1) categories of spend with smaller (Level-2) categories that are directed by the OMB government-wide (Figure 1).

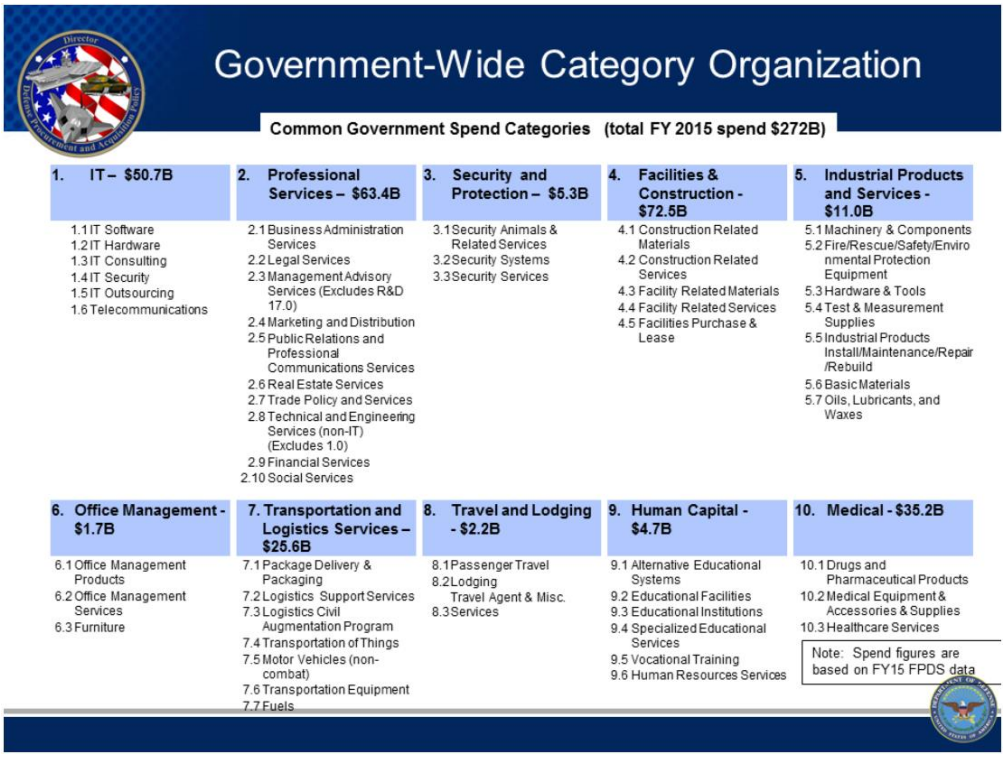


Figure 1. OMB Level-1 and Level-2 Categories (DPAP, n.d.)

Rather than investigate all ten Level-1 categories, this thesis focused on USAF IT-related contracts since this category contained both products and services and was specifically identified by the GAO (2015) as an improvement category (discussed in

Chapter 2). Furthermore, it is assumed the methodology applied in this thesis will be applicable to the other nine categories. To avoid confusion, it is important to note the Level-1 and Level-2 categories depicted in Figure 1 were assigned by the OMB and the General Service Administration (GSA) without input from AFICA.

Investigative Questions/Research Method

To answer the aforementioned fundamental research question, the following investigative questions were developed:

IQ 1. What criteria determines a sub-category?

IQ 2. How will themes be identified?

IQ 3. How will themes be useful AFICA?

The answer to the IQ 1 should identify how AFICA could group a historical list of IT-related contracts into sub-categories. Text mining was chosen as a reliable method to develop sub-categories (Dooley, 2016) primarily because the contract descriptions in the data had not been explored.

The answer to IQ 2 will enhance the validity of this research and ensure that the findings remain practically applicable. To achieve this, feedback from AFICA subject matter experts was solicited.

The answer to IQ 3 will bolster AFICA's strategic sourcing initiatives by establishing "common threads" within the historical contract data. Furthermore, the thematic nature of products and services within the contract data may enable AFICA to proactively plan for shifts in supply or demand.

Implications and Research Organization

The aim of this research is to build upon existing supply chain management (SCM) literature through the utilization of text mining in a procurement environment. Specifically, this thesis was viewed through a Resource Orchestration Theory (ROT) (Sirmon, Hitt, Ireland, & Gilbert, 2011) lens and sought to enhance understanding of the strategic sourcing and category management overlap. Operationally, this research sought to develop a generalized and repeatable method for classifying categories through text analysis.

This thesis is divided into five chapters. Chapter Two thoroughly reviews the relevant literature and explains why this study is relevant and useful to the Air Force. Chapter Three discusses the methodology for this research. Chapter Four applies the methodology to an example case. Chapter Five discusses the analysis, highlights strategic implications, and offers recommendations for action and future research.

II. Literature Review

Chapter Overview

One purpose of this chapter is to provide relevant background on Resource Orchestration Theory (ROT). In addition, an examination of the Federal Government's mandate of strategic sourcing, the role of USAF category management within the mandate, the role of both category management and strategic sourcing in ROT, and the need classify categories objectively will be provided. Finally, the chapter concludes with the introduction of text mining as a possible method for category creation. This review underpins the framework of the research and suggests a way forward.

Resource Orchestration Theory

Sirmon, Hitt, Ireland & Gilbert (2011) introduced ROT as a combination of the asset orchestration (Helfat et al., 2007) and resource management (Sirmon et al., 2007) frameworks. The primary thrust of this combination was that each framework's similarities and differences were complimentary (Sirmon et al., 2011). Specifically, ROT draws attributes from asset orchestration and resource management and focuses on how managers affect resource-based competitive advantage (Sirmon et al., 2011). A detailed comparison of the two frameworks is beyond the scope of this research, but an overview of the ROT is provided in **Figure 2** (Sirmon et al., p. 1395, 2011).

As an extension to Resource-Based View of the firm (RBV), ROT focuses on the actions of the firms' managers to create a competitive advantage. ROT attempts to explain why firms with similar resources perform differently. It is not enough to simply *have* advantageous resources, but a firm must *orchestrate* them to achieve a competitive

advantage (Sirmon et al., 2011). Hansen et al. (2004) summarized this concept when

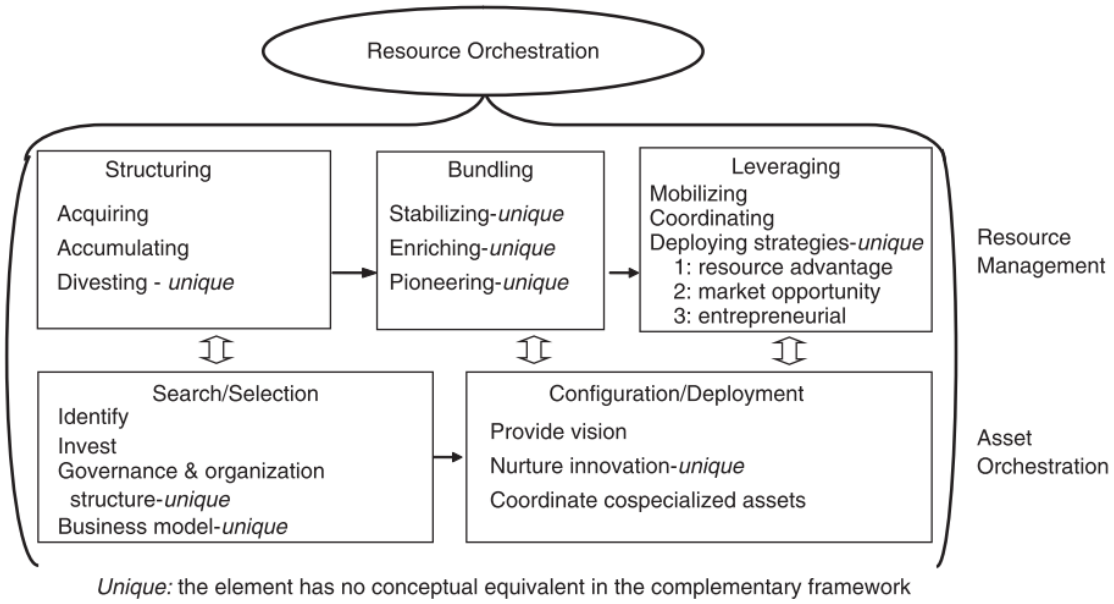


Figure 2. Resource Orchestration Overview (Sirmon et al., p. 1395, 2011)

they concluded their empirical results with: "What a firm does with its resources is at least as important as which resources it possesses" (p. 1280).

Strategic Sourcing in the Federal Government

In May, 2005 the United States Office of Management and Budget (OMB) formally directed the federal government to implement strategic sourcing initiatives in an effort to maximize taxpayer value (U.S. Office of Management and Budget, 2005). The memorandum defined strategic sourcing as "the collaborative and structured process of critically analyzing an organization's spending and using this information to make business decisions about acquiring commodities and services more effectively and efficiently" (U.S. Office of Management and Budget, 2005). It is understood that the terms "strategic sourcing" and "enterprise sourcing" are synonymous in their intent to

maximize stakeholder (taxpayer) value and acquire commodities (products and services) more efficiently.

Despite the OMB mandate, some government agencies have been slow to implement strategic sourcing initiatives, thus they have squandered opportunities to shape consumption and maximize taxpayer value. In 2012, the United States Government Accountability Office (GAO) reported to Congress during Fiscal Year (FY) 2011 selected agencies managed five percent or $25.8 billion through strategic sourcing efforts and although they reported a savings of $1.8 billion, the savings represented less than one-half of one percent of the selected agencies' budgets of $537 billion (GAO, 2012). This report (GAO, 2012) highlighted government agencies' need to bolster strategic sourcing initiatives and leverage internal procurement opportunities. Overall, the GAO identified that the Federal Government could save "billions in annual procurement costs" through the implementation of strategic sourcing initiatives (GAO, 2012).

Category Management in the USAF

Category Management is defined as "management of spend across an organization by category" (Muir et al., p. 9, 2004). Muir et al. (2004) further defines a category as "sensibly bound pockets of requirement type where future spend is expected to occur" (Muir et al., p. 9, 2004). However, many other definitions of "category" are used in academic literature (Hesping & Schiele, 2015) which has caused some confusion. For example, commodity groups (Schiele et al., 2007), material groups (Horn, Schiele, & Werner, 2013), and product groups (Luzzini & Ronchi, 2011) have all

been used to describe families of purchased products and services. Implicit to the aforementioned descriptions is that these "groups" belong to a similar group of suppliers and are similar in nature. For the sake of consistency, "category" as defined by Muir et al. (2004), was adapted as it was developed within the context of the USAF.

In response to the 2012 GAO report, *Category Management: A Concept of Operations (CONOPS) for Improving Costs at the Air Force Installation* (Muir et al., 2014) was published and recommended a detailed framework for reducing Air Force installation-support spend. The CONOPS framework specifically identified that AFICA was in a unique position to reduce installation support spend due to their above-the-wing-level centralization of contracts (Muir et al., 2014). As such, AFICA could contribute significantly to the OMB's overall effort to reduce federal spend through strategic sourcing initiatives.

In 2015, a second GAO report identified that selected agencies managed between 10 and 44 percent of their Information Technology (IT) services in FY 2013, which led to potentially hundreds of duplicative service contracts (GAO, 2015) that reduced the agencies' buying power and failed to capitalize on spend reduction opportunities. Again, AFICA was in a unique position to reduce IT installation support spend since all installation contract vehicles are "rolled up" to AFICA.

Role of Strategic Sourcing and Category Management in ROT

Recall that three components of ROT are structuring, bundling, and leveraging (Sirmon et al., 2011). Structuring are the processes in which firms acquire, accumulate, and divest resources (Sirmon et al., 2011, 2007). Bundling are the processes in which

firms stabilize (create minor improvements to existing capabilities), enrich (extend current capabilities) and pioneer (create new capabilities) resources to form capabilities (Sirmon et al., 2011, 2007). Leveraging are the processes in which firms mobilize (plan), coordinate (integrate capability configurations) and deploy (exploit market opportunities) capabilities to take advantage of specific market opportunities (Sirmon et al., 2011, 2007). Strategic sourcing and category management activities are prevalent in some if not all of ROT processes.

ROT is an appropriate theoretical lens to view this research. As stated before, ROT is an extension of RBV. Hunt & Davis (2012) argued that purchasing strategy should be grounded in RBV, and generally, supply chain management. Therefore, strategic sourcing (collaborative and structured process of analyzing organizational spend) activities and category management (management of spend across categories) activities are resource-related *processes* used to achieve a competitive advantage.

Strategic Sourcing Via Category Management

Although category management is a method to source strategically, it is also a logical first step to categorize products into similar categories to identify opportunities that may exist amongst products within the category. Category Management is a process, rooted in retailing, that seeks to identify "interrelatedness of products within a category" and focuses on the performance of the whole category vice individual brands (Basuroy, Mantrala, & Walters, p. 1, 2001). Category Management theory development and evolution is beyond the scope of this research. Instead, the assumption that category management is a beneficial and practical organizational process that facilitates strategic

sourcing initiatives is made. However, it should be noted that within the field of strategic sourcing there is much debate on *how* to create categories.

Muir et al. (2014, p. 9) defines a category as "sensibly bound pockets of requirement type where future spend is expected to occur". Conversely, many authors in academia have categorized products within Purchasing Portfolio Models (PPMs) based on supply risk (Kraljic, 1983), profit impact (Kraljic, 1983; Trautmann, Turkulainen, Hartmann, & Bals, 2009), organizational power position (Cox, 2015), purchase novelty and strategic importance (Cox, 2015; Luzzini, Caniato, Ronchi, & Spina, 2012; Olsen & Ellram, 1997; Trautmann et al., 2009). Some scholars argue that categories should be "defined by the Portfolio Manager" (Muir et al., p. 25, 2014). Furthermore, in a multilevel review of purchasing strategy (Hesping & Schiele, 2015) the authors noted that "literature was lacking in theoretically sound and empirically based classifications of sourcing categories" (p. 147). To be clear, the practical classification process of *both* requirement type and sourcing categories is subjective and is often contingent upon the purchasing function's interpretation of sourcing strategy.

It is important to note that this research does not seek to classify strategic sourcing categories as much research has been devoted to this objective (Cox, 2015; Gelderman, Cees J; van Weele, 2005; Kraljic, 1983; Olsen & Ellram, 1997; Trautmann et al., 2009). Instead, this research focuses on specific themes that may be present within an array of goods and services because purchase categories are domain-specific (Luzzini et al., 2012). In other words, the creation of goods and services groups *determines* the placement of goods and services groups within a strategic sourcing model. To this extent, the focus is placed on the grouping (categorization) of goods and services and it

is assumed AFICA will leverage the groups via strategic sourcing models that will provide the most value to the taxpayer.

Conclusion and Way Forward

It is imperative to recognize the following:

1. Significant opportunities still exist for the USAF to realize savings from strategic sourcing initiatives, specifically within the IT-related installation support spend Level-1 category.
2. To leverage strategic sourcing strategies, the USAF must first objectively group interrelated products and services into sub-categories.

However, the question remains: *How can AFICA group a historical list of IT-related contracts into sub-categories?* Big data has emerged as valuable resource in the context of SCM (Simpson et al., 2015). The use of data analytics can glean insights that might not have been possible before, and predictive analytics in the context of SCM are needed in literature (Waller & Fawcett, 2013). Specifically, the use of text analysis, when integrated with analytics, can yield unique insights about the content of a manifest (Dooley, 2016). Therefore, this research fills both an operational and research gap by:

1. providing an objective, repeatable methodology to identify themes of products and services from a historical list of IT-related contracts (manifest content analysis) and
2. objectively grouping similarly themed products and services into practical categories using domain expertise (human feedback loop) to facilitate AFICA strategic sourcing efforts.

The combination of latent and manifest content analysis can be a reliable and valid approach to study modern problems with a voluminous data set (Dooley, 2016). Manifest content analysis can uncover potential themes (product and service groups) that would be useful in the creation of categories. To increase the validity of this research, latent manifest analysis by subject-matter experts with domain expertise will be conducted. Together, these two approaches will enable AFICA to create more detailed categories for use in strategic sourcing initiatives.

III. Methodology and Data Description

Chapter Overview

This section introduces a generalized methodology overview and an expanded methodology used to conduct the analysis.

Methodology Overview

Recall the purpose of this thesis is to help AFICA group a historical list of IT-related contracts into sub-categories to bolster strategic sourcing activities. The following investigative questions from Chapter 1 were developed:

IQ 1. What criteria determines a sub-category?

IQ 2. How will themes be identified?

IQ 3. How will themes be useful to AFICA?

Unique insights can be uncovered through text analysis, in general, and manifest content analysis, specifically. (Dooley, 2016; Waller & Fawcett, 2013). The aim of this thesis is to identify themes or characteristics within the data, which may be used to determine sub-categories below the Level-2 (IT – Hardware, IT – Software, etc.) categories (Figure 1). It is important to reiterate the fact that the pre-existing levels were defined by OMB and GSA and may not be optimal or efficient. However, since the categorical levels were pre-defined by a higher hierarchical organization, AFICA would be best served by aligning sub-categories with those in existence. Therefore, the assumption that sub-categorical alignment to pre-defined levels will bolster strategic sourcing activities is made. From the text mining framework, content analysis (CA) and latent semantic analysis (LA) can be coupled with topic models to draw insights (themes

or characteristics) from unstructured text (contract text descriptions). The methodology overview (Figure 3) provides a strategy to extract insights and uncover the hidden structure in the data.

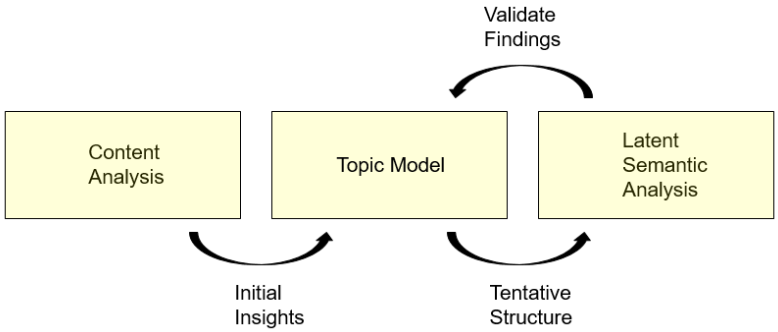


Figure 3. Methodology Overview

The methodology overview (Figure 3) is a generalizable guide for structure discovery due to its flexibility. There are copious types of CA, topic models, and LA, which allows for augmentation or substitution as the analysis progresses. Furthermore, the exact processes contained in each block have different meanings across academic literature. Therefore, an expanded methodology was developed to highlight the processes used within each block of the methodology overview (Figure 4). The expanded methodology will serve as a guide for the remainder of the chapter.

Get Data

The first step in the expanded methodology is simply to acquire data. Prior to this research, AFICA team members compiled an authoritative Microsoft Excel file of every contract on record from FY 2012 through April 13, 2017. All of the contracts belonged

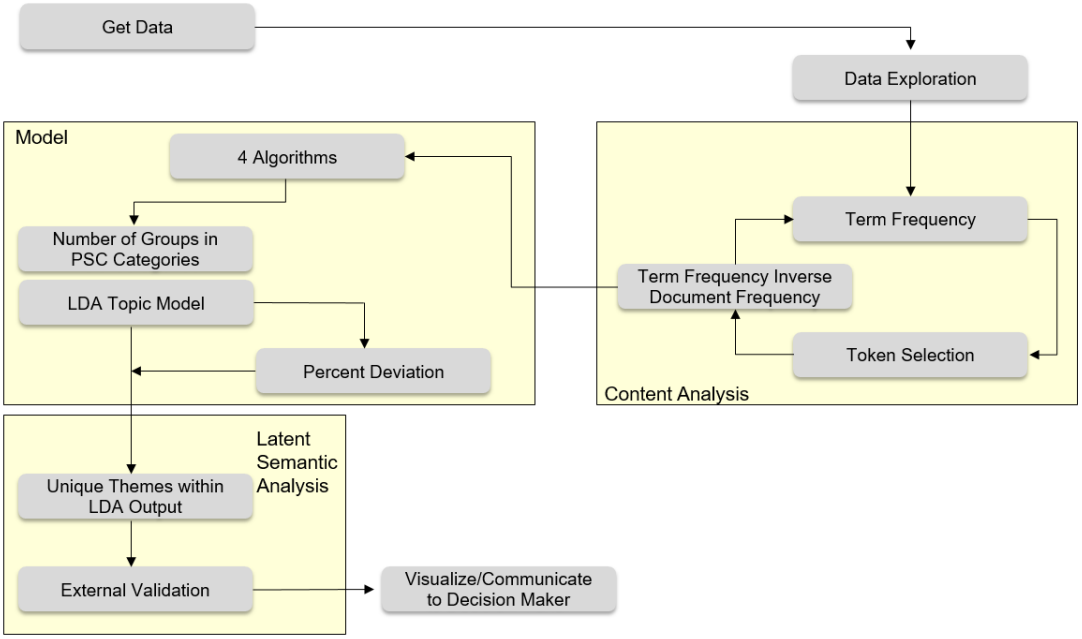


Figure 4. Expanded Methodology

to the IT (Level-1) category. The file contained 107,589 rows and 148 columns and was compiled using the Federal Procurement Data System – Next Generation (FPDSNG). Each row was representative of a contract action.

The FPDSNG data system is an interface that leverages multiple Federal Procurement data systems. For the purpose of this thesis, only AFICA-relevant sources of data were pulled from USA Spending or the Office of Management and Budget data systems.

Statistical Software

Although statistical software is not a step in the expanded methodology, it is necessary to elaborate on software selection. The statistical software used on the next and all subsequent steps was R programming software version 3.4.1. R is particularly useful in pre-processing, manipulating, modeling, and communicating complex data sets. Furthermore, the R code facilitates reproducible research, which is important if this

methodology is applied to additional Level-1 categories in the future. In addition, the code used to conduct the analysis can be found in Appendix A. The list of packages used to analyze the data set is shown in Table 1.

Table 1. R Packages

Package	Author(s)	URL
topicmodels	B. Grün and K. Hornik (2017)	https://CRAN.R-project.org/package=topicmodels
lubridate	Garrett Grolemund and Hadley Wickham (2011)	http://www.jstatsoft.org/v40/i03/
magrittr	Stefan Milton Bache and Hadley Wickham (2014)	https://CRAN.R-project.org/package=magrittr
tidyverse	Hadley Wickham (2017)	https://CRAN.R-project.org/package=tidyverse
tidytext	J. Silge and D. Robinson (2016)	http://dx.doi.org/10.21105/joss.00037

Data Exploration

Due to the sheer size of the data, it was necessary to reduce the number of variables. Through discussion with AFICA SMEs, it was determined that only three of the 148 variables were relevant this research (Table 2) for the following reasons:

1. The text description field was the only field that contained descriptive language of the contracts.
2. The analysis should incorporate the inherent constraints of the PSCs since it was the current system used to categorize contracts.
3. The Level-2 structure was an organization initiative from a higher management level.

