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

DISTRIBUTION STATEMENT A.
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.
The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.
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

DISTRIBUTION STATEMENT A.
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.
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
AFIT-ENS-MS-18-M-128

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

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>iv</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>vii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>x</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xi</td>
</tr>
<tr>
<td><strong>I. Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>Air Force Installation Contracting Agency</td>
<td>2</td>
</tr>
<tr>
<td>Research Question, Purpose, and Scope</td>
<td>3</td>
</tr>
<tr>
<td>Investigative Questions/Research Method</td>
<td>5</td>
</tr>
<tr>
<td>Implications and Research Organization</td>
<td>6</td>
</tr>
<tr>
<td><strong>II. Literature Review</strong></td>
<td>7</td>
</tr>
<tr>
<td>Chapter Overview</td>
<td>7</td>
</tr>
<tr>
<td>Resource Orchestration Theory</td>
<td>7</td>
</tr>
<tr>
<td>Strategic Sourcing in the Federal Government</td>
<td>8</td>
</tr>
<tr>
<td>Category Management in the USAF</td>
<td>9</td>
</tr>
<tr>
<td>Role of Strategic Sourcing and Category Management in ROT</td>
<td>10</td>
</tr>
<tr>
<td>Strategic Sourcing Via Category Management</td>
<td>11</td>
</tr>
<tr>
<td>Conclusion and Way Forward</td>
<td>13</td>
</tr>
<tr>
<td><strong>III. Methodology and Data Description</strong></td>
<td>15</td>
</tr>
<tr>
<td>Chapter Overview</td>
<td>15</td>
</tr>
<tr>
<td>Methodology Overview</td>
<td>15</td>
</tr>
<tr>
<td>Get Data</td>
<td>16</td>
</tr>
<tr>
<td>Statistical Software</td>
<td>17</td>
</tr>
<tr>
<td>Data Exploration</td>
<td>18</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>OMB Level-1 and Level-2 Categories</td>
<td>4</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Resource Orchestration Overview</td>
<td>8</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Methodology Overview</td>
<td>16</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Expanded Methodology</td>
<td>17</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Contract Actions by PSC and Level-2 Category</td>
<td>30</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Uni-grams IT Security by Frequency</td>
<td>32</td>
</tr>
<tr>
<td>Figure 7</td>
<td>IT Security Bi-grams by TF-IDF Weight</td>
<td>34</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Four-Algorithm Optimal Topic (IT Security)</td>
<td>35</td>
</tr>
<tr>
<td>Figure 9</td>
<td>LDA Output (Unnamed Topics) IT Security</td>
<td>35</td>
</tr>
</tbody>
</table>
List of Tables

Table 1. R Packages ........................................................................................................... 18
Table 2. Relevant Variables and Brief Description ............................................................ 19
Table 3. Percent Deviation Similar Groups (IT Security) ..................................................... 36
Table 4. IT Security Topic Word List .................................................................................. 37
Table 5. Summary of Results ............................................................................................ 37
Table 6. Summarized Topic Descriptions .......................................................................... 38
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
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).

![Government-Wide Category Organization](image)

**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 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
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**

<table>
<thead>
<tr>
<th>Package</th>
<th>Author(s)</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>lubridate</td>
<td>Garrett Grolemund and Hadley Wickham (2011)</td>
<td><a href="http://www.jstatsoft.org/v40/i03/">http://www.jstatsoft.org/v40/i03/</a></td>
</tr>
<tr>
<td>tidyverse</td>
<td>Hadley Wickham (2017)</td>
<td><a href="https://CRAN.R-project.org/package=tidyverse">https://CRAN.R-project.org/package=tidyverse</a></td>
</tr>
<tr>
<td>tidytext</td>
<td>J. Silge and D. Robinson (2016)</td>
<td><a href="http://dx.doi.org/10.21105/joss.00037">http://dx.doi.org/10.21105/joss.00037</a></td>
</tr>
</tbody>
</table>

**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**

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>text_describe</td>
<td>Brief description of goods or services bought (for award) or available (for IDV).</td>
</tr>
<tr>
<td>PSC</td>
<td>Product Service Code</td>
</tr>
<tr>
<td>lvl_2_category</td>
<td>Category the contract is assigned by OMB and GSA.</td>
</tr>
</tbody>
</table>

**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
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 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)

