Text Data Analytics: A Methodological Review and Demonstration

Beth-Anne Schuelke-Leech,[a] Betsy Barry,[b] and Clayton Darwin [b]

[a]John Glenn College of Public Affairs, Ohio State University
[b]BDataSmart, Athens, Georgia
[b]Broad River Data, Athens, Georgia

Abstract

The massive amount of information being produced and released by governments, regulatory agencies and other public entities afford policy and public administration scholars opportunities to investigate a range of new questions and relationships. However, these opportunities are accompanied by the challenges of learning and applying new methodologies of data collection and analysis. This paper looks at how Big Data analytics can be applied to public administration and policy research through an interdisciplinary approach. The paper begins with a discussion of the overview of text data analytics and current applications to public policy, public administration, and political science research. The paper then demonstrates how text data analytics using corpus linguistics and computational linguistics can be applied to policy research using an example of discussions surrounding sustainable transportation in the United States Congress using a 5.5 billion word U.S. legislative corpus spanning the years 1981 to 2012.

Keywords

Text Data Analytics; Big Data; Research Methods; Sustainable Transportation

Introduction

With the rise of electronically produced and stored information, there has been a substantial and sustained increase in the quantity of publicly available information in the past twenty years. Much of this information is available for perusal and investigation and can provide significant insight into the roles and actions of elected officials, public administrators, and regulators. However, this information is underutilized, and insights remain unrealized, without reliable methodologies for collecting, organizing and analyzing these vast stores of data.

Traditionally, there have been three dominant methodologies for social science research: statistical regression (econometrics); observations, surveys, and interviews; and experiments (see for example, Weathington, Cunningham et al., 2010; Cresswell, 2014). These methodologies have recently been joined by Big Data Analytics. While Big Data Analytics would seem to be just an extended form of statistics, it is actually more of a combination of modeling, evaluation, quantitative, and qualitative techniques on datasets so sufficiently large that is impossible to manage them with simple data management tools, such as excel. There have been many such datasets in science, engineering, medical, and mathematics fields; however, dealing with the sheer scale and scope of Big Data has been far less common in the social sciences.

It is only in the past several years that Big Data has become part of the mainstream conversations of research methodologies in social science. In contrast, managing, storing and mining Big Data has been an industry staple since before the nomenclature pervaded the lexicon. Private corporations like Google, Amazon, Netflix and Microsoft have been processing and mining Big Data for sales and marketing, as well as research and development-related enterprises. Government intelligence agencies like the National Security Administration (NSA) have also become experts at gathering, managing and analyzing vast amounts of data.

Generally speaking, data can be classified into three types: numeric, audio-visual, and text. Numeric data is comprised of numbers, quantities and binary code. Audio-visual is made up of images, recordings and videos. Text data is the graphical representation of language (Barry, 2008). Another useful distinction between data types is structured versus unstructured data. Structured data resides in fixed fields (columns and rows) in a file or a record. Numeric data is considered structured data. It has the advantage of being easily entered, organized, queried and analyzed. Conversely, unstructured data is not easily organized neatly into pre-determined fields or data models. Text and multimedia content is unstructured data. It is estimated that unstructured data, the bulk of which is text-based data, make up 80 to 90% of all of the data produced by all organizations. (Holzinger, Stocker et al., 2013). Despite the fact that unstructured data, specifically, unstructured text-based data, constitutes a large percentage of what we refer to as Big Data, much of the focus of data analytics has been on structured, numeric data (Chen, Chiang et al., 2012). This is in part due to the nature of the data itself: Relatively speaking, mining and exploiting structured numeric data is a more straightforward task. On the other hand, unstructured, text data is considered more complex and therefore more difficult to manage, process and analyze.

The focus of this paper is unstructured text data. Before this is can be discussed in detail, it is necessary to address the idea of "complexity" with respect to the nature of the data itself. The complexity of language is discussed in detail in Schuelke-Leech and Barry (forthcoming). First, unstructured text data originates from a wide variety of sources, such as email, reports, press releases, social media, newspapers, essays, books, web pages, or any place where written language is used to express ideas or communicate information. Also, unstructured text may exist in a variety of file structures (pdf, txt, html, doc, rtf, etc). These disparities mean that the act of transforming unstructured text into an analyzable dataset requires a range of technical expertise.[1] Simply put, it is not as easy as importing numbers into a database and then querying the database. The technical "complexity" of unstructured text data is significant and poses a challenge for text analysis, especially in the era of Big Data where datasets originate from disparate sources, across a variety of technology-mediated environments.

The second aspect to the complexity of unstructured text is due to the fact that text-based data is natural language data. As text is the graphical representation of language (Barry, 2008), it is also beholden to all of the linguistic principles that govern language and language use. Language is complex in all of its representations, both spoken and written. The linguistic complexity of unstructured text is not to be confused with the technical complexity. Technical complexities are much easier to address and overcome with the right software or tool. Linguistic complexity is much more difficult to accommodate for non-linguist professionals whose object of study is unstructured text-based natural language.

Linguistic complexity is due to the fact that Language is innovative, infinitely varied, and changes over time (see Siemund, 2011). Language varies in form and function, depending on many factors. Every linguistic style has particular linguistic characteristics and specialized lexicons, as does every genre. For example, an informal, personal Instant Message (IM) conveys information very differently than a formal business memorandum. Likewise, every industry has specific linguistic characteristics and specialized lexicons that form the linguistic habits of the industry, habits extant in their business communications and documentation (Stubbs, 1996). For example, if you are investigating a collection of unstructured text data from the automotive industry and you are interested in the concept of collisions, you will notice a range of linguistic forms that impart similar

[1] *Technical expertise* refers to the range of natural language processing and computational linguistic techniques used in data preparation methods, as well as general computational requirements for processing and storage.

lexical-semantic function in the discourse, forms such as accidents, crashes, wrecks, etc. Some forms will occur more than others, but to be sure there is rich linguistic variation used to concretely express the concept. This is an empirical fact about Language (ibid).

Linguistic complexity is the outcome of linguistic form and function that are dependent on context (Stubbs, 1996). That is, meaning is carried via the linguistic context. For instance, if you have a collection of text documents from the automotive industry, the presence of the form “crash” is not likely to refer to the stock market crash. Likewise, if you have a dynamic, varied collection of unstructured text from disparate data sources, conveying varied content covering a spectrum of different themes, styles and genres, the form “crash” may take on a range of different meanings, depending on the linguistic context in which it is used. Crash may mean a physical collision, a metaphorical plummet, a loud noise, or to enter a gathering without an invitation. Crash in the presence of “vehicles” will constrain the range of possible interpretations, just as crash in the presence “stock market” or “computer.” Thus, it is the relationship between linguistic forms and concepts that inform interpretation and meaning. When dealing with natural language text, one cannot assume a one-to-one correspondence between form and function without understanding and accommodating linguistic context (Stubbs, 2001b). Unlike numbers in which the symbol “1” can be taken always to represent the quantity of “1”, words can represent multiple things depending on the context.