The overarching rationale from the AFICA SME perspective was that any analysis should be conducted with existing constraints in mind. In other words, it was unlikely that a drastic change to the PSCs or Level-2 categories would be accepted

since it would constitute a dramatic change in federal procurement processes. Although a brief description is provided from the PSC Manual (2015), it is necessary to expand on the variables for clarity.

Table 2. Relevant Variables and Brief Description

Variable Name	Description
text_describe	Brief description of goods or services bought (for award) or available (for IDV).
PSC	Product Service Code
lvl_2_category	Category the contract is assigned by OMB and GSA.

text_describe

According to the data dictionary provided by AFICA, the text description field should contain a brief description of the goods and services bought or available. However, some of the text description entries contained a description of the rationale for funds obligation or de-obligation.

There were two entries that contained no text descriptions. These values were left untouched since they were a small proportion of the total. The text descriptions are the focus of this thesis, because they have not been used in any analysis prior to this research. The text descriptions may contain insight into the contract action beyond the information contained in other variables.

PSC

The PSC is a four-digit alpha numeric code that "indicate WHAT was bought for each contract action reported in the Federal Procurement Data System (FPDS)" (US GSA Product Service Manual, p. 5, 2015). There were 74 PSCs in the data set and no missing values.

The PSCs are identified as either a product or a service. The product PSCs are numeric only, and the service PSCs have a letter designator in the first character position.

lvl_2_category

As stated in Chapter 1, the Level-2 category is a sub-category defined by OMB or GSA. In this data set there were six Level-2 categories (Figure 1) and no missing values.

Content Analysis

Content Analysis (CA) is an important component of text mining with the purpose of transforming unstructured text (contract text descriptions) into formatted data using techniques such as tokenization, n-gram analysis, and removing words that do not add context (stop words).

The purpose of this is to transform the data into a malleable format suitable for term frequency (TF), term frequency – inverse document frequency (TF-IDF) analyses, and topic models.

Term Frequency

Term Frequency analysis (TF) is a natural starting point in CA as it simply returns the frequency of words in the manifest. TF can, in itself, relay what the manifest content is about since it is a tally of the occurrence of individual words. Drawing from the study of Natural Language Processing (NLP), it is assumed that words are descriptive of the manifest content. When TF is combined with bi-gram or tri-gram "tokens" (two-word sets, three-word sets), more contextual information is returned. The idea is such that the

more contextual information that is retrieved, likely themes will emerge. In this sense, a researcher can increase the granularity of the context by increasing the length of the token (bi-gram, tri-gram, quad-gram, etc.). However, there is a point of diminishing returns. Extending the tokens outward indefinitely will return an entire sentence, which defeats the purpose of text mining.

A limitation of TF analysis is its inability to discern words that are unique to the analyzed document. Homogenous documents will likely return similar frequently appearing words, which does little to establish "uniqueness" of the words in the document relative to other documents in the corpus. This limitation can be mitigated by the use of TF-IDF. However, the length of the n-gram must first be chosen.

Token Selection

As stated in the previous section, token selection is a tradeoff between granularity and interpretability. The goal is to strike a balance between the two in terms of domain expertise. In other words, the token length must contain enough terms for someone unfamiliar with the data to understand, but small enough to reduce the time it takes to digest the result (more on this in Chapter four).

TF-IDF

TF-IDF is a statistic that incorporates Zipf's Law (Zipf, 1932) that summarizes, within a group of documents the importance of a word is inversely proportional to its rank in the frequency table. To summarize, the more often a word occurs in a document, the less important it is in describing the context of that document. TF-IDF extends Zipf's Law and takes the product of TF and IDF, which will always be a number

between 0 and 1. The general idea is to find words that are common in only the specified document vice the entire collection of documents. Similarly, TF-IDF can be combined with bi-grams to increase the granularity of important word sets within specified documents.

To be clear, a document is a generic term in the text mining context. For example, a document would be one chapter out of many chapters of the same book. The book would be viewed as a collection of chapters about the same story, and a series of books would be viewed as a corpus of documents. However, the term "document" may be any incremental unit for analysis, as long as it is consistent across the analysis. In this case, a document is the text description associated with a Product Service Code (PSC). The PSCs exist in some capacity within the chapter (Level-2 category) of the book (Level-1 category) (Figure 1).

A limitation of TF-IDF is its sensitivity to anomalous words in any document. If a document has an obscure set of words relative to the other documents, TF-IDF will undoubtedly identify those words as "unique". While this is the intent of TF-IDF, the words may not necessarily capture what the document is "about", but merely what is different from the other documents. This limitation can be offset with the use of probabilistic topic models, because the words relative to other documents are normalized to a probability of occurrence within a topic despite their relative use.

Content Analysis (CA) as a whole, is particularly useful for identifying patterns and themes within a body of data (Leedy & Ormrod, 2013). Since it is unknown whether or not themes exist in the data, CA is an appropriate inductive methodology to identify and extract themes from the text-description column.

Latent Dirichlet Allocation (LDA) Topic Model

The aforementioned CA tools can help researchers determine what the document is "about" relative to all other documents in the corpus, but these tools do not detect multiple themes within a corpus of documents since they only return the descriptive terms in which describes the document relative to all other documents. Essentially, CA tools return a single theme which is subjective to interpretation of what the terms describe. Thus, a technique is needed to reveal multiple groups that exist naturally in the data without subjectivity.

Topic models are probabilistic models which infer a hidden structure that naturally exists in the text itself. The LDA topic model is the simplest topic model and relies primarily on two principles (Blei, Carin, & Dunson, 2010):

1. Every document is a mixture of topics.
2. Every topic is a mixture of words.

The LDA assumes a generative process where topics are created before the topics (Blei et al., 2010). This assumption is consistent with typical writing styles where the author identifies a topic and then proceeds to use words to add context to the topic. LDA topic models seek to infer the unobserved *hidden structure* that exists in the corpus of documents by using the observed documents. Furthermore, the LDA is an algorithm that seeks to reverse the generative process (Blei et al., 2010). In other words, given the text of multiple documents, what topics are being described?

LDA topic models use the Dirichlet Allocation process to assign a "beta" probability that a token belongs to some unnamed topic relative to all other topics. It is important to note that the beta score is a relative measure and speaks only about the

likelihood of the token belonging to the topic. Based on the aforementioned assumptions, it is possible to see similar mixtures of words (tokens) as well as "documents" that have a similar mixture of topics. This is a distinction between LDA models and classification algorithms, in that the LDA model does not seek to "assign" a token to only one topic, but return tokens-per-topic probabilities. Likewise, LDA returns a mixture of topics per document.

The output of the LDA topic model is essentially, a list of tokens with an associated beta probability ranked in descending order. Drawing on the field of topic recognition (Newman D., Lau J. H., Grieser K., 2010), a 10-token list of the most probable tokens would be adequate to convey the subject of a topic and distinguish one topic from another. Therefore, the top-10 tokens (by beta probability) will be used to describe the topic of which they belong.

LDA topic models are applicable to this problem since contract text descriptions (observed) could be used to infer their inherent thematic nature (unobserved, hidden structure). Thus, the previously hidden structure becomes an organized structure which aligns with Dr. Muir's concept of "sensibly bound pockets of spend" (p. 9, 2014).

One limitation of LDA topic models is the inability to determine what, specifically, the topic is. A list of likely tokens can be mathematically calculated, but the "theme" of membership into the topic is undefined. This limitation is overcome by the use of LA (discussed in a later section).

Another limitation is the number of themes must be established *a priori* in a LDA topic model. The model "fits" the probability of each token to a pre-defined topic. The results will vary based on the number of themes chosen before the model is run. The

aim of this thesis is to discover the themes below a certain level, the very limitation inherent to LDA topic models. To overcome this limitation, four algorithms were used as a guide to determine the mathematically optimal number of themes in each data set.

Four Algorithms

The *a priori* determination of the optimal number of groups is a well-known issue with LDA topic models (Arun, Suresh, Veni Madhavan, & Narasimha Murthy, 2010; Cao, Xia, Li, Zhang, & Tang, 2009; Deveaud, SanJuan, & Bellot, 2014; Griffiths & Steyvers, 2004), but the detailed study of such is beyond the scope of this research. However, a surface-level description of each algorithm is provided to bolster the validity of their use in this thesis.

Arun 2010 & Cao 2009

The two algorithms developed in their respective papers (Arun et al., 2010; Cao et al., 2009) use dissimilarity (as measured by distance) of groups. When the distance is greatest, the groups are "most dissimilar" and the inverse is true as well. These algorithms are particularly useful in the context of this problem since it is dissimilar groups can be identified as distinct themes.

Griffiths 2004 & Deveaud 2014

Griffiths' (2004) Markov chain Monte Carlo algorithm in conjunction with Bayesian inference to determine the optimal number of groups. Hence, the probability of a word given a topic is used to infer the topic given a word over a set number of topics. When the log probability is the highest, the corresponding number of topics (groups) is chosen.

Deveaud's (2014) algorithm is entirely unsupervised and uses a weighting scheme from an LDA topic model output to define the optimal number of groups. In other words, the algorithm "learns" from itself through multiple iterations of model creation.

Again, the intent of these algorithms is to provide a guide for the optimal number of groups within the data set. It is unlikely that all algorithms will identify the same optimal number of groups, but it is possible that they point to an approximate number range of groups. Each algorithm uses a different approach to determine group membership, and the distance measure is normalized on a scale from 0 to 1 (chapter four) to guide LDA topic model input selection.

Percent Deviation

Because the top-10 tokens per topic are used to convey the subject of the topic and distinguish one topic from another, it is possible that the token lists will be very similar (if not identical in closely related topics). Therefore, a measure of "uniqueness" is needed to distinguish similar topics.

Euclidean distance techniques were initially explored as a possible distinguishing method. However, it was observed that the frequency of text descriptions caused the contracts with the most words to be grouped together. In other words, those contracts that had a high proportion of actions, and consequently a high proportion of text descriptions would always form a cluster. These clusters only revealed the high contract actions relative to the rest of the contracts, which was already known.

As mentioned before, TF-IDF seeks to establish what tokens are unique to one group relative to all other groups. The use of TF-IDF as a second-layer post LDA model

was not possible as there was no way to tell how often the tokens occurred in the topic (only the beta probability) after the LDA model was executed.

Given a set vocabulary (tokens in a group), the LDA model output is such that the same tokens are present in every topic, but the beta probabilities are different (if in fact the groups are different). The mean average similar groups' beta probabilities are taken and the tokens which have the largest percent deviation are the tokens that are the "most unique" relative to the group. A token list was created by adding the top three identical tokens in the topics with the top seven (by percent deviation) to make a 10-token list that adequately addresses the similarities of the topic (top three) and the differences of the topic (top seven). The 10-token list was given to subject matter experts to determine what themes were identified.

Latent Semantic Analysis

The aforementioned tools will be used to a word list that can be supported mathematically. However, it is important to understand that domain expertise is needed to increase the validity and practical significance of this methodology. Therefore, the word lists (from each Level-2 category) were distributed to personnel familiar with the data to establish topics (themes). Research has shown that LA increases validity as it provides a "human in the loop" to corroborate findings (Dooley, 2016; Luca, Kleinberg, & Mullainathan, 2016). For this reason, subject-matter experts (SME) were utilized from AFICA to "label" the output of the LDA topic model and percent deviation word lists (chapter four).

Investigative Questions Revisited

Given the information provided in the PSCs and text descriptions, it is possible to answer all of the investigative questions. If the PSCs were treated as "documents" and their associated text descriptions as "words", then TF-IDF analysis could conceivably establish what words describe individual PSCs (IQ 1). The combination of the four algorithms and LDA topic model would mathematically establish not only how many themes are present within each Level-2 categories, but what words are used to describe the themes with a degree of certainty. SMEs could then identify what the themes (topics) are being described by the associated words (IQ 2). Finally, since the PSCs exist in a fixed capacity under each Level-2 category, a hierarchy could be formed to further strategic sourcing initiatives (IQ 3).

Chapter Summary

This chapter discussed the variables used in this thesis and presented a generalized and expanded methodology. In addition, the methodology's relationship with the investigative questions was established. Chapter four discusses the application of the methodology, results, and answers the investigative questions in turn.

IV. Analysis & Results

Chapter Overview

The purpose of this chapter is to present the results of the analysis using the methodology in Chapter three. IT Security will serve as an example of the methodology and the five remaining Level-2 categories (Figure 5) will be summarized.

Data Exploration

In an effort to extract insights from the data, the contract actions were counted and visualized by PSC and Level-2 category (Figure 5). It should be noted that no single PSC appeared in more than one Level-2 category, which suggested a forcing function within the classification system. The AFICA SMEs were unaware of how the PSCs were assigned to the Level-2 category, but it was evident from Figure 5 that a structure naturally existed.

During the next recurring meeting, the AFICA SMEs indicated that while the information in Figure 5 was useful, it was unreliable due to the PSC assignment process. More specifically, in their opinion, the PSCs were not necessarily indicative of "what was bought". When a contract is originated, an analyst reviews contract in detail and assigns the PSC based on the "predominant product or service being purchased" (US GSA Product Service Manual, p. 6, 2015). In other words, if a contract contains multiple products or services, the item with the largest amount of spend "wins" the PSC

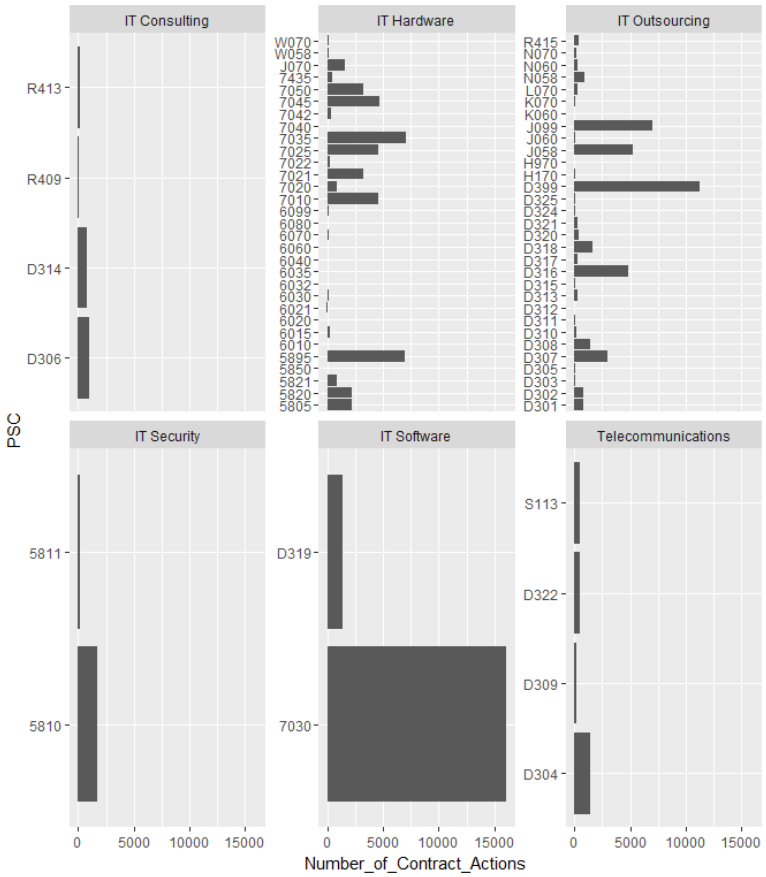


Figure 5. Contract Actions by PSC and Level-2 Category

assignment. The importance of this guidance cannot be overstated because, essentially *the PSCs cannot reliably identify the products and services contained within the associated contract action, only the product or service representing the largest proportion of spend.* Furthermore, once the analyst determines a majority spend item(s), they must make a determination on what PSC "best" matches. Presumably, the analyst would consult the PSC Manual (2015), which is 332 pages, contains hundreds of PSCs and PSC descriptions to make a determination. If PSC codes are somewhat arbitrary, then *what criteria determines a sub-category (IQ 1)?*

Through meetings with the AFICA SMEs, it was determined that AFICA would benefit more from creating groups as an extension of the existing structure (Figure 5) for the following reasons:

1. The temporal length of the data suggests that the PSCs are stable. The data was approximately five years and no PSCs were replicated in more than one Level-2 category.
2. Since the PSCs already existed in some fixed capacity, discovered sub-categories could be related to the existing structure.
3. Level-2 categories and PSC assignment was prevalent in all other Level-1 categories. Therefore, expanding upon the current structure would increase generalizability across other Level-1 categories.

Thus, a definitive description of what the PSC is "about" (using text descriptions) would identify what products and services are contained within each PSC and simultaneously identify what products and services are unique to each PSC (IQ 1). In addition, the issue of "predominant spend PSC labeling" would be mitigated since contract spend is not considered.

Term Frequency (TF)

As mentioned in Chapter Three, TF analysis can provide some context to what the data set is "about". Term Frequency (TF) analysis was conducted on the IT Security subset. The text descriptions were separated by word and stripped of all numbers, special characters, and stop words (i.e. "the", "and", "is", etc.). The top 10 words were returned (Figure 6) with their associated frequency to get a general insight of the words used in the text descriptions.

From the data in Figure 6, one of the top words is "igf". This word is used to identify contracts that must not be performed by a contractor. Hence, Inherently

Government Functions (IGF) appear as IGF::XX::IGF in the raw data. The "XX" portion of this designator is used to identify what type of IGF. For example,

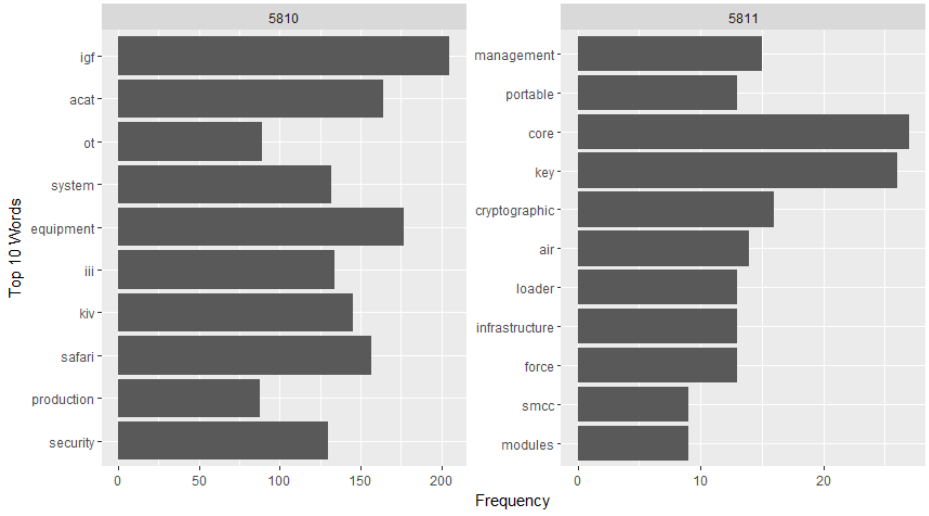


Figure 6. Uni-grams IT Security by Frequency

IGF::CT::IGF designates a contract that must be performed by a federal employee (IGF) that is of a critical (CT) nature. The raw data contains many IGF entries that are associated with the following two-letter designations: (CT) critical, (OT) other, (CL) closely associated, and (CT, CL) a combination of the two. As such, the words "igf", "ct", "cl", and "ot" do not describe specific products and services, but are descriptors of the contract type. Furthermore, these words were considered “domain stop words” as feedback from the AFICA SMEs indicated that the words did not add context to the PSC description. Finally, no other domain stop words were removed from the Level-1 data set due to the time constraints of the research and the risk associated with removal of words that could provide context to the SMEs.

Token Selection

The TF analysis was conducted again with bi-grams and tri-grams in order to increase the level of granularity of the data (Appendix A). Without domain expertise, it was difficult to determine if the words returned were practically significant, due to the various acronyms or abbreviations. However, it was apparent that use of bi-grams added context to associated acronyms. For example, "kiv" was returned in Figure 6 under the PSC 5810. Without domain expertise, this word added little context to the description of products and services contained within the PSC. When the same data set was analyzed using bi-grams (Appendix A), a proximity word was returned that added more context similar to "kiv production". Even though the definition of "kiv" is unknown, its proximal location to "production" added more context. Therefore, bi-grams (in this scenario) added more context than uni-grams. When the TF analysis was conducted with tri-grams (Appendix A), the results were more granular, but required more time to digest the results. Hence, the determination was made to use bi-grams as the basic unit of analysis on all data sets.

IT Security

TF-IDF analysis was conducted on the IT Security subset to determine what bi-grams were unique to the PSCs relative to all other PSCs (Figure 7). The PSC description from the PSC manual (2015) are:

5810 (COMMUNICATIONS SECURITY EQUIPMENT AND COMPONENTS)

5811 (OTHER CRYPTOLOGIC EQUIPMENT AND COMPONENTS)

The bi-grams found in each PSC suggest the products are closely related to the descriptions, but it is difficult to make such a statement without domain expertise. The two PSCs (5810, 5811) appear to contain the top-10 words that match the PSC descriptions (at least the predominant spend assignments). However, since IQ 1 seeks to

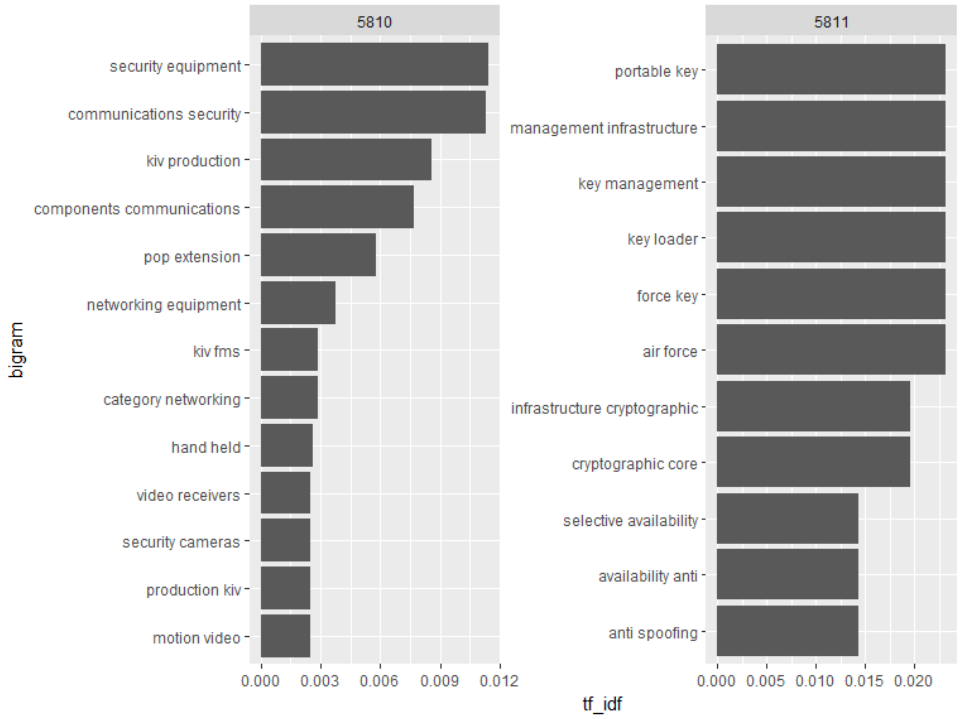


Figure 7. IT Security Bi-grams by TF-IDF Weight

determine what criteria is in a sub-category, it is necessary to investigate the existence of categories beyond the PSC. The four algorithms were applied to the IT Security data to

determine if any other groups existed. The algorithms converge on either four or five optimal topics (Figure 8). Four topics were chosen as the input to the LDA topic model.

