Brigitte Hogan (bwh5v@virginia.edu) and Jason Tiezzi (jbt5am@virginia.edu)
DS 5001: Exploratory Text Analytics
29 April 2020

Tables Manifest

This document provides an overview of the data files containing F2 through F5 data extracted from the corpus. The full OHCO for this digital analytic edition is : book identifier ('book_id'), volume number ('vol_num'), chapter number ('chap_num'), recipe number ('recp_num'), paragraph number ('para_num'), sentence number ('sent_num'), and token number ('token_num').

Core F2 Tables

Table	Description
DOC.csv (17,176 x 6)	Standard documents (DOC) table with OHCO columns to paragraph level and an additional column for paragraph string ('para_str'). There is one row for each paragraph in the corpus.
LIB.csv (20 x 7)	Library (LIB) table. Columns include standard features (book_id, author_last, author_full, book_year, book_title, book_file) as well as an added variable of "period" which reflects the general time-period of the cookbook. There is one row per book in the corpus.
TOKEN.csv (1048576 x 12)	Standard TOKEN table. Columns include full OHCO as well as part of speech tagging (pos_tuple and pos), token string (token_str), term string (term_str) and term identifier (term_id). There is one row per token in the corpus.
VOCAB.csv (16,786 x 16)	Vocabulary (VOCAB) table. Includes term_id, term_str, word frequency (n), a number dummy variable (num), a stop-word dummy variable (stop), stems (stem_porter and stem_snowball), two term rank calculations (term_rank and term_rank2), term percentage (p), three Zipf k measures (zipf_k, zipf_k2, zipf_k3) and three separate TFIDF sums based on different bags (TFIDF_sum_period, TFIDF_sum_book, and TFIDF_sum_recipe). There is one row per term in the corpus.

Embeddings

Table	Description
Embeddings_mid1800s.csv (845 x 19)	Word embeddings for the corpus of cookbooks written in the mid-1800s. Includes the columns from the VOCAB tables, as well as a vector column representing the embeddings generated from Word2Vec and an x and y coordinate generated by T-SNE. Links to VOCAB table via "term_str."
Emeddings_late1800s.csv (918 x 19)	Contains word embeddings for the corpus of cookbooks written in the late-1800s. Includes the columns from the VOCAB tables, as well as a vector column representing the embeddings

	generated from Word2Vec and an x and y are the coordinate generated by T-SNE. Links to VOCAB table via "term_str."
Embeddings_1900s.csv (942 x 19)	Contains word embeddings for the corpus of cookbooks written in the early-1900s. . Includes the columns from the VOCAB tables, as well as a vector column representing the embeddings generated from Word2Vec and an x and y are the coordinate generated by T-SNE. Links to VOCAB table via "term_str."

Sentiment

Table	Description
Emolex_sentiment.csv (3688 x 11)	The emolex lexicon, with columns for term_str, NRC sentiment type (nrc_anger, nrc_anticipation, nrc_disgust, nrc_fear, nrc_joy, nrc_sadness, nrc_suprise, nrc_trust) and NRC sentiment direction (nrc_negative and nrc_positive). This was not generated by us, but is necessary for our code to run. Links to the VOCAB table through "term_str."
Sentiment_book.csv (20 x 24)	Sentiment scores for each book. NRC values come from the emolex lexicon, while the VADER scores come from the VADER engine Columns include period, book_year, full OHCO, NRC sentiment types, NRC sentiment direction, VADER sentiment direction (VADER_pos, VADER_neg, VADER_neu) and overall scores (overall_NRC and VADER_compound.) There is one row per book in the corpus.
Sentiment_period.csv (3 x 24)	Sentiment scores for each time period. NRC values come from the emolex lexicon, while the VADER scores come from the VADER engine. Column values are the same as for Sentiment_book. There is one row per time period in the corpus.

