Humanities Digital Archives Project

L. Over, S. Gupta, L. Sooter

April 21, 2020

Contents

1 Introduction

The Humanities Data Librarian for the University Library at the University of Pittsburgh, Tyrica Terry Kapral, provided three data sets for an analysis of the digital collections in the Humanities department.

This was an exploratory, unsupervised learning project. The high-level goal was to investigate which topics are present within the humanities digital collection, and how those topics vary over time. Specifically, Mrs. Terry Kapral was interested in answers to the following questions about the data:

1. What are the latent topics across the digital items?
2. What items are related by topic?
3. How do topics change over time with respect to the time period covered by the items within each topic?
4. Are there any problems with the data?

These questions were answered through data exploration, including word embeddings and t-SNE plots, and topic modeling, using the unsupervised learning algorithm, Latent Dirichlet Allocation (LDA). Data exploration revealed problems in the data, some of which were mitigated. The final LDA model revealed 19 latent topics from the titles and abstracts in the metadata for the 124,517 digitized items that had a title.

2 Data

The data sets included the following metadata for both published and unpublished archival items:

- *Islandora_items_metadata.csv*: All digital item metadata at the item level (124,539 digitized items)
- *Finding_Aids_metadata.csv*: All finding aid metadata at the collection level (2210 finding aids)
- *Digital_Collections_List.csv*: All site and collection memberships at the item level (specifies to which site and collection an item belongs)

2.1 Preprocessing Part 1a: Basic Data Cleaning

The three files were imported and the contents explored. Then three new data sets were created with only the fields relevant to the project goals.

There were a lot of missing data in the *Islandora_items_metadata.csv* file, which was the main file to be used in the analysis. Almost all of the items (99.99%) included a 'Title', but only 40.78% included an 'Abstract.' The abstract field is the most relevant to the analysis.

Initially, the plan was to connect the item level metadata with the collection level finding aids metadata to add the higher level information to the items that were missing the Abstract field. In the end, however, the finding aids metadata was not added to the items metadata because the additional data would be too broad in scope as each item could belong to more than one collection.

Therefore, the data used in the topic modeling analysis were the Title and Abstract fields for the digital items, *Islandora_items_metadata.csv*.

2.2 Preprocessing Part 1b: Model Preparation

Further data preprocessing involved preparing the data for two different models:

- Word embeddings, using the two-layer neural network 'word2vec' function, and a t-SNE plot of the word vectors output from 'word2vec'
- Topic modeling with Latent Dirichlet Allocation (LDA)

For both models, the 'Abstract' column was cleaned to fix obvious spelling errors, to remove strange characters, and to remove web site URLs. It was possible to fix some spelling errors and strange characters. However, any remaining spelling or character errors should be addressed and descriptive titles or abstracts should be added to the items that are lacking that information.

For the word embeddings and t-SNE plot, the data was further cleaned at the record level. Each abstract was cleaned, separately, to expand contractions, remove stop words (common words such as 'is' and 'are'), punctuation, and special characters, and to lemmatize the text. Lemmatization involves converting a word to a form found in the dictionary. For example, 'sitting,' 'sits,' 'sat,' and 'sit' would all be converted to 'sit.' All entries without abstracts were then removed.

For the LDA topic modeling, a new field, Ttle_Abs, was created with the contents of the Title and Abstract fields concatenated together. LDA expects a text corpus, not separate blocks of text from individual records, so the new field was compiled from each record to form the corpus.

This corpus was gathered into a list and then cleaned to remove stop words, punctuation, and special characters. Words that often appear together were placed in groups of two or three, called bigrams and trigrams. For example united states would be converted to united_states. Finally, the words in the list were lemmatized.

3 Method

3.1 Word Embeddings and t-SNE Plot

Only those records that had data in the Abstract column were used for the word embeddings and t-SNE plot. First, a list of lists of words from each abstract was created. Then, word vectors were created using the two-layer neural network word2vec function. The word2vec function captures the context of a word in a

document, i.e., the semantic and syntactic similarity among words, and assigns each word a vector that represents its location in relation to other words by their vectors. These vectors were then used to create t-SNE models with plots. The t-SNE plot appears in Appendix C (section 10).

The t-SNE models could be used to find words that were found frequently with other words. Table 1 below shows words that are often found with the word 'steel.' The word 'jone' represents the Jones and Laughlin Steel Company. The table also shows the probability that the word is associated with 'steel.'