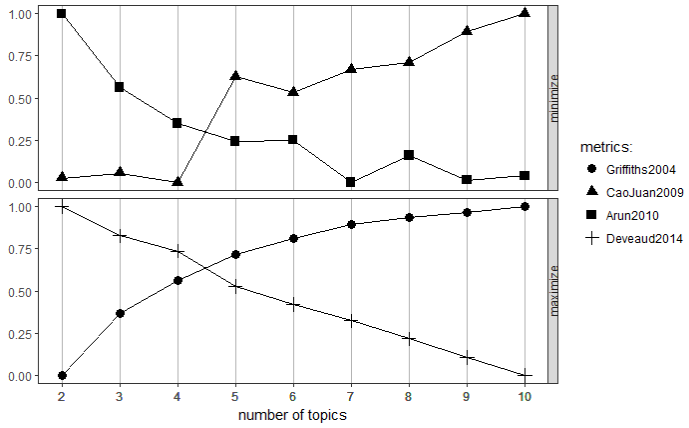


Figure 8. Four-Algorithm Optimal Topic (IT Security)

The lower number was chosen for two reasons. First, the analysis was to be replicated across all six Level-2 categories. Under the assumption that the product "groups" would

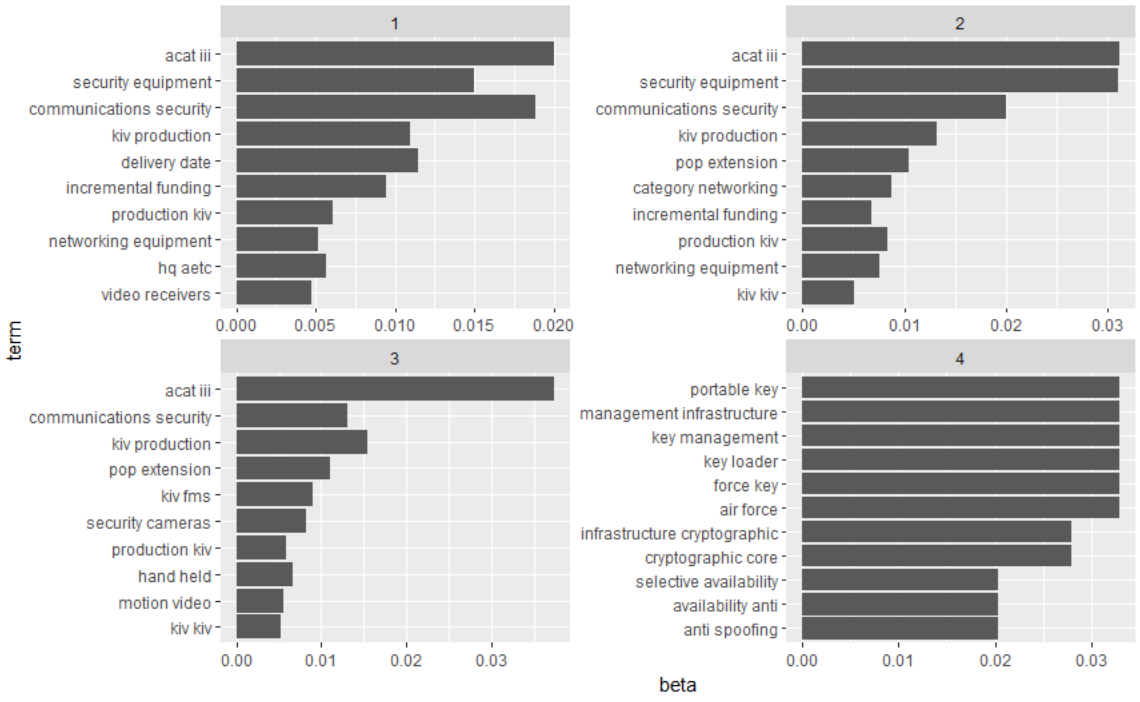


Figure 9. LDA Output (Unnamed Topics) IT Security

be managed by a portfolio manager, it would benefit the manager to have a smaller

number of groups to monitor. Second, the tradeoff between four or five topics was minimal (if not zero). Therefore, in all scenarios where the number range of groups appeared to be mathematically equivalent, and no other information about the number of groups was available, the lower number was chosen as the "optimal" number of groups for the LDA model.

The optimal number was then used to set the number of topics for the LDA topic model (Figure 9). The model suggests topic four is distinct due its exact representation of the TF-IDF for PSC 5811. Furthermore, when the LDA model was executed using a higher numbers of groups (up to 20) as an input, topic four always remained intact and the model further divided topics one through three. Topics one, two, and three contained many of the same bi-grams so the percent deviation was calculated (Table 3) and a list of bi-grams was compiled (Table 4) into a word list that would be distributed to the SMEs for topic assignment. The bi-grams in topic four were left untouched due to the assumption that it was a distinct group relative to the other three. Collectively, the word lists represent the actual terms that best describe the four mathematically supported unnamed topics (Table 4). Topic four is most likely PSC 5811, and the other three topics are sub-categories of PSC 5810.

Table 3. Percent Deviation Similar Groups (IT Security)

Topic 1	% Dev	Topic 2	% Dev	Topic 3	% Dev
adapter plate	198.3278	portable avenger	182.7984	predator elite	187.1427
model kg	198.0879	gfp correction	180.4616	capability study	179.7542
plate av	198.0339	diesel generator	172.1421	additional days	179.2613
noun core	197.9061	ldc audio	171.3365	clin overrun	170.9577
lightning strike	194.2480	lmr motorola	169.5146	linux production	170.3135
lightning strikes	193.7888	recaro seat	169.4502	apx digital	168.8644
av conference	193.6806	imaging technology	167.4894	afb option	166.3707
spoofing modules	191.1107	clin transfer	166.8580	mountain home	164.5851
mod clin	190.6411	extension clin	166.1154	acts crypto	164.4502
availability anti	190.4911	auto acquire	164.3151	communication equip	162.3807

Table 4. IT Security Topic Word List

Topic 1	Topic 2	Topic 3	Topic 4 (PSC 5811)
acat iii	acat iii	acat iii	air force
communications security	communications security	communications security	force key
security equipment	security equipment	kiv production	key loader
adapter plate	portable avenger	predator elite	key management
model kg	gfp correction	capability study	management infrastructure
plate av	diesel generator	additional days	portable key
noun core	ldc audio	clin overrun	cryptographic core
lightning strike	lmr motorola	linux production	infrastructure cryptographic
lightning strikes	recaro seat	apx digital	anti spoofing
av conference	imaging technology	afb option	availability anti

Level-2 Categories

The methodology was applied to the remaining five Level-2 categories. In categories where the LDA output showed distinct topics, the percent deviation was not applied. The results are summarized in Table 5 and can be viewed in Appendix A.

Table 5. Summary of Results

Level 2 Category	PSCs	Optimal Topics	Percent Deviation
IT Security	2	4	Yes
IT Consulting	4	3	Yes
IT Hardware	31	6	No
IT Outsourcing	31	7	No
IT Software	2	5	Yes
Telecommunications	4	4	Yes

Latent Semantic Analysis

As discussed in Chapter three, Latent Semantic Analysis (LA) is useful to validate findings as it provides a "human in the loop" (Dooley, 2016; Luca et al., 2016). Furthermore, topic designation by those familiar with the contract vocabulary would increase practical significance and reduce the perception of bias. For these reasons, SMEs were tasked assign the topics from the word lists created in the previous section (Table 4).

The word lists were sent via email to the IT Business Analytics Office located at

Table 6. Summarized Topic Descriptions

IT Security	**IT Consulting**
Topic 1 : Physical, Equipment	*Topic 1: Training, Mission, Sustainment, Development*
Topic 2 : Maintenance, Continuity	*Topic 2: Compliance, Sustainment*
Topic 3: Data, Platform	*Topic 3: Integration, Infrastructure, Funds, Sustainment*
Topic 4: Encryption, Authentication, Access	*Topic 4: Sustainment*
IT Hardware	
Topic 1: Computer, Asset, Data	
Topic 2: Network, Infrastructure	
Topic 3: Peripheral, Storage, Equipment	
Topic 4: Configuration, Communication	
Topic 5: Infrastructure, Data, Voice, Devices	
Topic 6: Workstation, Computer Components, Infrastructure, Sustainment	
IT Outsourcing	
Topic 1: Physical, Plant, Maintenance	
Topic 2: Engineering, Sustainment, Life Cycle	
Topic 3: Communication	
Topic 4: Communication, Telecommunication, Wideband	
Topic 5: Maintenance, Telecommunication, Configuration	
Topic 6: Telecommunication, Installation, Command & Control	
Topic 7: Network, Telecommunications, Infrastructure, Solutions	
IT Software	**Telecommunications**
Topic 1: Configuration, Infrastructure	*Topic 1: Television, Intranet*
Topic 2: Maintenance, Sustainment	*Topic 2: Television, Infrastructure*
Topic 3: License, End User, Governance	*Topic 3: Internet, Digital, Network, Access*
Topic 4: Network, License, Verification	*Topic 4: Cellular, Phased*
Topic 5: License, Renewals, Subscription	

Maxwell-Gunter Air Force Base with directions to annotate the topic that the words likely described (Appendix B). The responses were limited to three words or less and the results of the SME topic assignments can be found in Appendix C.

Five SME topic assignment sheets for each Level-2 category were returned (Appendix C) from five different SMEs familiar with the contracting vocabulary. The results were summarized by extracting specific nouns or verbs that were distinct relative to the topic. For example, under the topic IT Security, the noun "security" does little to describe the topic as all topics are under the subject "security". Therefore, only nouns and verbs that were descriptive of the topics were used to summarize topic assignments (Table 6).

The topic descriptions (Table 6) were extracted with the inclusion of product and service PSCs in mind. In other words, since IT Security only contained product PSCs, the topics are presumably descriptive of products. For the same rationale, IT Consulting, IT Outsourcing, and Telecommunications' topics are presumably descriptive of service PSCs. IT Hardware and IT Software contained both PSC types which suggests the topics could be descriptive of either products or services.

Summary

This chapter applied the methodology to IT Security and summarized the remaining five categories. Chapter five provides conclusions and significance of the research, and recommends action and future research.

V. Conclusions and Recommendations

Chapter Overview

The following chapter discusses findings and significance of the research. In addition, recommendations for immediate action and future research will be made.

Findings

IQ 1 – What criteria determines a sub-category? The TF-IDF analysis for each PSC effectively establishes criteria for sub-categories because the "preponderance of spend" PSC assignment is mitigated by the words used to describe the contract action. Overall, the goods identified in the product PSCs appeared to "align" to the PSC descriptions (Appendix A). Admittedly, some tokens were undecipherable due to the complexity of the acronyms or missing context, but a complete list of TF-IDF-weighted

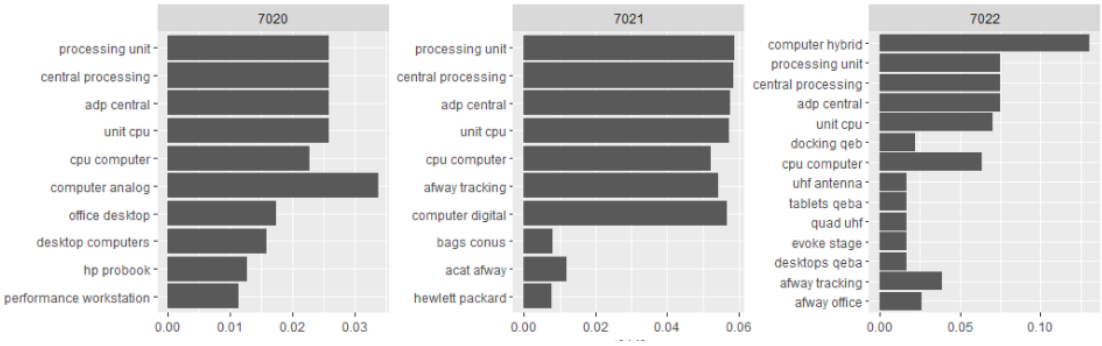


Figure 11. Computer PSCs Weighted by TF-IDF

tokens would likely provide a SME with a comprehensive list of words that are descriptive of goods (Appendix A). For example, PSCs 7020, 7021, and 7022 are described as Information Technology Central Processing Unit (analog, digital, and hybrid respectively) in the PSC Manual (2015). The TF-IDF (Figure 11) indicate computers (office/desktop/workstation) and tablets are prevalent in all three PSCs and

by definition are distinct from other PSCs. In addition, the tokens "processing unit", "central processing", and "adp central" are tokens used to describe the PSC groupings (Group 70) and are redundant.

IQ 2 – How will themes be identified? The intent of the SME topic assignment responses was to identify "themes" (topic labels) within the Level-2 categories. The summarized responses were more representative of "what" topics in product-centric topics and "why" descriptions in service-centric topics. For example, the topics identified in the four product-specific IT Security groups (Table 5) were all descriptive of "what" types of security items (i.e. "physical", "data", "maintenance" or "encryption"). Conversely, topics identified in service-centric groups (IT Consulting) were descriptive of "why" the service was acquired (i.e. "training; mission; sustainment;

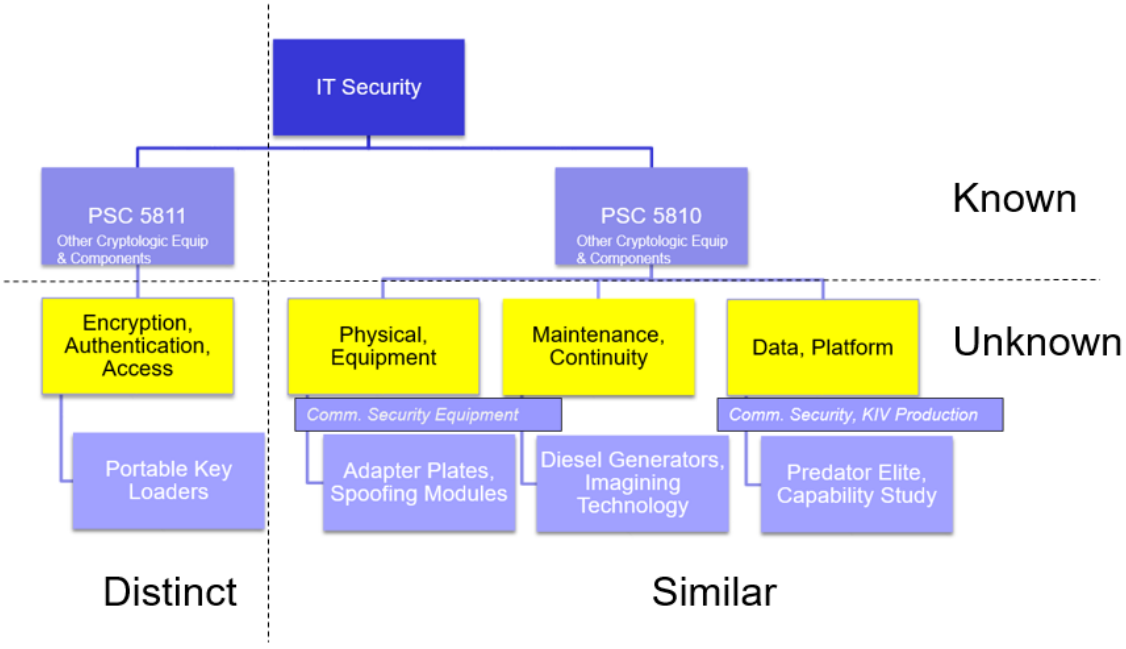


Figure 12. Example Hierarchal Breakdown of IT Security

development", compliance; sustainment", "integration; infrastructure; sustainment", and "sustainment").

IQ 3 – How will themes be useful to AFICA? The thematic topic assignment of products and services offers insights into historical contract data that was previously unidentified. Furthermore, the hidden structure of the data becomes apparent (Figure 12) and is extended beyond the partial PSC structure. The themes (highlighted) and their subsequent products/services are sensibly bound pockets linked to PSCs (which already have associated spend data). In this sense, the hierarchy provides a top-down view of contract data from Level-1 category down to products and services (with thematic membership).

Limitations

The methodology is generalizable to other organizations that seek to categorize historical purchase contract data and reproducible with the programming language (Appendix A), but it was limited by the lack of domain expertise and implementation of SME topic assignment sheets. For clarity, it is important to note that the AFICA SMEs were personnel that attended meetings and had a general knowledge of the data. Topic assignment SMEs were contract analysts that worked at the IT Business Analytics Office (ITBAO) at Maxwell-Gunter AFB, AL and had in-depth knowledge of the data and contracting language.

Although bi-monthly meetings were scheduled with AFICA SMEs, there were unforeseen personnel changes which impacted the ability to understand domain-specific tokens. The selection of bi-grams was made independent of SME inputs. As such, it

was not clear if token selection was optimal. Uni-grams were not used because the abbreviations and acronyms returned were not decipherable without domain expertise. It's possible that uni-grams would have provided more insight since they would have been independent of proximal words. In other words, the results were contingent on words that were situated in the text next to each other. Thus, a different token size would likely have yielded different results, but it is unknown which token size is the most useful without domain expertise.

The number SME topic assignment responses (Appendix B) were also problematic. First, due to time constraints and specificity of expertise required, there were only five respondents available, which came from the same unit. It would be desirable to have more respondents, but it was unclear how many SMEs were familiar with the IT contract purchase data (or for that matter, how many existed). The ITBAO (Maxwell-Gunter AFB, AL) was in the process of manually creating a similar hierarchy with similar data used in this research. The respondents were not only familiar with the data, but would likely be the beneficiaries of any insights gleaned from the research. Thus, the small number of respondents is partially offset by the SME's explicit familiarity with the data.

Second, the topic assignment sheets asked the SMEs to assign a "topic" from the aforementioned word lists (Appendix B). The interpretation of the word "topic" could have affected the responses from the ITBAO SMEs since it was unknown to them whether the words were descriptive of a product or service. The respondents were only given the Level-2 category from which the words were extracted via the LDA model/percent deviation (if required). Therefore, the "topic" they identified may have

been an attempt to encompass both products and services. This might explain why the SMEs used "why" topic assignments for service-centric word lists.

Finally, due to geographic separation the word lists were administered without any oversight. If the respondents had any questions about their task, they did not receive any clarification other than the directions provided on the sheet (Appendix B). It is possible that the respondents discussed the topic assignment sheets prior to topic assignment which would affect their independent assessment of the word lists. However, the intent of this research is to offer an empirical methodology that includes domain expertise and external validation. As such, the implementation of topic assignment sheets may not be necessary if the agency (or organization) has a collaborative design in regard to analysis and domain expertise.

Discussion

The overarching intent of this thesis was to provide an empirical methodology for AFICA to categorize a historical list of IT-related contracts. Although the data is specific to the AFICA, the methodology is generalizable to any large organization that possesses purchase contract data. The sample hierarchy (Figure 12) is specific to AFICA, but can be viewed as a proxy for any semi-structured purchase data. Furthermore, the internal validity of this research is bolstered as the products aligned with the themes and PSCs. In other words, the products and services found beneath each tier in the hierarchy appeared to "align". Moreover, when the four algorithms were applied to IT Hardware and IT Outsourcing (31 PSCs each) they converged on six and seven topics, respectively. This is important because the PSC groups (Group 60XX,

70XX, etc.) were equivalent without any manipulation (Appendix A). For this reason, the optimal number of topics for the LDA model were set to these numbers vice the lower number in other Level-2 categories. This suggests that although the PSCs are assigned via predominant spend, the words used in the text description are distinct enough to be detected by the LDA model.

Significance of Research

This research provides a foundational methodology to gain insights on historic contract data text descriptions. Moreover, the hidden structure of the Level-2 categories enables AFICA decision makers to understand not only what items are present in discovered categories, but what makes them unique compared to other categories. Furthermore, the identified themes provide context into the functional purpose of the Level-2 categories.

This thesis is a significant contribution to text mining literature. The unique military contract language identified potential pitfalls in text mining analysis that would not have been known otherwise. Although the literature review cannot be considered exhaustive, no other text mining study on military purchase data was found.

Recommendations for Action

The following recommendations are offered to strengthen AFICA strategic sourcing initiatives. First, the purpose of the "text description" field should be well defined to analysts. The contract text descriptions often contained words that were descriptive of an analyst action and NOT the products or services contained within. It is unclear whether or not the "text description" field was used for this purpose.

Furthermore, the words used to describe the PSC were often found in the text descriptions which diluted the descriptive information available. Assuming the intent of the text description field is to describe the items or services, there should be more words that are descriptive of a product or service and less words about obligations, contract size, government affiliations, installations, etc. An additional field for internal communication would help keep descriptions and internal communication data separate.

Second, group-level PSCs should be assigned to contracts with eclectic products and services. The "preponderance" of spend allocation introduces uncertainty into the PSC and diminishes its purpose. However, it is likely that the analyst is able to assign a PSC at the group level much quicker and more accurately than searching the PSC manual for a more granular PSC. Furthermore, if products and services are within the same PSC grouping, it would be unnecessary to determine the predominant spend amount to determine which PSC "wins" the assignment. This could have vast implications to decision makers as the PSC assignment process could be drastically reduced and ultimately "free-up" resources for other business activities. If a granular level of detail is required, the agency could use techniques contained in this thesis to identify specific products and services within the group. In addition, the rapidly-evolving nature of technology presents an extremely difficult task of continually updating PSC definitions to match the products and services. PSC assignment at the group level (60XX, 70XX, D3XX) would offer some buffer against antiquated technology (hence antiquated descriptive words).

Third, the assumption that PSCs are not representative of the contracts should be dispelled. The contract data (at least in IT) indicate the PSC designations are aligned

with the products and services contained within. Furthermore, the groups identified in IT Hardware and IT Outsourcing were representative of the pre-determined PSC groupings.

Fourth, the remaining Level-2 analysis (Appendix A) should be reviewed by SMEs to create a structure similar to Figure 12. The data in IT Security coincidentally contained words that were naively interpretable. However, the remaining categories would be better translated by SMEs.

Finally, and most importantly, AFICA decision makers should consider collaborative approaches to contract analysis. SMEs are the experts at the content of the data. Analysts can expertly apply quantitative techniques to data sets. Management should harness the synergistic effect of collaboratively analyzing contract data by co-locating SMEs and analysts or at least merging analysis functions with domain expertise.

