21 April 2020
LING 5990
Maia Petee

Annotated Bibliography: Computational Textual Analysis of Movie Synopses

Alegre, M., & Gordon, P. (1999). Frequency effects and the representational status of regular inflections. Journal of memory and language, 40(1), 41-61.

In this paper, Alegre and Gordon seek to examine how morphologically inflected words are accessed by individuals when read: are they accessed as whole, distinct lexemes, or are their root and inflection accessed separately and then mentally combined? One model that the authors propose is that more frequent words are more likely to have whole-word representations for common inflected forms. A noticeable effect on processing was only found for very frequent inflected forms: inflected forms with a frequency of over 6 per million were processed as a separate lexeme, while forms with a frequency below this were processed using a rule-based model that accessed the lexeme and its inflection morpheme separately. Since word frequency is a pivotal metric that I will use in my study, this paper was useful in understanding how words are processed at varying frequency levels. The wordlist I use (taken from Google's trillion-word corpus) is not lemmatized, so Alegre and Gordon's findings can be used to justify an additional preprocessing step to combine all of a lexeme's forms for purposes of frequency calculation.

Brysbaert, M., and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41:4, 977-990

Brysbaert and New (2009) seek to evaluate a very commonly used measure of word frequency developed by Kučera and Francis in 1982, determine its shortcomings, and propose relevant improvements. They argue that, since word frequency is commonly used in the literature as a metric for designing and evaluating new linguistic and behavioral studies, how it is calculated should not remain

unexamined due to the inconvenience and cost of improving it. They use several recently-released studies to prove that lexical decision times (as recently measured) do not match those published in Kučera and Francis, and propose several improvements. These improvements include adding spoken-language frequencies and individual word-form frequencies, on the basis that lemma frequencies were not superior. Their results were released as the SUBTL frequency norms and made publicly available. This study's emphasis on the critical role of word frequency in the mental representations of lexemes and lexical processing further justifies my own use of frequency as a primary independent variable. Brysbaert & New's empirical examination of the importance of different aspects of measuring word frequency has informed my own treatment of the subject. I will look into the SUBTL frequency norms and see whether any of the knowledge they represent can be incorporated into my own research.

Chen, T., Lu, A., & Hu, S. M. (2012). Visual storylines: Semantic visualization of movie sequence. Computers & Graphics, 36(4), 241-249.

Chen, Lu, and Hu have developed a new method of automatic video visualization: in this study, they seek to summarize a movie storyline into a single, easy-to-understand, and visually-pleasing image. This is essentially an extension of the NLP task of content summarization, as after the content is summarized, the researchers use good visualization principles to represent the summaries in visual form. They use a one-shot clustering algorithm to create the summary, then represent it in storyboard form with images of characters and backgrounds taken from the film in question that are meant to represent a sequence of events. They give examples of successful (counted as instances where the plot could generally be understood from the visualization) and unsuccessful visualizations. When undertaking a visualization project, especially one that attempts to convey complex linguistic information in visual form, it is important to gather research on some of the more unconventional visualizations being created. This study succeeds in disrupting one common concept of visualization as a matplotlib chart, and encourages

me to think outside the box (plot) when considering other possibilities. While I am not seeking to summarize video content, I certainly appreciated Chen, Lu, & Hu's creative approach to automatic storyboarding.

Collins, C., Carpendale, S., & Penn, G. (2009). DocuBurst: Visualizing document content using language structure. Computer Graphics Forum, 28(3), 1039-1046. doi:10.1111/j.1467-8659.2009.01439.x

Collins, Carpendale & Penn sought to examine and expand upon the current norms of document and textual visualization (which often use simple, measurable metrics like word frequency) by incorporating a semantic dimension into their visualizations. Their visualization software, called DocuBurst, summarizes a text across word frequency *and* hyponymic relationships (e.g. *dog* IS-A *animal*) by using burst-like radial visualizations that emanate from a center point. Collins, Carpendale, & Penn also include an interactive element that allows the user to zoom into any part of the visualization, focusing on a subgrouping of words or even a single word. This is a very creative way to approach textual visualization, and one that I will consider when visualizing movie synopsis data. Since my study will focus on word frequency in movie synopses and compare it to box office performance of the films in question, both of these can be incorporated into a radial visualization along with the semantic content of each summary.