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

<table>
<thead>
<tr>
<th>Level 2 Category</th>
<th>PSCs</th>
<th>Optimal Topics</th>
<th>Percent Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT Security</td>
<td>2</td>
<td>4</td>
<td>Yes</td>
</tr>
<tr>
<td>IT Consulting</td>
<td>4</td>
<td>3</td>
<td>Yes</td>
</tr>
<tr>
<td>IT Hardware</td>
<td>31</td>
<td>6</td>
<td>No</td>
</tr>
<tr>
<td>IT Outsourcing</td>
<td>31</td>
<td>7</td>
<td>No</td>
</tr>
<tr>
<td>IT Software</td>
<td>2</td>
<td>5</td>
<td>Yes</td>
</tr>
<tr>
<td>Telecommunications</td>
<td>4</td>
<td>4</td>
<td>Yes</td>
</tr>
</tbody>
</table>

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 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).

<table>
<thead>
<tr>
<th>IT Security</th>
<th>IT Consulting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1: Physical, Equipment</td>
<td>Topic 1: Training, Mission, Sustainment, Development</td>
</tr>
<tr>
<td>Topic 2: Maintenance, Continuity</td>
<td>Topic 2: Compliance, Sustainment</td>
</tr>
<tr>
<td>Topic 3: Data, Platform</td>
<td>Topic 3: Integration, Infrastructure, Funds, Sustainment</td>
</tr>
<tr>
<td>Topic 4: Encryption, Authentication, Access</td>
<td>Topic 4: Sustainment</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IT Hardware</th>
<th>IT Outsourcing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1: Computer, Asset, Data</td>
<td>Topic 1: Physical, Plant, Maintenance</td>
</tr>
<tr>
<td>Topic 2: Network, Infrastructure</td>
<td>Topic 2: Engineering, Sustainment, Life Cycle</td>
</tr>
<tr>
<td>Topic 3: Peripheral, Storage, Equipment</td>
<td>Topic 3: Communication</td>
</tr>
<tr>
<td>Topic 4: Configuration, Communication</td>
<td>Topic 4: Communication, Telecommunication, Wideband</td>
</tr>
<tr>
<td>Topic 5: Infrastructure, Data, Voice, Devices</td>
<td>Topic 5: Maintenance, Telecommunication, Configuration</td>
</tr>
<tr>
<td>Topic 7: Network, Telecommunications, Infrastructure, Solutions</td>
<td>Topic 7: Network, Telecommunications, Infrastructure, Solutions</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IT Software</th>
<th>Telecommunications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1: Configuration, Infrastructure</td>
<td>Topic 1: Television, Intranet</td>
</tr>
<tr>
<td>Topic 2: Maintenance, Sustainment</td>
<td>Topic 2: Television, Infrastructure</td>
</tr>
<tr>
<td>Topic 3: License, End User, Governance</td>
<td>Topic 3: Internet, Digital, Network, Access</td>
</tr>
<tr>
<td>Topic 4: Network, License, Verification</td>
<td>Topic 4: Cellular, Phased</td>
</tr>
<tr>
<td>Topic 5: License, Renewals, Subscription</td>
<td></td>
</tr>
</tbody>
</table>
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 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

![Figure 11. Computer PSCs Weighted by TF-IDF](image-url)
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](image)

41
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)

```r
# Filter by IT Security and see top 10 words by PSC

## 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() %>%
  bind_tf_idf(word, lvl_2_category, n) %>%
  arrange(desc(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 Unigrams in IT Security by PSC",
       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) %>%
  unnest_tokens(word, text_describe) %>%
  dplyr::count(word, sort = TRUE) %>%
  anti_join(stop_words) %>%
  anti_join(common_terms) %>%
  ungroup() %>%
  bind_tf_idf(word, lvl_2_category, n) %>%
  arrange(desc(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 Unigrams in IT Security by PSC",
       y = "tf-idf",
       title = "Top 10 Unigrams in IT Security by PSC",
       subtitle = "Weighted by tf-idf") +
  coord_flip() +
  theme(legend.position = "none")
```

---

### Top 10 Unigrams in IT Security by PSC

#### Weighted by tf-idf

- "sic"
- "scenario"
- "mkv"
- "cql"
- "slick"
- "mics"
- "jm"
- "grep"
- "sqlitze"
- "bea"
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) %>%
```r
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),
              "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) %>%
  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) %>%
  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>

54
# Percent Deviation Topic 2