The technical and linguistic complexity of unstructured text is magnified exponentially when dealing with the sheer volume or quantity in today's era of Big Data. The quality and the quantity of data make it extremely difficult for a person (or even a team of people) to collect, process and analyze them effectively. Large collections of data are only useful if there is some ability to extract useful information, discover interesting trends, patterns and correlations that can inform the decision-making process. ***Text Data Analytics*** is the overarching term for analysis of large datasets of unstructured text and is used to generally describe processing and analyzing text-based natural language data.

Text Data Analytics is really a collection of various techniques, tools and methodologies that have been developed in different fields with the aim of analyzing text for some specific purpose. Some of these tools and techniques developed from disciplines that focus on the rich content of text, with text being anything from a small passage of a speech, to a collection of speeches spanning decades. The text itself is the object of study. Researchers in these fields evaluate meaning and context in great detail, often manually with little or no technological intervention. Other researchers in different fields have developed computer technologies and algorithms that allow for the processing and analysis of large, disparate language corpora. The spectrum of Text Data Analytics is considerable.

Different tools and techniques have different intellectual origins and applications. Figure 1 presents the different areas of text data analytics, divided according to whether the methodology requires computer-assistance for analysis, and depending on whether the methodology considers the smallest unit of analysis (linguistic forms or "words") as discrete entities regardless of linguistic context, or whether content and context (linguistic form and function) are studied *in toto*, considering the relationship a central aspect in conveying and interpreting meaning.

Figure 1: Disciplinary Foundations for Text Data Analytics

The upper left-hand quadrant represents the space in which words are considered discrete datum, but the analysis is done manually, without computer assistance. There are no commonly used techniques in this space. The upper right-hand quadrant is for the techniques in which computer assistance is used to analyze the text using techniques that consider words as data. The majority of techniques in this quadrant come from computer science and mathematics, such as natural language processing and data mining. The lower left-hand quadrant contains more traditional methods of manual analysis of texts, such as literary analysis and discourse analysis. The lower right-hand quadrant also assumes the coupling of content and context, but it uses computer-assistance for analysis.

Not all of the methodologies in the typology are used in policy, public administration, or political science research. We will now discuss the more commonly used ones, including looking at their strengths and weaknesses.

Text Data Mining and Words as Data

Text is the graphic representation of language, just as numerals and symbols represent quantities and formulae in mathematics (Barry, 2008). Once we understand the patterns of the representations, we understand that each components within the representation has meaning. One of the challenges in analyzing and evaluating text is that everyone has an intuitive sense of language, which often means that researchers will impose meaning and value unconsciously on familiar words and grammatical structures. Domain experts run the risk of taking for-granted that words have a consistent meaning throughout text.

Data mining techniques developed with the intention of classifying, clustering, and analyzing large repositories of information and data in order to find patterns and trends that are not necessarily obvious with a manual analysis (Fayyad, 1997; Fayyad and Uthurusamy, 1999). Text data mining allows for a much faster organization and analysis of text data than could be done manually since it uses semi-automated data mining techniques. Information and data retrieval is heavily dependent on similarity of content, identified by computer algorithms that match content to specified search terms (Hand, Mannila et al., 2001) or linguistic forms. Typically, the text is tokenized[2] and reduced to an algebraic vector with a magnitude and direction, essentially imposing a mathematical structure on unstructured data. Similarity is based on the relative proximity of the data to a reference one (Hand, et al., 2001). This conversion of text into numeric data is common. Typically electronic "documents"[3] are reduced[4] and converted into a binary format that a computer can then decipher and transmit.

Text data mining originated primarily in the fields of library and computer science (Hand, et al., 2001). These fields were generally interested in data management and the classification, retrieval, and summarization of information (Feldman and Sanger, 2007). One of the primarily goals of text data mining is the discovery of patterns, especially the distribution, proportion, and frequency of words, as well as the associations and co-locations of words (ibid). Text Data Mining uses tools from computer science, information systems, and mathematics, particularly machine learning, data mining, information retrieval, natural language processing, knowledge management (ibid), as well as case-based reasoning and statistics (Spinakis and Peristera, 2004). These tools provide a means for developing computer algorithms to divide the text into meaningful components, tag the parts of speech, and syntactically parse the words, sentences, and phrases (Feldman and Sanger, 2007). This allows for categorizing text and extracting information automatically once the computer has been "trained."

Natural Language Processing (NLP) is a foundational field of text data mining. NLP processes texts in order to understand human language usage, and in turn, program computers to use language, to impart and interpret meaning, in the same way people impart and interpret meaning.

[2] Tokenizing a document means that the text is broken down into individual tokens. This typically means using the punctuation and white spaces in a text to delineate the tokens (Miner, Elder IV et al., 2012). In linguistic terms, a token is approximately the length of a word. The average length of a book is 64,000 words (Baeza-Yates and Riberio-Neto, 1999)

[33] ***Document*** is a term that refers to any individual file that contains unstructured text, not simply a word processing document.

[4] Reduced because the entire document is truncated to the first 100 – 500 tokens

One goal of NLP is to create "domain-independent linguistic features" (Feldman and Sanger, 2007, p. 58) to develop software that can analyze unstructured text-based natural language regardless of context, style or genre. In order to do this, NLP techniques analyze content and context independent of one another, as well as the interdependency of content and context. NLP is a subset of the larger field of artificial intelligence and machine learning. The goals of these fields are broader and investigate ways computers and machines can simulate and demonstrate intelligence (Grishman, 2010), a significant aspect of which is natural language usage.

Grimmer and Stewart (2013) present an excellent overview of the application of text data analytics to political science in their paper "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." Grimmer and Stewart place themselves into the realm of text as data within the content analysis sphere. They focus on the automation of text analysis at the document level employing a "bag-of-words" technique, in which the root word is used in the analysis (called stemming). The noise or stop words (e.g., the, to, from, and, for, etc.)[5] and punctuation are discarded, as are very uncommon words (i.e., those that occur at a raw frequency of one in each documents). A small set of documents is hand-coded or a coding dictionary created and then this is used to train a computer to classify the rest of the documents.