Table 1: Words Often Found with the Word 'Steel'

Word	Probability
jone	0.6476
corporation	0.6083
furnace	0.5624
mill	0.5594
produce	0.5573
work	0.5291
iron	0.4713

3.2 Latent Dirichlet Allocation (LDA)

LDA is a probabilistic approach to discovering latent topics in documents. It is based on the probability of words with respect to topics and not just word frequencies as with other topic modeling algorithms. The LDA model is a statistical algorithm for detecting context, identifying similarities, and extracting meaning from a set of text documents. The algorithm works on the basis that documents are probability distributions over latent topics and that topics are probability distributions over words. The LDA algorithm represents the corpus as a mixture of topics and generates a set of words and their probabilities for each topic, i.e., the probability the word is representative of that topic.

Initially, data was modeled using both the Gensim and Mallet implementations of LDA using Num_Topics=20. The coherence score, the degree of semantic similarity between high scoring words in a topic, was used to evaluate the model. Mallet, a Javabased package for statistical natural language processing, produced better results than Gensim with a coherence score of 0.6415 verses 0.5343.

Four strategies were implemented using Mallet, each with several different values for the num_topics hyperparameter.

1. Strategy 1 was run using only the records that contained an abstract (50,789 items) and with 7 different values for num_topics.

2. Strategy 2 was run on a new field created with the title and abstract concatenated together and where the title was given punctuation (".") to

be treated like a sentence. This model was run using 20 different values for num_topics on only records that contained an abstract (50,789 items). After concatenating the title and abstract, records without an abstract were removed.

3. Strategy 3 was run using the concatenated title and abstract fields on all records that had a title (124,517 items). This model was run using 20 different values for num_topics. For this model, the following words were added to the list of stop words to be removed: collection, view, man, woman, north, south, east, west, northeast, northwest, southeast, southwest so the substance of the data and not the gender or direction was considered.

4. Strategy 4 was run using the concatenated title and abstract fields on all records that had a title (124,517 items). This model was run using 14 different values for num_topics. For this model, the words added to the stop words included 'collection', 'view', 'amp', 'leave', 'hold', 'scene', 'show', 'panel', 'sit', 'stand', 'part', 'parts', 'day', 'image', 'photo', 'photograph'.

In Strategy 3, removing gender and direction words directed the topic designation incorrectly. Many short titles included Pittsburgh area neighborhoods such as Perry North and Perry South. When north and south were removed, LDA assigned many of these records to a topic about school (school's are named after famous people such as Commodore Perry) when they were better assigned to topics about geography or neighborhoods.

In Strategy 4, the words that were removed mostly described the membership of, the location of, or a common activity expressed in the item, which resulted in many ambiguous topic assignments.

4 Analysis

The best model for all strategies yielded a coherence score between 0.59 and 0.66. The decision for choosing the final strategy was based on both the coherence score and on how well the items were clustered from a human perspective. Strategy four produced the best separation of topics, which was determined by skimming the titles and abstracts in each topic. The best model in Strategy 4 had one of the highest coherence scores as well.

Table 2 below shows the coherence scores for the first three strategies.

Table 3 below shows the coherence scores for five of the models from Strategy 4.

Figure 1 below shows the coherence plot for Strategy 4 with num_topics 2 through 32 at increments of 4.

Figure 2 below shows the coherence plot for Strategy 4 with num_topics 18 through 25 at an increment of 1.

An interactive graph, created using pyLDAvis, includes two parts: *Intertopic Distance Map* and *Top-30 Most Salient Terms* bar chart.

Table 2: Coherence Scores for Strategies 1-3

Num_topics	Strategy 1	Strategy 2	Strategy 3
8	0.6315	0.5880	0.6094
14	0.6511	0.5541	0.6076
20	0.6245	0.6075	0.6483
26	0.6361	0.5864	0.6187
32	0.6172	0.5650	0.6394