Coxhead, A. (2000). A New Academic Word List. TESOL Quarterly; 34, 213 - 238. DOI: 10.2307/3587951

Averil Coxhead spent the late 1990s compiling an academic word list from 3.5 million words of written academic text. An academic word list, from a TESOL perspective, is defined as a moderate-frequency vocabulary list that ESL learners need to know as they move into postsecondary education in English. This study was a huge undertaking at the time, and the resulting wordlist is still used as a standard that ESL instructors use to develop educational materials for learners who are in an English-dominant higher education setting. To develop her wordlist, Coxhead used several frequency measures: first, she used West's (1953) General Service List to exclude the top 2,000 most frequently occurring words. She also

defined an academic-specific vocabulary by measuring its keyness against a non-academic corpus. By doing so, Coxhead identified 570 word families that, together, made up 10% of the academic corpus but only 1.4% of a non-academic corpus of the same size. This is a good example of how word frequency can be used as a metric to define and curate a corpus and as an intelligent textual filter.

Fortuna, B., Grobelnik, M., & Mladenic, D. (2005). Visualization of text document corpus. Informatica, 29(4).

Fortuna, Grobelnik, & Mladenic embark on a textual visualization project that utilizes techniques from linear algebra to synthesize background information about each text. This background information will then be used as input to a visualization system. The researchers' goal is to explain the fact that texts can have similar semantic content even where inter-text similarity between word forms is not found, and they accomplish this with a process termed LSI (Latent Semantic Indexing) to extract semantic information about each text. Once LSI, which is a form of dimensionality reduction, is accomplished, they visualize the results in a cloud-like semantic representation that includes density (frequency) of terms and their proximity to each other. Reading enough studies on creative textual visualizations to get an overview of how researchers across various fields (this article originally appeared in *Informatica)* have visualized document-level word frequency allows me to more intentionally craft my own frequency metric. The labeling of my own data in preprocessing according to lexical frequency groups is a very important process to get right, since frequency is one of my independent variables.

Francis, W. N., Kučera, H., & Mackie, A. W. (1982). Frequency analysis of English usage: Lexicon and grammar. Houghton Mifflin Harcourt (HMH).

This seminal work examining the frequency of lexical items and word families in English was undertaken before computational methods and analysis became widespread. Therefore, its design and analysis represent a huge decades-long effort (the original study, published in 1967, was eventually expanded into this book) that, once complete, enabled contemporaries to use word frequency norms in their own

research. Many works that implicitly and explicitly use Kucera and Francis as a foundational study in their own creation and development of frequency metrics would not have been possible without it, and its word frequency categories were for many years norms in the field. My own understanding of the history of word frequency norms and their development was expanded by this foundational work. It provides important context for the many expansions and refutations that were published in the 1990s and beyond.

Green, C., and Lambert, J. (2018). Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families for eight secondary subjects. Journal of English for Academic Purposes, 35, 105-115.

Green and Lambert, in this relatively recent offering, expanded upon Coxhead (2000) by acknowledging the need of pre-tertiary English learners to have discipline-specific academic word lists that they can focus on learning in order to understand academic English across a number of disciplines. Academic vocabulary used for scientific studies and that used for publications in the humanities can vary significantly, and a single general academic wordlist might not be sufficient to prepare ESL learners for studies in their chosen disciplines. To remedy this, Green & Lambert create eight subject-specific wordlists, compiled and referred to as the SVL (Secondary School Vocabulary Lists). They use frequency, as well as range and dispersion across source texts, to decide which words are of significant utility to ESL learners in an academic context and which words, although specific to a particular academic context, can safely be ignored for pedagogic purposes. This study helped to solidify my current understanding of how lexical keyness is understood in terms of a corpus: frequency is, as we have seen in other studies, an important measure of keyness, but it is especially helpful when used in combination with other metrics.