```r
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

<table>
<thead>
<tr>
<th>term</th>
<th>percent_dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>portable avenger</td>
<td>182.7984</td>
</tr>
<tr>
<td>gfp correction</td>
<td>180.4616</td>
</tr>
<tr>
<td>diesel generator</td>
<td>172.1421</td>
</tr>
<tr>
<td>ldc audio</td>
<td>171.3365</td>
</tr>
<tr>
<td>lmr motorola</td>
<td>169.5146</td>
</tr>
<tr>
<td>recaro seat</td>
<td>169.4502</td>
</tr>
<tr>
<td>imaging technology</td>
<td>167.4894</td>
</tr>
<tr>
<td>clin transfer</td>
<td>166.8580</td>
</tr>
<tr>
<td>extension clin</td>
<td>166.1154</td>
</tr>
<tr>
<td>auto acquire</td>
<td>164.3151</td>
</tr>
</tbody>
</table>

# Percent Deviation Topic 3

```r
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

<table>
<thead>
<tr>
<th>term</th>
<th>percent_dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>portable avenger</td>
<td></td>
</tr>
<tr>
<td>gfp correction</td>
<td></td>
</tr>
<tr>
<td>diesel generator</td>
<td></td>
</tr>
<tr>
<td>ldc audio</td>
<td></td>
</tr>
<tr>
<td>lmr motorola</td>
<td></td>
</tr>
<tr>
<td>recaro seat</td>
<td></td>
</tr>
<tr>
<td>imaging technology</td>
<td></td>
</tr>
<tr>
<td>clin transfer</td>
<td></td>
</tr>
<tr>
<td>extension clin</td>
<td></td>
</tr>
<tr>
<td>auto acquire</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;chr&gt;</td>
</tr>
<tr>
<td>---</td>
<td>----------------------</td>
</tr>
<tr>
<td>1</td>
<td>predator elite</td>
</tr>
<tr>
<td>2</td>
<td>capability study</td>
</tr>
<tr>
<td>3</td>
<td>additional days</td>
</tr>
<tr>
<td>4</td>
<td>clin overrun</td>
</tr>
<tr>
<td>5</td>
<td>linux production</td>
</tr>
<tr>
<td>6</td>
<td>apx digital</td>
</tr>
<tr>
<td>7</td>
<td>afb option</td>
</tr>
<tr>
<td>8</td>
<td>mountain home</td>
</tr>
<tr>
<td>9</td>
<td>acts crypto</td>
</tr>
<tr>
<td>10</td>
<td>communication equip</td>
</tr>
</tbody>
</table>
Telecommunications

## TF-IDF

```r
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),
"Deveaud2014"),
method = "Gibbs",
control = list(seed = 1234),
mc.cores = 2L,
verbose = TRUE
)
FindTopicsNumber_plot(result_telecom)
## LDA Model

```r
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

```r
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

```r
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
```
```r
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]

## Percent Deviation Topic 1

term percent_dev
   correct line   192.4841
 telescope services  188.9750
    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]

## Percent Deviation Topic 2

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
```

---

60
## Percent Deviation Topic 4

```r
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]
```

<table>
<thead>
<tr>
<th>term</th>
<th>percent_dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>month funds</td>
<td>193.9340</td>
</tr>
<tr>
<td>cama trunk</td>
<td>192.8743</td>
</tr>
<tr>
<td>commercial ds</td>
<td>190.3175</td>
</tr>
<tr>
<td>center internet</td>
<td>189.5056</td>
</tr>
<tr>
<td>clin description</td>
<td>186.5030</td>
</tr>
<tr>
<td>service mbps</td>
<td>185.9766</td>
</tr>
<tr>
<td>internet phase</td>
<td>180.1063</td>
</tr>
<tr>
<td>fy internet</td>
<td>179.9482</td>
</tr>
<tr>
<td>sirius xm</td>
<td>178.2841</td>
</tr>
<tr>
<td>force base</td>
<td>176.2873</td>
</tr>
</tbody>
</table>
IT Outsourcing

## TF-IDF

```r
na.omit(relevant_itdata) %>%
  mutate(text_describe = str_replace_all(text_describe,
                                          pattern = c(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 %>%
  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 Outsourcing by PSC",
       subtitle = "Weighted by tf-idf") +
  coord_flip() +
  theme(legend.position = "none")
```
## Document Term Matrix

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

## Optimal Number of Topics

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

## LDA Model

```r
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

```r
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()
## TF-IDF

```r
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

