CONNECTING DOMAIN KNOWLEDGE ACROSS CONTEXTS: AN APPLICATION FOR MOVIE SEARCH TASKS

Mac Vogelsang
madvogel@iu.edu

Thesis Committee:

Peter Todd
Faculty Sponsor, Ph.D.

Tom Busey
Committee Member, Ph.D.

Robert Goldstone
Committee Member, Ph.D.

Submitted to the faculty of Cognitive Science in partial fulfillment of the requirements for departmental Honors in the degree of Bachelor of Science in the Cognitive Science Program,

Indiana University - May 2019

Abstract

Movies are organized into genres and rated for quality to help potential viewers search for something to watch. But when people actually decide what to watch, they may often use more specific criteria that are not reflected by typical genre labels and tags. The Cognitive Science Movie Index (CSMI) attempts to address this problem for the domain of cognition by providing a curated list of relevant movies. It provides potential viewers with domain-specific characteristics that allow users to browse movies by relevance to cognitive science, scientific accuracy, and tags indicating sub-areas in the field.

In this paper we develop and demonstrate a data driven approach to domain-specific movie categorization. We train a topic model on the MIT Encyclopedia of Cognitive Science which is treated as the source corpus of the domain. Then we apply the topic model to unseen documents, IMDb reviews of CSMI movies, to identify domain-relevant movie characteristics. We extend this technique to distinguish cognitive science and non-cognitive science documents from two target corpora, the Stanford Encyclopedia of Philosophy and IMDb movie reviews, in order to evaluate how well the source corpus generalizes to corpora from near and distant contexts. We aim to demonstrate that this method can be used to a) meaningfully characterize a target corpus by using a source corpus from a potentially distant context and b) identify a domain-specific topology of movies that provides search utility to the CSMI.

Introduction

Modern movies are given genre labels and rated for quality by critics and audiences to help potential viewers decide what to watch and to provide descriptive information on the content, performances, and production quality. Many catalogues and aggregators exist for this purpose, such as the Internet Movie Database (IMDb), Metacritic, Rotten Tomatoes, and The Movie Database. As movie choices continue to increase year over year, aggregate metrics such as star ratings or simple proportions like the “Tomatometer” may not be helpful enough. Movie streaming companies like Netflix or Amazon have begun to employ techniques in machine intelligence to recommend movies for the customer based on their historic viewing patterns. These recommendation systems are often considered successful--even if the user (or the streaming provider) does not receive much of an explanation *why* particular movies are recommended.

Movie rating services provide and rely on a multitude of data. Features like keywords, genre tags, plot synopses, and trailers, can be used by the user to some extent to evaluate the quality and content of a movie. These features are used by streaming providers combined with viewing history, browsing data, and the patterns of customers at individual, local, and global scales. Still, searching and recommending movies that meet very specific criteria can be difficult if those criteria are not mainstream and the typical classification tags are not useful. The Cognitive

Science Movie Index (CSMI) attempts to meet this need for the domain of cognition (Motz, 2013). The CSMI is a curated list of films where cognitive science themes are central to the plot. It provides field-specific keyword tags for each movie (such as Memory, Language, Social Cognition, etc), and a unique three-scale rating system, where users can rate and sort movies on *quality* of the overall movie, *accuracy* in how the film portrays its central cognitive science theme, and *relevance* to the field of cognitive science. The goal of the CSMI is both pragmatic and educational: it catalogues the movies in which complex topics within cognitive science are depicted through media, and it facilitates outreach for the field in an enjoyable way, hopefully increasing engagement and collaboration (Motz, 2013).

Research on media pertinent to a particular discipline is not new. Past work has addressed how other domains have appeared in film, such as biology (Glassy, 2005), archaeology (Hall, 2004), and philosophy (Litch, 2010). These books and studies have explored the scientific accuracy of domain-relevant movies, their cultural impact, and how they can be used in an educational context. However, the CSMI's curated list of movies gives audiences an opportunity to better explore the topology and characteristics of films about cognition. We take the CSMI movie list as providing training category labels that can be used to help determine and evaluate the features of cognitive science movies that make them domain-relevant.

In this paper, we demonstrate a natural language processing framework that uses the CSMI along with several other corpora for identifying and evaluating cognitive science movies. The framework draws upon multiple data sources:

1. The MIT Encyclopedia of Cognitive Science (MITECS) (Wilson & Keil, 2001) serves as the source corpus and as a known index of articles about cognitive science.
2. The Stanford Encyclopedia of Philosophy (SEP) (Zalta, Nodelman, Allen, & Perry, 2003) serves as a nearby corpus to test if the representation of cognitive science (via MITECS) generalizes sufficiently to another definitional source.
3. The Cognitive Science Movie Index provides a set of movies deemed relevant to cognitive science, each accompanied by relevance, quality, and accuracy ratings.
4. The Internet Movie Database (IMDb) is our source of user reviews for each movie, giving us bodies of text to represent the movies and a distant context to compare to the encyclopedia articles.

We represent each document in the source corpus (MITECS) in a vector space model (VSM), such as bag-of-words, and additionally as a topic model, using latent dirichlet allocation (LDA). These document vectors give us a base representation of cognitive science that all other documents are compared to. Vectors of the two target corpora (movie reviews and SEP articles), are created with the terminology of the source corpus, and this allows us to see how well the source corpus can characterize the targets. We do this for both a near and far context: within the two targets, we divide documents into cognitive science (CS) and non-cognitive science (NCS) groups, allowing us to compare the two groups in terms of the MITECS articles.

This text-based evaluation of document features has many potential applications for exploring the topology of the CSMI movies or for comparing the characteristics of different corpora. However, the present study and analysis are restricted to two main goals:

a) To examine the strengths and weaknesses of using a source corpus from a potentially distant context to meaningfully represent a target corpus.
b) To add novel utility to the CSMI website by producing a cognition-relevant characterization of each movie and their movie-to-movie relationships.

We hope to accomplish these goals through two main analyses: Analysis I focuses on the distances of target documents to the source documents across corpora, and between CS and NCS document sets. Analysis II examines the movies themselves, how they are related to each other, and the cognitive science articles they are most related to.

All analyses rely on a variety of vector representations for the documents. A dictionary of terms is created from the source corpus, and we use that as a basis for a bag-of-words model representing each target corpus. A tf-idf transformation is applied to this vector to weight terms by relative importance. Our final representation is in the topic space: we use latent dirichlet allocation (LDA) to create document-topic and topic-word distributions. The topics are trained on the source corpus, and fold-in query sampling is applied to each of the target documents to get their representation in the topic space. The methods section will discuss the available levels of parameterization for these three vector representations, and the results will examine the effects of the parameters and limitations of the models.

Related Work

Documents are frequently processed and compared with vector space models (Salton, Wong, & Yang, 1975) such as tf-idf or topic models like LDA, especially in the field of information retrieval. Venkatesh (2010) includes a comprehensive overview of several models for modern information retrieval systems (see Chapter 2), including the use of tf-idf as a VSM. Several surveys of the usefulness of these models for semantic representation have also been done (M. Jones, Gruenfelder, & Recchia, 2011; Turney & Pantel, 2010). A primary advantage of these representations is that they are fundamentally simple and quick to compute, and therefore are often used in search tasks and similarity queries. When comparing vector space models to semantic topic models, Stone et al. (2010) found that VSMs often outperformed other models when estimating human ratings of similarity in a paragraph comparison task.

General Methodology

Corpora Selection

MIT Encyclopedia of the Cognitive Sciences

We needed a reasonable source corpus that represents a wide range of cognitive science, is trusted by researchers, and maintains a consistent writing and article structure. We chose the MIT Encyclopedia of the Cognitive Sciences (MITECS), a comprehensive reference covering six major areas of cognitive science: philosophy, psychology, neurosciences, computational intelligence, linguistics and language, and culture, cognition, and evolution (Wilson & Keil, 2001). It consists of 471 articles each written by leading experts in the field, and six extended essays for each of the described sections above that summarize the broad concepts. Although it was released in 2001 and some topics may be outdated, MITECS remains an accessible and well-cited (1000+ citations) resource for cognitive science.

All the entries are available online, and were extracted with a python web scraper. We excluded the extended essays because they mostly discussed the contents of other articles. Additionally, two article links ("Situatedness/embeddedness" and "What-It's-Like") did not resolve, so we ended up with a 469 extracted articles. After tokenization, stopword removal, and bigramming (described below), the documents had token counts ranging from 284 to 1344, with an average document length of 623 tokens.

Stanford Encyclopedia of Philosophy

Our second reference corpus is the Stanford Encyclopedia of Philosophy (SEP), a large peer-reviewed encyclopedia with over 1500 articles (Zalta et al., 2003). The encyclopedia is well-maintained and regularly updated by a board of editors. This corpus serves as a nearby context to MITECS; both are encyclopedias written by researchers that utilize an informational tone. Based on a list of articles and their respective subject areas, we selected a subset of articles most pertinent to cognitive science (CS). We identified eight subject areas related to cognitive science, and selected an article if it was tagged with at least one of these areas. See Appendix A for a list of all the subject areas and the eight used for the CS set selection.

This initial selection returned 227 articles for the CS group. From the remaining set of articles that did not have a relevant subject area, we randomly selected an equivalent number of articles to be our NCS group. We then attempted to retrieve each article from the Spring 2018 archive of the SEP. We manually downloaded any new articles not contained in the archive, but some in our list were not yet published on the live site and were ultimately excluded. We ended up with 209 CS articles and 210 NCS articles. One article, *Computational Linguistics*, was then excluded from the CS group for being an extreme outlier in length (containing over 100,000 words after

preprocessing). See Figure 1 for a histogram comparison of document lengths between the two sets, and Table 1 for a summary of the size of the two sets. The document length distributions are comparable between sets, with the greatest difference in length present in articles under 10,000 tokens.

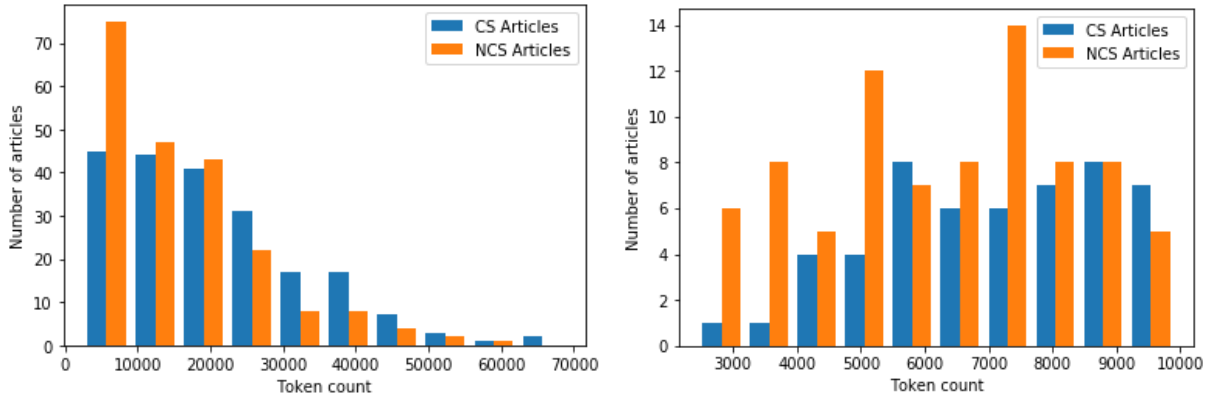

Figure 1: Distribution of article token counts for the CS and NCS SEP articles. The left histogram shows the entire dataset, while the right histogram is zoomed in on the subset of documents with less than 10,000 tokens

Cognitive Science Movie Index

The data from the CSMI came in the form of a series of tables from a March 2019 database export. Merging the tables on movie id gave us a list of 244 movies, the year they were released, their IMDb id, the average, standard deviation, and count for each of the three user rating scales, a short description on why the movie is related to cognitive science, and up to four domain-specific keyword tags set by the curators of the site. The IMDb ids were used to fetch additional data from IMDb such as their feature type (film, TV Short, document, etc), IMDb star rating, and their user reviews.

Internet Movie Database

The Cognitive Science Movie Index gives us a listing of movies relevant to cognitive science but it does not provide any textual information on each movie beyond a one sentence description of its relation to cognitive science. We explored a couple of options for incorporating outside information on these movies before settling on reviews, including the usage of Wikipedia plot synopses or the entire movie scripts themselves. Wikipedia plot synposes were too short, ranging from 200-600 words in length on average, and movie scripts only existed for a small portion of the CSMI movies. Through a quick skimming of the movie scripts that did exist, we found that they did not contain as many content-related words compared to reviews. This intuitively makes sense, as it would be rare for characters within a movie to explicitly discuss its own themes, whereas a movie critic may talk about such topics in their review.

We decided to use reviews from the Internet Movie Database (IMDb). Unlike MITECS or SEP, IMDb reviews are in no way scientific and can be written by any user of the site passionate enough about cinema. The reviews also serve an entirely different purpose: to express subjective thoughts about a piece of media rather than to educate an audience about a scientific research area. Although this target corpus is quite distant from the others we selected, its extremity helps evaluate how well characteristics of the source corpus generalize.