Table 3: Coherence Scores for Strategies 4

Num_topics	10	14	18	19	20
Coherence	0.6163	0.6496	0.6552	0.6656	0.6497

Figure 1: Coherence Plot Num_Topics 2 to 32 by 4

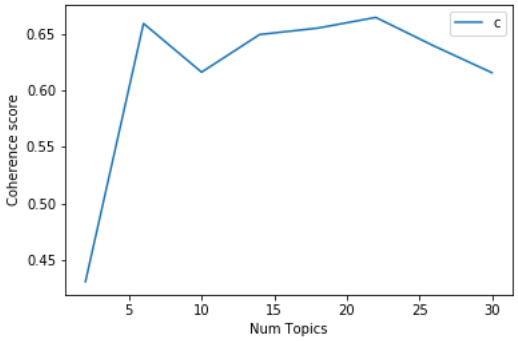
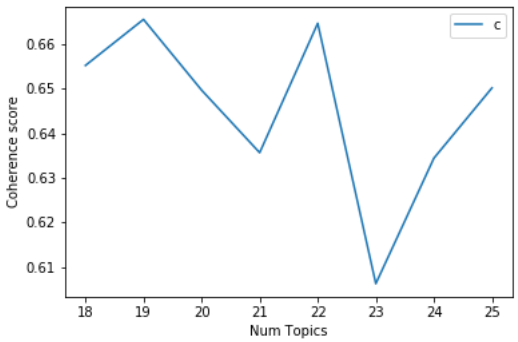


Figure 2: Coherence Plot Num_Topics 18 to 25 by 1



The *Intertopic Distance Map* represents topics as circles and each topic's prevalence in the document as the area of its circle. In the case of the digital humanities items, the size of the circle represents the number of items assigned to the respective topic with respect to the whole list. The location of circles on the grid is determined, by pyLDAvis, using the Jensen-Shannon divergence, which is a measure of the similarity between the probability distributions of the words found by the Latent Dirichlet Allocation (LDA) algorithm. Then, pyLDAvis employs multidimensional scaling to project these inter-topic distances onto a two-dimensional space.

The *Top-30 Most Salient Terms* bar chart shows the top-30 most relevant terms for each topic and the estimated frequency of each term within the whole document as well as within the given topic. Hovering over a word in this bar chart shows the topics that include that word in its top-30 list. While hovering, the circles of the topics that include the word appear in the *Intertopic Distance Map*, while the circles of topics that don't include the word, disappear.

The *Intertopic Distance Map* shows that most topics are clearly separate from the others; there is some overlap of topics but not a significant amount given the number of topics. This graph is located in the file *lda-mallet_vis.html*. ***Note:*** The Mallet model had to be converted to the Gensim format in order to use pyLDAvis. The topic numbers for Mallet did not match the Gensim topic numbers. New fields were added to the output files to map the Mallet topic numbers to Gensim for use with the graphic.

The final model was from Strategy 4 with Num_Topics=19 and a coherence score of 0.6656.

5 Results

Dominant topics were assigned to each item in the digital items list based on the probabilities of the words located in the title and abstract. That is, if the sum of the probabilities of the words associated with one topic was greater than the sum of the probabilities of the words associated with another topic, the topic with the highest sum of probabilities became the dominant topic for the given item. The file with dominant topics assigned was output to a CSV file named *items_dominant_topics.csv*.

The total number of items assigned to each topic as well as the proportion of items within each topic were written to the file *items_topics_summ.csv*. This file also contains the top-10 keywords. Topic names were added manually to this file after reviewing the keywords and titles and abstracts in *items_dominant_topics.csv*.

5.1 Latent Topics Across the Digital Items

Topic names were assigned manually by reviewing the *Top-30 Most Salient Terms* bar chart and the titles and abstracts. The top-10 words and any of the top-30 words that appeared only (or mostly only) in the topic were considered.

Many topics had 2 or more clusters within the single topic. Topics are listed in Appendix A (section 8).

5.2 Items Related by Topic

The file *items_dominant_topics.csv* contains all items used in the model with fields from the original *Islandora_items_metadata.csv* file, including but not limited to the item identifier, title, sort date, and abstract. It also includes new fields created from the model that relate to topic assignment. New topic-related fields include the following:

- Dominant_Topics is the topic number assigned from the optimal Mallet model.
- Perc_Contribution is the topic's percentage contribution for each item's Title and Abstract combination.
- Topic_Keywords is the top-10 keywords for the topic.
- Dominant_Topic_Vis is the dominant topic number that corresponds to the interactive pyLDAvis graph: *ldamallet_vis.html*.

5.3 Time Series Plots for Each Topic

The items data included the field 'Sort Date,' which is the date of the event depicted in or by the item. The file *topics_date_cnt.csv* includes a count of items for each date.