Content analysis is one of the few areas that crosses between manual and computer-assisted analysis and between the coupling and decoupling of content and context. The borders of content analysis are somewhat porous. Originally, content analysis focused on a quantitative analysis of the text, such as the frequency of words, but it quickly expanded to include syntactic, semantic, and

[5] Stop words are grammatical words that occur at high frequencies and make up large percentages of the total number of words in text. They are called "function" words in linguistics as they provide grammatical function, as opposed to content words, which impart ideas or actions, etc.

pragmatic functions.[6] Thus, content analysis incorporated statistical analysis relatively early. Content analysis now includes automated content analysis.

This technique has been the most commonly used in policy and political science research. It has been employed in analyzing political agendas (Laver, Benoit et al., 2003), legislative agendas (Grimmer, 2010), committee hearings (Jones, Wilkerson et al., 2009), treaties (Spirling and McLean, 2007), political manifestos (Gabel and Huber, 2000; Hansen, 2008; Volkens, Bara et al., 2009), comparative politics (Lucas, Nielsen et al., 2015), political speeches (Laver and Benoit, 2002) and other discernible genres within the policy realm. As computer programs and information technologies make it increasingly easier to gather text, the applications will grow.

Language Context and Content Coupled

The "words as data" techniques are only used with computer-assisted analysis. In their paper looking at the analysis of political agendas, Laver, Benoit, and Garry (Laver, et al., 2003) contrast this method of words as data with some of the other types of text analysis. They note that their technique deviates from the manual text analysis methods where texts are carefully read for meaning and the content and context are coupled. Similarly, Grimmer and Stewart (2013) acknowledge that the automated analysis of texts cannot completely replace the manual analysis of texts.

In contrast, much of the work in Critical Text Analysis requires a careful, close reading of texts. Thus, many techniques remain in the realm of manual analysis and have long-standing scholarly tradition. Texts have been analyzed to understand the author's intent and message (e.g., literary, literature, and drama analysis, rhetorical analysis), how the writer creates meaning (e.g.,

[6] Content analysis originated in the field psychology [referred to as statement analysis as well] to study individuals through their language usage, not necessarily to generally study language usage in and of itself, and extract meaning from language decoupled from what it tells us about an individual. Its applications include psychiatry, psychology, history, anthropology, education, philosophy, literary analysis, political science, and linguistics (Krippendorff, 1980)

narrative analysis) (see Riessman, 1993), the cultural and social context in which the text was written (e.g., content and discourse analysis) (see Neuendorf, 2002), and the structure of language (e.g., linguistics) (see Coulthard, 1985; Biber, Conrad et al., 1998; Hunston, 2002; McEnery and Hardie, 2012). Texts have also been analyzed in light of different critical frameworks, such as feminist critique, gender identity, or racial identity.

These different techniques come from different academic fields. Discourse Analysis and Critical Discourse Analysis, for instance, originated in sociology, psychology, philosophy and linguistics. Discourse analysis views language as a form of social interaction. It explicitly links language and social structure (Fairclough, 1989; Wodak, 1989; van Dijk, 1992; Schiffrin, 1994; Stubbs, 2001a). Discourse analysis is concerned with texts as objects of analysis in the context of culture, social structures, and historic dialogues. Instead of looking at the linguistic structure of language, discourse analysis is interested in the texts' cultural meaning and the power relationships that exist there, in line with the work of Foucault and Bourdieu (Jorgensen and Phillips, 2002). For discourse analysts, language and society are intertwined. Language shapes society and culture, which in turn shape the way that language is used. Analyzing texts provides some insight into society, power, relationships, ideology, and psychology. For researchers in this area, individual texts are always interposed within an ongoing historical dialogue and discourse.

The close reading and analysis of texts is necessarily a human undertaking and computer-assistance has generally been insignificant. However, computer assistance does not necessarily require a decoupling of content and context, the way that 'text as data' or 'bag-of-words' does. For instance, in the "bag-of-words" technique, the punctuation, "noise" words, and "infrequently" used words are removed. Thus, the words are removed from their original context. This simplification of words is helpful when the goal of the research is document classification, frequency counts, or the clustering of words as it reduces processing time and complexity. However, if the context in which the words are used is an important component of the text, removing words from this context eliminates important components of the data. For example, removing the punctuation from "1.2345" and "12,345," creates the expression "12345" but the original meaning of the numbers has been removed.

There are some situations, however, in which methodologies that do not simplify the data allow for a much more nuanced and complete understanding of the data. Linguistic methods, for instance, are developed with an understanding of the importance of the relationship between content and context, form and function, within the structure of language (Biber, et al., 1998). Linguistic methodologies are the foundation of many of the analytical tools used in text data mining (Kuechler, 2007). These methodologies look for patterns and trends in the text corpus.[7] Again, they rely on the context, as well as the content, in order to understand meaning. Instead of limiting the analysis to classifications, clusters, and frequency, corpus-based and computational linguistics allow for both a quantitatively rigorous analysis and a principled, qualitative investigation of the context and patterns in language (see Biber, et al., 1998; Hand, et al., 2001; Feldman and Sanger, 2007; Chattamvelli, 2009; McEnery and Hardie, 2012). In essence, qualitative, data-driven linguistic methodologies can validate quantitative methodologies, where qualitatively investigative frameworks inform what forms or patterns will be measured, or quantified, in analyzing text-based natural language data. When dealing with large quantities of data, the analytical techniques and algorithms used must scale to incorporate computer-assisted processing and management (Barry, 2008; Darwin, 2008; Barry, Smith et al., forthcoming).

Corpus linguistics is the computer-mediated study of language structure and use text-based corpora representing natural language in “real world” contexts. Computational linguistics explicitly integrate computer systems for the purpose of understanding the nature of language as a

[7] Corpus refers to a collection of text-based natural language data.

phenomenon (Grishman, 2010). Both fields incorporate technology to extract and understand meaning and patterns of linguistic behavior in a large text datasets. Using linguistics, it is possible to investigate the systematic associations of words, historical trends, meanings of words, and sentiment analysis, in addition to the frequency, clusters, and classifications that are used in the "words as data" analysis (Biber, et al., 1998). Furthermore, consulting linguistics makes it possible to address linguistic complexity in valid and reliable ways, rather than relying on "reducing" or "simplifying" natural language data to facilitate analytical endeavors. Rather than simply being a frequency count of the words, both corpus-based and computational linguistic analysis facilitate an investigation of the context and content of the data.

In all research, the methodology employed is determined by the research question of interest and the data. This overview demonstrates that there are many tools and techniques, not simply one methodology, and that the researcher must determine the methodology, based on the data and specific research question. At the same time, the tools and techniques available will determine what research questions can reasonably be asked.