```r
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() %>%
```r
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

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

## Optimal Number of Topics

```r
result_software <- FindTopicsNumber(
clean_ITSoftware_bigram_dtm, 
topics = seq(from = 2, to = 10, by = 1), 
"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

```r
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

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

```r
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

<table>
<thead>
<tr>
<th>term</th>
<th>percent_dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>final deobligation</td>
<td>261.0584</td>
</tr>
<tr>
<td>d.o pop</td>
<td>251.6756</td>
</tr>
<tr>
<td>integration test</td>
<td>250.4637</td>
</tr>
<tr>
<td>wkc pk</td>
<td>250.2471</td>
</tr>
<tr>
<td>facilitate payment</td>
<td>249.7149</td>
</tr>
<tr>
<td>licenses voice</td>
<td>248.3278</td>
</tr>
<tr>
<td>add logo</td>
<td>246.3301</td>
</tr>
<tr>
<td>market patriot</td>
<td>245.0462</td>
</tr>
<tr>
<td>starteam enterprise</td>
<td>239.9419</td>
</tr>
<tr>
<td>windows server</td>
<td>239.6439</td>
</tr>
</tbody>
</table>

## Topic 2

```r
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")
```

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

<table>
<thead>
<tr>
<th>term</th>
<th>percent_dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>security software</td>
<td>254.8172</td>
</tr>
<tr>
<td>sight support</td>
<td>254.1522</td>
</tr>
<tr>
<td>support monarch</td>
<td>253.1624</td>
</tr>
<tr>
<td>cals telerik</td>
<td>253.0215</td>
</tr>
<tr>
<td>infrastructure division</td>
<td>251.0456</td>
</tr>
<tr>
<td>renewal pc</td>
<td>249.0562</td>
</tr>
<tr>
<td>oscilloscope mobile</td>
<td>248.5901</td>
</tr>
<tr>
<td>av xamarin</td>
<td>246.2859</td>
</tr>
<tr>
<td>cables mcafee</td>
<td>245.4927</td>
</tr>
<tr>
<td>equipment netowl</td>
<td>244.1933</td>
</tr>
</tbody>
</table>

## Topic 3

```r
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")
```
### Topic 3

<table>
<thead>
<tr>
<th>term</th>
<th>percent_dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>microsoft software</td>
<td>298.7376</td>
</tr>
<tr>
<td>clause hardware</td>
<td>295.2698</td>
</tr>
<tr>
<td>seymour johnson</td>
<td>294.7687</td>
</tr>
<tr>
<td>host center</td>
<td>293.9748</td>
</tr>
<tr>
<td>advanced customer</td>
<td>291.1535</td>
</tr>
<tr>
<td>ms ts</td>
<td>290.7404</td>
</tr>
<tr>
<td>computer service</td>
<td>290.6979</td>
</tr>
<tr>
<td>subscription base</td>
<td>290.0651</td>
</tr>
<tr>
<td>mortuarys mission</td>
<td>288.1281</td>
</tr>
<tr>
<td>asapk description</td>
<td>287.9197</td>
</tr>
</tbody>
</table>

### Topic 4

<table>
<thead>
<tr>
<th>term</th>
<th>percent_dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>switchview kvm</td>
<td>270.7886</td>
</tr>
<tr>
<td>cost bilateral</td>
<td>268.6854</td>
</tr>
<tr>
<td>license ida</td>
<td>256.3139</td>
</tr>
<tr>
<td>standard sw</td>
<td>256.3136</td>
</tr>
<tr>
<td>solarwinds licenses</td>
<td>252.9405</td>
</tr>
<tr>
<td>life storage</td>
<td>249.3739</td>
</tr>
<tr>
<td>bpel licenses</td>
<td>248.5678</td>
</tr>
<tr>
<td>support emergency</td>
<td>242.7162</td>
</tr>
<tr>
<td>support imagine</td>
<td>240.3896</td>
</tr>
<tr>
<td>support linux</td>
<td>237.3325</td>
</tr>
</tbody>
</table>
## TF-IDF

```r
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")
```