We first needed two comparable lists of CS and NCS movies. Using the CSMI, we obtained a list of 244 CS movies. To select the corresponding NCS movie for each CS movie, we controlled for release year, popularity, and feature type (documentary, feature film, TV short, etc). For each CS movie, we found its position in the list of most popular movies[1] for its release year , and selected an NCS movie of the same type next to it in the ranking. For example, *Planet of the Apes* is the 7th most popular feature film released in 1968, so its NCS pairing was *Romeo and Juliet* which is the 8th most popular for that year. After pairs were selected, manual inspection was done to ensure that movies readily identifiable as cognitive-related (e.g., *Dr. Jekyll and Mr. Hyde*) were not included in the non-cognitive training set merely by coincidence. These movies were replaced with the next item in the popularity list.

There are 244 CS movies and 240 NCS movies. The number of reviews for each movie varies drastically, with more popular movies tending to have a bigger pool of reviews. For this reason, we capped the number of reviews per movie at 1000. We concatenated all reviews by movie to treat the movies as documents and not the individual reviews or the users reviewing them. Combining reviews also lessens the impact of short or uninformative reviews, because more often than not there are enough useful reviews added to the total document to make it a comprehensive representation of the movie. Figure 2 shows the distribution of document length for both sets of films, and Table 1 shows the size in comparison to the other corpora. As with the SEP article lengths, the review documents have comparable length distributions. The biggest difference is among CS and NCS articles with less than 10,000 tokens. There are 23 movies with a combined review word count of less than 500, 13 from the CS set and 10 from the NCS set. These documents were included, because as we will see later, even smaller documents can have an interesting vector characterization.

[1] IMDb uses a combination of user ratings, votes, and browsing activity to calculate popularity. https://help.imdb.com/article/imdbpro/industry-research/faq-for-starmeter-moviemeter-and-companymeter/GSPB7HDNPKVT5VHC#

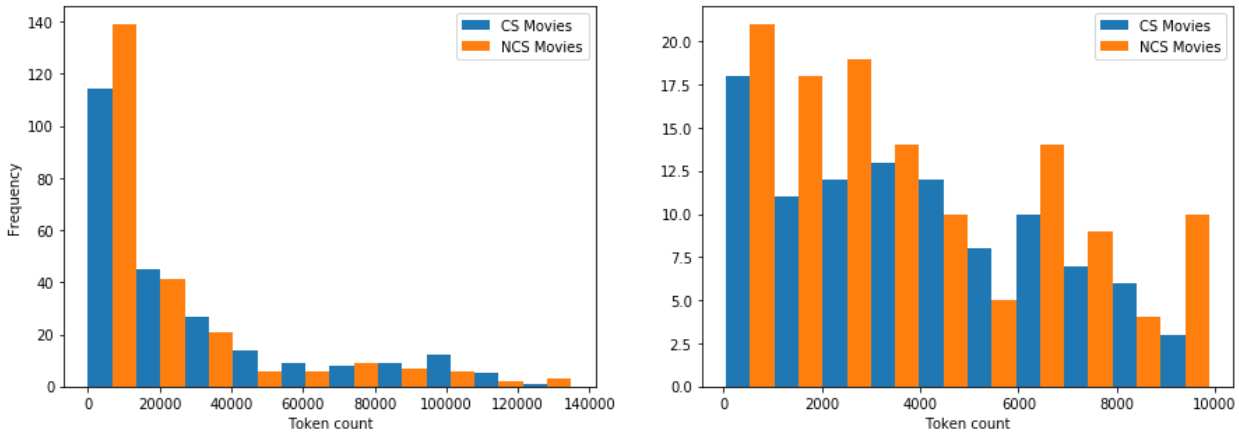

Figure 2. CS and NCS comparison of document lengths. The left histogram shows the entire dataset, while the right histogram is zoomed in on the subset of documents with less than 10,000 tokens.

Corpus	Number of documents	Min token count	Max token count	Mean token count
MITECS	469	284	1344	623
SEP CS	208	2614	68958	20460
SEP NCS	210	2444	57258	15755
IMDB CS	244	45	125698	27793
IMDB NCS	240	70	134891	22207

Table 1. An overview of all five document sets and their size.

Data Preprocessing

All corpora were preprocessed in the exact same way for consistency. The pipeline consists of four main steps: text cleaning, tokenization, stopword removal, and bigramming. Our text cleaning stage removed excess whitespace and HTML and LateX tags from the documents before tokenization. We then used spaCy, a popular industrial-grade python framework for natural language processing, to tokenize words and assign them part of speech tags (Honnibal & Montani, 2017). During the tokenization process, proper nouns were flagged and combined with the preceding word if the preceding word was also a proper noun. This heuristic catches a lot of first name, last name pairs and other named concepts which are frequent in articles discussing scientific theories or the researchers and philosophers that devise them.

At the tokenization stage we have the option to stem or lemmatize the words, but our current pipeline omits this step. Lemmatization involves a morphological analysis of the word to reduce it to its dictionary form, called a lemma. For example, “running” would become “run” and “am”,

"are", and "is" would all reduce to "be". Stemming is hastier form of word reduction, that can truncate plurals to singulars and leave longer words with only their stem (i.e. "animation" and "animator" would both be reduced to "anim"). While both lemmatization and stemming reduce the feature set by a large amount and are useful for improving recall and speed, a lot of contextual information and meaning is lost. One example in our specific application is "animate" versus "animation". Both words become "anim", yet the original words can have very different meaning in a philosophical context or when one is discussing an animated movie. In early versions of our preprocessing pipeline that did stem and lemmatize, we also saw undesirable behavior with regards to names, where they would be erroneously stemmed in ways that didn't make sense ("Turing" becoming "ture"). Furthermore, for the use case of topic modelling, it has been shown that stemming and lemmatization has little impact on the efficacy of the models, and in some cases even leads to worse performance (Schofield & Mimno, 2016). When developing our pipeline, it was also important to consider the end goal. We want a pragmatic outcome of this framework, such as giving users of the CSMI a list of relevant terms from the reviews, and these terms would be harder to interpret if stemming had been applied.

Stopwords were removed from our final list of tokens. The stopword list contains 377 of the most common words in the English language and was obtained from Stone, Dennis, and Kwantes (2011).

The final step of the pipeline was bigramming. We used Gensim's (Řehůřek & Sojka, 2011) bigram model to identify common pairings of words within each document. Words needed to co-occur at least 20 times at the individual document level to be paired as a bigram and added to the token list for that document. Proper noun pairings were not added again if they were already identified via our heuristic above. Including bigrams bolsters the document length for smaller documents because bigrams are added in addition to their existing constituent unigrams, and they help the models pick up on concepts like "artificial intelligence" or "embodied cognition".

Vector Representation of Documents

Each document needs to have a vector representation for distance comparisons, and there are three classes of representations we will use: bag-of-words, term frequency–inverse document frequency, and an LDA topics model. The first step of characterizing one corpus based on the statistical structure of another is to generate a list of terms that all corpora share. For example, distinguishing SEP articles and movies in the domain of cognitive science requires eliminating the words in these target corpora that do not exist in the source corpus (the MIT Encyclopedia of the Cognitive Sciences). By restricting all the models to a single vocabulary, only the structure of domain-related terms is used to analyze each document and terms specific to the other corpora like actor names are filtered out. We applied three basic filters on our source dictionary: we removed tokens that appeared in over 80% of the documents (too frequent), tokens that appeared in only one document (too rare), and a tokens that contained punctuation (mostly noise from the tokenization process). This reduced the vocabulary of MITECS from 20,152 unique words to 10,690 unique words.

This is still a fairly large vocabulary, containing common words spanning concepts outside of cognitive science. One option to handle this is to lower the high-end filtering threshold from 80% to a value like 50%. However, since the vast majority of terms appear in less than 10% of the documents, this method is only effective in removing words like “cognitive” or “human” which are common across all documents but still relevant to cognitive science. Adjusting the lower bound is difficult too, as rare terms like “metaphysics”, which only appears in 6 documents, are often the most useful in depicting fine-grained details about the domain. A more sophisticated approach for relevant term identification is required to encapsulate cognitive science as tightly and comprehensively as possible, and one such filtering method is discussed later within the target corpora.

Bag-of-Words

For each document in all corpora, we created a bag-of-words (BoW) vector space model based off of the pruned MITECS vocabulary. A BoW model is simply a raw count of each of the words in the dictionary, giving each document a 10690-dimensional vector of its individual word frequencies.

Term Frequency–Inverse Document Frequency

Using Gensim’s default term frequency-inverse document frequency model (tf-idf), we transformed the BoW vector into a weight representation of each term. The tf-idf weight for term i in document j was calculated as such:

$$weight_{i,j} = frequency_{i,j} * log_2 \frac{D}{document\_freq_i}$$

where $frequency_{i,j}$ is the raw count from the BoW model, D is the total number of documents in the corpus, and $document\_freq_i$ is the number of documents that term i appears in. Variants of this equation weight the term frequency component or inverse document frequency component in different ways, but we opted for the simplest approach in this transformation.

Tf-idf is designed to weight terms by their relative importance (K. S. Jones, 1972), so frequent words in a document will mean less if they are common across many other documents, and rare words overall will become more important if they have a high individual document frequency. Figure 3 shows the result of the tf-idf transformation on the top 20 review terms for the movie *WALL-E*.

```
[('film', 1033.) ('story', 842.) ('robot', 837.) ('like', 812.)
 ('love', 708.) ('humans', 549.) ('robots', 521.) ('time', 512.)
 ('little', 511.) ('good', 472.) ('best', 466.) ('earth', 465.)
 ('space', 459.) ('characters', 453.) ('movies', 439.) ('great', 439.)
 ('animated', 427.) ('life', 424.) ('people', 417.) ('human', 415.)]
```

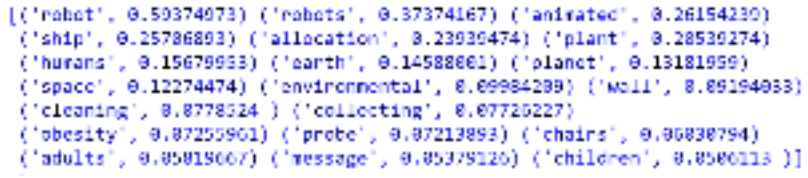

Figure 3. The top 20 highest weighted terms for both the BoW (top) and tf-idf (bottom) models of the movie *WALL-E*.

Latent Dirichlet Allocation

Latent dirichlet allocation is a generative probabilistic model that produces two distributions: a document by topic distribution θ and a topic by words distribution ϕ. The important thing about LDA topic modelling is that it is a form of soft-clustering documents. Each topic can be viewed as a cluster, and each document's distribution θ_d indicates the degree of membership the document has in each cluster. Similarly, the topic by words distribution is also mixed-membership, and the same word can belong to multiple topics (Blei, Ng, & Jordan, 2003; Steyvers & Griffiths, 2007; Xie & Xing, 2013). This property is especially useful for our purpose as it accommodates homographs, and lets the same word appear in multiple contexts (such as "animated" in the philosophical sense and "animated" in the movie sense).

We chose to do topic modelling with LDA for several reasons:

1. Topic models can significantly reduce the dimensionality of a bag of words input by inserting a probabilistic topic layer between the words and the documents. The topics often have some semantic meaning and serve to cluster words that appear in like-contexts together (Crain, Zhou, Yang, & Zha, 2012).
2. Due to the probabilistic nature of topic models, they are relatively insensitive to document lengths. This is useful because there is a large variance in our average document size across our corpora.
3. One goal is to provide new descriptive information to the CSMI, and if topic models trained on a source corpus can represent a distant target corpus well, those topic distributions and their most characteristic words can help users decide what to watch.
4. Topic models have been frequently used in document retrieval and similarity queries (Allen & Murdock, 2016; Greene, O'Callaghan, & Cunningham, 2014; Wei & Croft, 2006; Yi & Allan, 2009). In our case, we are measuring and retrieving similar documents across corpora, but nonetheless topic models are well-studied and are a reasonable starting point.

The LDA model takes several important parameters. The number of topics k must be chosen up front, and different values of k can affect how the documents are clustered. It also takes an alpha and beta value to create the underlying Dirichlet prior that θ and ϕ are sampled from, respectively. An alpha value of less than 1 creates a Dirichlet prior that causes the the documents to have a lower number of highly activated topics. An alpha value greater than 1 generates a Dirichlet prior which causes documents to have a more uniform activation across all

topics. An alpha of exactly 1 creates a uniform Dirichlet prior distribution, meaning the documents are equally likely to have any distribution of topics. The beta parameter behaves similarly to alpha, except it affects the topic by words distribution ϕ.

Later in our results, we discuss the effects of varying the number of topics k and alpha for our analyses. However, unless stated otherwise, we used k = 100, alpha = 1/k, and beta = 0.01. These are common values for alpha and beta in the literature, and they make sense for our purpose since we want documents to be distinct from each other in the topic space. Higher alpha values lead to trivial results where all documents are extremely similar.

The topic model is only trained once on the source corpus. Using the fold-in query sampling method described in (Hofmann, 2017), we apply the existing model to the unseen target documents. This technique gives each new document a topic representation in terms of the topics found in the source corpus, and is an efficient way to compute these representations across corpora (Lu, Mei, & Zhai, 2011). An alternative approach would have been to extend the source corpus with the documents from the target corpus, and train a model on both at once. But because our corpora are so different in structure and tone (especially the movie reviews compared with the encyclopedia articles), an LDA topic model would likely pick up on these core stylistic differences and generate topics more indicative of the the type of corpus rather than its content (Murdock, 2019).