Example of Text Data Analytics using Linguistics

A rudimentary demonstration of text data analytics using corpus and computational linguistics is presented here. The data comes from the official record of the U.S. Congress from 1981 through 2014. The Congressional corpus is comprised of speech, debate, and Hearing transcripts, as well as Congressional reports and documents. The transcripts are a particularly robust source of data since they provide a record of the give-and-take conversations in both chambers, as well as in the committee hearings and meetings. Though the transcripts are edited for grammar, spelling, and clarity, they are not edited for political correctness or ideology. They represent the ongoing context and quality of the conversation between various stakeholders in the policy process.

Every topic that has been taken up by Congress over the last three and a half decades is extant in this corpus. Thus, the content is highly variable.

As discussed above, the methodology employed is determined by the research question of interest and the data. For example, when dealing with technically complex, variable dataset of unstructured text like the Congressional data, one that is variable in file type, file size and file quality, it requires an additional methodological layer in transforming the original data into a dynamic, analyzable corpus.[8] Additionally, a linguistically complex data set like this that is comprised of a wide variety of content, styles and genres, requires comprehensive qualitative investigation and assessment as the foundation of any quantitative endeavors. The reason is due to the linguistic complexity characteristic of large, variable collections of text. It is not enough to rely on one's own intuition about how different concepts are concretely expressed in the face of such linguistic variability. It is necessary to investigate and validate the language used to express concepts and themes under investigation, *before* one can measure these concepts and themes. Furthermore, it is not enough to simply verify the existence of concepts qua natural language usage, but it is also important to validate the extent of variation that comprise the concepts or themes of interest. To put it succinctly, you cannot measure what you do not know. Thus, it is both the technical and linguistic complexity of large, varied corpora that informs not only what kinds of research questions you can ask of your data, but also what sorts of processing and analytical methodologies are appropriate in order to work with, rather than work around, these complexities in order to address your research questions in the most valid and reliable way possible.

[8] The corpus for the 97th through the 113th Congress consists of 89,528 files with a total of 5.516 billion tokens. The size of the corpus makes computer assistance in the analysis imperative. Most of the files were originally pdfs, though there were also html and text files. The files were first converted to text files, then converted to utf-8. They were then tokenized and organized according to Congressional terms to preserve the original organization of the archive from which the data was collected.

Sustainable Transportation

The specific example presented looks at how the concept of sustainable transportation is discussed in the United States Congress. Since the purpose of this example is to demonstrate the methodology, the literature, theory, and implications of the results are necessarily brief. Instead, the focus is on how the methodology can be used to investigate a topic of interest.

Sustainability has become an important topic of conversation in many areas of research and society. It is now coupled with our food system, transportation, natural resource development, energy, economic development, and urban planning, (etc.). Creating a sustainable and resilient system requires addressing some of the unsustainable aspects of that system. In the United States, there are 253.6 million registered vehicles for highway usage (U.S. Department of Transportation, 2012). Transportation-related uses account for 28 percent of all energy consumed in the United States (approximately 27 Quadrillion BTUs of energy) (U.S. Energy Information Administration, 2012b) and 71 percent of all petroleum used in the country (U.S. Energy Information Administration, 2012a).

Much of the sustainability debate started in the 1970s and 1980s with scholars and activists grappling with the constraints and limitations of the natural environment. In 1987, the United Nations issued their report on the Environment and Development, more commonly called the Brundtland Report (United Nations, 1987). The Brundtland report defined sustainable development broadly as "…development that meets the needs of the present without compromising the ability of future generations…" Since that report was issued, scholars and practitioners alike have struggled with what the really means and how to redesign and change existing systems to achieve the goal of sustainability.

For transportation, sustainability is tied to addressing three problems that make the current system unsustainable: the use of fossil fuels as the primary fuel source; traffic congestion; and safety.

Thus, a sustainable transportation system requires developing and adopting new technologies, as well as changing human behavior, public policies, and economic factors (Richardson, 1999).

Public policies, regulations, and government directives are important components of greater sustainability. However, it is not always clear whether federal policymakers are committed to sustainability, or even believe that sustainability is a significant issue to Congress. The United States Congress fulfills important functions in determining the priorities and allocation in the budget of the United States government. Thus, the conversation in the United States Congress about sustainable transportation is an important indicator of whether federal policies will support the goal of a transition to a more sustainable transportation system.

Methodology

Text data mining and analysis are necessarily dependent on the decisions made during the research process in support of specific research goals and questions. There is no standard methodology, process, or program for analysis. Instead, the tools and techniques used in text analysis are dependent on the research paradigm, ideologies, and data (Schuelke-Leech and Barry, forthcoming).

To begin a linguistic analysis of text, a report of the statistically salient language was generated for the U.S. Congressional session. A research study begins with defined a linguistic marker set (A fuller description of the methodology and the development of marker sets is outlined in Schuelke-Leech and Barry (forthcoming)). Once a marker set is defined, queries can be conducted either on an individual marker set, or in combination to look at the overlap. This overlap looks at the proximity of the words, or the linguistic context, and allows for the investigation of the relationship between the markers as it impacts both form and meaning. Thus, it is possible to investigate how tightly coupled, or associated, concept such as sustainability is with transportation.

To assess the overlap of the conversations, the marker sets are layered by specifying the proximity (i.e., constraining the linguistic context) of the results from the one marker set relative to the other. The closer the proximity between marker sets, the closer the lexical-semantic relationship. For example, specifying proximity of 5 tokens between marker sets means that you are looking for a very tight coupling of categories, or a close lexical-semantic relationship, since the words of each respective set are often collocated, or in the same phrase or sentence, modifying one another. Proximity of 15 or 25 tokens obviously means that there is less direct coupling of the categories, and a looser lexical-semantic relationship, even though the linguistic markers are still relatively close (often within the same paragraph or excerpt of discourse). Proximity of 50 or 100 tokens is wider, and although the categories are on the same document page or contained within the same file, they may actually be part of completely separate and distinct discussions, where no lexical-semantic relationship exists between them.

Table 1 presents results of searches with marker sets where transportation is narrowly defined and where it is more broadly defined. The narrowly defined marker set is: automo*; car; cars; motor coach; motorcycle; motor vehicle; motor vehicles; passenger vehicle; passenger vehicles; sport utility; suv; suvs; tractor trailer; truck; trucks. For the more broadly defined marker set, the following words were added: bus; buses; transport; transportation; vehicle; vehicles.