TFIDF

Table	Description
TFIDF_book (20 x 16,786)	TFIDF with bag of book. Columns include period, book_year, book_id, and a column for each term string. There is one row per book in the corpus. TFIDF_sums were added to VOCAB table with term_str.
TFIDF_recipe (5,631 x 16,786)	TFIDF with bag of recipe. Columns include period, book_year, OHCO to the recipe level, and a column for each term string. There is one row per recipe in the corpus. TFIDF_sums were added to VOCAB table with term_str.
TFIDF_period (3 x 16,786)	TFIDF with bag of period. Columns included period and a column for each term string. There is one row per period in the corpus. TFIDF_sums were added to VOCAB table with term_str.

Topic Model

Table	Description
TOPICS.csv (25 x 14)	A table of the top topics found in the corpus. Columns include topic identifier (topic_id), top ten words in the topic (0 – 9), combined string of top words (label), an human-generated topic name (name) and the sum of THETA per topic (doc_weight_sum). There is one row per preset number of topics in the corpus. Can be bound to PHI or THETA using topic_id.
PHI.csv (25 x 5001)	A TOPIC-WORD language model indicating how much a topic likes a word. Columns consist of top 5,000 most frequent TOKEN strings as well as the topic_id. There is one row per preset number of topics in the corpus.
THETA.csv (14,846 x 29)	A DOC-TOPIC language model indicating how much a document likes a topic. Columns include the OHCO to paragraph level and the topic_id of each topic (0 – 24). There is one row per paragraph in the corpus.
PARAS (14,846 x 1)	F1 corpus and reduced version of DOC table with only regular nouns. Columns include partial OHCO to paragraph level and a paragraph string (para_str). There is one row per paragraph in the corpus. Can be matched to DOC using OHCO, but not all DOC rows will have a match if they did not contain regular nouns.
LDA_AUTHOR.csv (25 x 14)	TOPIC table using author as bag. Columns include the topic_id, 12 individual author names, and the human-generated topic labels (names). There is one row per preset number of topics in the corpus.
LDA_PERIOD.csv (25 x 6)	TOPIC table using period as bag. Columns include topic_id, time period (1900s, late1800s, mid 1800s), top terms in the time period (topterms), and human-generated topic labels (names). There is one row per preset number of topics in the corpus.

PCA

Table	Description
DCM_book.csv (20 x 10)	Document-Content Matrix for book as a bag. Columns include book_id, book_year, period, author last name (author), publication year (year), book title (title), a label string for plots (doc), and the three Principal Components (PC0, PC1, and PC2). There is one row per book.
DCM_recipe.csv (5631 x 10)	Document-Content Matrix for recipe as a bag. Columns are the same as for DCM_book, but there is one row per recipe.

EIGPAIR_book.csv (4999 x 5001)	Eigen pairs using book as bag for term_id components. Columns include term_id, eigen value (eig_val), explained variance (exp_var) , and a column per term_id for the top ~5000 significant terms. There is one row per term for the top ~5000 significant terms.
EIGPAIR_recipe.csv (4999 x 5001)	Eigen pairs using recipe as bag for term_id components. Columns are the same as for EIGPAIR_book. There is one row per term for the top ~5000 significant terms.
PCACOMPS_book.csv (3 x 5001)	Top three principal components using book as bag. Columns include principal component name (index), eig_val, exp_var, and a column for each term_id for the top ~5000 most significant terms. There is one per principal component per row.
PCACOMPS_recipe.csv (3 x 5001)	Top three principal components using recipe as bag. Columns and rows are the same as for PCACOMPS_book.
PCALOADINGS_book.csv (4999 x 5)	PCA Loadings for the top 3 components using book as a bag. Columns include term_id, term_str, and three principal components (PC0, PC1, PC2). There is one row per term for the top ~5000 significant terms.
PCALOADINGS_recipe.csv (4999 x 5)	PCA Loadings for the top 3 components using recipe as a bag. Columns include term_id, term_str, and three principal components (PC0, PC1, PC2). There is one row per term for the top ~5000 significant terms.
COV_book.csv (4999 x 4999)	A covariance matrix of features for book. There is one row and one column for each of the top ~5000 most significant terms.
COV_recipe.csv (4999 x 4999)	A covariance matrix of features for recipe. There is one row and one column for each of the top ~5000 most significant terms.