Recommendations for Future Research

It is clear that much insight can be gained through text mining analysis of historical contract data. A well-defined structure aligns and enables Resource Orchestration Theory (ROT) principles of structuring, bundling, and leveraging. In fact, it is arguable that the discovery of hidden structures serves as a catalyst to ROT principles. A robust visualization of hidden structures within the data enables efficient acquisition processes by identifying products and services in multiple Level-2 categories (structuring), which could be redundant. Moreover, the minor improvements in acquisition processes translates into an improved capability (bundling). Furthermore,

the structure itself fosters strategic sourcing initiatives via category management and postures AFICA to exploit market opportunities (leveraging). To achieve this, the following recommendations for future research are offered.

First, future analysis should leverage the PSCs' product- and service-types. The realization that PSCs existed in two different capacities occurred late in the analysis and could not be separated due to time constraints. TF-IDF analysis would benefit from comparing PSC-types relative to each other vice all PSCs. In this thesis, descriptive nouns were used with products and descriptive verbs were used with activities (services). TF-IDF analysis combined the two and identified words (both nouns and verbs) unique to each PSC relative to all PSCs. It would likely be insightful to compare "apples with apples" to see how the analysis changed (if at all). Furthermore, the pre-existing PSC groups were not collapsed (60XX, 70XX, D3XX) in this thesis. There may be more insights from treating the PSC "groups" as a unit of analysis vice individual PSCs.

Second, future research should focus on PSCs' evolution over time. PSCs are deleted, merged, or updated with the publication of new PSC manuals. According to the PSC manual (2015) PSC S113 was merged with D304. There may be other revisions that explain why some products and services occur in multiple PSCs. A temporal analysis may well identify PSCs that are volatile or stable, which could serve as an indicator of rapid technological change.

Finally, and most importantly, the methodology would be well suited for actual contracts as opposed to the contract actions contained within the data. It is likely that original contracts would yield different insights than contract actions. The inclusion of

language to describe actions like obligation or de-obligation of funds was a barrier to extracting precise context.

Summary

AFICA is in a key position to reduce enterprise-wide spend and shape consumption (Muir et al., 2014). If the Air Force is to retain its competitive advantage, it needs to structure, leverage, and bundle the resources it possesses. This research will help the Air Force remain in the top position by discovering value in the data we already possess. The importance of IT acquisition has been acknowledged by the CSAF; "We're looking at a holistic view on how to acquire information technology because it's so central to our future as we look at networking capabilities together" (Serbu, 2017).

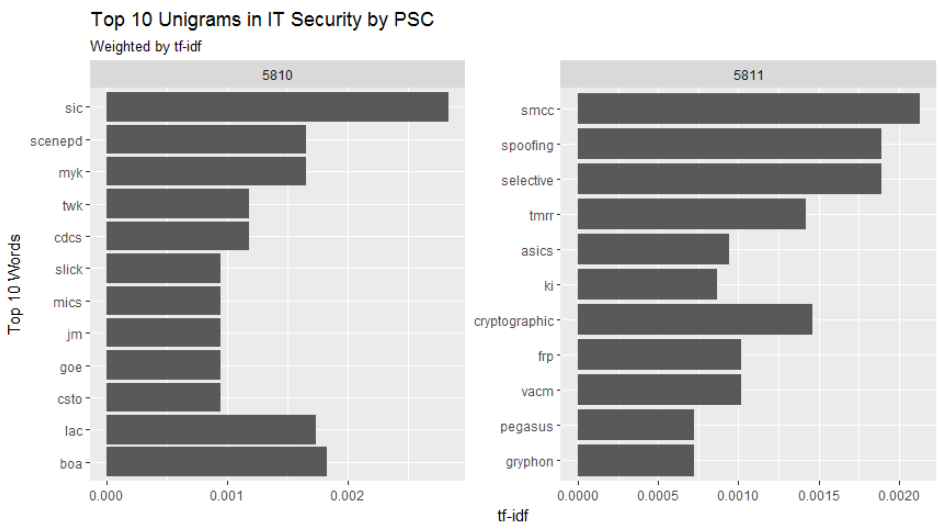
This thesis proposed an empirical methodology for the categorization of IT contracts to facilitate AFICA strategic sourcing initiatives. Through the use of Content Analysis, LDA models, and Latent Semantic Analysis tools, sub-category creation was achieved. Furthermore, an example top-down hierarchy of one Level-2 category was developed.

Appendix A. R Programming Code

IT Security

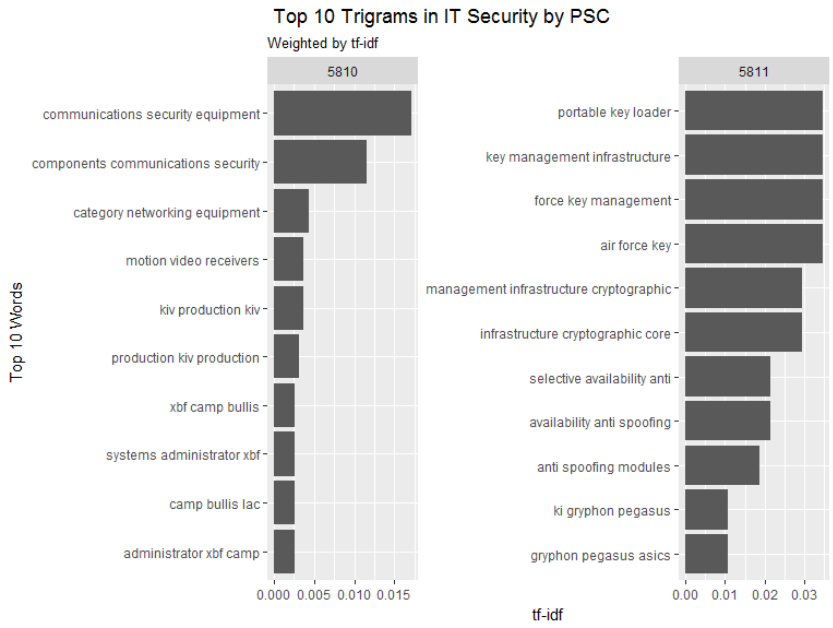
```
library(tidyverse)
library(tidytext)
library(topicmodels)
library(lubridate)
library(ldatuning)
library(magrittr)

## Selects and renames variables from a larger data frame.

itdata <- read_csv("IT_FPDSNG.csv")
relevant_itdata <- itdata[, c(1, 2, 24, 124, 142, 144)]
colnames(relevant_itdata)[c(1, 2, 3)] <- c("trans_ID", "spend",
"text_describe")

## Extracted the first 13 characters from transaction ID variable.
relevant_itdata %>%
  mutate(trans_ID = substr(trans_ID, start = 1, stop = 13)) ->
relevant_itdata

## Shows the transaction count by PSC within the Level-2 Category
(Figure 5)

relevant_itdata %>%
  group_by(PSC, lvl_2_category) %>%
  summarise(Number_of_Contract_Actions = n()) %>%
  ggplot(aes(x = PSC, y = Number_of_Contract_Actions, fill = NULL)) +
    geom_bar(stat = "identity") +
    facet_wrap(~ lvl_2_category, nrow = 2, scales = "free_y", shrink =
TRUE) +
    coord_flip() +
    theme(legend.position = "none") +
    ggtitle("Number of PSCs in Pre-existing Category")

## Non-value added terms
common_terms <- tibble(word = c("ot", "ct", "cl", "igf"))

## Clean and unnest tokens and take the top 10 (by count) terms

na.omit(relevant_itdata) %>%
  mutate(text_describe = str_replace_all(text_describe, pattern = "[0-
9]", replacement = "")) %>%
  group_by(PSC, lvl_2_category ) %>%
  unnest_tokens(word, text_describe) %>%
  dplyr::count(word, sort = TRUE) %>%
  anti_join(stop_words) %>%
  #anti_join(common_terms) %>%
  top_n(10) %>%
  ungroup() -> top_clean_terms_PSC
```

```
## Filter by IT Security and see top 10 words by PSC (Figure 6)

top_clean_terms_PSC %>%
  mutate(word = reorder(word, n)) %>%
  filter(lvl_2_category == "IT Security") %>%
  ggplot(aes(word, n)) +
    geom_bar(stat = "identity") +
    facet_wrap(~ PSC, scales = "free") +
    labs(x = "Top 10 Words", y = "Frequency")+
    coord_flip() +
    theme(legend.position = "none")


## TF-IDF with unigrams

na.omit(relevant_itdata) %>%
  mutate(text_describe = str_replace_all(text_describe, pattern = "[0-
9]", replacement = "")) %>%
  group_by(lvl_2_category, PSC) %>%
  unnest_tokens(word, text_describe) %>%
  dplyr::count(word, sort = TRUE) %>%
  anti_join(stop_words) %>%
  anti_join(common_terms) %>%
  ungroup() -> clean_terms

## Plot
clean_terms %>%
  bind_tf_idf(word, lvl_2_category, n) %>%
  arrange(desc(tf_idf)) %>%
  mutate(word = reorder(word, tf_idf)) %>%
  filter(lvl_2_category == "IT Security") %>%
  group_by(PSC) %>%
  top_n(10, wt = tf_idf) %>%
  ungroup() %>%
ggplot(aes(word, tf_idf)) +
    geom_bar(stat = "identity") +
    facet_wrap(~ PSC, scales = "free") +
    labs(x = "Top 10 Words", y = "tf-idf",
         title = "Top 10 Unigrams in IT Security by PSC",
         subtitle = "Weighted by tf-idf") +
    coord_flip() +
    theme(legend.position = "none")

## TF-IDF with bigrams
```



```
na.omit(relevant_itdata) %>%
  mutate(text_describe = str_replace_all(text_describe,
         pattern = "[0-9]", replacement = "")) %>%
  group_by(lvl_2_category, PSC) %>%
  filter(lvl_2_category == "IT Security") %>%
  unnest_tokens(bigram, text_describe, token = "ngrams", n = 2) %>%
  separate(bigram, c("word1", "word2"), sep = " ") %>%
  filter(!word1 %in% stop_words$word, !word2 %in% stop_words$word) %>%
  filter(!word1 %in% common_terms$word, !word2 %in% common_terms$word)
%>%
  unite("bigram", c(word1, word2), sep = " ") %>%
  dplyr::count(PSC, bigram, sort = TRUE) %>%
  ungroup() -> clean_ITSecurity_bigram

## Plot (Figure 7.)

clean_ITSecurity_bigram %>%
  bind_tf_idf(bigram, PSC, n) %>%
  arrange(desc(tf_idf)) %>%
  mutate(bigram = reorder(bigram, tf_idf)) %>%
  group_by(PSC) %>%
  top_n(10, wt = tf_idf) %>%
  ungroup() %>%
ggplot(aes(bigram, tf_idf)) +
    geom_bar(stat = "identity") +
    facet_wrap(~ PSC, ncol = 2, scales = "free") +
    labs(x = "Top 10 Words", y = "tf-idf",
         title = "Top 10 Bigrams in IT Security by PSC",
         subtitle = "Weighted by tf-idf") +
    coord_flip() +

    theme(legend.position = "none")

## TF-IDF with trigrams

na.omit(relevant_itdata) %>%
  mutate(text_describe = str_replace_all(text_describe,
           pattern = "[0-9]", replacement = "")) %>%
  group_by(lvl_2_category, PSC) %>%
  filter(lvl_2_category == "IT Security" ) %>%
  unnest_tokens(trigram, text_describe, token = "ngrams", n = 3) %>%
  separate(trigram, c("word1", "word2", "word3"), sep = " ") %>%
  filter(!word1 %in% stop_words$word, !word2 %in% stop_words$word,
!word3 %in% stop_words$word) %>%
  filter(!word1 %in% common_terms$word, !word2 %in% common_terms$word,
!word3 %in% common_terms$word) %>%
  unite("trigram", c(word1, word2, word3), sep = " ") %>%
  count(PSC, trigram, sort = TRUE) %>%
  ungroup() -> clean_trigram

clean_trigram %>%
  bind_tf_idf(trigram, PSC, n) %>%
  arrange(desc(tf_idf)) %>%
  mutate(trigram = reorder(trigram, tf_idf)) %>%
  group_by(PSC) %>%
```

```
  top_n(10, wt = tf_idf) %>%
  ungroup() %>%
ggplot(aes(trigram, tf_idf)) +
    geom_bar(stat = "identity") +
    facet_wrap(~ PSC, ncol = 2, scales = "free") +
    labs(x = "Top 10 Words", y = "tf-idf",
         title = " Top 10 Trigrams in IT Security by PSC",
         subtitle = "Weighted by tf-idf") +
    coord_flip() +
    theme(legend.position = "none")
```




```
## Document Term Matrix

cast_dtm(clean_ITSecurity_bigram, PSC, bigram, n) ->

clean_ITSecurity_bigram_dtm

## Optimal Number of Topics

result_sec <- FindTopicsNumber(
  clean_ITSecurity_bigram_dtm,
  topics = seq(from = 2, to = 10, by = 1),
  metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010",
"Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 1234),
  mc.cores = 2L, #make sure this is appropriate number of cores you
wish to use
  verbose = TRUE)

## (Figure 8)
FindTopicsNumber_plot(result_sec)

## LDA Model
```

```
it_security_lda <- LDA(clean_ITSecurity_bigram_dtm, k = 4, control =
list(seed = 1234))

it_security_topics <- tidy(it_security_lda, matrix = "beta")

## Top 10 words by topic

top_it_security_topics <- it_security_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

## (Figure 9)
top_it_security_topics %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  labs(x = "Top 10 Words", y = "beta",
         title = "Top 10 Bigrams by LDA",
         subtitle = "Weighted by Term Frequency") +
  coord_flip()

## Percent Deviation Topic 1

var_sec <- it_security_topics %>%
  filter(topic < "4") %>%
  spread(topic, beta)
colnames(var_sec)[c(2, 3, 4)] <- c("t1","t2","t3")

var_sec %>%
  mutate(mean = (t1+t2+t3)/3, avdev_t1 = (t1 - mean), avdev_t2 = (t2 -
mean), avdev_t3 = (t3 - mean)) -> var_sec

var_sec %>%
  mutate(per_t1 = (avdev_t1/mean) * 100, per_t2 = (avdev_t2/mean) *
100, per_t3 = (avdev_t3/mean) * 100) -> var_sec

sec_terms_t1 <- var_sec %>%
  gather("per_t1", "per_t2", "per_t3", key = "topic", value =
"percent_dev") %>%
  group_by(topic) %>%
  top_n(10, percent_dev) %>%
  ungroup() %>%
  arrange(desc(percent_dev)) %>%
  filter(topic == "per_t1")



sec_terms_t1[1:10,c("term", "percent_dev"), drop=FALSE]
## # A tibble: 10 x 2
##                  term percent_dev
##                <chr>       <dbl>
```

```
##  1     adapter plate    198.3278
##  2          model kg    198.0879
##  3          plate av    198.0339
##  4         noun core    197.9061
##  5  lightning strike    194.2480
##  6 lightning strikes    193.7888
##  7     av conference    193.6806
##  8  spoofing modules    191.1107
##  9          mod clin    190.6411
## 10 availability anti    190.4911
```

```
## Percent Deviation Topic 2

sec_terms_t2 <- var_sec %>%
  gather("per_t1", "per_t2", "per_t3", key = "topic", value =
"percent_dev") %>%
  group_by(topic) %>%
  top_n(10, percent_dev) %>%
  ungroup() %>%
  arrange(desc(percent_dev)) %>%
  filter(topic == "per_t2")


sec_terms_t2[1:10,c("term", "percent_dev"), drop=FALSE]
```

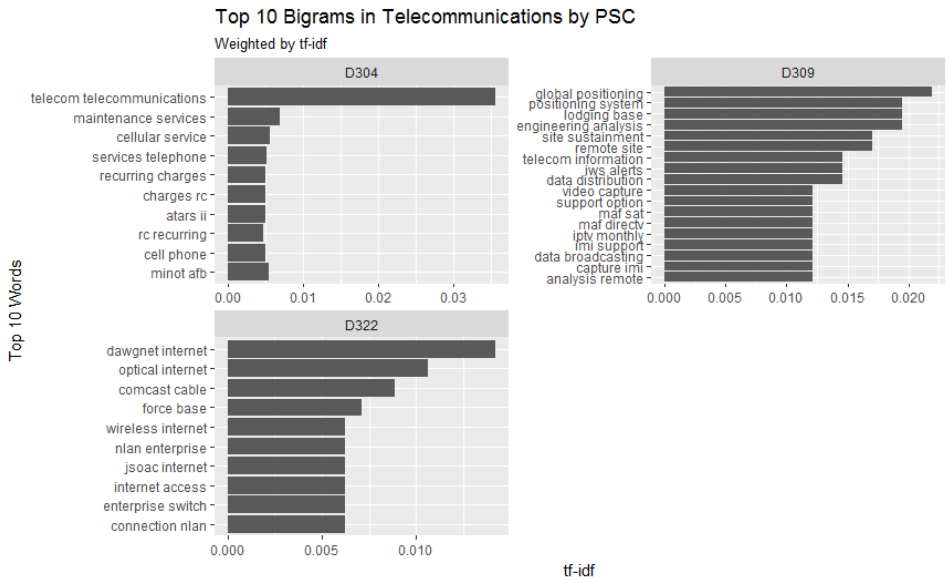
```
## # A tibble: 10 x 2
##                  term percent_dev
##                 <chr>       <dbl>
##  1   portable avenger    182.7984
##  2     gfp correction    180.4616
##  3   diesel generator    172.1421
##  4          ldc audio    171.3365
##  5       lmr motorola    169.5146
##  6        recaro seat    169.4502
##  7 imaging technology    167.4894
##  8      clin transfer    166.8580
##  9     extension clin    166.1154
## 10       auto acquire    164.3151
```

```
## Percent Deviation Topic 3

sec_terms_t3 <- var_sec %>%
  gather("per_t1", "per_t2", "per_t3", key = "topic", value =
"percent_dev") %>%
  group_by(topic) %>%
  top_n(10, percent_dev) %>%
  ungroup() %>%
  arrange(desc(percent_dev)) %>%
  filter(topic == "per_t3")

sec_terms_t3[1:10,c("term", "percent_dev"), drop=FALSE]
## # A tibble: 10 x 2
##                   term percent_dev
```

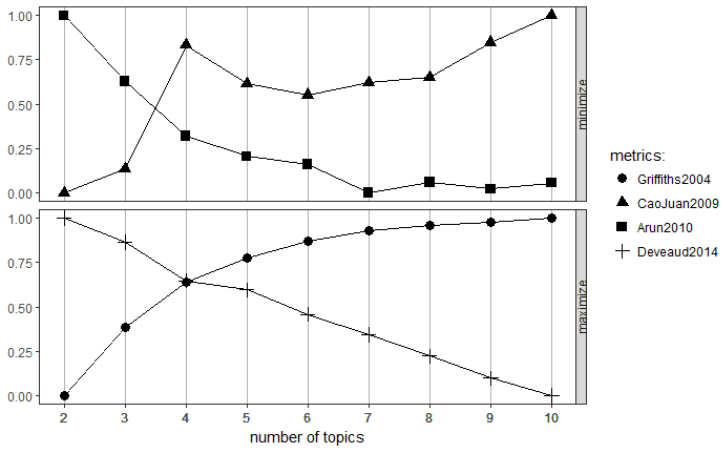
```
##                   <chr>       <dbl>
##  1      predator elite    187.1427
##  2    capability study    179.7542
##  3     additional days    179.2613
##  4        clin overrun    170.9577
##  5    linux production    170.3135
##  6         apx digital    168.8644
##  7          afb option    166.3707
##  8       mountain home    164.5851
##  9         acts crypto    164.4502
## 10 communication equip    162.3807
```

Telecommunications

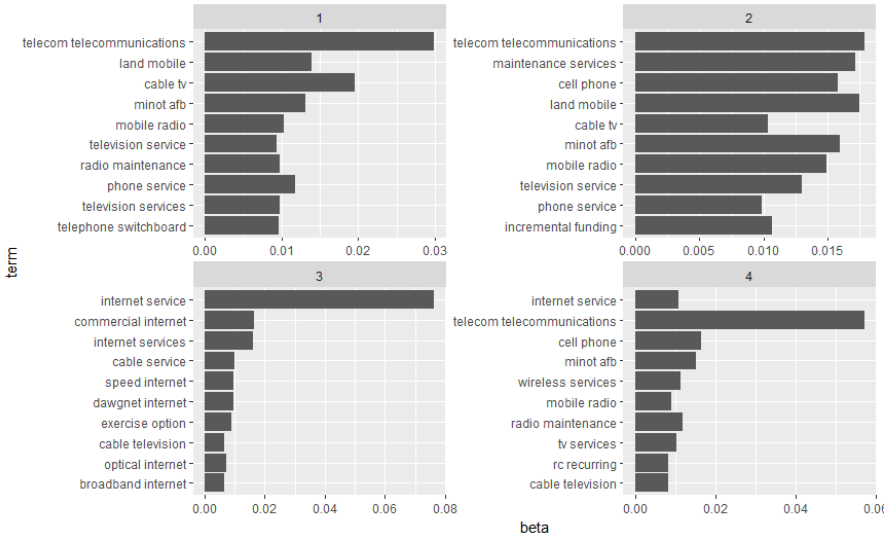
```
## TF-IDF

na.omit(relevant_itdata) %>%
  mutate(text_describe = str_replace_all(text_describe,
          pattern = "[0-9]", replacement = "")) %>%
  group_by(lvl_2_category, PSC) %>%
  filter(lvl_2_category == "Telecommunications") %>%
  unnest_tokens(bigram, text_describe, token = "ngrams", n = 2) %>%
  separate(bigram, c("word1", "word2"), sep = " ") %>%
  filter(!word1 %in% stop_words$word, !word2 %in% stop_words$word) %>%
  filter(!word1 %in% common_terms$word,
         !word2 %in% common_terms$word) %>%
  unite("bigram", c(word1, word2), sep = " ") %>%
  count(PSC, bigram, sort = TRUE) %>%
  ungroup() -> clean_Telecom_bigram

## Merge S113 and D304 (PSC manual Page 319)
clean_Telecom_bigram$PSC[clean_Telecom_bigram$PSC == "S113"] <- "D304"

clean_Telecom_bigram %>%
  bind_tf_idf(bigram, PSC, n) %>%
  arrange(desc(tf_idf)) %>%
  mutate(bigram = reorder(bigram, tf_idf)) %>%
  group_by(PSC) %>%
  top_n(10, wt = tf_idf) %>%
  ungroup() %>%
ggplot(aes(bigram, tf_idf)) +
    geom_bar(stat = "identity") +
    facet_wrap(~ PSC, ncol = 2, scales = "free") +
    labs(x = "Top 10 Words", y = "tf-idf",
         title = "Top 10 Bigrams in Telecommunications by PSC",
         subtitle = "Weighted by tf-idf") +
    coord_flip() +
    theme(legend.position = "none")
```




```
## Create Document Term Matrix

cast_dtm(clean_Telecom_bigram, PSC, bigram, n) ->
clean_Telecom_bigram_dtm

## Optimal Number of Topics

result_telecom <- FindTopicsNumber(
  clean_Telecom_bigram_dtm,
  topics = seq(from = 2, to = 10, by = 1),
  metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010",
"Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 1234),
  mc.cores = 2L,
  verbose = TRUE
)
FindTopicsNumber_plot(result_telecom)
```



```
## LDA Model

it_Telecom_lda <- LDA(clean_Telecom_bigram_dtm, k = 4, control =
list(seed = 1234))

it_Telecom_topics <- tidy(it_Telecom_lda, matrix = "beta")

## Plot LDA Output

top_it_Telecom_topics <- it_Telecom_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

top_it_Telecom_topics %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
```




```
## Percent Deviation Topic 1

var_tele <- it_Telecom_topics %>%
  filter(topic == "1" | topic == "2" | topic == "4" ) %>%
  spread(topic, beta)
colnames(var_tele)[c(2, 3, 4)] <- c("t1", "t2", "t4")

var_tele %>%
  mutate(mean = (t1 + t2 + t4)/3, avdev_t1 = (t1 - mean), avdev_t2 =
(t2 - mean), avdev_t4 = (t4 - mean)) -> var_tele

var_tele %>%
  mutate(per_t1 = (avdev_t1/mean) * 100, per_t2 = (avdev_t2/mean) *
100, per_t4 = (avdev_t4/mean) * 100) -> var_tele
```