Insert Table 1: Transportation Returns

The results for Transportation-related returns are presented in Table 1. The average number of files for each session is 5,333 files with 334.5 million tokens. The broader definition of transportation returns approximately twice as many results as the narrower definition. Transportation returns using a broader definition is an average of 0.4 percent of the total tokens, while a narrower definition returns about 0.2 percent of the total tokens. An important measure of the intensity of discourse is the hits per million tokens. This measure standardizes the returns. The results of returns per million tokens do show that the level of transportation-related discussions have remained fairly consistent in the past three decades. These returns are essentially frequency counts. Using the absolute returns numbers, it is difficult to know whether transportation is an important topic to Congress since there are so many discussions and topics. The relative discussions are often more revealing. This requires investigating the specific topics of transportation more carefully using associations and proximity of concepts. In this case, we will use sustainable transportation as the concept of interest.[9]

Once a marker set is defined, queries can be conducted either on an individual marker set, or in combination to look at the overlap. This overlap looks at the proximity of the words, or the *linguistic context*, and allows for the investigation of the relationship between the markers as it impacts both form and meaning. Thus, it is possible to investigate how tightly coupled, or associated, concept such as sustainability is with transportation. To assess the overlap of the conversations, the marker sets are layered by specifying the proximity (i.e., constraining the linguistic context) of the results from the one marker set relative to the other. The closer the proximity between marker sets, the closer the lexcial-semantic relationship. For example, specifying proximity of 5 tokens between marker sets means that you are looking for a very tight coupling of categories, or a close lexical-semantic relationship, since the words of each respective set are often collocated, or in the same phrase or sentence, modifying one another. Proximity of 15 or 25 tokens obviously means that there is less direct coupling of the categories, and a looser lexical-semantic relationship, even though the linguistic markers are still relatively close (often within the same paragraph or excerpt of

[9] The sustainability marker set includes: conservation; eco-friendly; environmental; non-renewable; renewable; sustainable; and sustainability. When the associations are investigated, additional sustainability terms that are specifically related to transportation were included: alternative fuel; green; hybrid.

discourse). Proximity of 50 or 100 tokens is wider, and although the categories are on the same document page or contained within the same file, they may actually be part of completely separate and distinct discussions, where no lexical-semantic relationship exists between them.

Thus, the proximity range has a substantial effect on the returns, as demonstrated in Tables 2 and 3.

Insert Table 2: Sustainable Transportation Returns with a Narrower Definition of Transportation

Insert Table 3: Sustainable Transportation Returns with a Broader Definition of Transportation

One of the important aspects of investigating is to validate the returns to see whether they are truly indicative of the desired research topic. When the proximity is small and the association is close, then a broader definition can be used since the returns have few false positives. That is, the vast majority of the returns are directly related to the concept of sustainable transportation. On the other hand, once the proximity is specified to be large (starting at 25 tokens), many more false positives are included in the returns. So, returns like the following are included, which are not part of the target discussion:

- …cars without seatbelts or airbags; Or maybe we recall times when we travel throughout our community and we notice not only a heavy fog but polluted (113th Congressional Session)
- …natural gas (LNG) bulk tank cars, LNG locomotive tenders, and technologies suitable to retrofit tank cars (discussing railroad cars, which was not the target conversation). (113th Congressional Session)
- …Environmental Response, Compensation, and Liability Act of 1980 included lead oxide in the list of chemicals subject to the tax. The typical automobile. (98th Congressional Session)
- …Clean Air Act. As citizens we are concerned that clean air be maintained; as players in the automotive aftermarket industry we are concerned that government rules not arbitrarily upset the competitive structure in the marketplace. (98th Congressional Session)
- Our legal system thus says: -- Thou shalt not paint thy truck green while, at the same time, permitting trucks painted with an equal mixture of blue and yellow paints. Unfortunately, liability today hinges in no small measure on whether one applies a "green" (98th Congressional Session)

The other challenge in investigating language is that the meaning of words and phrases changes over time. So, for instance, in the early 1980s, discussions of “green” coupled with “trucks” are about military vehicles, not sustainability:

- military-green pickup trucks of a new type. They bear official Air Force markings, hut they are Volkswagens (97th Congressional Session)
- A row of green-and-black camoflauged trucks and jeeps are lined up outside the training center (97th Congressional Session)

Thus, a qualitative assessment of the associations and context must be part of the empirical research, as well as a quantitative analysis of the relative frequency, norms, and[10] subcorpus statistics. The proximity and layering are not clustering or associating document topics. It is investigating the actual usage of the language and concepts of interest. It is possible through this method of investigation to identify topics associated with the main research question and to look at the saliency and sentiment of discussions.

Fewer results are returned with a smaller proximity than a larger one, just as fewer results are returned when relying on narrowly defined marker sets. The more tightly constrained the linguistic context, and the more narrowly defined categories, the smaller the query returns. This means that the research protocol must consider how closely associated the categories should be and why. Again, it is not simply enough to establish the existence of co-occurrence or overlapping categories according to some randomly assigned proximity. It is important to investigate the relationship of the co-occurrence of the overlapping categories to make sure that the language content and context contribute to the overall research objectives.

The results show that there is a consistent increase in the conversations about sustainability in automotive transportation from the early 1980s to the 2010s, though there is quite a bit of volatility. The interpretation of the connections of the conversations of sustainability with

[10] Demonstrating these techniques would require more space than is available here and, thus, this will be left for other papers.

transportation in the U.S. Congress is heavily dependent on the marker sets and the specified proximity. When a narrow definition of transportation is used and the proximity is small (i.e., when the concepts are very tightly couple) at 5 tokens, there appears to be a very small overlap of between sustainability and transportation with an average of 2.8 returns per million tokens (compared with 190.5 for transportation) or approximately 1.6 percent of the transportation returns. On the other hand, when transportation is broadly defined and the proximity is specified to be much larger (within 50 tokens), the concepts of sustainability and transportation appear to be more frequently linked with an average of 84.1 returns per million tokens, or 21.0 percent of the transportation returns. That is, how much of the overall conversation about transportation appears to be related to sustainability depends on how broadly transportation is defined, how broadly sustainability is defined, and how closely the association between the two is specified. Using a narrow definition of sustainable transportation, the proportion of the transportation conversation connected to sustainability does not exceed 15 percent of the transportation returns. This is also true if a broader definition is used, but the proximity is specified at 25 or fewer tokens. The broader definition of sustainable transportation, coupled with the larger proximity of 50 tokens, generally has too many results that are not directly related to sustainable transportation. Thus, these larger proximities would not be acceptable. Sustainability is certainly a part of the conversations of transportation, but definitely not dominant.