Interactive time series plots were created from these counts using plotly. There is one graph per topic and each includes a slider that allows users to display smaller windows of time. These plots are listed in Appendix B (section 9).

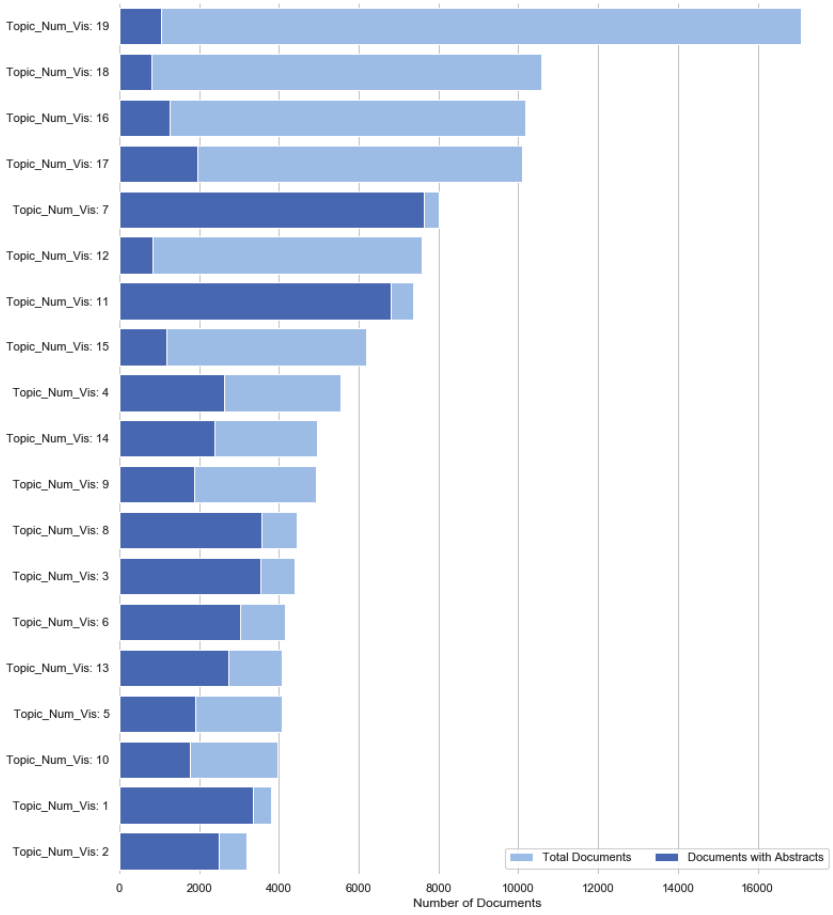
5.4 Problems with the Data

There are several problems with the data that, if corrected, would improve the topic modeling results.

- There are a lot of items with missing abstracts.
- There are many spelling errors and strange characters that were probably punctuation or special characters that Excel did not translate properly from the metadata.
- Not all titles are descriptive. Many records have very short, non-descriptive titles, many of which contain codes or cryptic words rather than words that can be lemmatized. Items with such titles and no abstract may have negatively influenced the model. There were enough good titles to warrant their inclusion.

Figure 3 below illustrates the distribution of items per topic and the proportion of documents with abstracts per topic. The light blue bar shows the total number of items for the given topic. The dark blue bar shows the number of documents with abstracts for the given topic.

Figure 3: Distribution of Items by Topic (Total and with Abstracts)



The topic 'Topic_Num_Vis: 19', 'University of Pittsburgh News and Press Releases; Pittsburgh Neighborhoods,' has very few abstracts. This is an example of how the short titles, many of which included the phrase 'Pitt News,' were clustered generally by 'Pitt News' instead of by the topic of the news item.

6 Conclusion

Several strategies of data preprocessing and models with varying values of Num_Topics resulted in a model with a coherence score of 0.6656 and 19 topics.

These results could be improved if the problems with the data were addressed. However, the questions posed in this study were answered using the data and the graphs created from the optimal model.

The *Distribution of Items by Topic* bar chart, *topics.png*, shows the number of documents per topic and reveals where there is missing data (abstracts).

The *Intertopic Distance Map*, *lda-mallet_vis.html*, shows that the topics are separate from one another with only some overlap. The *Top-30 Most Salient Terms* bar chart in the same file provides a condensed way to view the terms that are important to each topic.