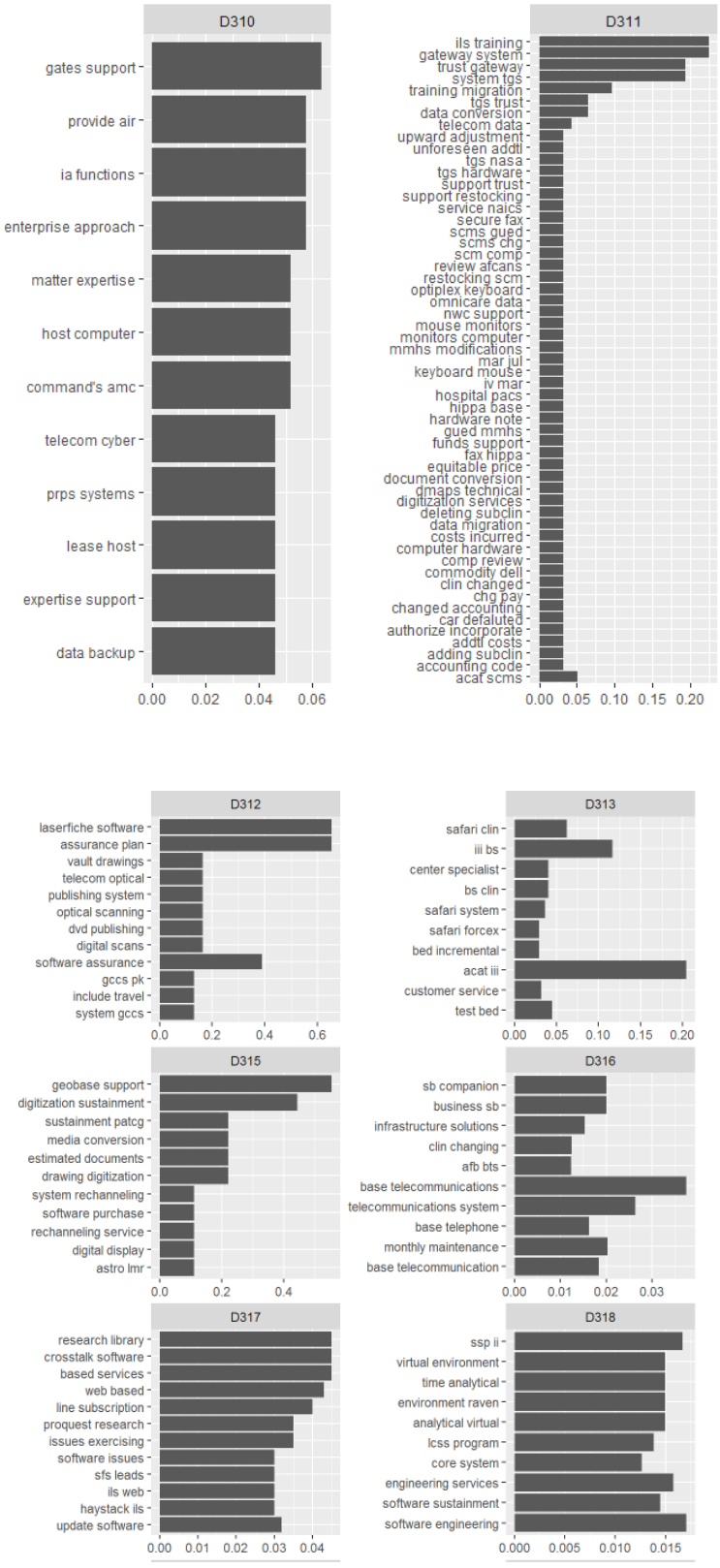
```
tele_terms_t1 <- var_tele %>%
  gather("per_t1", "per_t2", "per_t4", key = "topic", value =
"percent_dev") %>%
  group_by(topic) %>%
  top_n(10, percent_dev) %>%
  ungroup() %>%
  arrange(desc(percent_dev)) %>%
  filter(topic == "per_t1")

 tele_terms_t1[1:10,c("term", "percent_dev"), drop=FALSE]

term                          percent_dev

correct line                  192.4841
television services           188.975
system maintenance            185.5604
modification changing         181.8469
changing unit                 177.9699
excess funding                173.0266
communication telephone       172.2993
drop modification             172.2735
missing wage                  171.7364
cg fund                       169.5004

## Percent Deviation Topic 2

tele_terms_t2 <- var_tele %>%
  gather("per_t1", "per_t2", "per_t4", key = "topic", value =
"percent_dev") %>%
  group_by(topic) %>%
  top_n(10, percent_dev) %>%
  ungroup() %>%
  arrange(desc(percent_dev)) %>%
  filter(topic == "per_t2")

 tele_terms_t2[1:10,c("term", "percent_dev"), drop=FALSE]

term                          percent_dev
cable distribution            199.8745
sw cable                      199.7230
cable outlets                 199.7212
price adjustment              199.6292
tv requirement                199.5778
usaf fhc                      199.5538
government's obligation       199.5370
life circuit                  199.4969
ot:igf base                   199.4447
annual dsl                    199.4156
```

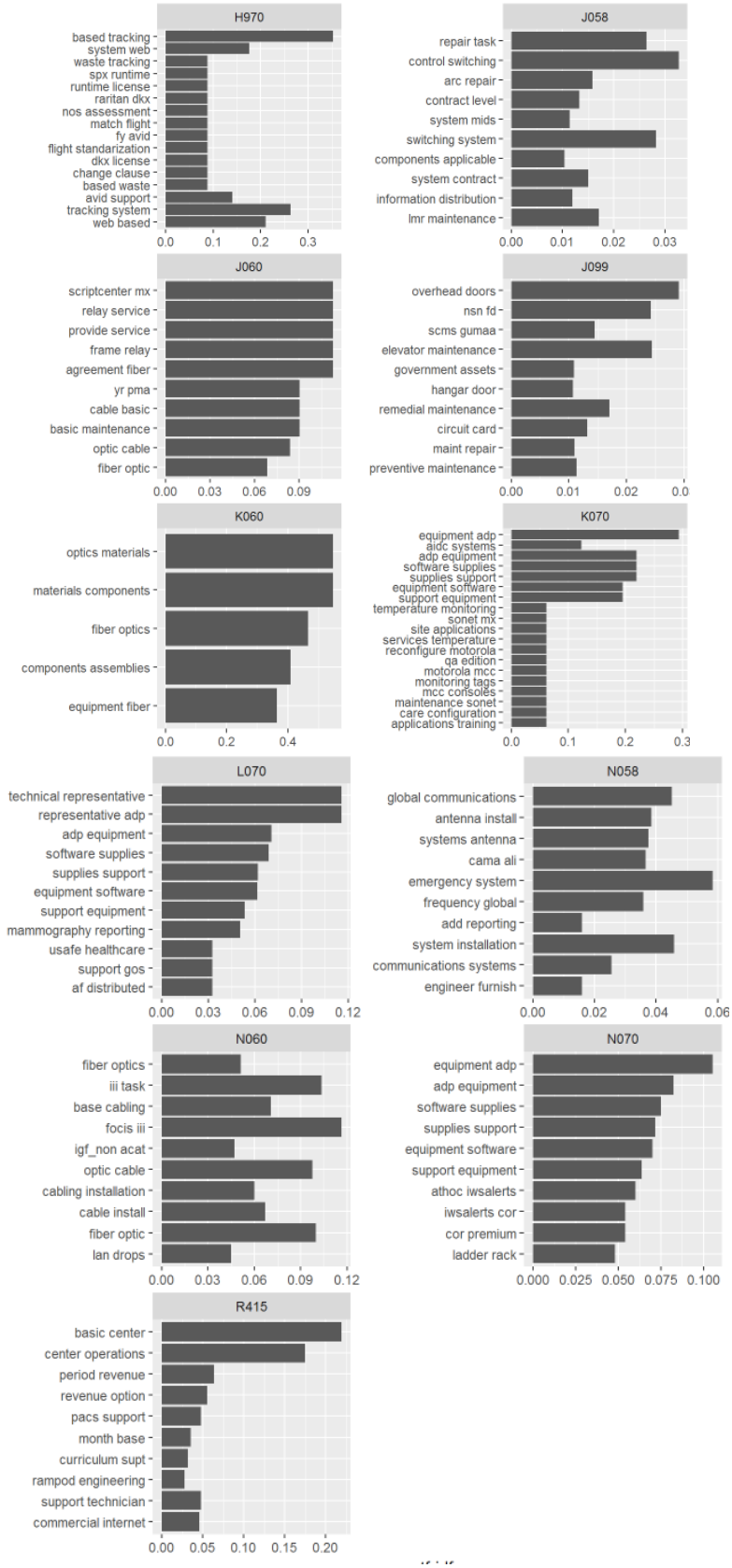
```
## Percent Deviation Topic 4

tele_terms_t4 <- var_tele %>%
  gather("per_t1", "per_t2", "per_t4", key = "topic", value =
"percent_dev") %>%
  group_by(topic) %>%
  top_n(10, percent_dev) %>%
  ungroup() %>%
  arrange(desc(percent_dev)) %>%
  filter(topic == "per_t4")



tele_terms_t4[1:10,c("term", "percent_dev"), drop=FALSE]



term              percent_dev

month funds       193.9340
cama trunk        192.8743
commercial ds     190.3175
center internet   189.5056
clin description  186.5030
service mbps      185.9766
internet phase    180.1063
fy internet       179.9482
sirius xm         178.2841
force base        176.2873
```

IT Outsourcing

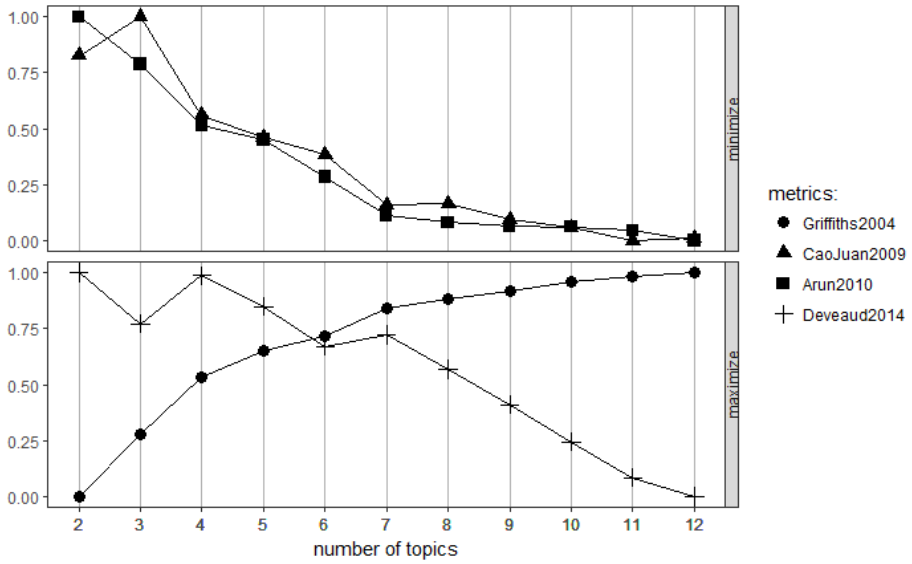
```
## TF-IDF

na.omit(relevant_itdata) %>%
  mutate(text_describe = str_replace_all(text_describe,
          pattern = "[0-9]", replacement = "")) %>%
  group_by(lvl_2_category, PSC) %>%
  filter(lvl_2_category == "IT Outsourcing") %>%
  unnest_tokens(bigram, text_describe, token = "ngrams", n = 2) %>%
  separate(bigram, c("word1", "word2"), sep = " ") %>%
  filter(!word1 %in% stop_words$word, !word2 %in% stop_words$word) %>%
  filter(!word1 %in% common_terms$word, !word2 %in% common_terms$word)
%>%
  unite("bigram", c(word1, word2), sep = " ") %>%
  count(PSC, bigram, sort = TRUE) %>%
  ungroup() -> clean_ITOut_bigram

clean_ITOut_bigram %>%
  bind_tf_idf(bigram, PSC, n) %>%
  arrange(desc(tf_idf)) %>%
  mutate(bigram = reorder(bigram, tf_idf)) %>%
  group_by(PSC) %>%
  top_n(10, wt = tf_idf) %>%
  ungroup() %>%
  #filter(PSC > D299 & PSC < D400) %>%
ggplot(aes(bigram, tf_idf)) +
    geom_bar(stat = "identity") +
    facet_wrap(~ PSC, ncol = 2, scales = "free") +
    labs(x = "Top 10 Words", y = "tf-idf",
         title = "Top 10 Bigrams in IT Outsourcing by PSC",
         subtitle = "Weighted by tf-idf") +
coord_flip() +
theme(legend.position = "none")
```


Top 10 Bigrams in IT Outsourcing by PSC
Weighted by tf-idf
D301
av level
dcapes sustainment
dcapes development
training management
vtc av
systems equipment
replace defective
products equipment
maintain systems
maintenance pmi
pmi support
technical training
level maintenance
0.00 0.02 0.04 0.06
D302
ffp loe
requisite engineering
ares ii
network requisite
ii ares
process standardization
net developers
services ffp
services ii
force network
0.00 0.01 0.02 0.03 0.04
D303
entry clerk
migrated data
service bulk
data analyst
unknown migrated
fte data
taskmaster capt
symantec training
pn eaxtll
ms symantec
ibm datacap
frc data
eaxtll ibm
datacap taskmaster
bulk transactional
data entry
personal service
0.00 0.05 0.10 0.15 0.20
D305
svcs support
services bundle
rmrc services
private cloud
commercial private
amazon web
afrs modification
web services
technical support
incrementally fund
0.00 0.05 0.10 0.15 0.20 0.25
D307
core services
directorate increment
circuit maintenance
enterprise systems
defense enterprise
business enterprise
enterprise accounting
system deams
systems bes
bes directorate
0.000 0.005 0.010 0.015 0.020
D308
range drone
gulf range
drone control
aps spiral
site supp
psdt sp
support grdcs
ordnance disposal
explosive ordnance
operational system
maintenance operational
0.00 0.01 0.02
D320
annual hardware
telecom annual
service plans
technical maintenance
vtc technical
aetc video
aetc installations
hardware maintenance
service plan
maintenance service
0.00 0.05 0.10 0.15
D321
admin engineer
engineer network
desk technician
aqs engineering
network admin
enterprise information
fte systems
services eis
command afspc
afspc enterprise
microsoft premier
desk support
0.00 0.01 0.02 0.03 0.04 0.05
D324
integration adci
acquisition domain
domain capabilities
telecom business
business continuity
domain capabilites
capabilites integration
affsc financial
tellabs standard
standard service
services technicians
service technicians
overtime funding
manager funding
financial services
financial service
additional overtime
capabilities integration
0.0 0.1 0.2 0.3 0.4
D325
provide data
network san
entry system
processing node
node ipn
installation processing
ipn pathfind
missile database
ipn pathfinder
integrated missile
san support
0.00 0.05 0.10
D399
support services
spectrum management
system remis
maintainability information
incremental funding
acat afrl
acat aflcmc
maintenance sustainment
air force
management system
0.000 0.005 0.010 0.015
H170
services charge
docave premier
gsl support
fy gsl
adp equipment
wireless inteternet
monthly services
premier maintenance
avionics simulator
network performance
performance monitor
solarwinds network
0.0 0.1 0.2


D310
gates support
provide air
ia functions
enterprise approach
matter expertise
host computer
command's amc
telecom cyber
prps systems
lease host
expertise support
data backup
0.00 0.02 0.04 0.06
D311
ils training
gateway system
trust gateway
system tgs
training migration
tgs trust
data conversion
telecom data
upward adjustment
unforeseen addtl
tgs nasa
tgs hardware
support trust
support restocking
service naics
secure fax
scms qued
scms chg
scm comp
review afcans
restocking scm
optiplex keyboard
omnicare data
nwc support
mouse monitors
monitors computer
mmhs modifications
mar jul
keyboard mouse
iv mar
hospital pacs
hippa base
hardware note
qued mmhs
funds support
fax hippa
equitable price
document conversion
dmaps technical
digitization services
deleting subclin
data migration
costs incurred
computer hardware
comp review
commodity dell
clin changed
chg pay
changed accounting
car defaluted
authorize incorporate
addtl costs
adding subclin
accounting code
acat scms
0.00 0.05 0.10 0.15 0.20
D312
laserfiche software
assurance plan
vault drawings
telecom optical
publishing system
optical scanning
dvd publishing
digital scans
software assurance
gccs pk
include travel
system gccs
0.0 0.2 0.4 0.6
D313
safari clin
iii bs
center specialist
bs clin
safari system
safari forcex
bed incremental
acat iii
customer service
test bed
0.00 0.05 0.10 0.15 0.20
D315
geobase support
digitization sustainment
sustainment patcg
media conversion
estimated documents
drawing digitization
system rechanneling
software purchase
rechanneling service
digital display
astro lmr
0.0 0.2 0.4
D316
sb companion
business sb
infrastructure solutions
clin changing
afb bts
base telecommunications
telecommunications system
base telephone
monthly maintenance
base telecommunication
0.00 0.01 0.02 0.03
D317
research library
crosstalk software
based services
web based
line subscription
proquest research
issues exercising
software issues
sfs leads
ils web
haystack ils
update software
0.00 0.01 0.02 0.03 0.04
D318
ssp ii
virtual environment
time analytical
environment raven
analytical virtual
lcss program
core system
engineering services
software sustainment
software engineering
0.000 0.005 0.010 0.015

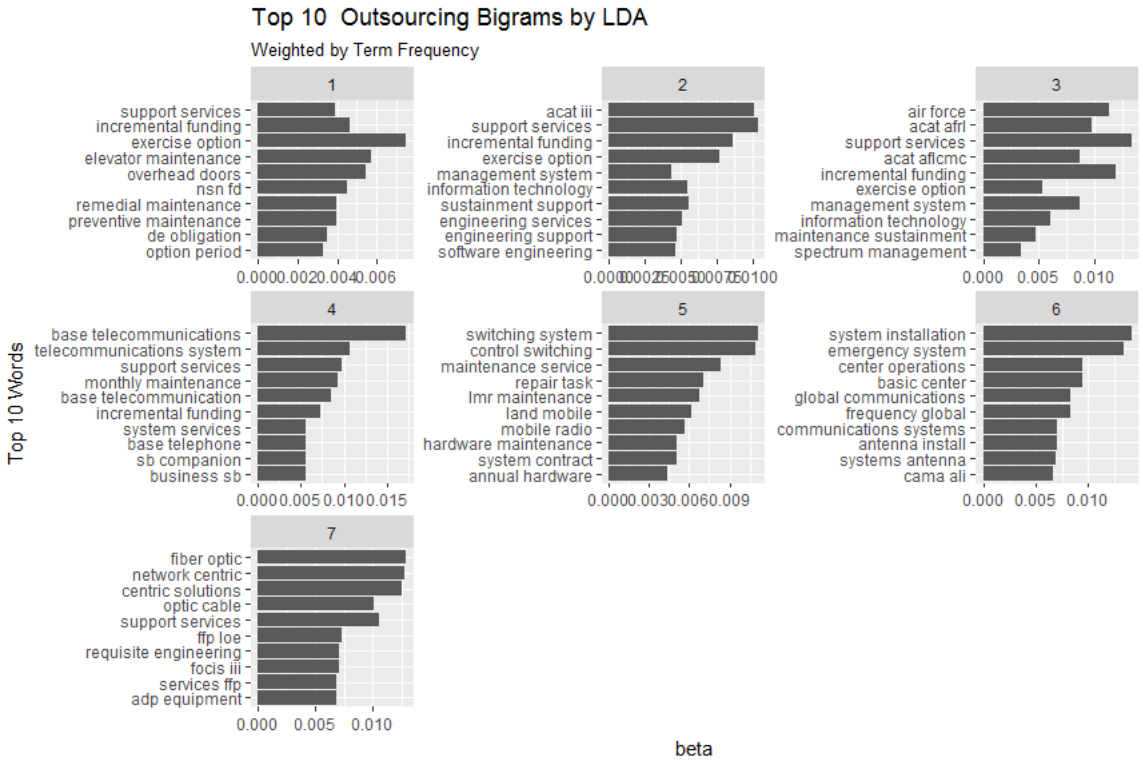

H970
based tracking
system web
waste tracking
spx runtime
runtime license
raritan dkx
nos assessment
match flight
fy avid
flight standarization
dkx license
change clause
based waste
avid support
tracking system
web based
0.0 0.1 0.2 0.3
J058
repair task
control switching
arc repair
contract level
system mids
switching system
components applicable
system contract
information distribution
lmr maintenance
0.00 0.01 0.02 0.03
J060
scriptcenter mx
relay service
provide service
frame relay
agreement fiber
yr pma
cable basic
basic maintenance
optic cable
fiber optic
0.00 0.03 0.06 0.09
J099
overhead doors
nsn fd
scms gumaa
elevator maintenance
government assets
hangar door
remedial maintenance
circuit card
maint repair
preventive maintenance
0.00 0.01 0.02 0.0
K060
optics materials
materials components
fiber optics
components assemblies
equipment fiber
0.0 0.2 0.4
K070
equipment adp
aidc systems
adp equipment
software supplies
supplies support
equipment software
support equipment
temperature monitoring
sonet mx
site applications
services temperature
reconfigure motorola
qa edition
motorola mcc
monitoring tags
mcc consoles
maintenance sonet
care configuration
applications training
0.0 0.1 0.2 0.3
L070
technical representative
representative adp
adp equipment
software supplies
supplies support
equipment software
support equipment
mammography reporting
usafe healthcare
support gos
af distributed
0.00 0.03 0.06 0.09 0.12
N058
global communications
antenna install
systems antenna
cama ali
emergency system
frequency global
add reporting
system installation
communications systems
engineer furnish
0.00 0.02 0.04 0.06
N060
fiber optics
iii task
base cabling
focis iii
igf_non acat
optic cable
cabling installation
cable install
fiber optic
lan drops
0.00 0.03 0.06 0.09 0.12
N070
equipment adp
adp equipment
software supplies
supplies support
equipment software
support equipment
athoc iwsalerts
iwsalerts cor
cor premium
ladder rack
0.000 0.025 0.050 0.075 0.100
R415
basic center
center operations
period revenue
revenue option
pacs support
month base
curriculum supt
rampod engineering
support technician
commercial internet
0.00 0.05 0.10 0.15 0.20

```
## Document Term Matrix

cast_dtm(clean_ITOut_bigram, PSC, bigram, n) -> clean_ITOut_bigram_dtm

clean_ITOut_bigram_dtm

## Optimal Number of Topics

result_out <- FindTopicsNumber(
  clean_ITOut_bigram_dtm,
  topics = seq(from = 2, to = 12, by = 1),
  metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010",
"Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 1234),
  mc.cores = 2L,
  verbose = TRUE
)
FindTopicsNumber_plot(result_out)
```