Filtered Bag-of-Words and Tf-idf

Since the original BoW model is an unweighted word count, words that are frequent have high activation. Figure 4 shows the most frequent words across all cognitive science target documents, both SEP articles (bottom) and reviews of the CSMI movies (top). For the movie reviews, most of these words are not relevant to our desired domain of cognition, and display unsurprising patterns of language use in reviews with words such as “story”, “effects”, “acting”, “good”, and “love”. In the SEP articles there are more relevant top words, but we still see a few highly-frequent non-relevant words such as “given”, “like”, and “example”.

```
-- Top words in CogSci Movie Reviews --
Matrix shape: (10690, 244), 9062 nonzero words
[('film', 128245.0), ('like', 60040.0), ('story', 42437.0), ('good', 41992.0),
 ('time', 39407.0), ('great', 30785.0), ('people', 28736.0), ('think', 25914.0),
 ('movies', 24886.0), ('character', 22738.0), ('action', 21828.0), ('best', 21758.0),
 ('life', 21633.0), ('watch', 19959.0), ('characters', 19576.0), ('world', 19548.0),
 ('seen', 18358.0), ('love', 18268.0), ('know', 17536.0), ('better', 15910.0),
 ('little', 15836.0), ('effects', 15590.0), ('real', 15212.0), ('human', 15172.0),
 ('scenes', 15093.0), ('acting', 14727.0), ('makes', 14393.0), ('going', 13734.0),
 ('scene', 13501.0), ('mind', 13497.0)]

-- Top words in CogSci SEP Articles --
Matrix shape: (10690, 208), 9717 nonzero words
[('theory', 22068.0), ('mental', 14410.0), ('true', 14061.0), ('example', 13849.0),
 ('language', 13439.0), ('different', 12661.0), ('truth', 12581.0), ('view', 12541.0),
 ('like', 12404.0), ('experience', 12034.0), ('content', 11697.0), ('case', 11666.0),
 ('logic', 11550.0), ('states', 11512.0), ('information', 11435.0), ('sense', 10986.0),
 ('properties', 10532.0), ('argument', 10362.0), ('consciousness', 9402.0),
 ('theories', 9378.0), ('terms', 9291.0), ('world', 8996.0), ('given', 8926.0),
 ('problem', 8824.0), ('account', 8765.0), ('physical', 8536.0), ('form', 8399.0),
 ('question', 8308.0), ('possible', 8078.0), ('knowledge', 8073.0)]
```

Figure 4. A listing of the top 20 word-frequency pairs in cognitive science (CS) documents for both the SEP corpus and the IMDb corpus.

Tf-idf does some filtering of frequent, non-diagnostic terms, reducing them non-zero but small tf-idf weights, so they are never truly removed from the document. We employ a strict filtering technique that attempts to remove non-CS-relevant terms, while maintaining a standard BoW frequency weighting on the remaining terms.

For each CS-NCS pair of within-corpora sets, we get the overlapping terms between them based on their total bag-of-words table (a BoW vector summed across all documents in each set). A word is considered to be in the overlap if it appears at or above the 80th frequency percentile in both the CS documents and the NCS documents. This creates a new list of stopwords that are commonly shared between the CS and NCS documents. We iterate through each of the documents and remove words that are in this stoplist—unless the word is in the top 5% most important words for the current document (determined via the tf-idf score). This technique utilizes both corpora-level and document-level information, letting us remove shared words without the loss of diagnostic document-level features.

Analysis I: Cross-Corpus Distance Comparison

We would like to accurately characterize documents in the domain of cognitive science using a vocabulary and topic distribution trained on a known cognitive science knowledge base. Past research has shown these three models (BoW, tf-idf, and LDA) can represent the trained corpus quite well, but are they able to generalize when applied to an unseen target corpus? We can measure the degree of generalization to both a nearby context (Stanford Encyclopedia of

Philosophy articles), and a distant context (IMDb movie reviews). For each of these target corpora, we get the distribution of their document distances to the MITECS documents. The diagnostic for generalization is the separation between the means of the distributions for the cognitive science documents and the non cognitive science documents if the model has characterized the target corpora in terms of this domain properly.

Methods

To calculate distances between documents, we primarily used cosine distance for the BoW and tf-idf vectors, and Jensen-Shannon distance (Lin, 1991) for the topic distributions. Cosine distance is simply the cosine of the angle between the two vectors. If the vectors are equal (and the angle between them is 0), then cosine distance is 0. If they are orthogonal to each other, the cosine distance is 1.

$$CD(\mathbf{X} \,||\, \mathbf{Y}) = 1 - \cos(\theta) = 1 - (\boldsymbol{X} \cdot \boldsymbol{Y} / \, ||\boldsymbol{X}|| \cdot ||\boldsymbol{Y}||).$$

Jensen-Shannon is a symmetric version of Kullback-Leibler divergence, a measure of relative entropy and how one probability distribution differs from a reference probability distribution (Kullback & Leibler, 1951).

$$KL(\mathbf{X} \,||\, \mathbf{Y}) = \sum_i X(i) \log (X(i)/Y(i))$$

Since $KL(\mathbf{X} \,||\, \mathbf{Y}) \neq KL(\mathbf{Y} \,||\, \mathbf{X})$, we use Jensen-Shannon:

$$JS(\mathbf{X} \,||\, \mathbf{Y}) = 1/2 \; KL(\mathbf{X} \,||\, \mathbf{A}) + 1/2 \; KL(\mathbf{Y} \,||\, \mathbf{A})$$
$$\text{where } \mathbf{A} = 1/2 \; (\mathbf{X} + \mathbf{Y})$$

For each of the model types (BoW, tf-idf, and topics), we computed the square distance matrix between all of the MITECS articles. The BoW and tf-idf vectors were normalized to unit length when distance computations other than cosine were used (such as manhattan distance), to ensure the large difference in document sizes between the source and targets did not affect the distance measure. The distribution of these distances is a measure of the average spread between all the source documents.

Then for each document in the target corpora [CS SEP (208 docs), NCS SEP (210 docs), CS Movies (244 docs), and NCS Movies (240 docs)], we calculated its average distance to either all of the MITECS articles, or to the k nearest articles. This gives us a frequency distribution of distances for each corpus that we can then take the mean of and show how it compares to the others.

Results

Primary Model Comparison

Figure 5 shows a comparison between the three models, LDA topics (top), BoW (middle), and tf-idf (bottom). The left column is a plot of the frequency distributions of average distance to the 10-nearest neighbors in the source corpus. The vertical dashed lines show the average of each distribution, and the right column shows these averages in an alternative way. The topic model was trained with 100 topics and alpha and beta of 0.01. Overall the topic model results were not very sensitive to changes in parameters: we saw the same relative ordering of the corpora by distance measure with many different numbers of topics, and values of alpha and beta (as long as they were not too high to make all documents inseparable), as well as stability in different runs of the same topic model parameters.

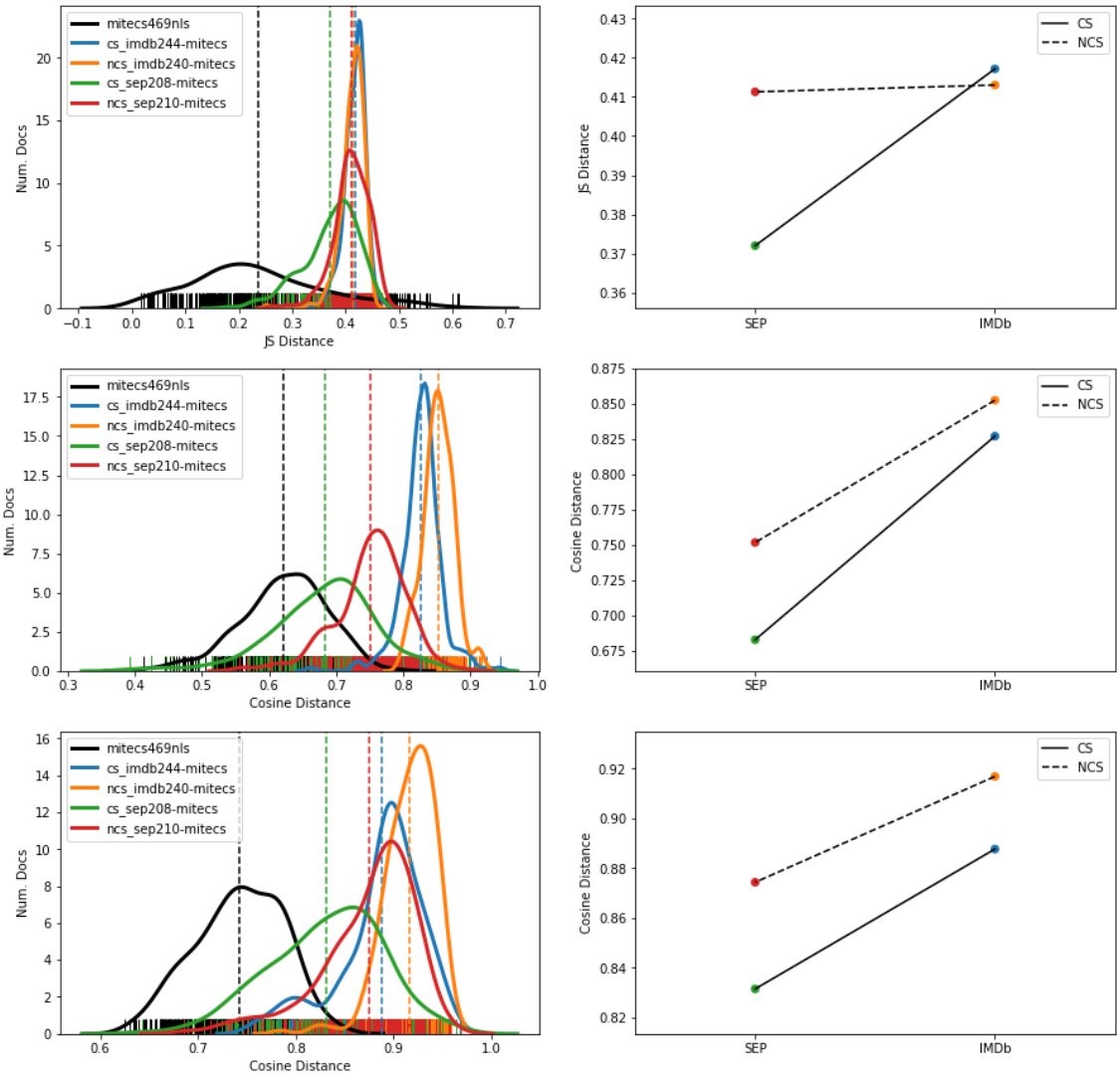

Figure 5. Average distances to the 10-nearest source documents for each of the three models, LDA topics (top), BoW (middle), and tf-idf (bottom). The left column shows the frequency distributions (with the means as dashed lines) and the the right column compares the near and far contexts for documents both inside the domain of the source (CS) and outside of the domain (NCS).

To better further visualize the spread of the documents in the three vector spaces, we ran multidimensional scaling (MDS) for two dimensions. Figure 6 shows these results below.

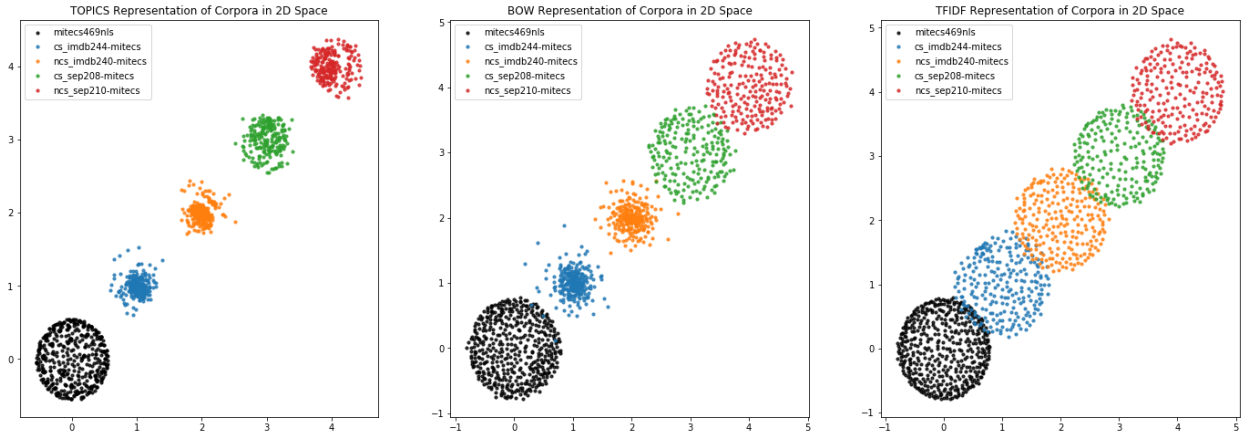

Figure 6. MDS for each of the models. The resulting dimensions were originally each centered at (0,0), but they were offset by (+1,+1) to separate the clouds for easier comparison.

Effect of K in Choosing Closest Source Documents

We also compared different values of k for taking the average distance from each target document to the k-nearest source documents. Figure 7 highlights the effect of reducing k from 469 (the total number of source documents) to 50 and then to 10 for the document distances with the topic model. The graphs reflect an intuitive result: lowering k reduces variance and the average document distance, as each target document is increasingly represented by its nearest neighbors in the source corpus. Not shown are the plots for the other two models, but changing k for the BoW distances leads to a slight reduction in average document distances but no real change in variance. The version with the tf-idf distances shows very similar patterns as the topic model. These plots led us to choose k=10 for the majority of analyses for a more precise

comparison to the source corpus.