Graua, B. C., Horrocksa, I., Motika, B., Parsiab, B., Patel-Schneiderc, P., & Sattlerb, U. (2008). OWL 2: The next step for OWL. Web Semantics: Science, Services and Agents on the World Wide Web, 6, 309-322.

This follow-up paper by Graua et al. to their wildly popular and widely used Web Ontology Language (commonly referred to as OWL) outlines the limitations of their originally published language. In this paper, they propose extensions to OWL that will address its limitations and make common workarounds unnecessary. The original OWL had acquired a huge number of users, many of which were researchers conducting computational projects that used complex semantic modeling. The demand for an extended version of the language, therefore, was undeniable. Creating and using a semantic ontology in my own research, especially if I eliminate the box office data element (using financial data greatly restricts the size of the final movie synopsis corpus, since only the top 800-grossing movies have this data available) would elevate my analysis to another level. Instead of relying on the content tags supplied in the corpus to serve as genre in my dataset, I could explore the semantic content of the corpus directly by creating a genre ontology. OWL in particular and the Semantic Web in general are two tools that I would love to incorporate into my research.

Le, T. M., & Lauw, H. W. (2016). Semantic visualization with neighborhood graph regularization. Journal of Artificial Intelligence Research, 55, 1091-1133.

Le & Lauw approach the visualization of textual semantics from a computer science standpoint. Their goal in this research is to discover how to go beyond common document- and sentence-level representations, such as semantic vectors, when visualizing multidimensional textual data. As a baseline to go beyond, Le & Lauw mention the common technique of visualizing word embeddings on a 2D plane as a scatterplot. This technique, however, doesn't take into account forms of textual ambiguity such as polysemy and synonymy. To address this, they propose a method of dimensionality reduction which will first perform topic modeling at the document level instead of the word level, and which will also reduce instances of semantic ambiguity such as polysemy. Once these two steps have been accomplished, visualizing high-level document information becomes a much more intuitive process. The model they create to accomplish this, named SEMAFORE, uses a neighborhood regularization framework developed

specifically for the task and outperforms previously-existing frameworks on a number of textual inputs. This study was fascinating and useful to me because of the frequency with which I have used BoW (bag-of-words) semantic representations in my own work; they are a very common and easy-to-extract feature for machine learning systems that address semantic issues. It is fascinating to see how these simple representations can be significantly outperformed by a different modeling approach.

Leech, G., & Rayson, P. (2014). Word frequencies in written and spoken English: Based on the British National Corpus. Routledge.

This book, written principally by distinguished professor emeritus Geoffrey Leech, approaches the topic of word frequency from both a written and a spoken perspective (most studies only deal with written textual frequency, possibly due to the relative ease with which these frequencies can be obtained). Reflecting the relative cost of obtaining spoken-language data, spoken data, which makes up only 10% of the book's source corpus (the British National Corpus), is given special attention via detailed conversational analysis. Much of the book is dedicated to tabulations of the word frequencies themselves. This book is much more than a simple wordlist, however: Leech & Rayson also include data and grammatical analysis regarding the distribution of grammatical constructs such as modals. This book has been praised as being the first of a particular kind of intelligent wordlist that both functions well as a reference text and can make recommendations about lexical choice based on frequency. This and other books have impressed upon me the importance of choice of source when calculating textual frequency; although in my pilot study I used a list compiled (not curated) from the Google Trillion-Word Corpus, I will consider the merits of this and other sources before choosing a wordlist.

Piantadosi, S. (2014). Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic Bulletin & Review; 21:5, 1112-30.

Zipf's Law, proposed originally by linguist George Zipf in 1935, is a statistical property of all natural languages that describes the distribution of word frequencies across the text of a given language. It

states that the *r*th most frequent word in a language has a frequency proportional to 1/r, meaning that the second most frequent word in that language is half as frequent as the first, the third is a third as frequent as the first, and so on. The extent to which human languages follow Zipf's Law has been studied extensively over the past several decades, but here Piantidosi aims to take this familiar understanding and expand it using statistical principles. He argues that what is usually treated as a simple statistical maxim is in actuality more complex (and far from solved). One point he brings up is that within the broad overall distribution, individual word frequencies are continually shifting from source to source based on conversational context. Word frequencies are also quite dynamic from a longitudinal perspective. Piantadosi also disputes a common explanation for the Zipfian distribution of natural languages (that it is the result of language change and evolution) by arguing that even novel words have also been shown to follow this distribution. This study helped me to think critically about a law of natural language that is taught in classes as an aphorism without critical examination. It will inform my understanding of word frequency laws as dynamic and multi-faceted.