**Top 10 Bigrams in IT Hardware by PSC**

Weighted by tf-idf

- 5905: telegraph equipment
- 5820: television communication
- 5821: telemetry support
- 5905: provide spares
- 5820: operational control
- 5821: support package
- 5905: control switching
- 5820: support package id
- 5905: switching system
- 5821: mode aifl
- 5905: cell phone
- 5905: network centric
- 5905: system delivery
- 5820: mobile radios
- 5905: switching system
- 5821: mobile radios
- 5905: telecommunication
- 5905: moteur vehicles
- 5905: vehicle
- 5905: telecommunication
- 5905: telecommunication
- 5905: telecommunication

73
## 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),
  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

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

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

## Plot LDA Output

```r
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()
```
## TF-IDF

```r
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),  
"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
```r
## Percent Deviation

```r
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 %>%
  mutate(per_t1 = (avdev_t1/mean) * 100, per_t3 = (avdev_t3/mean) * 100) %>%

## Topic 1

```r
c_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")

c_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
eeronet 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**

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

**Topic 4**

<table>
<thead>
<tr>
<th>Topic 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>air force</td>
</tr>
<tr>
<td>force key</td>
</tr>
<tr>
<td>key loader</td>
</tr>
<tr>
<td>key management</td>
</tr>
<tr>
<td>management infrastructure</td>
</tr>
<tr>
<td>portable key</td>
</tr>
<tr>
<td>cryptographic core</td>
</tr>
<tr>
<td>infrastructure cryptographic</td>
</tr>
<tr>
<td>anti spoofing</td>
</tr>
<tr>
<td>availability anti</td>
</tr>
</tbody>
</table>
**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**

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>application services</td>
<td>acat ii</td>
<td>netcents application</td>
</tr>
<tr>
<td>netcents application</td>
<td>engineering services</td>
<td>application services</td>
</tr>
<tr>
<td>development sustainment</td>
<td>incremental funding</td>
<td>integration installation</td>
</tr>
<tr>
<td>price increase</td>
<td>aar system</td>
<td>add ecmsra</td>
</tr>
<tr>
<td>dcc camera</td>
<td>acat iii</td>
<td>design emd</td>
</tr>
<tr>
<td>management security</td>
<td>dc sustainment</td>
<td>uft software</td>
</tr>
<tr>
<td>sp training</td>
<td>minor tape</td>
<td>service deomi</td>
</tr>
<tr>
<td>applications support</td>
<td>pop extension</td>
<td>clin acrn</td>
</tr>
<tr>
<td>telephony products</td>
<td>alg system</td>
<td>mar car</td>
</tr>
<tr>
<td>air program</td>
<td>smc support</td>
<td>acctg line</td>
</tr>
</tbody>
</table>

**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**

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>system configuration</td>
<td>networking equipment</td>
<td>adp components</td>
</tr>
<tr>
<td>adpe system</td>
<td>telegraph equipment</td>
<td>storage devices</td>
</tr>
<tr>
<td>afway tracking</td>
<td>acat iii</td>
<td>input output</td>
</tr>
<tr>
<td>acat afway</td>
<td>provide spares</td>
<td>adp input</td>
</tr>
<tr>
<td>office desktop</td>
<td>network centric</td>
<td>networking equipment</td>
</tr>
<tr>
<td>networking equipment</td>
<td>centric solutions</td>
<td>delivery date</td>
</tr>
<tr>
<td>hewlett Packard</td>
<td>voip phones</td>
<td>afway tracking</td>
</tr>
<tr>
<td>computer control</td>
<td>delivery date</td>
<td>hewlett packard</td>
</tr>
<tr>
<td>control devices</td>
<td>control switching</td>
<td>acat nasic</td>
</tr>
<tr>
<td>micro computer</td>
<td>switching system</td>
<td>office desktop</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic 4</th>
<th>Topic 5</th>
<th>Topic 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>information system</td>
<td>communication equipment</td>
<td>support equipment</td>
</tr>
<tr>
<td>system equipment</td>
<td>miscellaneous communication</td>
<td>adp support</td>
</tr>
<tr>
<td>office information</td>
<td>adp supplies</td>
<td>afway tracking</td>
</tr>
<tr>
<td>land mobile</td>
<td>netcentric products</td>
<td>processing unit</td>
</tr>
<tr>
<td>capabilities fy</td>
<td>networking equipment</td>
<td>central processing</td>
</tr>
<tr>
<td>contractor logistical</td>
<td>acat iii</td>
<td>adp central</td>
</tr>
<tr>
<td>logistical support</td>
<td>acat nasic</td>
<td>cpu computer</td>
</tr>
<tr>
<td>mobile radios</td>
<td>acat ctrl</td>
<td>unit cpu</td>
</tr>
<tr>
<td>communication equipment</td>
<td>nasic soxa</td>
<td>computer digital</td>
</tr>
<tr>
<td>television communication</td>
<td>pop extension</td>
<td>fiber optic</td>
</tr>
</tbody>
</table>
**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**