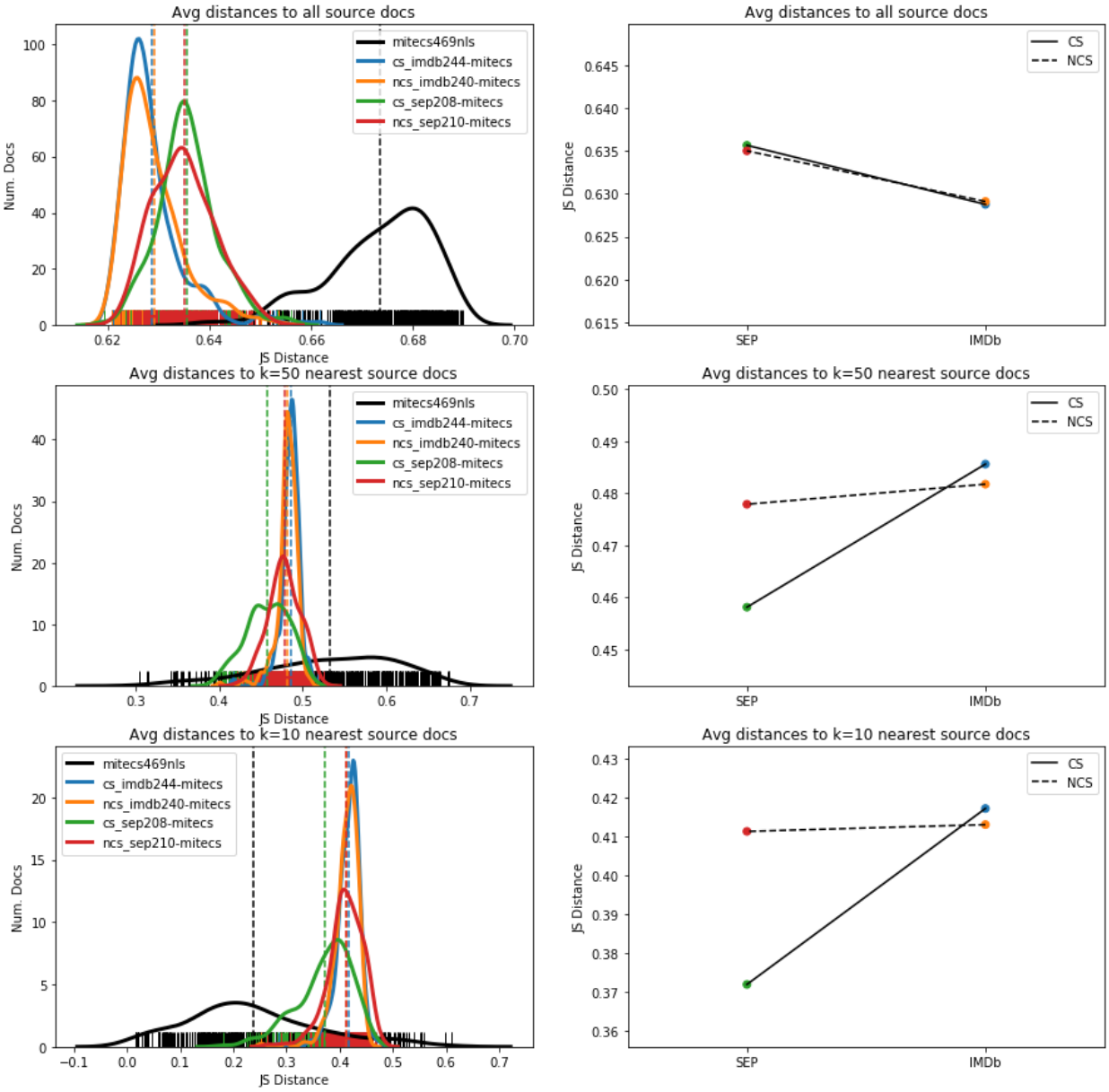

Figure 7. Changes in the topic distance distributions based on k-nearest source documents.

Bag-of-Words and Tf-idf Filtering

This comparison shows the effect of our strict overlap-based term filtering on the original BoW model. The top two plots of Figure 8 show the filtered BoW model, with the original unfiltered model shown in dim colors behind. The bottom two plots show the tf-idf-F transformation on the BoW-F model, and the original tf-idf in dim colors. Topic distributions were also recomputed using the filtered BoW, but no significant change was observed.

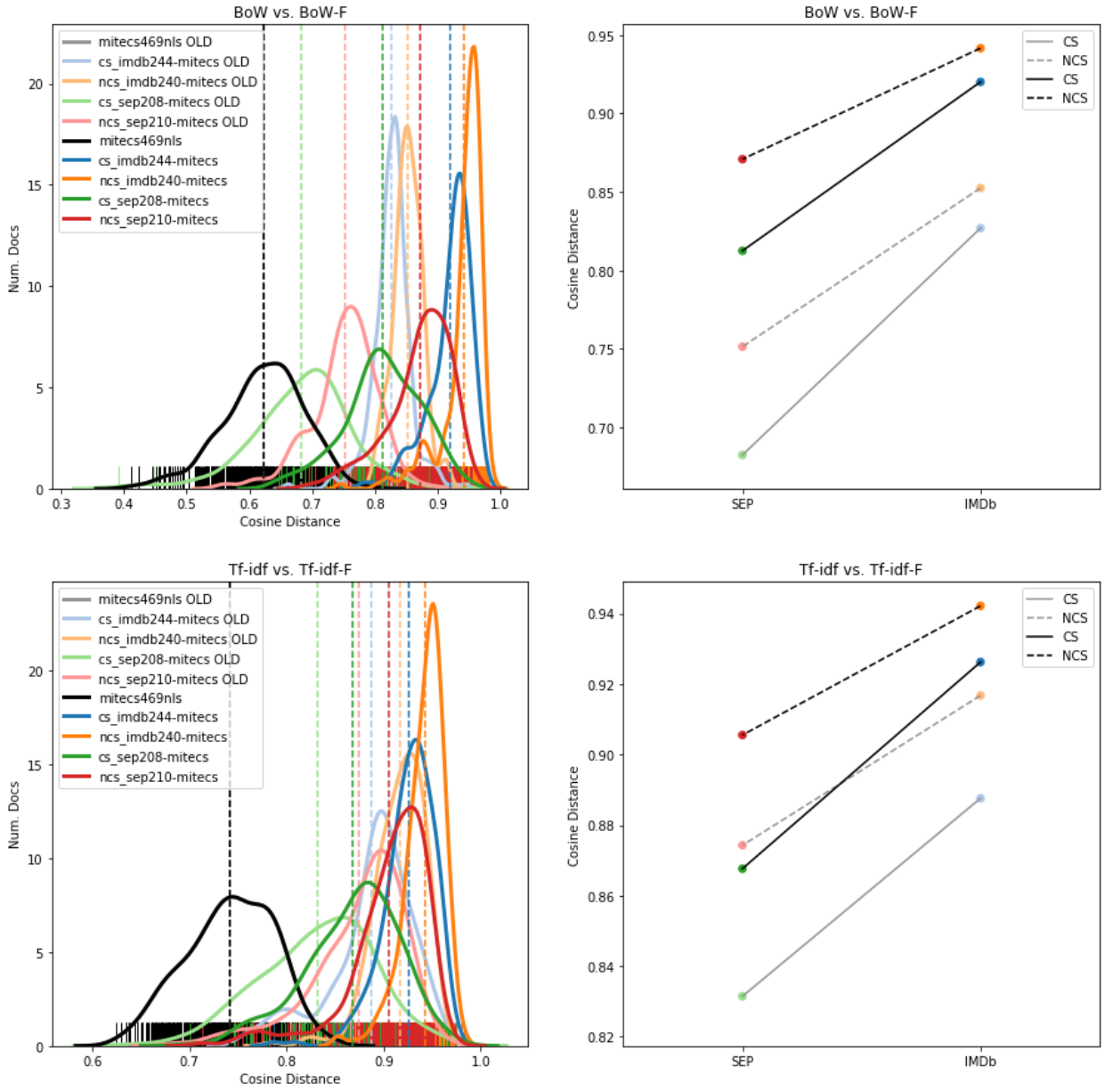

Figure 8. The effect of filtering on the distances between corpora. BoW and BoW-F is shown at the top, while tf-idf and tf-idf-F are shown at the bottom.

Sensitivity to Distance Measures

Our final plot show the BoW and tf-idf models with two different distance measures (Figure 9). Rows 1 and 3 are the same as Figure 5 and were calculated with the cosine distance. The distributions in rows 2 and 4 were calculated with the manhattan (a.k.a cityblock) distance, where MAN(X || Y) = $\sum_i |\mathbf{x}_i - \mathbf{y}_i|$. Vectors were normalized to unit length before taking the manhattan distance.

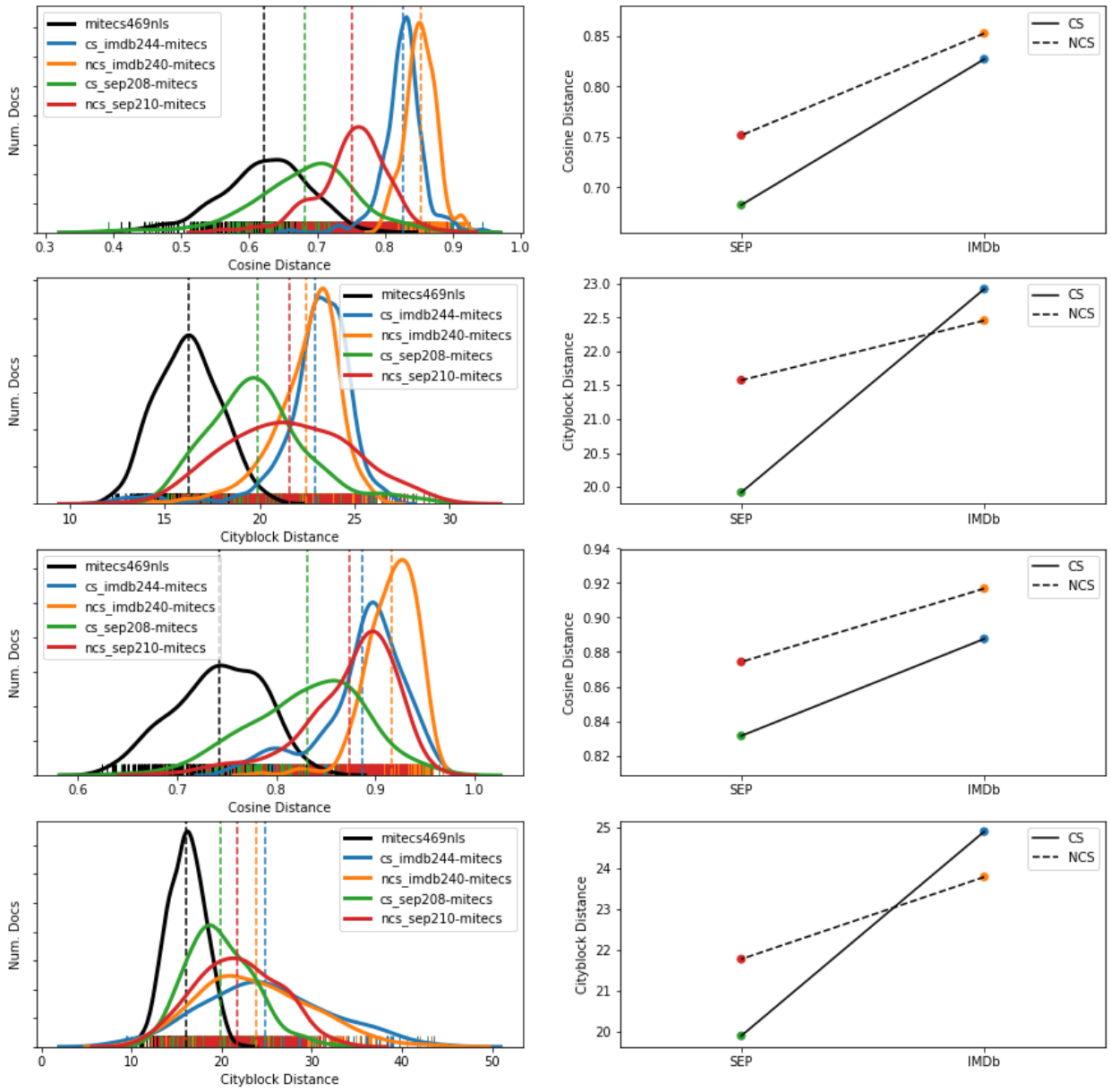

Figure 9. BoW (1st and 3rd row) and tf-idf (2nd and 4th row) are highly sensitive to the distance measure.

Discussion

Before discussing the distribution plots, we would like to make a note about the within-corpus distance distribution of MITECS (shown in black in all plots). Although it is plotted on the same axis, it should be interpreted differently than the colored curves. The MITECS curve describes the within-corpus spread of documents. If MITECS articles are far from the their k-closest neighbors on average, the distance values shown in this curve and the spread of the documents

will be high. If MITECS articles are close to their neighbors on average, the documents are more condensed in the model space and the distance values will be low. One may think they should compare the position of the colored curves to the black curve, but each of the colored curves were generated by taking the distance to the MITECS articles to begin with. With each of the colored curves, a value closer to zero tells us the documents are closer to the source corpus on average.