The interactive time series graphs for each topic show the distribution of items within topics with respect to the time periods covered within the different topics.

This project was interesting because it was a real-word problem with a local data set. The results may be implemented by the library to improve table of contents and search mechanisms with the goal of connecting users with the digital items. If the problems with the data are addressed, the results from improved models could be used to create a supervised learning model to classify new library items.

7 References

Bhatt, Bhavesh. "Latent Dirichlet Allocation (LDA) for Topic Modeling." (March 2020). https://www.youtube.com/watch?v=Cpt97BpI-t4.

Gensim. "Latent Dirichlet Allocation via Mallet." (March 2020). https://radimrehurek.com/gensim/models/wrappers/ldamallet.html.

Bl.ocks. "Visualizing topics as distributions over words." (March 2020). http://bl.ocks.org/AlessandraSozzi/raw/ce1ace56e4aed6f2d614ae2243aab5a5/.

Kaggle. "Visualizing Word Vectors with t-SNE." (March 2020). https://www.kaggle.com/jeffd23/visualizing-word-vectors-with-t-sne.

Machine Learning Plus. "Topic Modeling with Gensim (Python)." (March 2020). https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python.

Plotly. "Horizontal Bar Charts in Python." (March 2020). https://plotly.com/python/horizontal-bar-charts/.

Plotly. "Range Slider and Selector in Python." (March 2020). https://plotly.com/python/range-slider/.

Sullivan, Scott. "LDA Algorithm Description." (March 2020). https://www.youtube.com/watch?v=DWJYZq_fQ2A&list=TLPQMTIwNDIwMjDGp7F13334kw&index=1.

8 Appendix A: Topics

Table 4 lists the topics discovered from the LDA model using the titles and abstracts of the digital humanities items. The Topic_Num_Vis column maps each topic to the topic numbers that appear in the *Intertopic Distance Map* and *Top-30 Most Salient Terms* bar chart created with pyLDAvis.

Table 4: Topics Mapped to Topic_Num_Vis Number

Topic_Num_Vis	**Topic Name(s)**
1	Religious Structures, Practices, History (Christianity); Gothic Architecture; Windows and Porches: Views of People and Stores
2	Buildings: History, Construction, Restoration, and Renovation
3	Primary and Secondary Industry: People Working and Companies
4	Unions and Labor; Political and Economic Ideals and Revolutions
5	Neighborhood Institutions, Facilities, and Public Space
6	Allegheny County History, Geography, and Governemnt; Allegheny Mountains; Allegheny City; Allegheny River; Allegheny Avenue
7	City Roadways and Neighborhoods; Locations of City Establishments
8	Portraits; People: Living, Working, Learning, and Playing; Animals: Pets and Wild
9	Maps of Pittsburgh; Performing Arts; Gatherings: Clubs, Camps, Carnivals, Luncheons
10	Architecture: Homes, Buildings of Worship, Cathedral of Learning; Fine Arts; Celebrations, Marches, and Ceremonies; McCurdy Family of Oakmont
11	Civil Services and Structures Around Water: Development and Restoration; Financial Institution
12	Historical Documents: American History, Government Files/Notes, Literature, Correspondence
13	Geography: Neighborhoods and Communities
14	Regional, National, and World Histories: Military, Population, and Government
15	Historical Homes, Families, and Correspondence
16	Higher Education; Public Programs and Research: Awards and Funding
17	Health and Medicine; Sports: Events, Facilities, and Players; Coal Industry and Mine Safety

Continued on next page

Table 4 – continued from previous page

Topic_Num_Vis	Topic Name(s)
18	Government; Organizational Administration, Leadership, and Training
19	University of Pittsburgh News and Press Releases; Pittsburgh Neighborhoods

9 Appendix B: Deliverable Files

9.1 Graphs

Table 5 lists the graphs created from the t-SNE and LDA models.