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>exercise option</td>
<td>support services</td>
<td>support services</td>
</tr>
<tr>
<td>elevator maintenance</td>
<td>acs i</td>
<td>incremental funding</td>
</tr>
<tr>
<td>overhead doors</td>
<td>incremental funding</td>
<td>air force</td>
</tr>
<tr>
<td>incremental funding</td>
<td>exercise option</td>
<td>acs afii</td>
</tr>
<tr>
<td>non fd</td>
<td>sustainment support</td>
<td>acs aficm</td>
</tr>
<tr>
<td>preventive maintenance</td>
<td>information technology</td>
<td>management system</td>
</tr>
<tr>
<td>remedial maintenance</td>
<td>engineering services</td>
<td>information technology</td>
</tr>
<tr>
<td>support services</td>
<td>engineering support</td>
<td>exercise option</td>
</tr>
<tr>
<td>de obligation</td>
<td>software engineering</td>
<td>maintenance sustainment</td>
</tr>
<tr>
<td>option period</td>
<td>management system</td>
<td>spectrum management</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic 4</th>
<th>Topic 5</th>
<th>Topic 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>base telecommunications</td>
<td>switching system</td>
<td>system installation</td>
</tr>
<tr>
<td>telecommunications system</td>
<td>control switching</td>
<td>emergency system</td>
</tr>
<tr>
<td>support services</td>
<td>maintenance service</td>
<td>basic center</td>
</tr>
<tr>
<td>monthly maintenance</td>
<td>repair task</td>
<td>center operations</td>
</tr>
<tr>
<td>base telecommunications</td>
<td>imr maintenance</td>
<td>frequency global</td>
</tr>
<tr>
<td>incremental funding</td>
<td>land mobile</td>
<td>global communications</td>
</tr>
<tr>
<td>base telephone</td>
<td>mobile radio</td>
<td>antenna install</td>
</tr>
<tr>
<td>system services</td>
<td>hardware maintenance</td>
<td>communications systems</td>
</tr>
<tr>
<td>business sb</td>
<td>system contract</td>
<td>systems antenna</td>
</tr>
<tr>
<td>sb companion</td>
<td>annual hardware</td>
<td>cam a li</td>
</tr>
</tbody>
</table>

**Topic 7**

| Fiber optic |
| Network centric |
| Centric solutions |
| Support services |
| Optic cable |
| Ffp loc |
| Requisite engineering |
| Focs 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**

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>adp software</td>
<td>adp software</td>
<td>adp software</td>
</tr>
<tr>
<td>acat asrf</td>
<td>license renewal</td>
<td>software maintenance</td>
</tr>
<tr>
<td>software maintenance</td>
<td>software maintenance</td>
<td>software license</td>
</tr>
<tr>
<td>premium sw</td>
<td>agreement exercise</td>
<td>seat license</td>
</tr>
<tr>
<td>xenapp xendesktop</td>
<td>graphics pads</td>
<td>annual ccc</td>
</tr>
<tr>
<td>edo card</td>
<td>commons month</td>
<td>mapping toolbox</td>
</tr>
<tr>
<td>interface infrastructure</td>
<td>gfp clause</td>
<td>cie dren</td>
</tr>
<tr>
<td>server option</td>
<td>core system</td>
<td>tree software</td>
</tr>
<tr>
<td>tell wsn</td>
<td>wills software</td>
<td>database query</td>
</tr>
<tr>
<td>afame bias</td>
<td>emulex fiber</td>
<td>asf rymb</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic 4</th>
<th>Topic 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>acat asrf</td>
<td>software maintenance</td>
</tr>
<tr>
<td>software maintenance</td>
<td>annual software</td>
</tr>
<tr>
<td>software license</td>
<td>maintenance service</td>
</tr>
<tr>
<td>contract task</td>
<td>telecom annual</td>
</tr>
<tr>
<td>operational picture</td>
<td>service plans</td>
</tr>
<tr>
<td>update ombudsman</td>
<td>software support</td>
</tr>
<tr>
<td>taskings tss</td>
<td>maintenance renewal</td>
</tr>
<tr>
<td>network li</td>
<td>software license</td>
</tr>
<tr>
<td>vrg licensing</td>
<td>air force</td>
</tr>
<tr>
<td>trusted licenses</td>
<td>annual maintenance</td>
</tr>
</tbody>
</table>
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**