The separation between the CS documents and NCS documents is our diagnostic for how well the models distinguish between documents in and outside of the desired domain. When the diagnostic is consistent in both a near and far context, it indicates the vocabulary and models trained on the source corpus have effectively generalized to the targets. The right column in Figure 5 shows line plots that indicate this generalizability, and the closer the angle between the CS and NCS lines is to zero, the more consistent the separation is across contexts. We see that the topic model does a poor job at generalizing, but the vector space models perform much better.

When we represent the nearby context of SEP with all three models trained on MITECS, we see a significant separation between the CS and NCS articles, with the CS articles being closer to the source corpus on average. However, the topic model cannot effectively separate the distant context of movie reviews, while the vector space models can. The diagnostic between CS and NCS movies is noticeable in the BoW and tf-idf models, and the movie reviews are further away from the MITECS articles than the SEP articles on average. In the BoW model, this separation is smaller between movie reviews than SEP articles, showing the effect of a distant context less related to cognition when compared to the encyclopedias. The BoW model is less useful in distinguishing these documents from each other, but still effective. Interestingly, the tf-idf VSM does not exhibit this decrease in diagnosticity from articles to movie reviews, and we see a stronger ability for the model to generalize with more parallel lines.

These results indicate, in a couple of ways, the nature of movie reviews when compared to the other corpora. The context in which these reviews are written (and who they are written by) is very different, and we see this with the larger distance on average from the source corpus when compared to the SEP articles. Despite this, the tf-idf model (and to a lesser extent the BoW model) has a surprising ability to represent such an unrelated set of documents. These models perhaps have an advantage in their simplicity, as they are purely on a shared set of word frequencies and are not trained to find underlying latent structure like the LDA model is. The latent topics that the LDA produces seem to be specific to MITECS, and fail to create separable distributions when applied to unseen documents. The nature of the LDA model and its shortcomings in this task are explored more in Analysis II.

In Figure 7, we looked at how choosing k in the k-nearest-neighbors distance computations affected the models' generalizability. In the case of the topic model, reducing k gave us more separable distributions and more meaningful results. For the other models, setting k to the full source corpus size of 469 also gave us messier results, but smaller values had similar effects as

what we reported with k=10. It made sense to stick with a small value of k across the board to improve precision and lower average document distances.

Out filtering approaches did not have much of an impact on generalizability as we hoped. Figure 8 shows the change in distributions, and the lines are a little less parallel for tf-idf, but mostly we see the distances increase across the board for both models. The filtering technique described in the methods adds more zeros to the vectors and increases the average distance between documents overall, pushing the distributions towards the upper bound of 1 for cosine distance. This can have a compression effect, which could be why the diagnostic between CS and NCS movies decreases in the tf-idf-F model.

The last consideration for this analysis regards distance measures in high dimensionality. The "curse of dimensionality" is a well-known problem in computer science where as the number of dimensions *n* increases, the distances of the closest neighbors and the farthest neighbors to a given target point approach equality. Beyer et al. (1999) provide a very intuitive explanation for why this happens. It presents issues for our current study, and our high-dimensional vectors cause the distances to have low contrast. This explains the high degree of dissimilarity and restricted variance that we see in all of the models except the topics model. Dealing with this issue can be difficult when distance measures or the k-closest neighbors are no longer effective and meaningful. Aggarwal, Hinneburg, and Keim (2001) suggest that L^p-norm distance for values of $p <= 1$ is empirically and theoretically more effective in high-dimensions than Euclidean distance ($p = 2$).

Although we aren't comparing it to the Euclidean distance, we regenerated the BoW and tf-idf results using the manhattan distance (L^p-norm where $p = 1$) to evaluate the robustness of our models in a different distance space. Figure 9 shows the result, and the manhattan distance drastically changes the result for both BoW and tf-idf. The bound is no longer capped at 1, and the spread between the CS and NCS sets is much higher than before across the SEP articles. The CS and NCS movie distributions are now more similar to each other, and create a cross-over effect where the CS movies are strangely further away from the MITECS articles than the NCS movies. This is contrary to what we see in the original cosine-distance representation of the models. Without measures of significance and a bounded 0-1 distance scale, we cannot tell whether this cross-over effect is meaningful.

This analysis presents a set of results that show the efficacy of topic models compared to vector space models for representing target corpora with a source corpus in both near and far contexts. They also highlight the variable—and often conflicting—outcomes when changing parameters like k or the distance metric, and the greater issues with distances in extremely high dimensional spaces.

Analysis II: Movie-level Inspections

This first analyses showed that a BoW or tf-idf model trained on vocabulary from the MITECS corpus can characterize the movies of the CSMI to some degree (at least enough to separate them from non-cognitive movies). In this section, we explore the topology the movies to gain some insight into the nature of this characterization. We start by inspecting the vectors of individual movies and building word clouds that could add utility to a movie catalogue like the CSMI. Then we briefly discuss the efficacy of multidimensional scaling and similarity networks to cluster movies into latent clusters. Finally, we show exactly what the 10-nearest MITECS documents look like for a selection of movies and how the poor generalizability of LDA can influence the list.

Movie Inspection

We rely upon the BoW and tf-idf models presented earlier to inspect the content of the movie reviews, and qualitatively examine how accurately these models represent the cognitive science content within a movie. We include the filtered versions of the BoW and tfi-idf models for direct comparison, as well as the topic model to show its poor characterization. In Table 2 below, we display the top 20 terms by frequency/weight for each vector as a heuristic for evaluation. For the topic model, we show the top five topics along with the top five most relevant terms to represent each topic.

Frankenstein, BoW, len: 3694 [('film', 1275.) ('time', 394.) ('like', 378.) ('classic', 309.) ('life', 292.) ('great', 290.) ('scene', 289.) ('story', 282.) ('movies', 246.) ('good', 246.) ('creature', 243.) ('little', 225.) ('novel', 216.) ('girl', 186.) ('book', 177.) ('think', 173.) ('1931', 170.) ('scenes', 164.) ('people', 164.) ('character', 160.)]	**Frankenstein**, BoW Filtered, len: 1902 [('1931', 170.) ('laboratory', 60.) ('misunderstood', 40.) ('flowers', 38.) ('experiment', 35.) ('experiments', 25.) ('abnormal', 25.) ('universal', 24.) ('james', 24.) ('burning', 24.) ('1930', 24.) ('faithful', 21.) ('birth', 20.) ('influential', 19.) ('1930s', 19.) ('shadows', 17.) ('electrical', 16.) ('professor', 15.) ('neck', 15.) ('mentor', 15.)]
Frankenstein, tf-idf, len: 3694 [('1931', 0.747) ('creature', 0.33) ('assistant', 0.134) ('novel', 0.128) ('laboratory', 0.126) ('creation', 0.118) ('scientist', 0.104) ('makeup', 0.099) ('flowers', 0.096) ('lighting', 0.086) ('1930', 0.082) ('classic', 0.08) ('harris', 0.073) ('girl', 0.073) ('alive', 0.073) ('abnormal', 0.07) ('misunderstood', 0.065) ('doctor', 0.065) ('1935', 0.065) ('1930s', 0.065)]	**Frankenstein**, tf-idf Filtered, len: 1902 [('1931', 0.904) ('laboratory', 0.153) ('flowers', 0.116) ('1930', 0.1) ('harris', 0.088) ('abnormal', 0.085) ('misunderstood', 0.079) ('1935', 0.079) ('1930s', 0.079) ('james', 0.076) ('electricity', 0.048) ('electrical', 0.048) ('burning', 0.048) ('universal', 0.045) ('experiment', 0.045) ('seventy', 0.044) ('mentor', 0.044) ('experiments', 0.044) ('brute', 0.042) ('flower', 0.04)]
Frankenstein, Topics, len: 100 84: 0.176 | language,0.014; knowledge,0.008; states,0.007; mental,0.006; problem,0.005; 72: 0.097 | psychology,0.007; cognitive,0.007; depth,0.006; causation,0.006; behavior,0.006; 94: 0.070 | language,0.011; causal,0.008; infants,0.006; psychology,0.006; human,0.005; 86: 0.046 | theory,0.009; meaning,0.008; information,0.008; language,0.008; knowledge,0.007; 75: 0.044 | cognitive,0.013; play,0.010; example,0.006; cultural,0.005; movements,0.005;	
The Matrix, BoW, len: 4203 [('film', 1327.) ('action', 761.) ('like', 735.) ('effects', 729.) ('good', 590.) ('time', 573.) ('matrix', 550.) ('world', 504.) ('story', 497.) ('special', 483.) ('movies', 453.) ('great', 451.) ('people', 414.) ('think', 392.) ('best', 370.) ('seen', 367.) ('scenes', 296.) ('real', 294.) ('reality', 265.) ('know', 259.)]	**The Matrix**, BoW Filtered, len: 2344 [('matrix', 550.) ('machines', 157.) ('philosophical', 89.) ('virtual', 71.) ('computers', 39.) ('illusion', 35.) ('groundbreaking', 35.) ('artificial', 32.) ('parallels', 30.) ('concepts', 29.) ('mnemonic', 27.) ('rabbit', 24.) ('programs', 24.) ('simulation', 21.) ('sources', 19.) ('software', 19.) ('gibson', 18.) ('influential', 17.) ('code', 17.) ('parallel', 16.)]
The Matrix, tf-idf, len: 4203 [('matrix', 0.73) ('machines', 0.172) ('smith', 0.166) ('action', 0.164) ('effects', 0.147) ('agents', 0.124) ('1999', 0.119) ('arts', 0.111) ('fight', 0.102) ('virtual', 0.1) ('mnemonic', 0.098) ('wars', 0.084) ('philosophy', 0.081) ('special', 0.074) ('humans', 0.074) ('reality', 0.065) ('real_world', 0.064) ('fighting', 0.064) ('sequences', 0.058) ('religious', 0.058)]	**The Matrix**, tf-idf Filtered, len: 2344 [('matrix', 0.889) ('machines', 0.209) ('virtual', 0.121) ('mnemonic', 0.12) ('philosophical', 0.07) ('gibson', 0.068) ('simulation', 0.06) ('rabbit', 0.06) ('disciples', 0.055) ('groundbreaking', 0.052) ('programs', 0.051) ('computers', 0.051) ('illusion', 0.05) ('software', 0.049) ('revolutions', 0.043) ('religions', 0.043) ('parallels', 0.04) ('sources', 0.038) ('cave', 0.037) ('unconscious', 0.031)]
The Matrix, Topics, len: 100 84: 0.164 | language,0.014; knowledge,0.008; states,0.007; mental,0.006; problem,0.005; 93: 0.079 | information,0.006; social,0.005; human,0.005; focus,0.005; language,0.005; 72: 0.069 | psychology,0.007; cognitive,0.007; depth,0.006; causation,0.006; behavior,0.006; 83: 0.058 | theory,0.011; model,0.007; structure,0.007; example,0.006; models,0.005; 94: 0.058 | language,0.011; causal,0.008; infants,0.006; psychology,0.006; human,0.005;	
WALL·E, BoW, len: 4663 [('film', 1833.) ('story', 842.) ('robot', 837.) ('like', 812.) ('love', 798.) ('humans', 549.) ('robots', 521.) ('time', 512.) ('little', 511.) ('good', 472.) ('best', 466.) ('earth', 465.) ('space', 459.) ('characters', 453.) ('movies', 439.) ('great', 439.) ('animated', 427.) ('life', 424.) ('people', 417.) ('human', 415.)]	**WALL·E**, BoW Filtered, len: 2797 [('plant', 202.) ('machines', 76.) ('environmental', 71.) ('cleaning', 60.) ('load', 58.) ('collecting', 58.) ('allocation', 56.) ('robotic', 53.) ('voices', 44.) ('advanced', 42.) ('circuit', 41.) ('probe', 40.) ('curiosity', 40.) ('centuries', 39.) ('chairs', 37.) ('companion', 36.) ('programmed', 32.) ('mechanical', 32.) ('population', 29.) ('items', 29.)]
WALL·E, tf-idf, len: 4663 [('robot', 0.594) ('robots', 0.374) ('animated', 0.262) ('ship', 0.257) ('allocation', 0.239) ('plant', 0.205) ('humans', 0.157) ('earth', 0.146) ('planet', 0.132) ('space', 0.123) ('environmental', 0.1) ('wall', 0.092) ('cleaning', 0.078) ('collecting', 0.077) ('obesity', 0.073) ('probe', 0.072) ('chairs', 0.06) ('adults', 0.058) ('message', 0.054) ('children', 0.051)]	**WALL·E**, tf-idf Filtered, len: 2797 [('allocation', 0.547) ('plant', 0.469) ('environmental', 0.228) ('cleaning', 0.178) ('collecting', 0.177) ('obesity', 0.166) ('probe', 0.165) ('chairs', 0.138) ('collects', 0.107) ('machines', 0.099) ('centuries', 0.098) ('load', 0.097) ('directive', 0.095) ('circuit', 0.091) ('robotic', 0.087) ('voices', 0.074) ('binocular', 0.073) ('corporation', 0.068) ('ecological', 0.066) ('items', 0.057)]
WALL·E, Topics, len: 100 84: 0.187 | language,0.014; knowledge,0.008; states,0.007; mental,0.006; problem,0.005; 86: 0.076 | theory,0.009; meaning,0.008; information,0.008; language,0.008; knowledge,0.007; 72: 0.066 | psychology,0.007; cognitive,0.007; depth,0.006; causation,0.006; behavior,0.006; 75: 0.055 | cognitive,0.013; play,0.010; example,0.006; cultural,0.005; movements,0.005; 94: 0.053 | language,0.011; causal,0.008; infants,0.006; psychology,0.006; human,0.005;	

Table 2. A 5-way comparison table for each of the three example movies. The top 20 terms are shown for each vector, along with its total number of unique words. The top five topics (out of 100) along with their top five most relevant terms are shown for each movie topic vector.