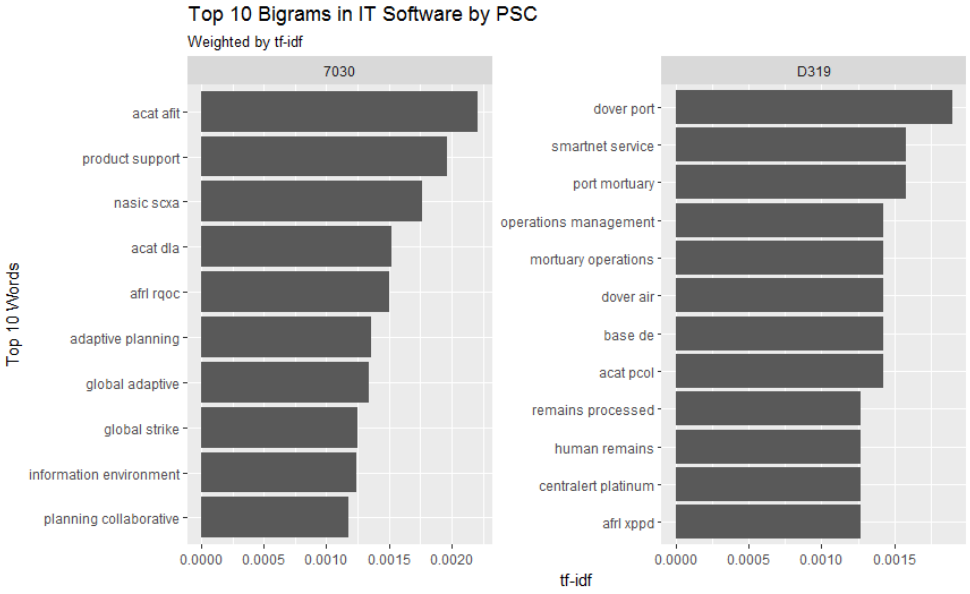

```
## LDA Model

it_out_lda <- LDA(clean_ITOut_bigram_dtm, k = 7, control = list(seed =
1234))

it_out_topics <- tidy(it_out_lda, matrix = "beta")

## Plot LDA Output

top_it_out_topics <- it_out_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
```

```
top_it_out_topics %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  labs(x = "Top 10 Words", y = "beta",
         title = "Top 10  Outsourcing Bigrams by LDA",
         subtitle = "Weighted by Term Frequency") +
  coord_flip()
```



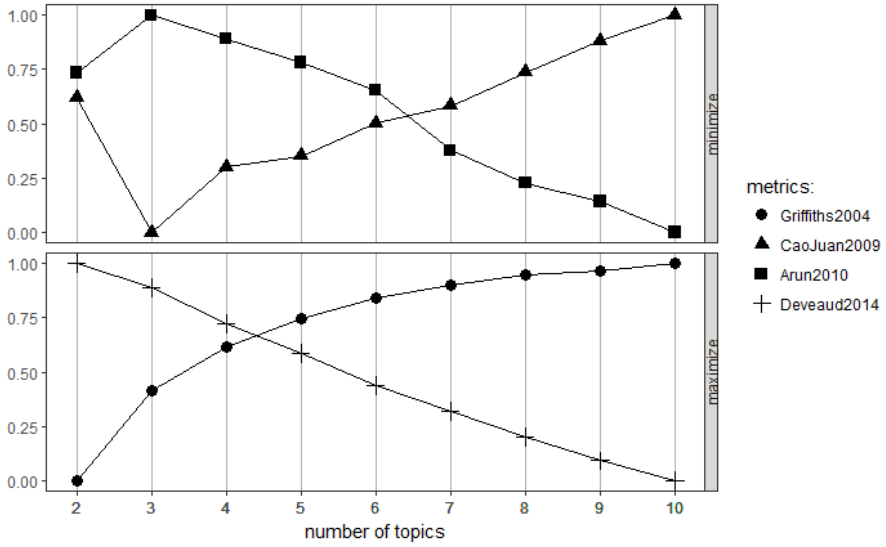
IT Software

```
## TF-IDF

na.omit(relevant_itdata) %>%
  mutate(text_describe = str_replace_all(text_describe,
         pattern = "[0-9]", replacement = "")) %>%
  group_by(lvl_2_category, PSC) %>%
  filter(lvl_2_category == "IT Software") %>%
  unnest_tokens(bigram, text_describe, token = "ngrams", n = 2) %>%
  separate(bigram, c("word1", "word2"), sep = " ") %>%
  filter(!word1 %in% stop_words$word, !word2 %in% stop_words$word) %>%
  filter(!word1 %in% common_terms$word, !word2 %in% common_terms$word)
%>%
  unite("bigram", c(word1, word2), sep = " ") %>%
  count(PSC, bigram, sort = TRUE) %>%
```



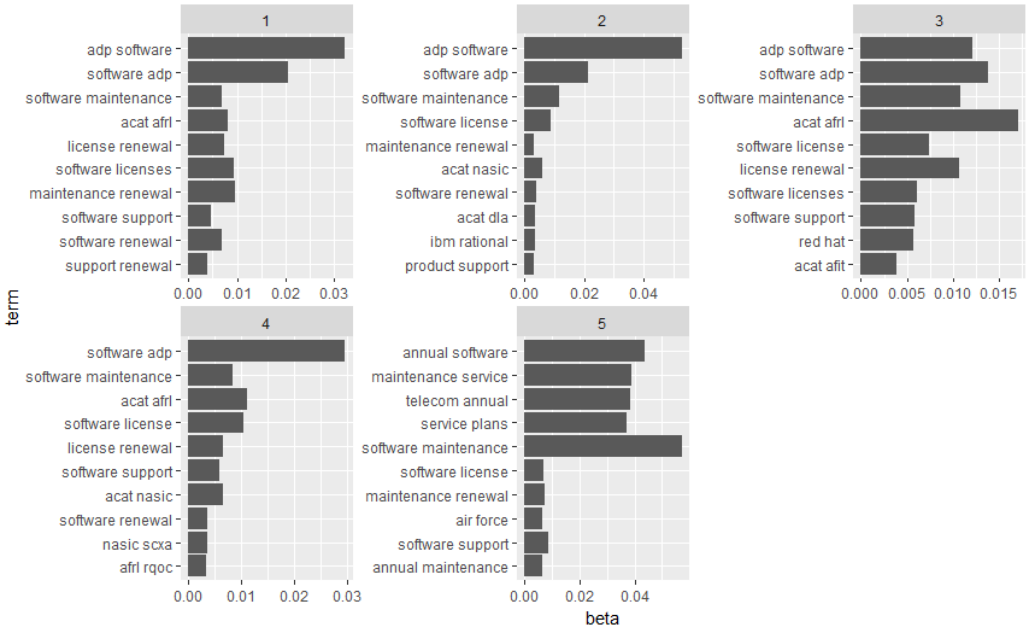

```
  ungroup() -> clean_ITSoftware_bigram

clean_ITSoftware_bigram %>%
  bind_tf_idf(bigram, PSC, n) %>%
  arrange(desc(tf_idf)) %>%
  mutate(bigram = reorder(bigram, tf_idf)) %>%
  group_by(PSC) %>%
  top_n(10, wt = tf_idf) %>%
  ungroup() %>%
ggplot(aes(bigram, tf_idf)) +
    geom_bar(stat = "identity") +
    facet_wrap(~ PSC, ncol = 2, scales = "free") +
    labs(x = "Top 10 Words", y = "tf-idf",
         title = "Top 10 Bigrams in IT Software by PSC",
         subtitle = "Weighted by tf-idf") +
    coord_flip() +
    theme(legend.position = "none")
```

```
## Document Term Matrix

cast_dtm(clean_ITSoftware_bigram, PSC, bigram, n) ->
clean_ITSoftware_bigram_dtm


## Optimal Number of Topics

result_software <- FindTopicsNumber(
  clean_ITSoftware_bigram_dtm,
  topics = seq(from = 2, to = 10, by = 1),
  metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010",
"Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 1234),
  mc.cores = 2L, #make sure this is appropriate number of cores you
wish to use
  verbose = TRUE
)
FindTopicsNumber_plot(result_software)
```




```
## LDA Model

it_software_lda <- LDA(clean_ITSoftware_bigram_dtm, k = 5, control =
list(seed = 1234))

it_software_topics <- tidy(it_software_lda, matrix = "beta")

## Plot LDA Output

top_it_software_topics <- it_software_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
```

```
top_it_software_topics %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
```



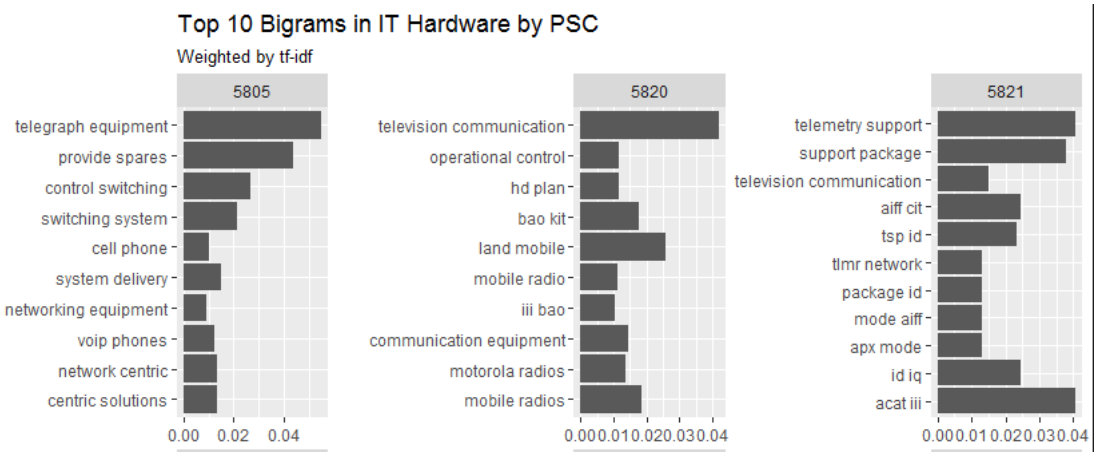

```
## Percent Deviation

var_soft <- it_software_topics %>%
  spread(topic, beta)
colnames(var_soft)[c(2, 3, 4, 5)] <- c("t1","t2","t3", "t4")

var_soft %>%
  mutate(mean = (t1 + t2 + t3 + t4)/4, avdev_t1 = (t1 - mean), avdev_t2
= (t2 - mean), avdev_t3 = (t3 - mean), avdev_t4 = (t4 - mean)) ->
var_soft

var_soft %>%
  mutate(per_t1 = (avdev_t1/mean) * 100, per_t2 = (avdev_t2/mean) *
100, per_t3 = (avdev_t3/mean) * 100, per_t4 = (avdev_t4/mean) * 100) ->
var_soft

soft_terms_t1 <- var_soft %>%
  gather("per_t1", "per_t2", "per_t3", "per_t4", key = "topic", value =
"percent_dev") %>%
  group_by(topic) %>%
  top_n(10, percent_dev) %>%
  ungroup() %>%
  arrange(desc(percent_dev)) %>%
  filter(topic == "per_t1")
```

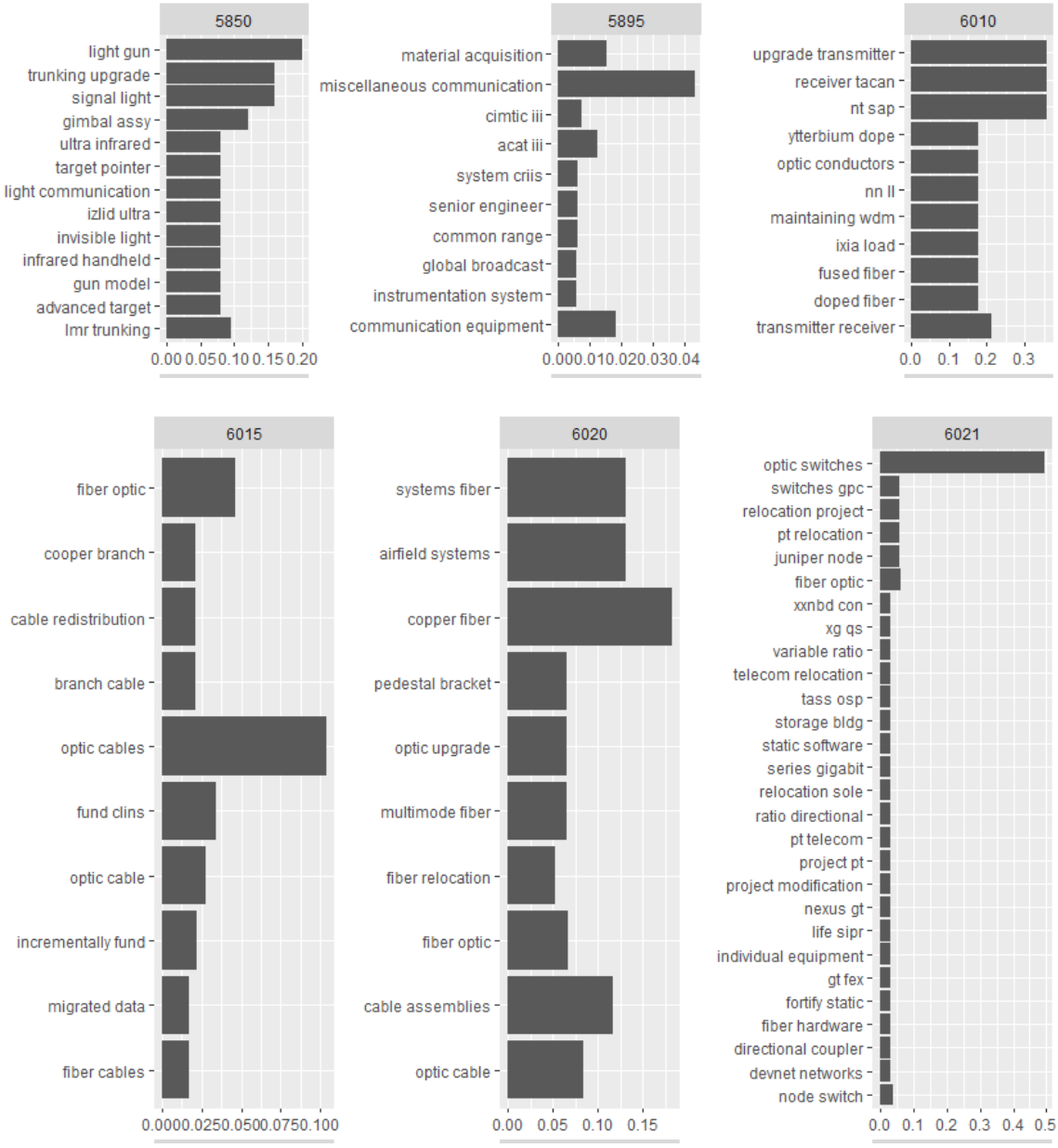
```
## Topic 1

term                    percent_dev

final deobligation      261.0584
d.o pop                 251.6756
integration test        250.4637
wkc pk                  250.2471
facilitate payment      249.7149
licenses voice          248.3278
add logo                246.3301
market patriot          245.0462
starteam enterprise     239.9419
windows server          239.6439

## Topic 2

soft_terms_t2 <- var_soft %>%
  gather("per_t1", "per_t2", "per_t3", "per_t4", key = "topic", value =
"percent_dev") %>%
  group_by(topic) %>%
  top_n(10, percent_dev) %>%
  ungroup() %>%
  arrange(desc(percent_dev)) %>%
  filter(topic == "per_t2")



soft_terms_t2[1:10,c("term", "percent_dev"), drop=FALSE]

term                    percent_dev

security software       254.8172
sight support           254.1522
support monarch         253.1624
cals telerik            253.0215
infrastructure division 251.0456
renewal pc              249.0562
oscilloscope mobile     248.5901
av xamarin              246.2859
cables mcafee           245.4927
equipment netowl        244.1933


## Topic 3

soft_terms_t3 <- var_soft %>%
  gather("per_t1", "per_t2", "per_t3", "per_t4", key = "topic", value =
"percent_dev") %>%
  group_by(topic) %>%
  top_n(10, percent_dev) %>%
  ungroup() %>%
  arrange(desc(percent_dev)) %>%
  filter(topic == "per_t3")
```

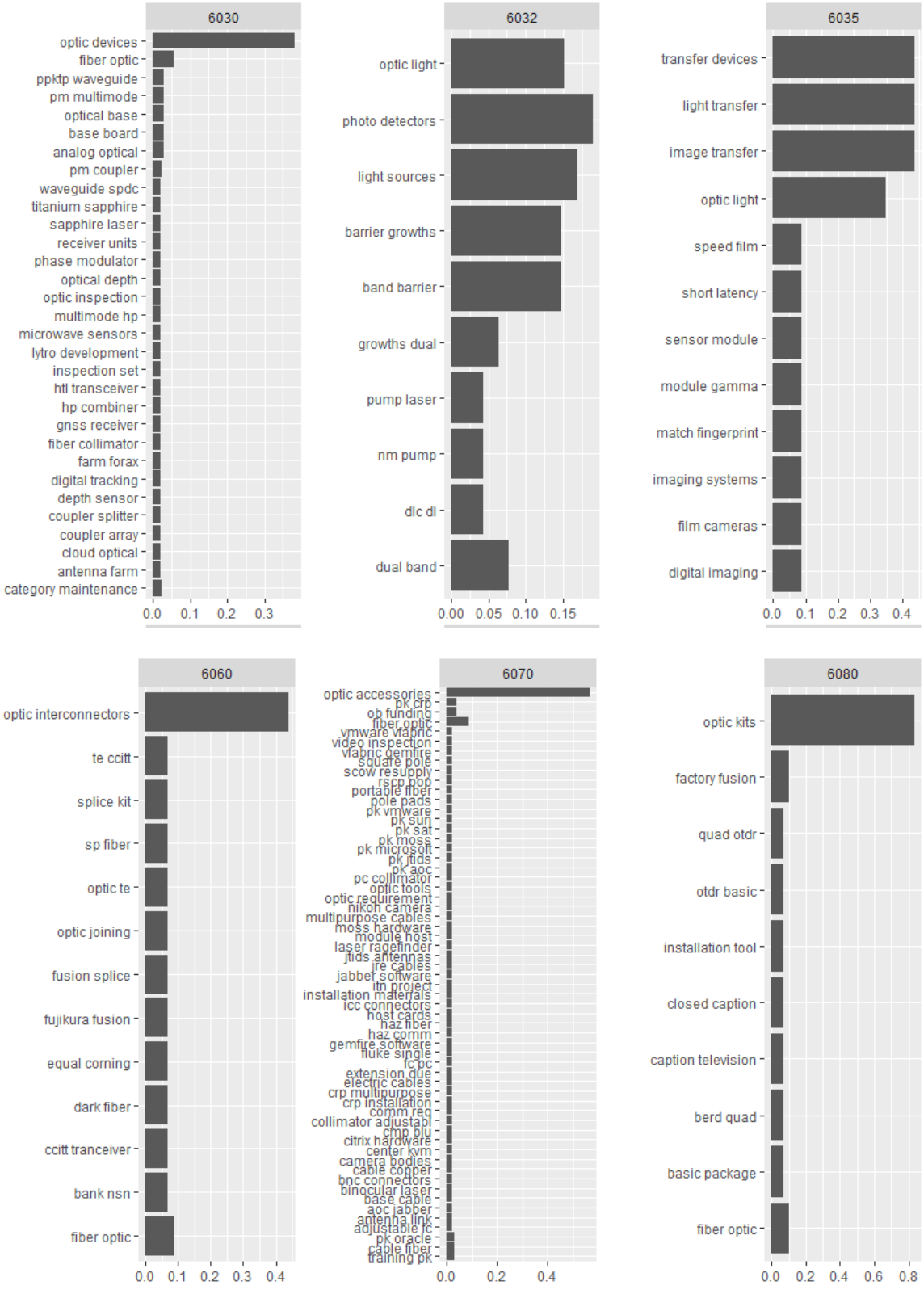
```
soft_terms_t3[1:10,c("term", "percent_dev"), drop=FALSE]

term                    percent_dev

microsoft softrware     298.7376
clause hardware         295.2698
seymour johnson         294.7687
host center             293.9748
advanced customer       291.1535
ms ts                   290.7404
computer service        290.6979
subscription base       290.0651
mortuarys mission       288.1281
asapk descriptio        287.9197

## Topic 4

soft_terms_t4 <- var_soft %>%
  gather("per_t1", "per_t2", "per_t3", "per_t4", key = "topic", value =
"percent_dev") %>%
  group_by(topic) %>%
  top_n(10, percent_dev) %>%
  ungroup() %>%
  arrange(desc(percent_dev)) %>%
  filter(topic == "per_t4")

soft_terms_t4[1:10,c("term", "percent_dev"), drop=FALSE]

term                    percent_dev

switchview kvm          270.7886
cost bilateral          268.6854
license ida             256.3139
standard sw             256.3136
solarwinds licenses     252.9405
life storage            249.3739
bpel licenses           248.5678
support emergency       242.7162
support imagine         240.3896
support linux           237.3325
```