Given the importance of public policies in addressing the issues of sustainable transportation, the relatively low portion of the transportation conversation related to sustainability (less than 15%) indicates that Congress may not be discussing and implementing substantive policies that will address these issues.

Conclusion

Text data analytics is an important emerging methodology in policy and social science research. It allows for empirical investigations of all sorts of new research questions using rigorous, principled methodologies. As this paper has outlined, there are different methods and techniques that are appropriate for different research questions and data. There is not one single way of doing text data analytics because there is not one type of text data or one type of research question. In our case, the complexity of the data and research question necessarily translate into using a methodology that can account for this complexity, rather than trying to reduce or eliminate it. Corpus and computational linguistics are predicated on the complexity of language, and thus, have developed suitable tools and techniques.

The example of sustainable transportation highlights some of the issues associated with the methodology to potential researchers and reviewers. It is possible to investigate complex concepts. However, how the linguistic marker set is constructed and the proximity used in any association investigations will influence the results, and therefore, the implications and conclusions coming from those results. This is one of the reasons that validation is absolutely crucial. Since it can be difficult to replicate the dataset and computer algorithm, research assumptions must be explicitly stated so as to allow reviewers to assess the methodology and conclusions.

References

Baeza-Yates, R., and B. Riberio-Neto (1999), *Modern Information Retrieval*, New York, NY: Addison-Wesley.

Barry, Betsy (2008), *Transcription as Speech-to-text data transformation*, PhD Dissertation, The University of Georgia, Athens, GA.

Barry, Betsy, Suzanne Smith, Beth-Anne Schuelke-Leech, and Clayton Darwin (forthcoming), "From Big Data to Better Data: Issues in Text-Based Analytics," *I/S: A Journal of Law and Policy for the Information Society,* Vol., No.

Biber, Douglas, Susan Conrad, and Randi Reppen (1998), *Corpus Linguistics: Investigating Language Structure and Use*, New York, NY: Cambridge University Press.

Chattamvelli, Rajan (2009), *Data Mining Methods*, Oxford, UK: Alpha Science International Ltd.

Chen, Hsinchun, Roger HL Chiang, and Veda C Storey (2012), "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS quarterly,* Vol. 36, No. 4, pp. 1165-1188.

Coulthard, Malcolm (1985), *An Introduction to Discourse Analysis*, New York, NY: Longman Group Limited.

Cresswell, John W. (2014), *Research Design: Qualitative, Quantiative, and Mixed Methods Approaches, Fourth Edition*, Thousand Oaks, CA: Sage Publications, Inc.

Darwin, Clayton M. (2008), *Construction and Analysis of the University of Georgia Tobacco Documents Corpus*, PhD Dissertation, The University of Georgia, Athens, GA.

Fairclough, Norman (1989), "Discourse and power," *Language and power. London: Longman,* Vol., No.

Fayyad, Usama (1997), "Editorial," *Data Mining and Knowledge Discovery,* Vol. 1, No. 1.

Fayyad, Usama, and R. Uthurusamy (1999), "Data Mining and Knowledge Discovery in Databases: Introduction to the Special Issue," *Communications of the ACM,* Vol. 39, No. 11.

Feldman, Ronen, and James Sanger (2007), *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, New York, NY: Cambridge University Press.

Gabel, Matthew J., and John D. Huber (2000), "Putting Parties in Their Place: Inferring Party Left-Right Ideological Positions from Party Manifestos Data," *American Journal of Political Science,* Vol. 44, No. 1, pp. 94-103.

Grimmer, Justin (2010), "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases," *Political Analysis,* Vol. 18, No. 1, pp. 1-35.

Grimmer, Justin, and Brandon M. Stewart (2013), "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis,* Vol., No.

Grishman, Ralph (2010), "Information Extraction," in Alexander Clark, Chris Fox and Shalom Lappin (Eds.), *The Handbook of Computational Linguistics and Natural Language Processing*, Malden, MA: Wiley-Blackwell Publishing Ltd., pp. 517-530.

Hand, David, Heikki Mannila, and Padhraic Smyth (2001), *Principles of Data Mining*, Cambridge, MA: The MIT Press.

Hansen, Martin Ejnar (2008), "Back to the Archives? A Critique of the Danish Part of the Manifesto Dataset," *Scandinavian Political Studies,* Vol. 31, No. 2, pp. 201-216.

Holzinger, Andreas, Christof Stocker, Bernhard Ofner, Gottfried Prohaska, Alberto Brabenetz, and Rainer Hofmann-Wellenhof (2013), "Combining HCI, Natural Language Processing, and Knowledge Discovery - Potential of IBM Content Analytics as an Assistive Technology in the Biomedical Field," in Andreas Holzinger and Gabriella Pasi (Eds.), *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data,* (Vol. 7947): Springer Berlin Heidelberg. pp. 13-24.

Hunston, Susan (2002), *Corpora in Applied Linguistics*, New York, NY: Cambridge University Press.

Jones, Bryan D., John Wilkerson, and Frank R. Baumgartner, (2009), The Policy Agenda Project, retrieved April 5, 2015, from http://www.policyagendas.org/page/about-project.

Jorgensen, Marianne, and Louise J. Phillips (2002), *Discourse Analysis as Theory and Method*, Thousand Oaks, CA: Sage Publications Inc.

Krippendorff, Klaus (1980), *Content Analysis: An Introduction to Its Methodology*, Newbury Park, CA: Sage Publications.

Kuechler, William L. (2007), "Business Applications of Unstructured Text," [Article], *Communications of the ACM,* Vol. 50, No. 10, pp. 86-93.

Laver, Michael, and Kenneth Benoit (2002), "Locating TDs in Policy Spaces: The Computational Text Analysis of Dáil Speeches," *Irish Political Studies,* Vol. 17, No. 1, pp. 59-73.

Laver, Michael, Kenneth Benoit, and John Garry (2003), "Extracting Policy Positions from Political Texts Using Words as Data," *The American Political Science Review,* Vol. 97, No. 2, pp. 311-331.

Lucas, Christopher, Richard Nielsen, Margaret Roberts, Brandon Stewart, Alex Storer, and Dustin Tingley (2015), "Computer assisted text analysis for comparative politics," *Political Analysis,* Vol. 23, No. 2015, pp. 254-277.

McEnery, Tony, and Andrew Hardie (2012), *Corpus Linguistics: Method, Theory, and Practice*, New York, NY: Cambridge University Press.

Miner, Gary, John Elder IV, Thomas Hill, Robert Nisbet, Dursun Delen, and Andrew Fast (2012), *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, Waltham, MA: Academic Press.