Immediately, it is clear that the topic model shown in the bottom left cell for each movie does not represent the movie well. The same topic 84 (with terms like language, knowledge, states, …) is the most activated topic for all movies, regardless of what that movie is about. *WALL-E* has no topic about robots in the top five, and *The Matrix* lacks anything about AI. Perhaps a better LDA parameterization is needed, but without that, these topics may represent MITECS or SEP well but fail to generalize to a completely unrelated set of movie reviews.

The other models are much better and have increasing levels of specificity with regards to the content of the reviews. The BoW model has a few frequent diagnostic words such as “creature” for *Frankenstein* or “space” for *WALL-E*, but the most frequent words are often noise that relate more to movie review terms than the movies or cognitive science. The tf-idf and BoW-F both do a good job at reducing this noise. There are slight differences between the two models: tf-idf is simply re-weighting terms based on relative document importance, but BoW-F is drawing upon

an external NCS set of movie reviews to filter out non-relevant terms. Still, there is a surprising amount of overlap in top terms between the two models. Tf-idf-F is gives us perhaps the *most* specific terms. As a reminder, tf-idf-F is just normal tf-idf applied on the BoW-F model. Many non-relevant terms are already removed in the BoW-F model, so the addition of tf-idf further accentuates the diagnostic words. Looking at the selected examples, we have some scarily specific words, such as “binocular” for *WALL-E* (referring to his binocular-shaped eyes), and “cave” for *The Matrix* (referring to the cave where the characters conceal themselves in the real world). As a whole, many of these terms are not describing concepts within cognitive science, but rather specific actions, settings, and plot elements of the movies. Glancing at *WALL-E’s* tf-idf-F list gives us a sense that the movie is about machines, allocation, a corporation, the environment, and cleaning. The BoW-F and tf-idf vectors seem to strike the best balance between extreme content specificity and broad concepts related to cognitive science.

Word Cloud Visualization

We turned these rather interesting movie characterizations into an interactive word cloud. The word cloud is hosted online with Microsoft Power BI, and can be viewed on page 1 of [this link](). We used the BoW-F representations to make the cloud because BoW-F both provides an easily-interpretable number (frequency) and a list of relevant terms with minimal noise. If no movie is selected for the word cloud, the BoW-F distribution is plotted summed across all documents. Figure 10 below shows a screenshot of this full-corpus CS movie review word cloud. “Matrix” is the most common term overall, and not because there are a lot of cognitive science films about linear algebra. Even a highly-filtered BoW model can not separate the context of an influential science fiction film and a mathematical construct, but a topic model trained on the review corpus (not the MITECS source) could maybe identify these contextual meanings.

Figure 10. A portion of the entire CS movie review BoW-F vector in the form of a word cloud.

This is simple and attractive method to see key terms in individual movies or the CSMI as a whole. This intuitive plot serves to condense reviews into a few dozen colorful terms, with the goal of making it easier to process movie features and decide what to what next.

Multidimensional Scaling

Methods

To make the multidimensional scaling (MDS) plots, we had to process the raw movie vectors into squareform distance matrices (244 x 244). The topic vectors in this analysis were trained using 40 topics, an alpha of 1/40, and a beta of 1/40. We used the topic vectors to take the Jensen-Shannon distance between each movie and every other movie. The same process was repeated for the BoW and tf-idf vectors except with cosine distance instead. These matrices were given to the MDS algorithm, which tries to reduce the dimensionality of a set of observations as much as possible while preserving the distance between the observations (Borg & Groenen, 2003). We used Sklearn's manifold MDS package (Buitinck et al., 2013) to run the algorithm with four SMACOF initializations, 300 max iterations, and a desired final dimensionality of two. The generated 2D coordinates for each movie were joined with other descriptive information about the movies such as their CSMI ratings, keyword tags, and short descriptions. We visualized the resulting scatter plots interactively with Tableau.

Discussion

The interactive MDS results can be viewed here. The first dashboard has a dropdown for switching between the three models for purposes of comparison. Each circle is a movie in 2D space, sized by its number of ratings in the CSMI. Selecting a keyword parameter highlights the movies with that keyword in yellow. Upon immediate inspection of this interactive plot, none of the data representations do a great job at clustering the movies, and this is verified by inspecting where movies with similar keyword labels appear in the space. The best cluster is seen with the "Robotics" keyword in the tf-idf MDS. Figure 11 displays these robotics movies in yellow, with tf-idf on the left and the topics models on the right.

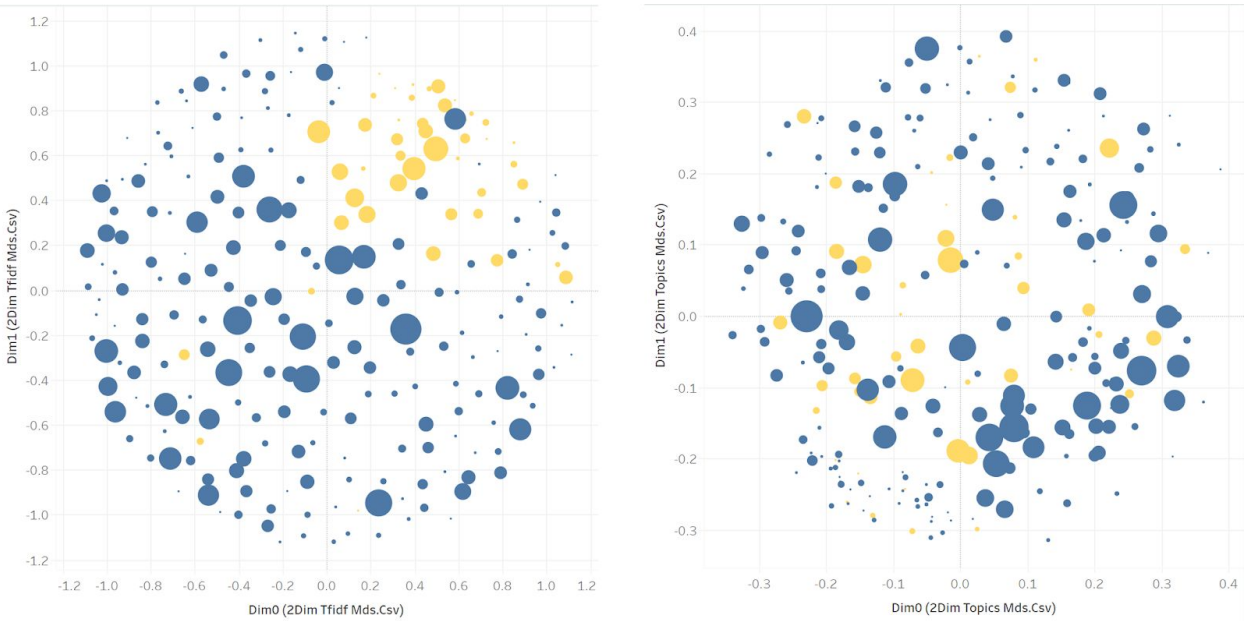

Figure 11. Comparison robot-related movie clustering in tf-idf MDS (left) and topics MDS (right). Movies containing the keyword "Robotics" are highlighted in yellow.

The tf-idf MDS does an especially good job at separating documents from each other, but neither it nor the BoW representation pick up on latent themes that are in common with the assigned keyword tags. In theory, an LDA topic model would do this, but because our topic model is trained on an entirely unrelated corpus, it has difficulty clustering the movies in a meaningful way. A practical extension of this result would be train a new topic model on the movie corpus alone.

Movie Similarity Network

Methods

Two movie-to-movie similarity networks were constructed using the 244 x 244 cosine distance matrix from the tf-idf vectors, the same input that we gave to the MDS algorithm. We chose tf-idf in particular because it produced the most reasonable MDS layout in two-dimensions, and is especially effective at spreading apart documents within the same corpus. In both networks, movies are nodes and the edges are the similarity between movies (1 - cosine distance). In the first network, edges between two nodes were only kept if their cosine similarity was >= 0.5. In the second network, each node kept exactly five edges to its five-closest movies, making it a nearly fully connected network. Both visualizations are interactive and were produced in Microsoft Power BI.

Discussion

The full interactive dashboard is located here. Page 1 is the interactive word cloud shown earlier, Page 2 is the first network with thresholded edges, and Page 3 is the second network with the 5-closest edges. Figure 12 is a static screenshot of the first network.

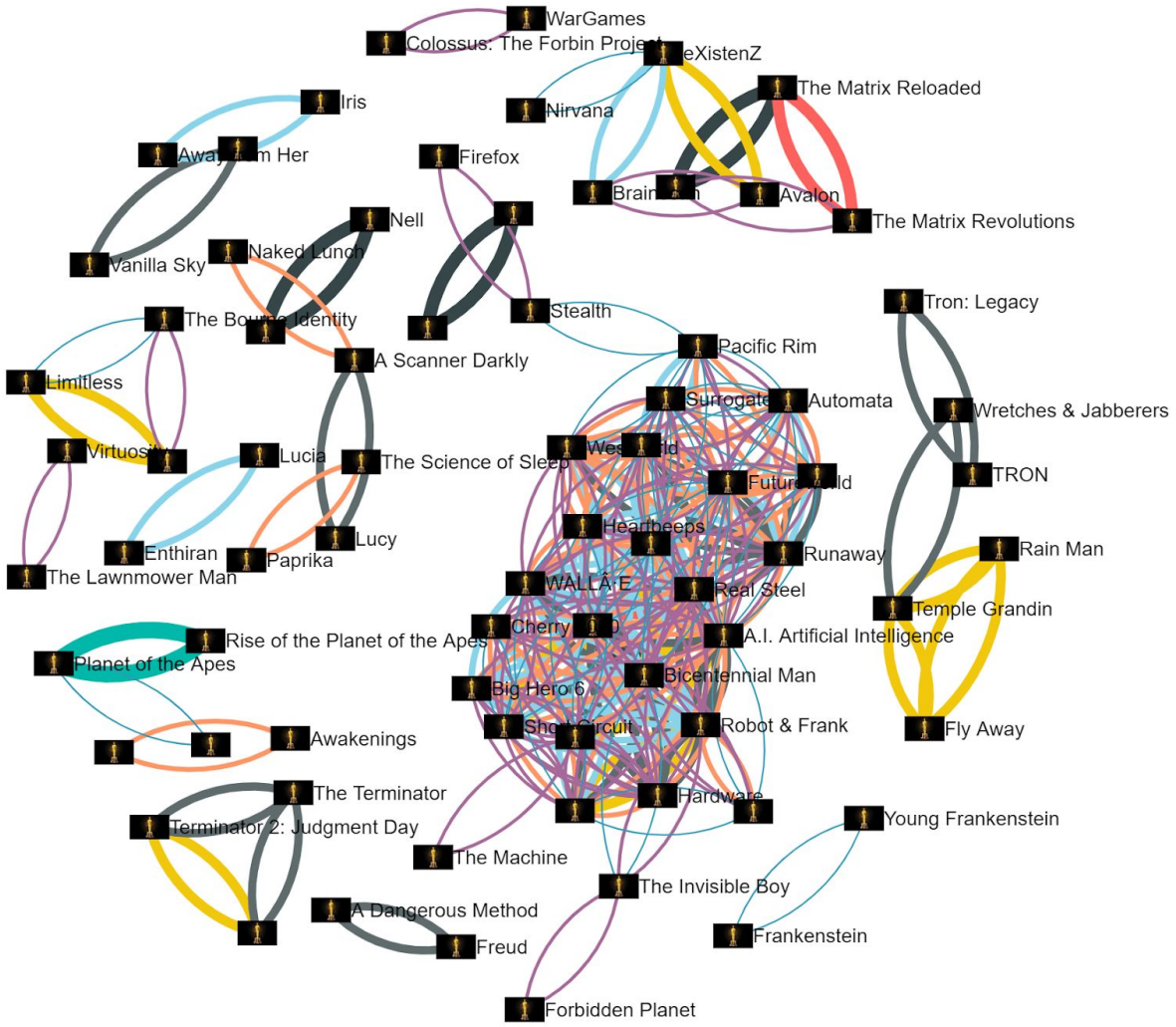

Figure 12. Movie similarity network with edge weights >= 0.5 (cosine similarity between the tf-idf vectors). The edge colors are categorical and depict edges with similar weights. Dark blue edges are the smallest (lest similar), and bright red edges are the largest (most similar).

As a result of the edge thresholding, we get many small connected components as the vast majority of movie-to-movie similarities are less than 0.5 and are filtered out. The largest connected component is that of the robotics/AI related movies. Outside of that, many of the smaller components connect sequels or trilogies together (i.e. *The Matrix, The Terminator,* or the *Tron* series). Where franchises are not involved, the other dyads and triads make sense (i.e.

Paprika paired with *The Science of Sleep*, and *The Stanford Prison Experiment* connecting to the other movie about the Stanford Prison Experiment, *The Experiment*).