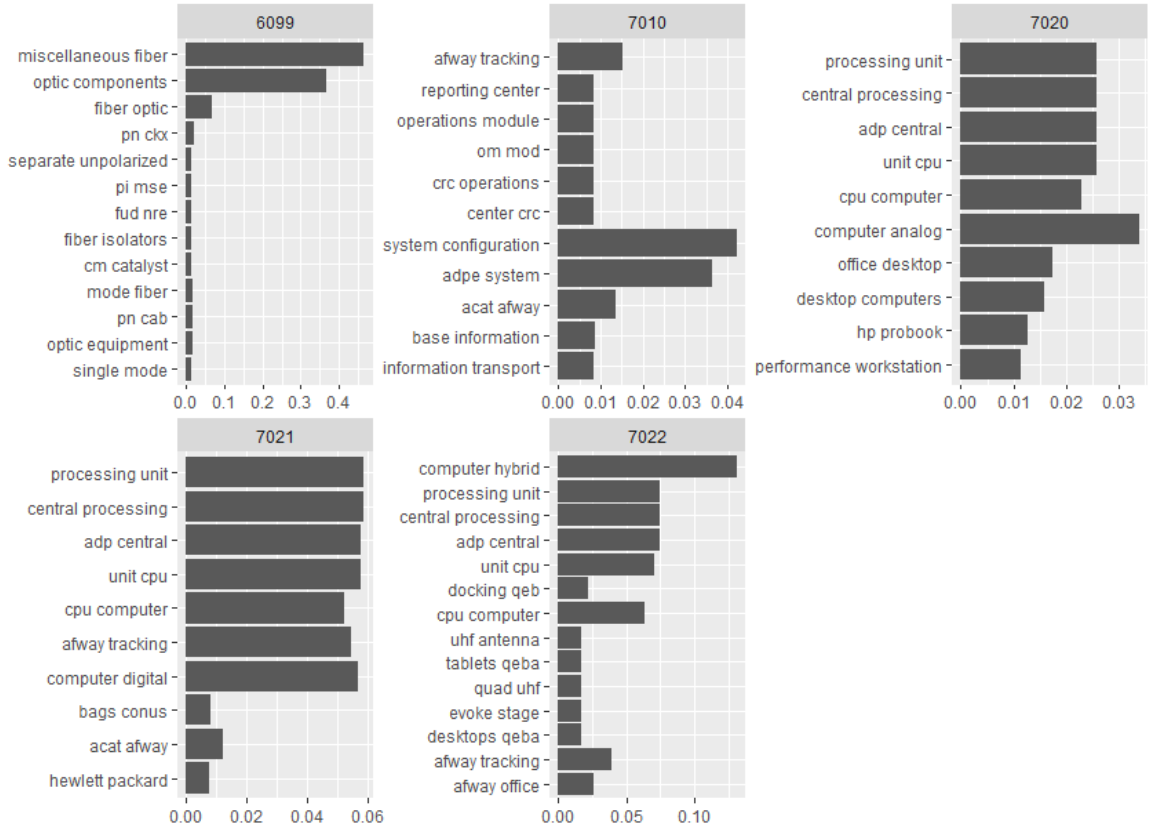
IT Hardware

```
## TF-IDF

na.omit(relevant_itdata) %>%
  mutate(text_describe = str_replace_all(text_describe,
          pattern = "[0-9]", replacement = "")) %>%
  group_by(lvl_2_category, PSC) %>%
  filter(lvl_2_category == "IT Hardware") %>%
  unnest_tokens(bigram, text_describe, token = "ngrams", n = 2) %>%
  separate(bigram, c("word1", "word2"), sep = " ") %>%
  filter(!word1 %in% stop_words$word, !word2 %in% stop_words$word) %>%
  filter(!word1 %in% common_terms$word, !word2 %in% common_terms$word)
%>%
  unite("bigram", c(word1, word2), sep = " ") %>%
  count(PSC, bigram, sort = TRUE) %>%
  ungroup() -> clean_ITHardware_bigram

clean_ITHardware_bigram %>%
  bind_tf_idf(bigram, PSC, n) %>%
  arrange(desc(tf_idf)) %>%
  mutate(bigram = reorder(bigram, tf_idf)) %>%
  group_by(PSC) %>%
  top_n(10, wt = tf_idf) %>%
  ungroup() %>%
  filter(PSC > 7019 & PSC < 7023) %>%
ggplot(aes(bigram, tf_idf)) +
    geom_bar(stat = "identity") +
    facet_wrap(~ PSC, ncol = 3, scales = "free") +
    labs(x = "Top 10 Words", y = "tf-idf",
         title = "Top 10 Bigrams in IT Hardware by PSC",
         subtitle = "Weighted by tf-idf") +
    coord_flip() +
    theme(legend.position = "none")
```

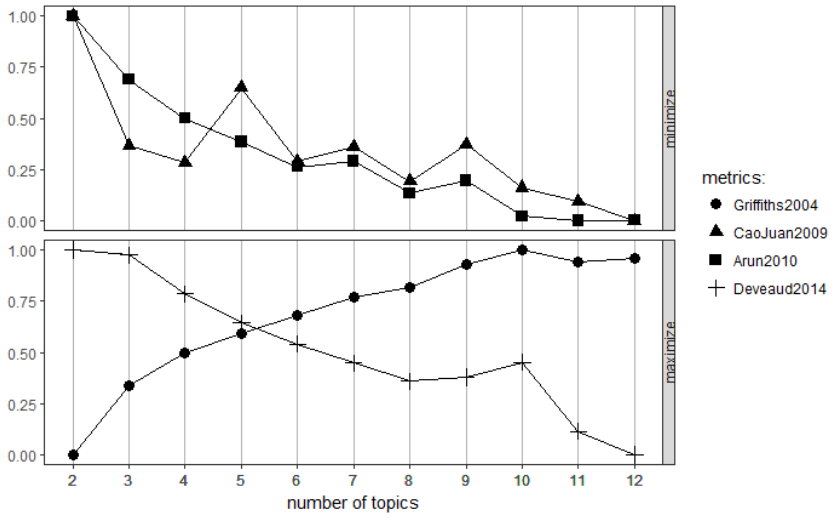



5850
light gun
trunking upgrade
signal light
gimbal assy
ultra infrared
target pointer
light communication
izlid ultra
invisible light
infrared handheld
gun model
advanced target
lmr trunking
0.00 0.05 0.10 0.15 0.20
5895
material acquisition
miscellaneous communication
cimtic iii
acat iii
system criis
senior engineer
common range
global broadcast
instrumentation system
communication equipment
0.000.010.020.030.04
6010
upgrade transmitter
receiver tacan
nt sap
ytterbium dope
optic conductors
nn ll
maintaining wdm
ixia load
fused fiber
doped fiber
transmitter receiver
0.0 0.1 0.2 0.3
6015
fiber optic
cooper branch
cable redistribution
branch cable
optic cables
fund clins
optic cable
incrementally fund
migrated data
fiber cables
0.0000.0250.0500.0750.100
6020
systems fiber
airfield systems
copper fiber
pedestal bracket
optic upgrade
multimode fiber
fiber relocation
fiber optic
cable assemblies
optic cable
0.00 0.05 0.10 0.15
6021
optic switches
switches gpc
relocation project
pt relocation
juniper node
fiber optic
xxnbd con
xg qs
variable ratio
telecom relocation
tass osp
storage bldg
static software
series gigabit
relocation sole
ratio directional
pt telecom
project pt
project modification
nexus gt
life sipr
individual equipment
gt fex
fortify static
fiber hardware
directional coupler
devnet networks
node switch
0.0 0.1 0.2 0.3 0.4 0.5

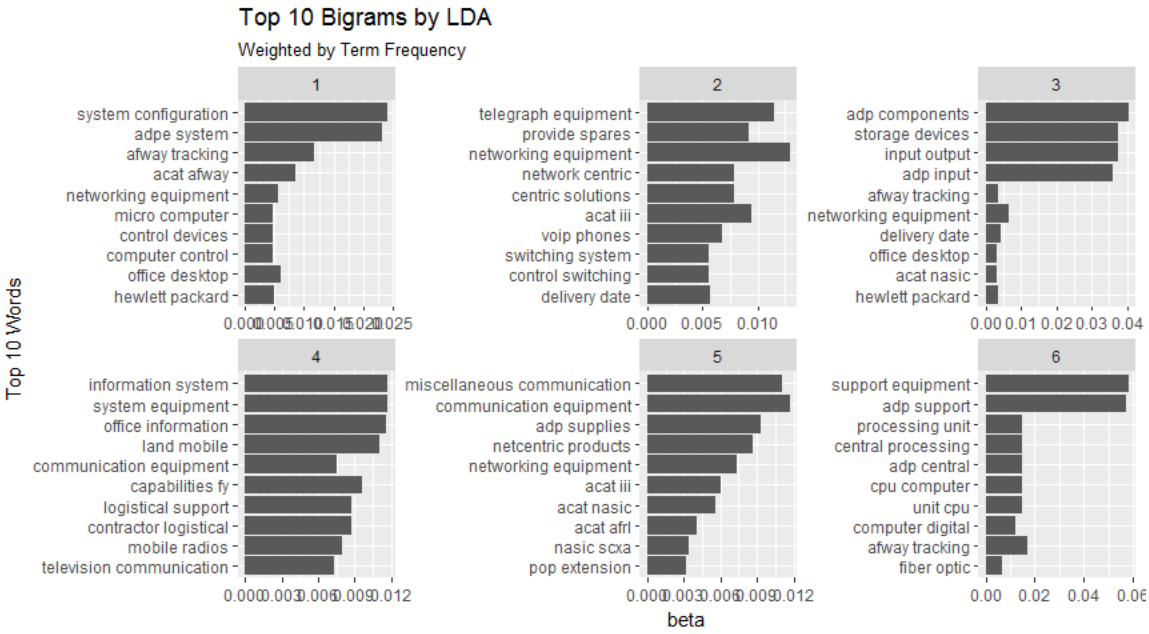

6030
optic devices
fiber optic
ppktp waveguide
pm multimode
optical base
base board
analog optical
pm coupler
waveguide spdc
titanium sapphire
sapphire laser
receiver units
phase modulator
optical depth
optic inspection
multimode hp
microwave sensors
lytro development
inspection set
htl transceiver
hp combiner
gnss receiver
fiber collimator
farm forax
digital tracking
depth sensor
coupler splitter
coupler array
cloud optical
antenna farm
category maintenance
0.0 0.1 0.2 0.3
6032
optic light
photo detectors
light sources
barrier growths
band barrier
growths dual
pump laser
nm pump
dlc dl
dual band
0.00 0.05 0.10 0.15
6035
transfer devices
light transfer
image transfer
optic light
speed film
short latency
sensor module
module gamma
match fingerprint
imaging systems
film cameras
digital imaging
0.0 0.1 0.2 0.3 0.4
6060
optic interconnectors
te ccitt
splice kit
sp fiber
optic te
optic joining
fusion splice
fujikura fusion
equal corning
dark fiber
ccitt tranceiver
bank nsn
fiber optic
0.0 0.1 0.2 0.3 0.4
6070
optic accessories
pk crp
ob funding
fiber optic
vmware vfabric
video inspection
vfabric gemfire
square pole
scow resupply
rscp hop
portable fiber
pole pads
pk vmware
pk sun
pk sat
pk moss
pk microsoft
pk jtids
pk aoc
pc collimator
optic tools
optic requirement
nikon camera
multipurpose cables
moss hardware
module host
laser rangefinder
jtids antennas
jre cables
jabber software
itn project
installation materials
icc connectors
host cards
haz fiber
haz comm
gemfire software
fluke single
fc pc
extension due
electric cables
crp multipurpose
crp installation
comm req
collimator adjustabl
chnp blu
citrix hardware
center kvm
camera bodies
cable copper
bnc connectors
binocular laser
base cable
aoc jaber
antenna link
adjustable fc
pk oracle
cable fiber
training pk
0.0 0.2 0.4
6080
optic kits
factory fusion
quad otdr
otdr basic
installation tool
closed caption
caption television
berd quad
basic package
fiber optic
0.0 0.2 0.4 0.6 0.8




```
## Document Term Matrix

cast_dtm(clean_ITHardware_bigram, PSC, bigram, n) ->
clean_ITHardware_bigram_dtm

## Optimal Number of Topics

result_hard <- FindTopicsNumber(
  clean_ITHardware_bigram_dtm,
  topics = seq(from = 2, to = 12, by = 1),
  metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010",
"Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 1234),
  mc.cores = 2L, #make sure this is appropriate number of cores you
wish to use
  verbose = TRUE
)
FindTopicsNumber_plot(result_hard)
```

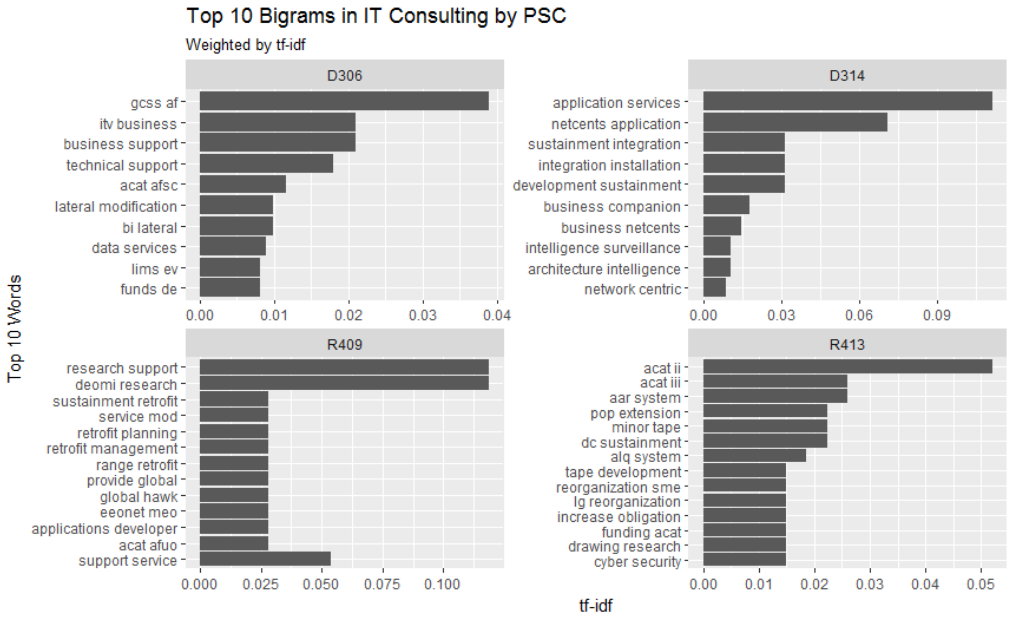



```
## LDA Model

it_hardware_lda <- LDA(clean_ITHardware_bigram_dtm, k = 6, control =
list(seed = 1234))

it_hardware_topics <- tidy(it_hardware_lda, matrix = "beta")

## Plot LDA Output

top_it_hardware_topics <- it_hardware_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

top_it_hardware_topics %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  labs(x = "Top 10 Words", y = "beta",
         title = "Top 10 Bigrams by LDA",
         subtitle = "Weighted by Term Frequency") +
  coord_flip()
```

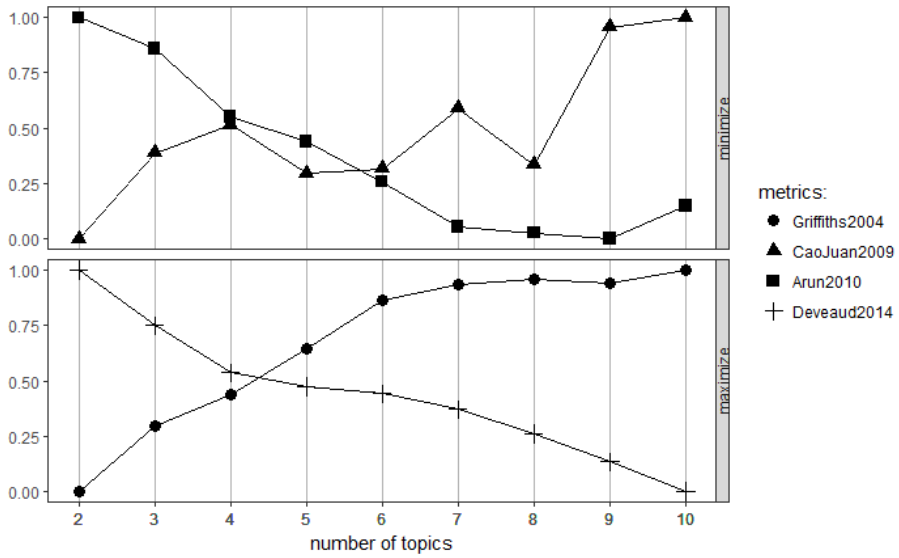

Top 10 Bigrams by LDA
Weighted by Term Frequency
1
system configuration
adpe system
afway tracking
acat afway
networking equipment
micro computer
control devices
computer control
office desktop
hewlett packard
2
telegraph equipment
provide spares
networking equipment
network centric
centric solutions
acat iii
voip phones
switching system
control switching
delivery date
3
adp components
storage devices
input output
adp input
afway tracking
networking equipment
delivery date
office desktop
acat nasic
hewlett packard
4
information system
system equipment
office information
land mobile
communication equipment
capabilities fy
logistical support
contractor logistical
mobile radios
television communication
5
miscellaneous communication
communication equipment
adp supplies
netcentric products
networking equipment
acat iii
acat nasic
acat afrl
nasic scxa
pop extension
6
support equipment
adp support
processing unit
central processing
adp central
cpu computer
unit cpu
computer digital
afway tracking
fiber optic
Top 10 Words
beta

IT Consulting

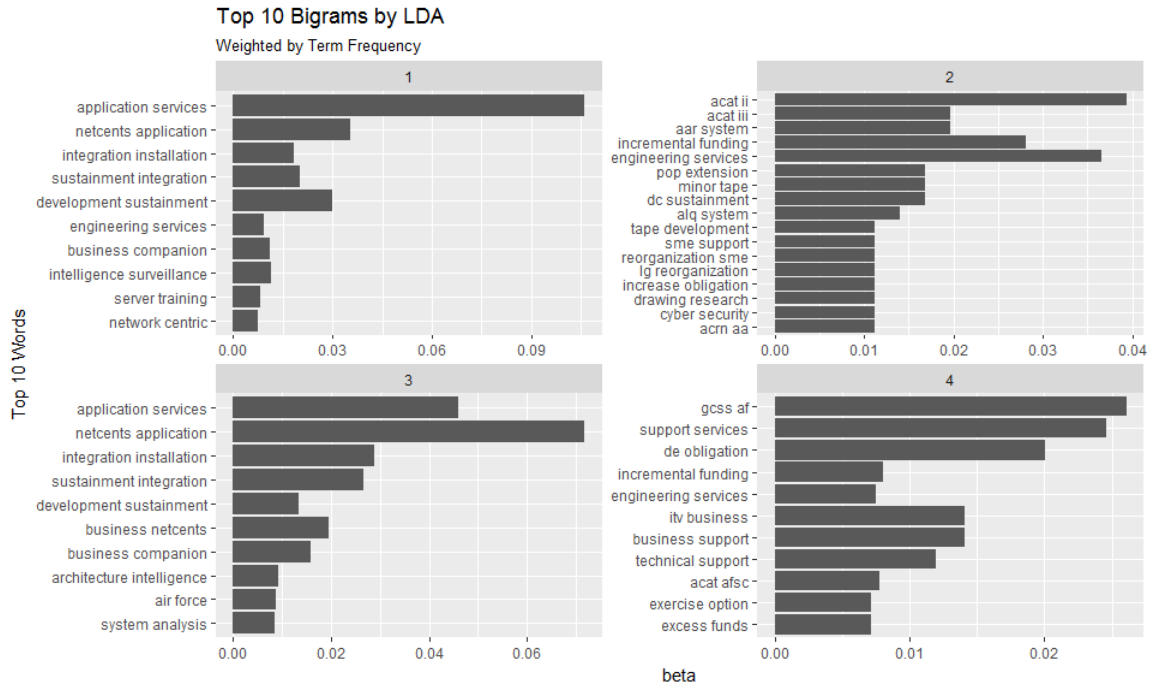
```
## TF-IDF

na.omit(relevant_itdata) %>%
  mutate(text_describe = str_replace_all(text_describe,
         pattern = "[0-9]", replacement = "")) %>%
  group_by(lvl_2_category, PSC) %>%
  filter(lvl_2_category == "IT Consulting") %>%
  unnest_tokens(bigram, text_describe, token = "ngrams", n = 2) %>%
  separate(bigram, c("word1", "word2"), sep = " ") %>%
  filter(!word1 %in% stop_words$word, !word2 %in% stop_words$word) %>%
  filter(!word1 %in% common_terms$word, !word2 %in% common_terms$word)
%>%
  unite("bigram", c(word1, word2), sep = " ") %>%
  count(PSC, bigram, sort = TRUE) %>%
  ungroup() -> clean_ITConsulting_bigram

clean_ITConsulting_bigram %>%
  bind_tf_idf(bigram, PSC, n) %>%
  arrange(desc(tf_idf)) %>%
  mutate(bigram = reorder(bigram, tf_idf)) %>%
  group_by(PSC) %>%
  top_n(10, wt = tf_idf) %>%
  ungroup() %>%
ggplot(aes(bigram, tf_idf)) +
    geom_bar(stat = "identity") +
    facet_wrap(~ PSC, ncol = 2, scales = "free") +
    labs(x = "Top 10 Words", y = "tf-idf",
         title = "Top 10 Bigrams in IT Consulting by PSC",
         subtitle = "Weighted by tf-idf") +
    coord_flip() +
    theme(legend.position = "none")
```



```
## Document Term Matrix

cast_dtm(clean_ITConsulting_bigram, PSC, bigram, n) ->
clean_ITConsulting_bigram_dtm

## Optimal Number of Topics

result_consult <- FindTopicsNumber(
  clean_ITConsulting_bigram_dtm,
  topics = seq(from = 2, to = 10, by = 1),
  metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010",
"Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 1234),
  mc.cores = 2L, #make sure this is appropriate number of cores you
wish to use
  verbose = TRUE
)
FindTopicsNumber_plot(result_consult)
```




```
## LDA Model

it_consulting_lda <- LDA(clean_ITConsulting_bigram_dtm, k = 4, control
= list(seed = 1234))

it_consulting_topics <- tidy(it_consulting_lda, matrix = "beta")


## Plot LDA Output
```




```
## Percent Deviation

var_cons <- it_consulting_topics %>%
  filter(topic == "1" | topic == "3") %>%
  spread(topic, beta)
colnames(var_cons)[c(2, 3)] <- c("t1","t3")

var_cons %>%
  mutate(mean = (t1 + t3)/2, avdev_t1 = (t1 - mean), avdev_t3 = (t3 -
mean)) -> var_cons

var_cons %>%
  mutate(per_t1 = (avdev_t1/mean) * 100, per_t3 = (avdev_t3/mean) *
100) -> var_cons

## Topic 1

cons_terms_t1 <- var_cons %>%
  gather("per_t1", "per_t3", key = "topic", value = "percent_dev") %>%
  group_by(topic) %>%
  top_n(10, percent_dev) %>%
  ungroup() %>%
  arrange(desc(percent_dev)) %>%
  filter(topic == "per_t1")



cons_terms_t1[1:10,c("term", "percent_dev"), drop=FALSE]
```

```
term                    percent_dev

price increase          99.87976
dcc camera              99.18703
management security     99.03239
sp training             99.00528
applications support    98.14040
telephony products      98.04043
air program             97.94898
asr router              97.85032
seesaw cd               97.60175
router itlc             96.28322
```

```
## Topic 3

cons_terms_t3 <- var_cons %>%
  gather("per_t1", "per_t3", key = "topic", value = "percent_dev") %>%
  group_by(topic) %>%
  top_n(10, percent_dev) %>%
  ungroup() %>%
  arrange(desc(percent_dev)) %>%
  filter(topic == "per_t3")



cons_terms_t3[1:10,c("term", "percent_dev"), drop=FALSE]
```

```
term              percent_dev

add ecmra         99.99762
design emd        99.99023
uft software      99.98874
service deomi     99.97399
clin acrn         99.97374
mar car           99.97175
acctg line        99.96564
eeonet extension  99.92233
laircm test       99.92132
mths funding      99.92008
```

Appendix B. Topic Assignment Sheets

Instructions

The following word lists come from IT Level 2 Category contract text descriptions. For each list of words type the "topic" the words are likely describing in the highlighted cell above each list. Please limit the topic responses to three words or less. If the word list does not define any topic, leave it blank.