Neuendorf, Kimberly A. (2002), *The Content Analysis Guidebook*, Thousand Oaks, CA: Sage Publications.

Richardson, Barbara (1999), "Toward a Policy on a Sustainable Transportation System," *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 1670, No. -1, pp. 27-34.

Riessman, Catherine Kohler (1993), *Narrative analysis* (Vol. 30), Newbury Park, CA: Sage.

Schiffrin, Deborah (1994), *Approaches to discourse* (Vol. 8): Blackwell Oxford.

Schuelke-Leech, Beth-Anne, and Betsy Barry (forthcoming), "Text Data Analytics for Innovation and Entrepreneurship Research," in Andreas Kurckertz and Elisabeth Berger (Eds.), *Complexity in Entrepreneurship, Innovation and Technology Research – Applications of Emergent and Neglected Methods*, New York, NY: Springer. pp. xxx-xxx.

Siemund, Peter (2011), "Universals and Variation: An Introduction," in Peter Siemund (Ed.), *Linguistic Universals and Language Variation (Trends in Linguistics. Studies and Monographs)*, Berlin, Germany: De Gruyter Mouton. pp. 1-22.

Spinakis, Antonis, and Parakevi Peristera (2004), "Text Mining Tools: Evaluation Methods and Criteria," in Spiros Sirmakessis (Ed.), *Text Mining and Its Applications: Results of the NEMIS Launch Conference*, Berlin: Springer-Verlag. pp. 131-150.

Spirling, Arthur, and Iain McLean (2007), "UK OC OK? Interpreting Optimal Classification Scores for the U.K. House of Commons," *Political Analysis,* Vol. 15, No. 1, pp. 85-96.

Stubbs, Michael (1996), *Text and Corpus Analysis: Computer-Assisted Study of Language and Culture*, Oxford, UK: Blackwell Publishers.

Stubbs, Michael (2001a), *Discourse Analysis: The Sociolinguistic Analysis of Natural Language*, Oxford, UK: Blackwell Publishers.

Stubbs, Michael (2001b), *Words and Phrases: Corpus Studies of Lexical Semantics*, Oxford, UK: Blackwell Publishers.

U.S. Department of Transportation, (2012), Number of U.S. Aircraft, Vehicles, Vessels, and Other Conveyances, retrieved May 11, 2015, from http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_statistics/html/table_01_11.html.

U.S. Energy Information Administration, (2012a), Annual Energy Review, retrieved May 11, 2015, from http://www.eia.gov/totalenergy/data/annual/pecss_diagram.cfm.

U.S. Energy Information Administration, (2012b), International Energy Statistics, retrieved May 11, 2015, from http://www.eia.gov/cfapps/ipdbproject/IEDIndex3.cfm?tid=44&pid=44&aid=2.

United Nations (1987), *Our Common Future: Report of the World Commission on Environment and Development, UN World Commission on Environment and Development, Switzerland.*

van Dijk, Teun A. (1992), "Discourse and the Denial of Racism," *Discourse & Society,* Vol. 3, No. 1, pp. 87-118.

Volkens, Andrea, Judith Bara, and Ian Budge (2009), "Data Quality in Content Analysis. The Case of the Comparative Manifestos Project," *Historical Social Research / Historische Sozialforschung,* Vol. 34, No. 1 (127), pp. 234-251.

Weathington, Bart L., Christopher J.L. Cunningham, and David J. Pittenger (2010), *Research Methods for the Behavioral and Social Sciences*, Hoboken, NJ: John Wiley & Sons, Inc.

Wodak, Ruth (1989), *Language, Power, and Ideology*, Amsterdam: Benjamins.

Figure 1: Disciplinary Foundations for Text Data Analytics

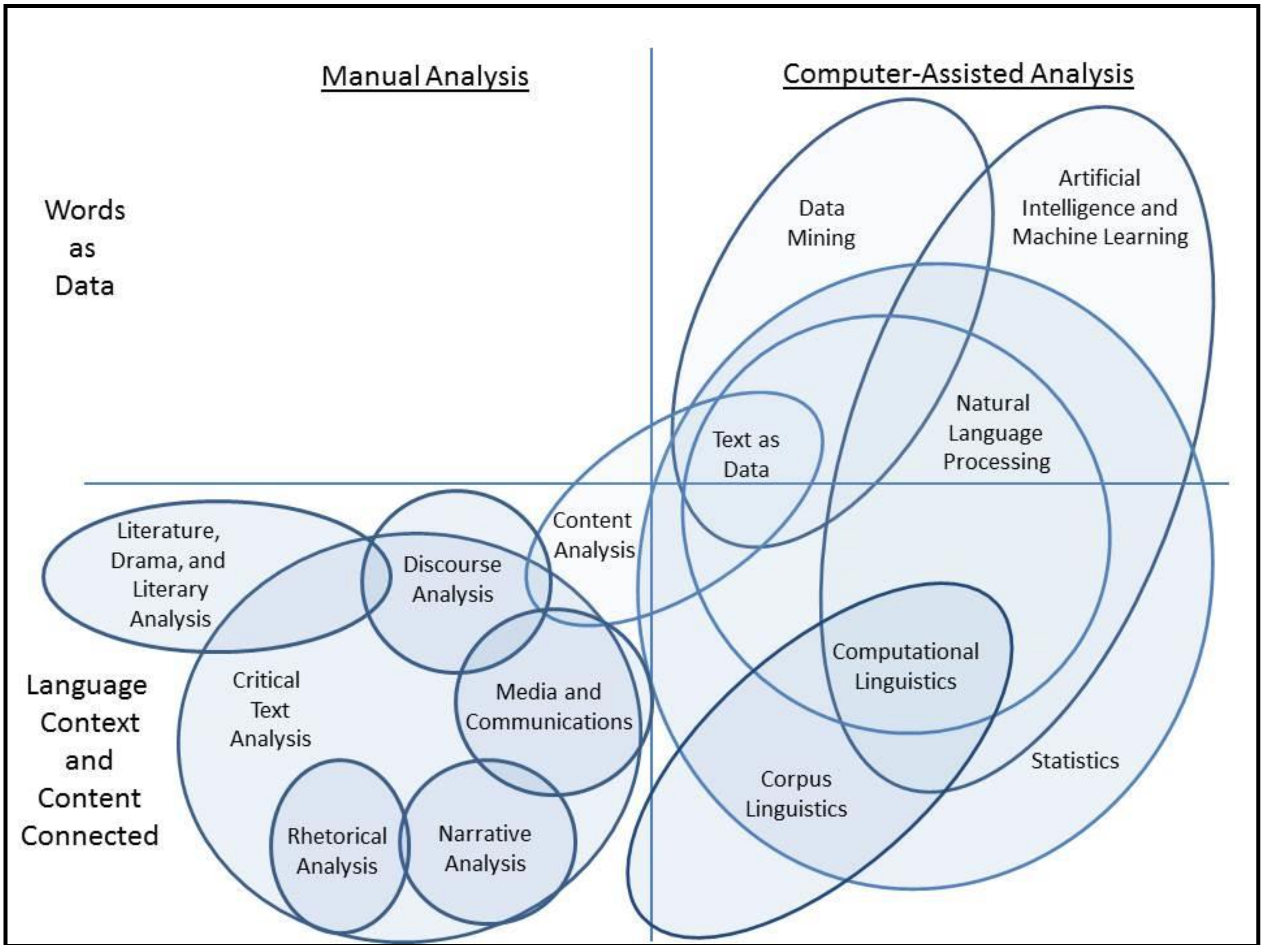
Manual Analysis
Computer-Assisted Analysis
Words as Data
Language Context and Content Connected
Data Mining
Artificial Intelligence and Machine Learning
Natural Language Processing
Text as Data
Content Analysis
Literature, Drama, and Literary Analysis
Discourse Analysis
Critical Text Analysis
Media and Communications
Rhetorical Analysis
Narrative Analysis
Computational Linguistics
Corpus Linguistics
Statistics