This network replicates our MDS result from earlier on the same data set, where the only real latent cluster contained the robotics and AI movies. Perhaps when attempting to assign a category to a cognitive science movie, the first question to ask is if the movie is robotics or AI related. While experimenting with a lower edge threshold would lead to larger connected components, the current result still says a lot about the dataset. These clustering methods may not be robust enough to pick up on latent themes that either do not exist in the original reviews or are too faint and lost in the noise. It is intuitively much easier for a casual movie-reviewer to notice and discuss themes surrounding artificial intelligence and robotics than it is for the more esoteric themes of cognitive science. The same applies to film makers too. The CSMI is proof of a wide representation of cognitive science concepts within movies, but perhaps themes outside of robotics and AI are more nuanced and represented in a variety of ways not as salient to the average reviewer.

The second similarity network is harder to view holistically since it is almost fully connected. To examine it, we take advantage of the interactive nature of the graph and filter by source nodes, which excludes others except their five closest neighbors. We highlight two example movies (*Inception* and *The Imitation Game*) in Figure 13 below.

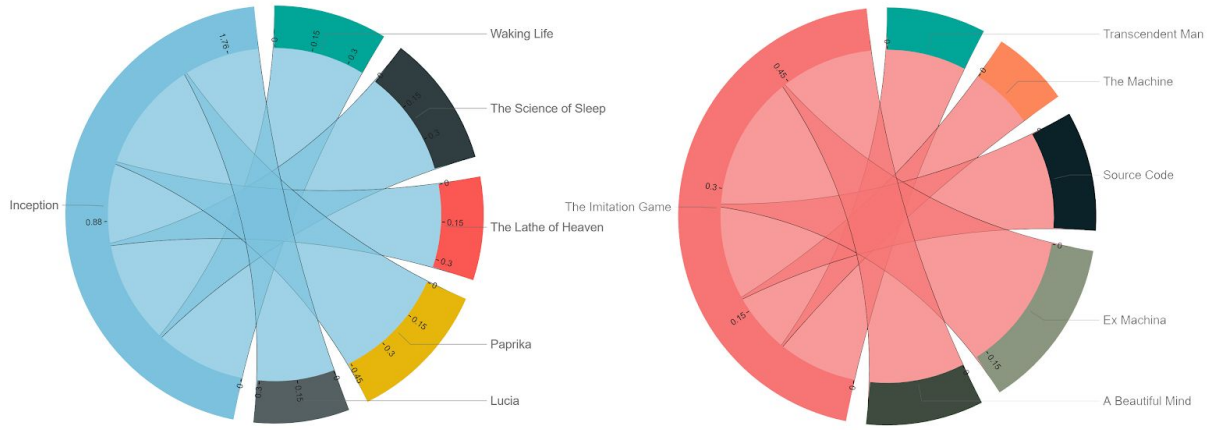

Figure 13. Movie similarity network with only the 5-closest edge weights. Most of the network is hidden except for two examples, *Inception* (left) and *The Imitation Game* (right).

The movies related to *Inception* consistently involve dreaming and surrealism, and the movies related to *The Imitation Game* touch on a wider range of content and styles. A downside of the reliance on terms without other context can be seen in the closest movie, *Ex Machina*. *The Imitation Game* is biographical and has very little to do with artificial intelligence, while *Ex Machina* is science fiction and deeply relates to artificial intelligence. The vectors are still similar to each other because concepts such as the Turing test are important in *Ex Machina*, and despite

the title, *The Imitation Game* has nothing to do with the Turing test aside from featuring its author, Alan Turing.

In general, the second network does a sensible job at presenting a list of similar movies to a query. Even if some items in the list do not hold up, holistically it should contain a couple good recommendations. Although it can't be used like the first network or the MDS to view the overall arrangement of the cognitive science movies, it adds utility at the small scale.

Related Encyclopedia Articles

Examining the articles from the source MITECS corpus that the movies are most related to gives us both a practical outcome for CSMI and a deeper understanding of the results in Analysis I. This is also an additional qualitative way to visualize the generalizability of the models.

We used the movie by article distance matrices (244 x 469) for tf-idf and topics to create Figure 14 below. It shows the top 10 most similar articles to three selected movies from the CSMI, based on the tf-idf model (left) and the topic model (right).

```
=== EX MACHINA | tfidf ===
0. (0.548) robotics-and-learning
1. (0.641) mobile-robots
2. (0.670) behavior-based-robotics
3. (0.810) planning
4. (0.817) situation-calculus
5. (0.826) intelligence
6. (0.832) artificial-life
7. (0.832) turing-alan-mathison
8. (0.843) machine-learning
9. (0.849) computation

=== INSIDE OUT | tfidf ===
0. (0.537) emotions
1. (0.794) ethnopsychology
2. (0.819) emotion-and-human-brain
3. (0.821) theory-of-mind
4. (0.825) emotion-and-animal-brain
5. (0.854) language-acquisition
6. (0.854) intersubjectivity
7. (0.859) psychoanalysis-history-of
8. (0.865) word-meaning-acquisition-of
9. (0.882) hume-david

=== INCEPTION | tfidf ===
0. (0.754) dreaming
1. (0.862) psychoanalysis-history-of
2. (0.909) connectionism-philosophical-issues
3. (0.918) human-navigation
4. (0.923) theory-of-mind
5. (0.926) computation-and-brain
6. (0.929) sleep
7. (0.930) language-acquisition
8. (0.931) comparative-psychology
9. (0.931) freud-sigmund
```

```
=== EX MACHINA | topics ===
0. (0.332) phonology
1. (0.407) generative-grammar
2. (0.410) language-acquisition
3. (0.411) taste
4. (0.414) knowledge-based-systems
5. (0.417) introspection
6. (0.419) schemata
7. (0.424) lexicon-neural-basis-of
8. (0.428) language-and-gender
9. (0.428) temporal-reasoning

=== INSIDE OUT | topics ===
0. (0.373) phonology
1. (0.423) taste
2. (0.434) logic
3. (0.436) schemata
4. (0.441) imitation
5. (0.443) generative-grammar
6. (0.453) language-acquisition
7. (0.457) cognitive-anthropology
8. (0.459) introspection
9. (0.462) knowledge-based-systems

=== INCEPTION | topics ===
0. (0.352) phonology
1. (0.413) generative-grammar
2. (0.422) taste
3. (0.425) language-acquisition
4. (0.427) functionalism
5. (0.433) lexicon-neural-basis-of
6. (0.438) knowledge-based-systems
7. (0.439) schemata
8. (0.442) introspection
9. (0.443) logical-form-origins-of
```

Figure 14. The closest 10 MITECS articles for *Ex Machina, Inside Out,* and *Inception*. The distances (shown in parenthesis) were calculated either with cosine for the tf-idf model (left) or Jensen-Shannon for the topic model (right).

The tf-idf model provides a set of related articles that may be useful for a visitor of CSMI. A lot of the articles are obvious matches, such as “Dreaming” and “Sleep” for *Inception,* but many others are not something the average viewer may know to look up on their own, or they may lack the technical knowledge to find the exact concepts they would like to read more about. In this way, more esoteric but still relevant article recommendations are useful such as “Ethnopsychology” for *Inside Out* or the article on Alan Turing for *Ex Machina*. On the other hand, the related articles based on the topic representations are non-diagnostic and quite useless for any relevant reading recommendations. The same set of articles appears for each movie, even though the movies span different subareas of cognition. There is very little overlap between what the tf-idf model recommends and what the topic model recommends. To explore this phenomenon further, we created the same figure for examining document relationship but this time looked at related MITECS articles to a given SEP article. Figure 15 below shows this result.

```
=== LINGUISTICS | tfidf ===
0. (0.675) generative-grammar
1. (0.683) formal-grammars
2. (0.713) linguistics-philosophical-issues
3. (0.733) computational-linguistics
4. (0.740) acquisition-formal-theories-of
5. (0.744) linguistic-universals-and-universal-grammar
6. (0.763) bloomfield-leonard
7. (0.766) cognitive-linguistics
8. (0.777) language-acquisition
9. (0.780) sapir-edward

=== COLOR | tfidf ===
0. (0.426) color-vision
1. (0.453) color-categorization
2. (0.529) color-neurophysiology-of
3. (0.729) language-and-thought
4. (0.847) gestalt-perception
5. (0.857) machine-vision
6. (0.859) psychophysics
7. (0.859) sensations
8. (0.883) linguistic-relativity-hypothesis
9. (0.897) surface-perception

=== COGNITIVE-SCIENCE | tfidf ===
0. (0.708) computational-neuroscience
1. (0.754) computation-and-brain
2. (0.774) mental-representation
3. (0.775) cognitive-modeling-symbolic
4. (0.791) connectionism-philosophical-issues
5. (0.801) rules-and-representations
6. (0.804) computation
7. (0.805) imagery
8. (0.814) modeling-neuropsychological-deficits
9. (0.819) natural-language-processing
```

```
=== LINGUISTICS | topics ===
0. (0.279) phonology
1. (0.306) generative-grammar
2. (0.332) lexicon-neural-basis-of
3. (0.335) formal-grammars
4. (0.352) nativism
5. (0.366) language-acquisition
6. (0.372) codeswitching
7. (0.389) theory-of-mind
8. (0.394) logical-form-origins-of
9. (0.398) language-and-gender

=== COLOR | topics ===
0. (0.150) logical-form-origins-of
1. (0.190) sensations
2. (0.211) color-vision
3. (0.236) qualia
4. (0.239) brentano-franz
5. (0.239) folk-psychology
6. (0.240) control-theory
7. (0.240) eliminative-materialism
8. (0.240) reductionism
9. (0.266) sense-and-reference

=== COGNITIVE-SCIENCE | topics ===
0. (0.383) theory-of-mind
1. (0.394) logical-form-origins-of
2. (0.403) simulation-vs.-theory-theory
3. (0.412) descartes-rené
4. (0.419) knowledge-based-systems
5. (0.423) problem-solving
6. (0.434) introspection
7. (0.434) schemata
8. (0.445) sense-and-reference
9. (0.448) metarepresentation
```

Figure 15. The closest 10 MITECS articles for the SEP articles on “Linguistics”, “Color”, and “Cognitive Science”. The distances (shown in parenthesis) were calculated either with cosine for the tf-idf model (left) or Jensen-Shannon for the topic model (right).

We see the tf-idf model also works for the SEP-to-MITECS article recommendation, which is not surprising given the results in Analysis I. We also see that the topic model recommendations are more accurate now, and it returns more articles in common with what the tf-idf produces. Although the topic model is performing better on this nearby context, it is still not perfect and has noise such as the recommendation of “Theory of Mind” for “Linguistics” or “Origins of Logical Form” for “Color”.

We show the topic model's failure to generalize in one last way with the bar plots in Figure 16. They count the number of times each MITECS article appears in the top 10 lists across all movies (top) or SEP articles (bottom). For the movie reviews, the topic model distances produce the same 8-10 closest MITECS articles for over half of the movies. The SEP articles have a more diverse set of closest MITECS articles, indicated by the uniform distribution rather than the extreme right skew. These graphs do not indicate anything about the quality or accuracy of these related articles, but the variety of recommendations is some indication of performance assuming the set of target documents is also varied in subject area. These results intuitively show that the LDA topic model can perform better on an unseen corpus if its closer in context to what it was initially trained on.

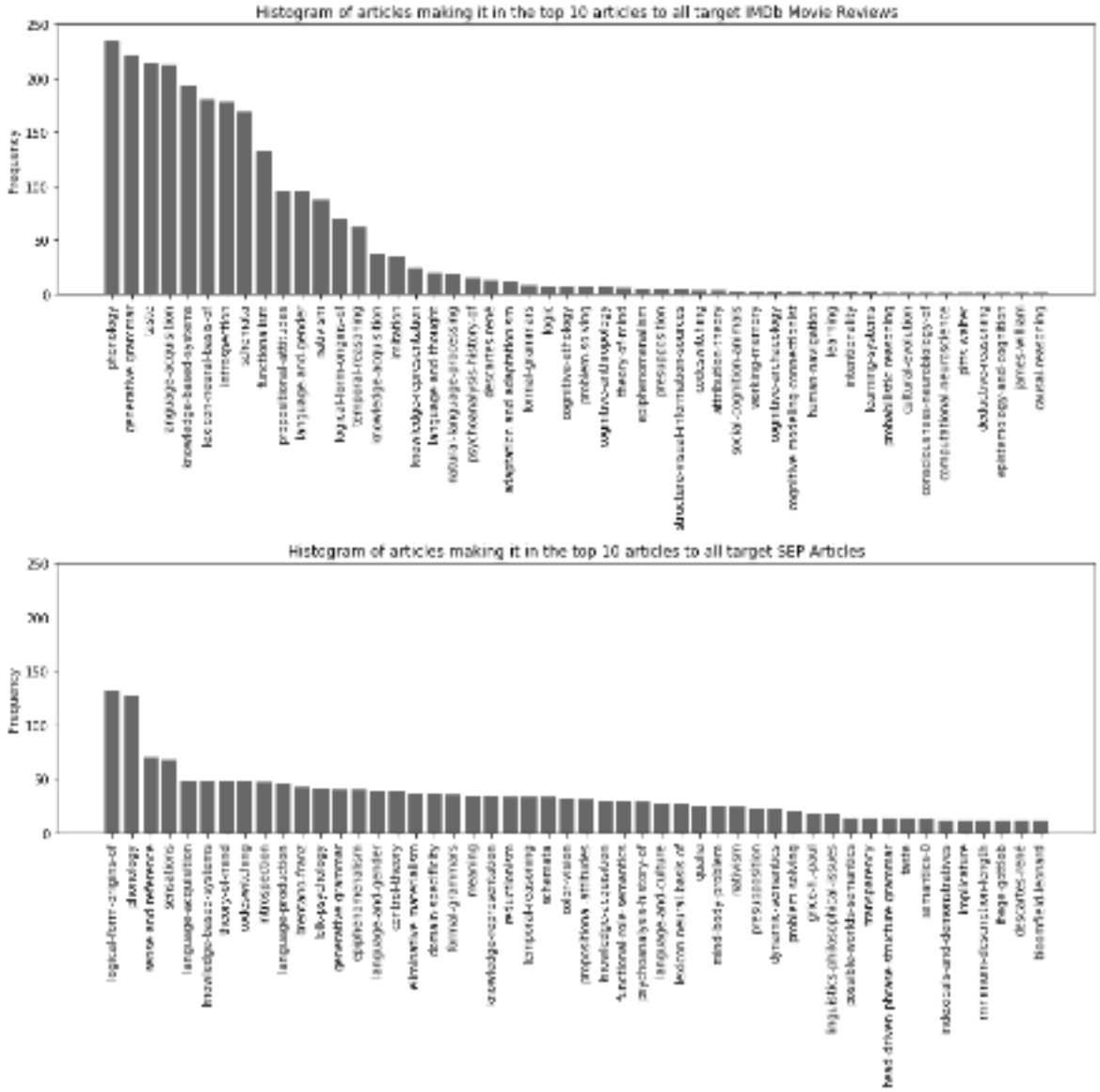

Figure 16. Frequencies of the MITECS articles that appear in the top 10 similarity lists to either target movie reviews (top) or target SEP articles (bottom).