Pradeep Reddy, K., Raghunadha Reddy, T., Apparao Naidu, G., & Vishnu Vardhan, B. (2018). Term weight measures influence in information retrieval. Int. J. Eng. Technol, 7(2), 832-836.

In this paper, Pradeep Reddy et al. use intelligent weighting of search terms to improve accuracy of information retrieval. They explore a number of methods for term weighting, eventually settling upon cosine similarity (a measure reflecting interdocumental distance) as the metric that yields the best results. The two calculations that are multiplied in the calculation of cosine similarity (allowing us to represent them as vectors) are term frequency, or the number of times that each word appears in a given document, and inverse document frequency, which applies weights to terms that emphasize infrequent search terms and deemphasize frequent terms. These two numbers, when multiplied together, yield an intelligent measure of a document's relevance to a particular search query. Understanding the mathematical underpinnings of terms (cosine similarity, tf-idf) that I've become

familiar with through my coursework allows me to understand more clearly how these concepts can be applied in my own research, and the ways in which frequency can be used as a starting point for developing more complex metrics for textual analysis. The application of these techniques can be accomplished using Python (Scikit-Learn has a good tf-idf vectorizer package), but examining them on a more granular level and reading about an application in which they were used successfully has helped expand the breadth of my understanding of complex textual frequency measures.

Qin, P., Xu, W., & Guo, J. (2016). A novel negative sampling based on TFIDF for learning word representation. Neurocomputing, 177, 257-265.

Qin, Xu & Guo propose a new model for encoding semantic textual information that directly addresses some identified shortcomings in a previous approach, termed NEG, that itself was an improvement on a model termed NCE (noise-contrastive estimation). These models and their improvements are created and refined in service to a few primary goals, among which are reducing training time and attaining accurate representations regardless of document sparsity or minimal training data. The current model, termed NEG-TFIDF, focuses on reducing the number of high-frequency words included in training and improving the sampling probability of medium-frequency words, which are often underrepresented due to their higher lexical content and more specific applications. Cosine similarities between document vectors trained using the resulting embeddings are used as an evaluative measure, which allows the authors to conclude that they have met their goal of creating a better-performing embedding set using their NEG-TFIDF sampling strategy. I wanted to include at least one computational paper of this difficulty in my research to demonstrate that frequency can be used as a pivotal metric in complex textual modeling. It also helps to illustrate the academic environment currently surrounding word, sentence, and document embeddings, namely the race to create incrementally better models that will yield greater performance from this valuable technology.

Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1995, October). Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In Proceedings of Visualization 1995 Conference (pp. 51-58). IEEE.

This somewhat older visualization paper explores ways to, as Wise et al. term it, visualize the non-visual. The researchers classify large bodies of text by arranging them spatially, by which they are attempting to mimic the experience of perceiving something intuitively as opposed to the conscious process of reading and comprehending. They aver that finding ways to arrange and present vast quantities of information in an intuitive manner is increasingly necessary with the availability of endless information at one's fingertips. To create and arrange these spatial visualizations, they needed a way to reduce dimensionality by distilling text down to its most important features. This effort, ongoing at the time of publication, was termed MVAB, or the Multidimensional Visualization or Advanced Browsing project; two specific types of visualizations developed under this initiative were termed "galaxies" and "themescapes." Both of these visualizations represent document corpora as spatially arranged dots (or "star systems") of varying sizes across a black landscape. These visualizations seem slightly quaint to the modern eye, but by effectively representing textual content spatially, they are useful for their intended purpose. I felt that this paper was worth inclusion despite its age because of its innovative approach to textual visualizations. This provides me with more inspiration in my own efforts to create unconventional visualizations that incorporate linguistic elements such as term frequency.