Table 5: Graphs

Filename	Description
tsne_800.png	t-SNE plot created from the 'word2vec' word embeddings
topics.png	Graph of the distribution of items by topic, total items and items with abstracts
lda-mallet_vis.html	Interactive graph with Intertopic Distance Map and Top-30 Most Salient Terms bar chart
Topic_Num_Vis 1.html	Interactive time series plot for topic 1 'Religious Structures, Practices, History (Christianity); Gothic Architecture; Windows and Porches: Views of People and Stores'
Topic_Num_Vis 2.html	Interactive time series plot for topic 2 'Buildings: History, Construction, Restoration, and Renovation'
Topic_Num_Vis 3.html	Interactive time series plot for topic 3 'Primary and Secondary Industry: People Working and Companies'
Topic_Num_Vis 4.html	Interactive time series plot for topic 4 'Unions and Labor; Political and Economic Ideals and Revolutions'
Topic_Num_Vis 5.html	Interactive time series plot for topic 5 'Neighborhood Institutions, Facilities, and Public Space'
Topic_Num_Vis 6.html	Interactive time series plot for topic 6 'Allegheny County History, Geography, and Governemnt; Allegheny Mountains; Allegheny City; Allegheny River; Allegheny Avenue'
Topic_Num_Vis 7.html	Interactive time series plot for topic 7 'City Roadways and Neighborhoods; Locations of City Establishments'
Topic_Num_Vis 8.html	Interactive time series plot for topic 8 'Portraits; People: Living, Working, Learning, and Playing; Animals: Pets and Wild'
Topic_Num_Vis 9.html	Interactive time series plot for topic 9 'Maps of Pittsburgh; Performing Arts; Gatherings: Clubs, Camps, Carnivals, Luncheons'

Continued on next page

Table 5 – continued from previous page

Filename	Description
Topic_Num_Vis 10.html	Interactive time series plot for topic 10 'Architecture: Homes, Buildings of Worship, Cathedral of Learning; Fine Arts; Celebrations, Marches, and Ceremonies; McCurdy Family of Oakmont'
Topic_Num_Vis 11.html	Interactive time series plot for topic 11 'Civil Services and Structures Around Water: Development and Restoration; Financial Institution'
Topic_Num_Vis 12.html	Interactive time series plot for topic 12 'Historical Documents: American History, Government Files/Notes, Literature, Correspondence'
Topic_Num_Vis 13.html	Interactive time series plot for topic 13 'Geography: Neighborhoods and Communities'
Topic_Num_Vis 14.html	Interactive time series plot for topic 14 'Regional, National, and World Histories: Military, Population, and Government'
Topic_Num_Vis 15.html	Interactive time series plot for topic 15 'Historical Homes, Families, and Correspondence'
Topic_Num_Vis 16.html	Interactive time series plot for topic 16 'Higher Education; Public Programs and Research: Awards and Funding'
Topic_Num_Vis 17.html	Interactive time series plot for topic 17 'Health and Medicine; Sports: Events, Facilities, and Players; Coal Industry and Mine Safety'
Topic_Num_Vis 18.html	Interactive time series plot for topic 18 'Government; Organizational Administration, Leadership, and Training'
Topic_Num_Vis 19.html	Interactive time series plot for topic 19 'University of Pittsburgh News and Press Releases; Pittsburgh Neighborhoods'

9.2 Data

Table 6 lists the output data files for the project. Table 7 lists the input data files for the project.

Table 6: Output Data Files

Filename	Description
Items_dominant_topics.csv	Lists the digital humanities items with fields that identify the dominant topic for each item

Continued on next page

Table 6 – continued from previous page

Filename	Description
Items_topics_summ.csv	Lists the topic numbers from both Mallet and Gensim models, keywords, and topic name for each topic
Topics_date_cnt.csv	Counts of items by 'Sort Date,' the date of the event depicted by the item
Data Dictionary.xlsx	Data dictionary for the data files

Table 7: Input Data Files

Filename	Description
Islandora_items_metadata.csv	Data set used for the t-SNE and LDA models: all digital item metadata at the item level (124,539 digitized items)
Finding_Aids_metadata.csv	All finding aid metadata at the collection level (2210 finding aids)
Digital_Collections_List.csv	All site and collection memberships at the item level (specifies to which site and collection an item belongs)

9.3 Code

Table 8 lists the code files for the project.

Table 8: Code

Filename	Description
Part 1a Data Prep.html	The code for the basic data cleaning.
Part 1b Data Prep.html	The code for the model preparation.
Part 2 Topic Modeling.html	The code for the word vectors and t-SNE plots and the LDA models.
Part 3 Data Visualization.html	The code for the *Distributions of Items by Topic* bar chart and the interactive time series graphs for each topic.

10 Appendix C: t-SNE Plot

Figure 4 shows the t-SNE plot created from the 'word2vec' word embeddings.

Figure 4: t-SNE Plot from 'word2vec' Word Embeddings