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>telecom telecommunications</td>
<td>telecom telecommunications</td>
<td>internet service</td>
</tr>
<tr>
<td>cable tv</td>
<td>land mobile</td>
<td>commercial internet</td>
</tr>
<tr>
<td>land mobile</td>
<td>maintenance services</td>
<td>internet services</td>
</tr>
<tr>
<td>correct line</td>
<td>cable distribution</td>
<td>cable service</td>
</tr>
<tr>
<td>television services</td>
<td>sw cable</td>
<td>speed internet</td>
</tr>
<tr>
<td>system maintenance</td>
<td>cable outlets</td>
<td>dialup internet</td>
</tr>
<tr>
<td>modification changing</td>
<td>price adjustment</td>
<td>exercise option</td>
</tr>
<tr>
<td>changing unit</td>
<td>tv requirement</td>
<td>optical internet</td>
</tr>
<tr>
<td>excess funding</td>
<td>usaf hq</td>
<td>broadband internet</td>
</tr>
<tr>
<td>communication telephone</td>
<td>government's obligation</td>
<td>cable television</td>
</tr>
</tbody>
</table>

**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

<table>
<thead>
<tr>
<th>IT Security</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
</tr>
<tr>
<td>Equipment</td>
</tr>
<tr>
<td>Secure Measures</td>
</tr>
<tr>
<td>Surveillance</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IT Consulting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
</tr>
<tr>
<td>Sustainment</td>
</tr>
<tr>
<td>Development</td>
</tr>
<tr>
<td>Training &amp; Support Services</td>
</tr>
<tr>
<td>Mission Support</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IT Hardware</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
</tr>
<tr>
<td>System</td>
</tr>
<tr>
<td>Infrastructure</td>
</tr>
<tr>
<td>Computer</td>
</tr>
<tr>
<td>Asset Management</td>
</tr>
<tr>
<td>Data Asset</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IT Offsourcing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
</tr>
<tr>
<td>Support</td>
</tr>
<tr>
<td>Contracts</td>
</tr>
<tr>
<td>Physical Plant Maintenance</td>
</tr>
<tr>
<td>Maintenance</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IT Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
</tr>
<tr>
<td>Software</td>
</tr>
<tr>
<td>Baseline Systems Configuration</td>
</tr>
<tr>
<td>Support</td>
</tr>
<tr>
<td>Infrastructure</td>
</tr>
<tr>
<td>Telecommunications</td>
</tr>
<tr>
<td>Topic 1</td>
</tr>
<tr>
<td>Connectivity</td>
</tr>
<tr>
<td>Television</td>
</tr>
<tr>
<td>Telecom Services</td>
</tr>
<tr>
<td>Internet Management</td>
</tr>
<tr>
<td>Global Communications</td>
</tr>
</tbody>
</table>
Appendix D. Quad Chart

**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 AFCEA using the Federal Procurement Data System – Next Generation (FPDSNG). The file contained 107,389 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.


### Title and Subtitle

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

### Authors

Holliger, Theodore S., Master Sergeant, USAF

### Funding Information

- **Contract Number**: 5a.
- **Grant Number**: 5b.
- **Program Element Number**: 5c.
- **Project Number**: 5d.
- **Task Number**: 5e.
- **Work Unit Number**: 5f.

### Performing Organization

**Name**: Air Force Institute of Technology
**Address**: Graduate School of Engineering and Management (AFIT/EN), 2950 Hobson Way, Building 640, WPAFB OH 45433-8865

### Sponsor/monitoring agency

**Name**: Air Force Installation Contracting Agency
**Address**: Roger H. Westermeyer, Director of Enterprise Sourcing Support, 1940 Allbrook Drive, Bldg. 1, Wright-Patterson AFB, OH 45433
darin.ashley@us.af.mil

### Distribution/Availability Statement

Approved for public release; distribution unlimited.

### Abstract

The United States Air Force can dramatically reduce resource consumption through strategic sourcing initiatives that leverage sensibly-bounded 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.

### Subject Terms

- Strategic Sourcing
- Category Management
- Content Analysis
- Latent Semantic Analysis
- Topic Model