Table 1: Transportation Returns

Session	Year 1	Year 2	Number of Files	Number of Tokens	Files with Transportation Returns	Number of Returns	% of Files that contain transportation return	Returns per million tokens
113	2013	2014	4208	164,306,646	1,683	20,760	40.00%	126.3
112	2011	2012	4561	164,236,662	1,740	23,907	38.15%	145.6
111	2009	2010	6110	247,243,388	2,610	41,630	42.72%	168.4
110	2007	2008	6940	263,816,803	2,825	47,821	40.71%	181.3
109	2005	2006	4517	232,074,372	2,244	42,766	49.68%	184.3
108	2003	2004	7104	319,856,928	2,801	53,784	39.43%	168.2
107	2001	2002	6026	301,627,593	2,393	56,838	39.71%	188.4
106	1999	2000	6039	275,512,343	2,042	44,536	33.81%	161.6
105	1997	1998	5071	252,394,411	1,550	41,866	30.57%	165.9
104	1995	1996	4858	417,460,481	3,012	67,171	62.00%	160.9
103	1993	1994	4117	375,886,264	2,590	70,544	62.91%	187.7
102	1991	1992	4153	335,393,629	2,377	84,272	57.24%	251.3
101	1989	1990	5134	422,590,786	3,169	94,910	61.73%	224.6
100	1987	1988	5622	460,083,483	3,446	90,371	61.29%	196.4
99	1985	1986	3332	328,261,343	2,107	57,311	63.24%	174.6
98	1983	1984	6506	530,712,147	3,804	123,033	58.47%	231.8
97	1981	1982	5230	424,917,856	3,155	109,254	60.32%	257.1
Average			5332.50	334,504,281	2,617	65,626	50.12%	190.5
Stdev			1072.14	97,449,853	615	27,683		33
total			89528.00	5,516,375,135	43,548	1,070,775		3174

Table 2: Sustainable Transportation Returns per Million Tokens with a Narrower Definition of Transportation

Session	Transportation Tokens	Sustainable markers and Transportation markers within 5 tokens of each other	Sustainable markers and Transportation markers within 10 tokens of each other	Sustainable markers and Transportation markers within 15 tokens of each other	Sustainable markers and Transportation markers within 25 tokens of each other	Sustainable markers and Transportation markers within 50 tokens of each other
113	126.3	3.4	6.6	9.6	15.7	30.7
112	145.6	3.8	7.1	10.0	16.5	31.5
111	168.4	3.1	6.8	10.4	17.1	33.7
110	181.3	5.5	10.6	15.5	25.4	49.7
109	184.3	3.3	6.2	9.1	15.4	30.6
108	168.2	3.3	6.4	9.4	15.5	30.5
107	188.4	3.3	6.6	10.0	16.7	33.1
106	161.6	3.7	7.2	10.8	17.9	35.4
105	165.9	1.8	3.2	4.7	8.1	15.9
104	160.9	1.9	3.5	5.1	8.2	15.4
103	187.7	2.0	3.9	5.8	9.3	17.7
102	251.3	3.0	5.9	9.3	15.6	31.4
101	224.6	4.5	8.8	12.9	20.7	40.0
100	196.4	2.1	3.9	5.7	9.3	17.6
99	174.6	1.3	2.5	3.6	6.0	11.6
98	231.8	1.5	2.8	4.0	6.4	11.7
97	257.1	4.5	7.8	10.9	17.1	32.0
Average	190.5	2.8	5.4	8.0	13.0	25.4

Table 3: Sustainable Transportation Returns per Million Tokens with a Broader Definition of Transportation

Session	Transportation Tokens	Sustainable markers and Transportation markers within 5 tokens of each other	Sustainable markers and Transportation markers within 10 tokens of each other	Sustainable markers and Transportation markers within 15 tokens of each other	Sustainable markers and Transportation markers within 25 tokens of each other	Sustainable markers and Transportation markers within 50 tokens of each other
113	260.6	9.8	19.2	29.2	48.3	94.3
112	361.4	9.3	17.4	25.6	41.0	79.0
111	380.1	12.0	22.9	34.5	55.9	107.9
110	412.3	15.9	29.6	43.3	70.9	137.0
109	416.2	14.2	25.2	36.2	59.0	114.5
108	461.3	15.3	26.2	37.2	59.7	114.9
107	457.0	13.4	24.3	35.4	57.9	111.0
106	359.7	11.5	21.3	31.5	52.5	102.8
105	438.2	7.6	14.0	19.5	33.0	64.9
104	321.3	6.1	11.6	16.9	27.5	52.3
103	354.0	7.8	14.2	20.5	32.9	63.2
102	399.3	10.3	19.7	29.6	49.3	98.6
101	441.8	15.9	29.2	42.6	69.4	134.3
100	335.5	4.4	8.9	13.6	23.2	46.1
99	329.7	2.7	5.7	8.9	15.5	29.9
98	397.8	2.7	5.8	8.9	14.8	28.7
97	433.7	6.3	12.5	18.6	30.5	60.5
Average	393.7	9.7	18.0	26.4	43.3	84.1