The Level 2 Category label applies to all topics on the page.

Level 2 Category: **IT Security**

Topic 1 :	Topic 2 :	Topic 3 :
acat iii	acat iii	acat iii
communications security	communications security	communications security
security equipment	security equipment	kiv production
adapter plate	portable avenger	predator elite
model kg	gfp correction	capability study
plate av	diesel generator	additional days
noun core	ldc audio	clin overrun
lightning strike	lmr motorola	linux production
lightning strikes	recaro seat	apx digital
av conference	imaging technology	afb option

Topic 4 :
air force
force key
key loader
key management
management infrastructure
portable key
cryptographic core
infrastructure cryptographic
anti spoofing
availability anti

Instructions

The following word lists come from IT Level 2 Category contract text descriptions. For each list of words type the "topic" the words are likely describing in the highlighted cell above each list. Please limit the topic responses to three words or less. If the word list does not define any topic, leave it blank.

The Level 2 Category label applies to all topics on the page.

Level 2 Category: **IT Consulting**

Topic 1 :	Topic 2 :	Topic 3 :
application services	acat ii	netcents application
netcents application	engineering services	application services
development sustainment	incremental funding	integration installation
price increase	aar system	add ecmra
dcc camera	acat iii	design emd
management security	dc sustainment	uft software
sp training	minor tape	service deomi
applications support	pop extension	clin acrn
telephony products	alg system	mar car
air program	sme support	acctg line

Topic 4 :
gcss af
support services
de obligation
business support
itv business
technical support
incremental funding
acat afsc
engineering services
excess funds
exercise option

Instructions

The following word lists come from IT Level 2 Category contract text descriptions. For each list of words type the "topic" the words are likely describing in the highlighted cell above each list. Please limit the topic responses to three words or less. If the word list does not define any topic, leave it blank.

The Level 2 Category label applies to all topics on the page.

Level 2 Category: **IT Hardware**

Topic 1 :	Topic 2 :	Topic 3 :
system configuration	networking equipment	adp components
adpe system	telegraph equipment	storage devices
afway tracking	acat iii	input output
acat afway	provide spares	adp input
office desktop	network centric	networking equipment
networking equipment	centric solutions	delivery date
hewlett Packard	voip phones	afway tracking
computer control	delivery date	hewlett packard
control devices	control switching	acat nasic
micro computer	switching system	office desktop

Topic 4 :	Topic 5 :	Topic 6 :
information system	communication equipment	support equipment
system equipment	miscellaneous communication	adp support
office information	adp supplies	afway tracking
land mobile	netcentric products	processing unit
capabilities fy	networking equipment	central processing
contractor logistical	acat iii	adp central
logistical support	acat nasic	cpu computer
mobile radios	acat afrl	unit cpu
communication equipment	nasic scxa	computer digital
television communication	pop extension	fiber optic

Instructions

The following word lists come from IT Level 2 Category contract text descriptions. For each list of words type the "topic" the words are likely describing in the highlighted cell above each list. Please limit the topic responses to three words or less. If the word list does not define any topic, leave it blank.

The Level 2 Category label applies to all topics on the page.

Level 2 Category: **IT Hardware**

Topic 1 :	Topic 2 :	Topic 3 :
exercise option	support services	support services
elevator maintenance	acat iii	incremental funding
overhead doors	incremental funding	air force
incremental funding	exercise option	acat afrl
nsn fd	sustainment support	acat aflcmc
preventive maintenance	information technology	management system
remedial maintenance	engineering services	information technology
support services	engineering support	exercise option
de obligation	software engineering	maintenance sustainment
option period	management system	spectrum management

Topic 4 :	Topic 5 :	Topic 6 :
base telecommunications	switching system	system installation
telecommunications system	control switching	emergency system
support services	maintenance service	basic center
monthly maintenance	repair task	center operations
base telecommunication	lmr maintenance	frequency global
incremental funding	land mobile	global communications
base telephone	mobile radio	antenna install
system services	hardware maintenance	communications systems
business sb	system contract	systems antenna
sb companion	annual hardware	cama ali

Topic 7 :
fiber optic
network centric
centric solutions
support services
optic cable
ffp loe
requisite engineering
focis iii
services ffp
adp equipment

Instructions

The following word lists come from IT Level 2 Category contract text descriptions. For each list of words type the "topic" the words are likely describing in the highlighted cell above each list. Please limit the topic responses to three words or less. If the word list does not define any topic, leave it blank.

The Level 2 Category label applies to all topics on the page.

Level 2 Category: **IT Software**

Topic 1 :	Topic 2 :	Topic 3 :
adp software	adp software	adp software
acat afrl	license renewal	software maintenance
software maintenance	software maintenance	software license
premium sw	agreement exercise	seat license
xenapp xendesktop	graphics pads	annual ccc
eda card	commons month	mapping toolbox
interface infrastructure	gfp clause	cie dren
server option	core system	itree software
tdl wsn	wills software	database query
aflcmc hia	emulex fiber	afrl rymh

Topic 4 :	Topic 5 :
acat afrl	software maintenance
software maintenance	annual software
software license	maintenance service
contract task	telecom annual
operational picture	service plans
update ombudsman	software support
taskings tss	maintenance renewal
network li	software license
vrsg licensing	air force
trusted licenses	annual maintenance

Instructions

The following word lists come from IT Level 2 Category contract text descriptions. For each list of words type the "topic" the words are likely describing in the highlighted cell above each list. Please limit the topic responses to three words or less. If the word list does not define any topic, leave it blank.

The Level 2 Category label applies to all topics on the page.

Level 2 Category: **Telecommunications**

Topic 1 :	Topic 2 :	Topic 3 :
telecom telecommunications	telecom telecommunications	internet service
cable tv	land mobile	commercial internet
land mobile	maintenance services	internet services
correct line	cable distribution	cable service
television services	sw cable	speed internet
system maintenance	cable outlets	dawgnet internet
modification changing	price adjustment	exercise option
changing unit	tv requirement	optical internet
excess funding	usaf fhc	broadband internet
communication telephone	government's obligation	cable television

Topic 4 :
telecom telecommunications
cell phone
minot afb
month funds
cama trunk
commercial ds
center internet
clin description
service mbps
internet phase

Appendix C. Topic Assignment Sheet Responses

IT Security						
Topic 1	*Topic 2*	*Topic 3*	*Topic 4*			
			Authentication			
Equipment	Maintenance Equipment	Proof of Concept	Crypto Security			
Physical Security	Continuity of Operations & Services	Data Security	Identity & Access Management			
Secure Measures	Security Resources	Defense Measures	Security Countermeasures			
Surveillance	Security	Platform	Encryption			
IT Consulting						
Topic 1	*Topic 2*	*Topic 3*	*Topic 4*			
Sustainment	Operation Support	System Integration	Services			
Development			Support			
	Service Categories	Service Types				
Training & Support Services	Sustainment Services	Infrastructure Sustainment	Sustainment Business Activities			
Mission Support	Compliance Services	Funds Management	Funding Outliers			
IT Hardware						
Topic 1	*Topic 2*	*Topic 3*	*Topic 4*	*Topic 5*	*Topic 6*	
System	Network	Storage	Communications	Infrastructures	Workstation	
	Infrastructure		Communication		Computer Components	
Computer	Network	Peripheral	Support	Products	Description	
Asset Management	Network Infrastructure	IT Equipment	Systems Configuration	Data / Voice Communications Support	Infrastructure Sustainment	
Data Asset	Connectivity	Support Systems	Info Sharing	Reach Devices	Centralization Equip	
IT Outsourcing						
Topic 1	*Topic 2*	*Topic 3*	*Topic 4*	*Topic 5*	*Topic 6*	*Topic 7*
	Support		Communication	Maintenance	Ops	Network
Contracts	Services	Management	Contract Categories	Description	Services Offered	Network Products
Physical Plant Maintenance	Engineering Support	Technical Support - Systems	Telecommunications	Telecommunications Support	Telecommunications Maintenance& Installation	Telecommunications Infrastructure
Maintenance	Sustainment Support	Communications	Telecommunications	Interconnection	Installation	Network Connection
System Support	Life Cycle Support	Management Support	Wideband Support	Configuration Management	Command Control Support	Solutions Management
IT Software						
Topic 1	*Topic 2*	*Topic 3*	*Topic 4*	*Topic 5*		
				License		
Software	Software Maintenance	Licensing	Software Licenses	Renewals		
Baseline Systems Configuration	Software Sustainment	Software License Management	Software Technical Support	Software Subscription Management		
Support	Sustainment	Governance Support	Verification Support	Long Term Support		
Infrastructure	System	End User	Network	Software Support		
Telecommunications						
Topic 1	*Topic 2*	*Topic 3*	*Topic 4*			
Connectivity	Comm Support	Digital Support	Phased Comm			
Television	Television Hardware	Internet				
Telecom Services	Telecom Infrastructure	Service Categories	Service Types			
Intranet Management	Infrastructure Management	Internet Access Management	Telecommunications Systems Management			
Global Communications	Services	Network Connection	Cellular			

Appnedix D. Quad Chart

STRATEGIC SOURCING VIA CATEGORY MANAGEMENT: HELPING AIR FORCE INSTALLATION CONTRACTING AGENCY EAT ONE PIECE OF THE ELEPHANT

Master Sergeant Theodore S. Holliger
Advisor: Bradley Boehmke, PhD
Reader: Col Matthew Douglas, PhD
Reader: Jeffrey Ogden, PhD
Reader: Edward White, PhD
Department of Operational Sciences (ENS)
Air Force Institute of Technology

Opportunity Statement:

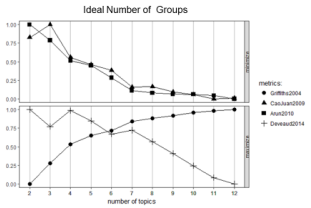
Significant opportunities exist for the USAF to realize savings from strategic sourcing initiatives, specifically within the IT-related installation support spend Level-1 category. To leverage strategic sourcing strategies, the USAF must first objectively group interrelated products and services into sub-categories.

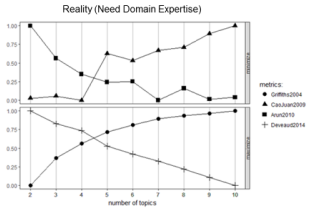
Purpose Statement:

This research seeks to develop an empirical methodology to uncover the hidden structure of products and services beyond the known Level-2 categories. In addition, this research will attempt to extract themes that may be present in the data. Visualizing commodities at a granular level will bolster strategic sourcing initiatives thereby strengthening the competitive advantage of the Air Force over adversaries.

Data:

An authoritative data file was compiled by AFICA using the Federal Procurement Data System – Next Generation (FPDSNG). The file contained 107,589 contract actions from October 2012 through April 2017. Each action contained and associated a text description of products and/or service acquired, Product Service Code, and Level-2 category membership.





Methodology:

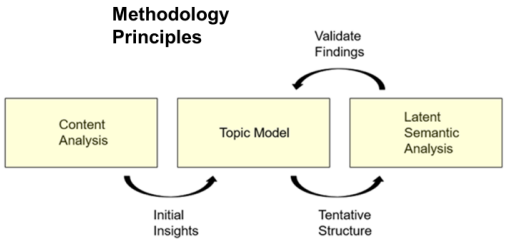
Content Analysis and Latent Semantic Analysis principles were utilized to steer the development of an expanded methodology that incorporated a host of text mining techniques. The methodology was applied to a Level-2 category to extend the categorical hierarchy beyond the known structure.

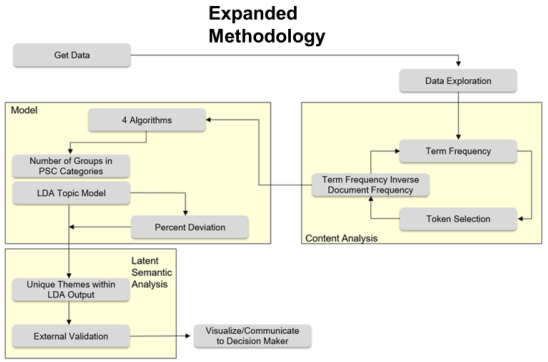
Conclusion:

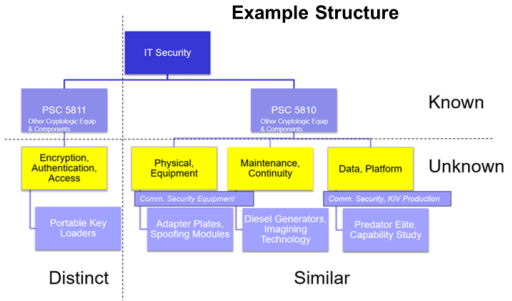
Text mining techniques can extract meaningful insights from historical purchase contract data, highlight group thematic nature, identify group product/service membership, and map the previously unknown structure down to a granular level. However, domain expertise is critical to practically significant analysis.

Recommendations:

Decision makers should harness the synergistic effect of collaboratively analyzing contract data by co-locating SMEs and analysts or at least merging analysis functions with domain expertise. In addition, leadership should encourage text mining techniques in regard to purchase contract analysis since it is both reliable and valid.







Bibliography

AFICA Flight Plan. (n.d.). Retrieved from https://cs.eis.af.mil/sites/10074/afcc/afica/KA/strat_comm/Flight Plan Worksite/FINAL - AFICA Flight Plan.pdf

Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. *Proceedings of the 14th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part I* (pp. 391–402). Berlin, Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-642-13657-3_43

Basuroy, S., Mantrala, M. K., & Walters, R. G. (2001). The Impact of Category Management on Retailer Prices and Performance: Theory and Evidence. *Journal of Marketing*, *65*(4), 16–32. https://doi.org/10.1509/jmkg.65.4.16.18382

Blei, D., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, *27*(6), 55–65. https://doi.org/10.1109/MSP.2010.938079

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A Density-based Method for Adaptive LDA Model Selection. *Neurocomputing*, *72*(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Cox, A. (2015). Sourcing portfolio analysis and power positioning: towards a "paradigm shift" in category management and strategic sourcing. *Supply Chain Management: An International Journal*, *20*(6), 717–736. https://doi.org/10.1108/SCM-06-2015-0226

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Revue Des Sciences et Technologies de l'Information - Serie Document Numerique, Lavoisier*, *17*(1), 61–84. https://doi.org/10.3166/dn.17.1.61-84

GSA (2015). Government-Wide Category Management, (May), 1–42. [Accessed March 1, 2017] https://hallways.cap.gsa.gov/information/Gov-wide_CM_Guidance_V1.pdf .

Dooley, K. J. (2016). Using manifest content analysis in purchasing and supply management research. *Journal of Purchasing and Supply Management*, *22*(4), 244–246. https://doi.org/10.1016/j.pursup.2016.08.004

DPAP (n.d.). Defense Procurement and Acquisition Policy. [Accessed March 4, 2017] https://www.acq.osd.mil/dpap/ss/index.html

GAO (2012). *Strategic Sourcing: Improved and Expanded Use Could Save Billions in Annual Procurement Costs*. *United States Government Accountability Office - Report* (Vol. GAO-12-919). Washington D.C.

GAO (2015). *Strategic Opportunities Exist to Better Manage Information Technology Services Spending Opportunities Exist to Better Manage Information (*Vol. GAO-15-549). Washington D.C.

Gelderman, Cees J & van Weele, A. J. (2005). Purchasing Portfolio Models-A critique and update. *Journal of Supply Chain Management*, *Summer* (August), 19–29.

Gibson, A. (2017). SECAF: The first interview. [Accessed April 1, 2017] http://www.af.mil/News/Article-Display/Article/1199661/secaf-the-first-interview/

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(Supplement 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Hansen, M. H., Perry, L. T., & Reese, S. (2004). A bayesian operationalization of the resource-based view. *Strategic Management Journal*, (25), 1279–1295. https://doi.org/10.1002/smj.432.

Helfat, C., Finkelstein, S., Mitchell, W., Peteraf, M., Singh, H., Teece, D. and Winter, S. (2007). *Dynamic Capabilities: Understanding Strategic Change in Organisations*. Blackwell Publishing, Malden.

Hesping, F. H., & Schiele, H. (2015). Purchasing strategy development: A multi-level review. *Journal of Purchasing and Supply Management*, *21*(2), 138–150. https://doi.org/10.1016/j.pursup.2014.12.005

Horn, P., Schiele, H., & Werner, W. (2013). The “ugly twins”: Failed low-wage-country sourcing projects and their expensive replacements. *Journal of Purchasing and Supply Management*, *19*(1), 27–38. https://doi.org/http://dx.doi.org/10.1016/j.pursup.2012.09.001

Hunt, S. D., & Davis, D. F. (2012). Grounding supply chain management in resource - advantage theory : in defense of a resource - based view of the firm . *Journal of Supply Chain Management*, *48*(2), 12. https://doi.org/10.1111/j.1745-493X.2012.03266.x

Kraljic, P. (1983). Purchasing must become supply management: a strategy for supply. *Harvard Business Review*, *Sept-Oct*, 109–117.

Leedy, P. D., & Ormrod, J. E. (2013). *Practical Research: Planning and Design* (10th edn). Upper Saddle River NJ: Pearson.

Luca, M., Kleinberg, J., & Mullainathan, S. (2016). Algorithms Need Managers, Too. *Havard Business Review*, 96–101.

Luzzini, D., Caniato, F., Ronchi, S., & Spina, G. (2012). A transaction costs approach to purchasing portfolio management. *Journal of Production Management*, *32*(9), 1015–1042. https://doi.org/10.1108/01443571211265684

Luzzini, D., & Ronchi, S. (2011). Organizing the purchasing department for innovation. *Operations Management Research*, *4*(1), 14–27. https://doi.org/10.1007/s12063-010-0042-2

Montgomery, R. T. (2015). *Using Multiple Objective Decision Analysis to Position Federal Product and Service Codes Within the Kraljic Portfolio matrix (Master's Thesis)*. Air Force Institute of Technology, Dayton OH.

Muir, W. A., Keller, R. S. & Knight, L. E. (2014). *Category Management: A Concept of Operations for Improving Costs at the Air Force Installation*. United States Air Force.

Newman D., Lau J. H., Grieser K., B. T. (2010). Automatic Evaluation of Topic Coherence In Human Language Technologies. *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 100–108). Stroudsburg PA.

Olsen, R. F., & Ellram, L. M. (1997). A portfolio approach to supplier relationships. *Industrial Marketing Management*, *26*(2), 101–113. https://doi.org/10.1016/S0019-

8501(96)00089-2

Serbu, J. (2017). Air Force plans June summit to scrub acquisitions reguations. [Accessed January 5, 2018] https://federalnewsradio.com/air-force/2017/06/air-force-plans-june-summit-to-scrub-acquisition-regulations/

Simpson, D., Meredith, J., Boyer, K., Dilts, D., Ellram, L. M., & Leong, G. K. (2015). Professional, Research, and Publishing Trends in Operations and Supply Chain Management. *Journal of Supply Chain Management*, *51*(3), 87–100. https://doi.org/10.1111/jscm.12078

Sirmon, D. G., Hitt, M. A., Ireland, R. D., & Gilbert, B. A. (2011). Resource Orchestration to Create Competitive Advantage: Breadth, Depth, and Life Cycle Effects. *Journal of Management*, *37*(5), 1390–1412. https://doi.org/10.1177/0149206310385695

Sirmon, D. G., Hitt, M. A. & Ireland, R. D. (2007). Managing Firm Resources in Dynamic Environments To Create Value : Looking Inside the Black Box. *Academy of Management Review*, *32*(1), 273–292.

Trautmann, G., Turkulainen, V., Hartmann, E., & Bals, L. (2009). Integration in the global sourcing organization - An information processing perspective. *Journal of Supply Chain Management*, *45*(2), 57–74. https://doi.org/10.1111/j.1745-493X.2009.03163.x

U.S. Office of Management and Budget. (2005). Implementing Strategic Sourcing. [Accessed February 7, 2017] www.uspto.gov/web/offices/ac/comp/proc/OMBmemo.pdf

US General Services Administration Federal Acquisition Services. (2015). Federal Procurement Data System: Product and Service Codes Manual. [Accessed November 7, 2017] https://www.acquisition.gov/sites/default/files/page_file_uploads/PSC Manual - Final - 9 August 2015_0.pdf

Waller, M. A., & Fawcett, S. E. (2013). Data Science , Predictive Analytics , and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics*, *34*(2), 77–84. https://doi.org/10.1111/jbl.12010

Zipf, G. (1932). *Selective Studies and the Principle of Relative Frequency in Language*. Cambridge MA.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
22-03-2018	Master's Thesis	October 2017 – March 2018

TITLE AND SUBTITLE

Strategic Sourcing Via Category Management: Helping Air Force Installation Contracting Agency Eat One Piece of the Elephant

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Holliger, Theodore S., Master Sergeant, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way, Building 640
WPAFB OH 45433-8865

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENS-MS-18-M-128

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Installation Contracting Agency
Roger H. Westermeyer, Director of Enterprise Sourcing Support
1940 Allbrook Drive
Bldg. 1
Wright-Patterson AFB, OH 45433
darin.ashley@us.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

AFICA/KA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRUBTION STATEMENT A. APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES
This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT

The United States Air Force can dramatically reduce resource consumption through strategic sourcing initiatives that leverage sensibly-bound pockets of spend via category management. However, category creation is a particularly daunting task due to the sheer magnitude of purchasing data in large organizations. Text mining is one way to identify categories. Specifically, term frequency analysis, term frequency-inverse document frequency analysis, and topic modeling can identify category membership, unique characteristics of categories, and thematic natures of the categories. This thesis developed an empirical, generalizable, reproducible methodology to analyze historical contract text descriptions to uncover the data's hidden structure. A sample case was transformed into a practical hierarchy, which was internally and externally validated. As a foundational methodology, the impact of token selection, domain expertise, and unique contracting language were identified as considerations for future research.

15. SUBJECT TERMS
Strategic Sourcing, Category Management, Content Analysis, Latent Semantic Analysis, Topic Model

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Col Matthew A. Douglas, Ph.d, AFIT/ENS
U	**U**	**U**	**UU**	108	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-3636, ext 4737 matthew.douglas@afit.edu

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18