As a final analysis, we show in Figure 17 how the recommendation system can be bi-directional. Below is a list of the 10 most similar movies to a selection of three MITECS articles. In addition to the utility of discovering movies related to an article one just read, these highly domain-specific article names could be used as category labels for movie aggregators like the CSMI. For example, instead of browsing an aggregator for the top rated drama films, interested viewers could get a list of movies most relevant to an interest area such as language acquisition.

```
=== EVOLUTIONARY-PSYCHOLOGY | tfidf ===
0. (0.891) 2001: A Space Odyssey
1. (0.894) Transcendence
2. (0.902) Transcendent Man
3. (0.903) Blade Runner
4. (0.913) Idiocracy
5. (0.916) Ex Machina
6. (0.917) Mr. Nobody
7. (0.917) Avatar
8. (0.918) Lucy
9. (0.922) Paycheck

=== LANGUAGE-ACQUISITION | tfidf ===
0. (0.716) The City of Lost Children
1. (0.812) The Wild Child
2. (0.848) Arrival
3. (0.854) Inside Out
4. (0.854) Code 46
5. (0.875) Ender's Game
6. (0.879) Jack
7. (0.880) Blade Runner
8. (0.882) Mr. Nobody
9. (0.882) The Science of Sleep

=== EMOTIONS | tfidf ===
0. (0.537) Inside Out
1. (0.838) Her
2. (0.870) Equilibrium
3. (0.875) A.I. Artificial Intelligence
4. (0.882) Blade Runner
5. (0.885) Transcendence
6. (0.889) Eternal Sunshine of the Spotless Mind
7. (0.892) Synecdoche, New York
8. (0.901) Terminator 3: Rise of the Machines
9. (0.902) The Music Never Stopped
```

Figure 17. MITECS articles to movie recommendations with the tf-idf model and cosine distance.

Conclusion

Through our first analysis, we developed a methodology for measuring how well the knowledge base of a source corpus (MITECS articles) generalizes to a near and far context. We applied this methodology to the domain of cognition, with the hopes of being able to successfully use academic literature to represent movies that explore scientific themes. We compared the results of two vector space models, BoW and tf-idf, with an LDA topic model and showed that the VSMs were able to generalize to foreign corpora, even one as removed from encyclopedia articles as IMDb movie reviews. The LDA model could separate the domain-relevant articles from the non domain-relevant articles in a the near context of SEP, but not the far context of IMDb reviews. This demonstrated the downsides of fold-in query sampling for topic modelling unseen

documents, and the low fidelity of this technique may not be worth the improved speed and simplicity. Future work can explore other methods for topic modelling across corpora, including corpus merging or directly comparing two independently trained models. Topic modelling aside, the BoW and tf-idf models performed well on the target corpora, which leads to a couple useful applications: first, they can help us use analogy to better understand complex datasets, and second, we could expand the size of a small dataset by pulling features from a nearby context.

Our second analysis is one such application of this more general problem. The CSMI has a small feature set, but by pulling in other corpora like MITECS and IMDb reviews, we successfully characterized the movies in new ways. The tf-idf, BoW-F, and tf-idf-F vector inspections, word clouds, movie similarity networks, and recommended related encyclopedia articles are all functional applications of our models that display the topology of cognitive science movies. The LDA topic model again did not have success in this application, and it is important to remember that the vector space models have downfalls in their simplicity. Even though we trained all the models on a distant source corpus (MITECS), the only part of MITECS that the movie review representations incorporate is the shared vocabulary, which is not necessarily tightly bound around cognitive science terms. The topic model would have been a more ideal representation of the documents in many ways, as it adds an extra semantic layer going beyond term frequencies. But the LDA model was trained on a separate corpus, and the applied topics on the unseen documents have no guarantee to be actually representative of the underlying latent structure of those documents. Both analyses highlight the issues with the topic space being applied in this way, and the vector space models perform well because they are more localized to the document set they are applied across.

We have made good progress with regards to goal (a), and to further answer the question we will conduct additional experimentation using different corpora, topic models, distance measures, and methods for determining statistically significant differences between distributions in a low-contrast distance space. Results for goal (b) and Analysis II are readily available and we believe our movie characterizations are practical for movie search tasks, even if they are not completely domain-relevant. In the future, a powerful approach for comparing two distant corpora could also lead us to machine learning algorithms that automatically predict and classify a movie's relevance to a domain or scientific accuracy.

Acknowledgements

I would like to thank Brad Rogers and Dr. Peter Todd for their extensive mentorship on this manuscript and project. Additionally, I must thank my visualization class team, Ana Hakhamaneshi, Evan Beall, Erelyn Apolinar, and Saurabh Shukla, for their assistance in all the interactive visualizations. Finally, many thanks to Jaimie Murdock and Dr. Colin Allen for their assistance with topic modelling methods and fetching data from the Stanford Encyclopedia of Philosophy.

References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. In *Database Theory — ICDT 2001* (pp. 420–434). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-44503-X_27

Allen, C., & Murdock, J. (2016). Hypershelf: A Multidimensional Browser for Exploring Content Across the Library. Retrieved from https://scholarworks.iu.edu/dspace/handle/2022/20975

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When Is "Nearest Neighbor" Meaningful? In *Database Theory — ICDT'99* (pp. 217–235). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-49257-7_15

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research: JMLR*, *3*(Jan), 993–1022. Retrieved from http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Borg, I., & Groenen, P. (2003). Modern Multidimensional Scaling: Theory and Applications. *Journal of Educational Measurement*, *40*(3), 277–280. https://doi.org/10.1111/j.1745-3984.2003.tb01108.x

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., … Varoquaux, G. (2013). *API design for machine learning software: experiences from the scikit-learn project*. *arXiv [cs.LG]*. Retrieved from http://arxiv.org/abs/1309.0238

Crain, S. P., Zhou, K., Yang, S.-H., & Zha, H. (2012). Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 129–161). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4614-3223-4_5

Glassy, M. C. (2005). *The Biology of Science Fiction Cinema*. McFarland. Retrieved from https://market.android.com/details?id=book--raJCgAAQBAJ

Greene, D., O'Callaghan, D., & Cunningham, P. (2014). How Many Topics? Stability Analysis for Topic Models. In *Machine Learning and Knowledge Discovery in Databases* (pp. 498–513). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-44848-9_32

Hall, M. A. (2004). Romancing the Stones: Archaeology in Popular Cinema. *European Journal of Archaeology*, *7*(2), 159–176. https://doi.org/10.1177/1461957104053713

Hofmann, T. (2017). Probabilistic Latent Semantic Indexing. *SIGIR Forum*, *51*(2), 211–218. https://doi.org/10.1145/3130348.3130370

Honnibal, M., & Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To Appear*.

Jones, K. S. (1972). A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Jones, M., Gruenfelder, T., & Recchia, G. (2011). In defense of spatial models of lexical semantics. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33). cloudfront.escholarship.org. Retrieved from https://cloudfront.escholarship.org/dist/prd/content/qt1d28323f/qt1d28323f.pdf

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics*, *22*(1), 79–86. Retrieved from http://www.jstor.org/stable/2236703

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory / Professional Technical Group on Information Theory*, *37*(1), 145–151. https://doi.org/10.1109/18.61115

Litch, M. M. (2010). *Philosophy through film*. Routledge. Retrieved from

https://content.taylorfrancis.com/books/download?dac=C2009-0-04765-9&isbn=9781135188252&format=googlePreviewPdf

Lu, Y., Mei, Q., & Zhai, C. (2011). Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval*, *14*(2), 178–203. https://doi.org/10.1007/s10791-010-9141-9

Motz, B. (2013). Cognitive science in popular film: the Cognitive Science Movie Index. *Trends in Cognitive Sciences*, *17*(10), 483–485. https://doi.org/10.1016/j.tics.2013.08.002

Murdock, J. (2019). *Topic Modeling the Reading and Writing Behaviors of Information Forager*. Indiana University.

Řehůřek, R., & Sojka, P. (2011). Gensim—statistical semantics in python. *Statistical Semantics; Gensim; Python; LDA; SVD*. Retrieved from https://www.fi.muni.cz/usr/sojka/posters/rehurek-sojka-scipy2011.pdf

Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220

Schofield, A., & Mimno, D. (2016). Comparing Apples to Apple: The Effects of Stemmers on Topic Models. *Transactions of the Association for Computational Linguistics*, *4*, 287–300. https://doi.org/10.1162/tacl_a_00099

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of Latent Semantic Analysis*, *427*(7), 424–440. Retrieved from https://books.google.com/books?hl=en&lr=&id=JbzCzPvzpmQC&oi=fnd&pg=PA427&dq=manual+on+latent+semantic+analysis+griffiths&ots=aNK4G0Q0NI&sig=h_NksorkpVbimZzOOAJ7kd6yKlw

Stone, B., Dennis, S., & Kwantes, P. J. (2011). Comparing methods for single paragraph similarity analysis. *Topics in Cognitive Science*, *3*(1), 92–122.

https://doi.org/10.1111/j.1756-8765.2010.01108.x

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188. https://doi.org/10.1613/jair.2934

Venkatesh, J. (2010). *Pairwise Document Similarity Using an Incremental Approach to TF-IDF*. Retrieved from https://market.android.com/details?id=book-LTKgnQAACAAJ

Wei, X., & Croft, W. B. (2006). LDA-based Document Models for Ad-hoc Retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 178–185). New York, NY, USA: ACM. https://doi.org/10.1145/1148170.1148204

Wilson, R. A., & Keil, F. C. (2001). *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press. Retrieved from https://market.android.com/details?id=book--wt1aZrGXLYC

Xie, P., & Xing, E. P. (2013). *Integrating Document Clustering and Topic Modeling*. *arXiv [cs.LG]*. Retrieved from http://arxiv.org/abs/1309.6874

Yi, X., & Allan, J. (2009). A Comparative Study of Utilizing Topic Models for Information Retrieval. In *Advances in Information Retrieval* (pp. 29–41). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-00958-7_6

Zalta, E. N., Nodelman, U., Allen, C., & Perry, J. (2003). Stanford encyclopedia of philosophy. Stanford University. The Metaphysics Research Lab. Retrieved from http://ruccs.rutgers.edu/images/personal-zenon-pylyshyn/class-info/Consciousness_2014/StanfordEncyclopedia/consciousness-temporal_sc.pdf

Appendix A

17th Century Philosophy
18th Century Philosophy
19th Century Philosophy
20th Century Philosophy
Aesthetics
African and African-American Philosophy
Ancient Philosophy
Ancient Philosophy [Aristotle]
Ancient Philosophy [Plato]
Arabic and Islamic Philosophy
Chinese Philosophy
Epistemology
Epistemology [Formal]
Ethics [Applied-Biomedical]
Ethics [Applied-Information Technology]
Ethics [Applied]
Ethics [History]
Ethics [Metaethics]
Ethics [Normative]
Feminism
Guest Area
Indian and Tibetan Philosophy
Japanese Philosophy
Judaic Philosophy
Kant
Latin American and Iberian Philosophy
Logic [Computation and Agency]
Logic [History - Contemporary]
Logic [Language]
Logic [Mathematical]
Logic [Philosophical]
Logic [Philosophy of]
Logic [and Language]
Medieval Philosophy
Metaphysics
Philosophy
Philosophy of Action
Philosophy of Biology
Philosophy of Cognitive Science
Philosophy of Language
Philosophy of Law
Philosophy of Mathematics
Philosophy of Mind
Philosophy of Physics [Quantum Mechanics]
Philosophy of Physics [Spacetime]
Philosophy of Religion
Philosophy of Science
Philosophy of Social Science
Renaissance and 16th Century Philosophy
Social and Political Philosophy
Women in the History of Philosophy

Figure 18: Listing of all SEP high-level subject areas. Cognitive science articles were selected from the bolded subject areas.