MAT 259 Assignment 1
Paula Wang

Concept Description

I am a student in the Communication department. In general, the term 'communication' is very broadly defined and can be interpreted in many ways. In the social science field, communication research encompasses a variety of areas such as interpersonal communication (e.g., parent-child, couples etc.), mass communication (e.g., social media, news etc.), organizational communication (e.g., workplace, political entities etc.), and computer-mediated communication (e.g., online messaging, video calling etc.). For this project, I wanted to explore which topic areas of communication are interesting to the general public.

Query Approach

To answer my question, I make the following assumptions:

- Dewey-classified items are generally non-fiction
- Communication information is generally relayed in book form
- Books with 'communication' in their title contain information related to communication topics
- Checkout counts can viably measure public interest in a given topic

My Query

I queried the checkout counts of books that have 'communication' in their title, separated by their Dewey classification into the ten topic areas of computer science (000), philosophy and psychology (100), religion (200), social sciences (300), language (400), pure science (500), technology (600), arts and recreation (700), literature (800), and history and geography (900). I look at the trends of communication-related books in Dewey categories over time.

SQL Query 1 (Dewey categories over time):

```
Unset
SELECT
      YEAR(cout) AS years,
      COUNT(IF(deweyClass>=000 AND deweyClass<=099, 1, NULL)) AS "00",
      COUNT(IF(deweyClass>=100 AND deweyClass<=199, 1, NULL)) AS "01",
      COUNT(IF(deweyClass>=200 AND deweyClass<=299, 1, NULL)) AS "02",
      COUNT(IF(deweyClass>=300 AND deweyClass<=399, 1, NULL)) AS "03",
      COUNT(IF(deweyClass>=400 AND deweyClass<=499, 1, NULL)) AS "04",
      COUNT(IF(deweyClass>=500 AND deweyClass<=599, 1, NULL)) AS "05",
      COUNT(IF(deweyClass>=600 AND deweyClass<=699, 1, NULL)) AS "06",
      COUNT(IF(deweyClass>=700 AND deweyClass<=799, 1, NULL)) AS "07",
```

```
      COUNT(IF(deweyClass>=800 AND deweyClass<=899, 1, NULL)) AS "08",
      COUNT(IF(deweyClass>=900 AND deweyClass<=999, 1, NULL)) AS "09"
FROM outraw
WHERE itemtype LIKE '%bk'
      AND LOWER(title) LIKE '%communication%'
GROUP BY years
```

After examining the data (see results section), it appears that communication-related books are most popular within the Dewey topics of technology, philosophy and psychology, social sciences, and language. To examine these further, I independently queried the distinct titles of all books that have ‘communication’ in their titles, within all four Dewey categories.

SQL Query 2 (Technology titles):

```
Unset
SELECT
    title, count(title) as copies
FROM outraw
WHERE itemtype LIKE '%bk'
    AND LOWER(title) LIKE '%communication%'
    AND (deweyClass >= 600 AND deweyClass <= 699) #technology
GROUP BY title
ORDER BY copies DESC
```

SQL Query 3 (philosophy and psychology titles):

```
Unset
SELECT
    title, count(title) as copies
FROM outraw
WHERE itemtype LIKE '%bk'
    AND LOWER(title) LIKE '%communication%'
    AND (deweyClass >= 100 AND deweyClass <= 199) #psychology
GROUP BY title
ORDER BY copies DESC
```

SQL Query 4 (social sciences titles):

```
Unset
SELECT
    title, count(title) as copies
FROM outraw
WHERE itemtype LIKE '%bk'
    AND LOWER(title) LIKE '%communication%'
    AND (deweyClass >= 300 AND deweyClass <= 399) #social sciences
GROUP BY title
ORDER BY copies DESC
```

SQL Query 5 (language titles):

```
Unset
SELECT
    title, count(title) as copies
FROM outraw
WHERE itemtype LIKE '%bk'
    AND LOWER(title) LIKE '%communication%'
    AND (deweyClass >= 400 AND deweyClass <= 499) #language
GROUP BY title
ORDER BY copies DESC
```

Data and Results

The query for the count of 'communication' titled books across Dewey categories over time (Query 1) returned an 18 x 10 (rows x columns) matrix with rows corresponding to years 2006-2023 and columns corresponding to Dewey categories. To visualize the data, I used the *matplotlib* and *seaborn* packages in Python 3.7.

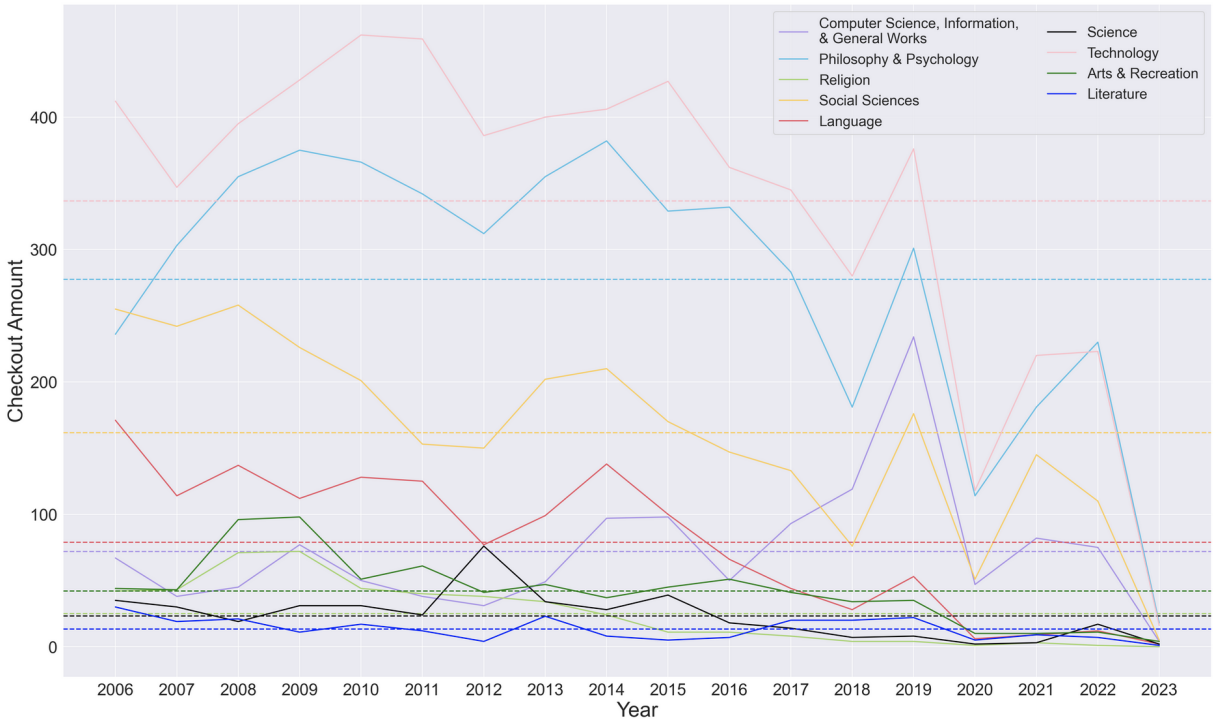
Python 1 (Plot Dewey categories over time):

```
Python
# import and process data
import pandas as pd
comm = pd.read_csv(path + '/communication.csv', index_col='years')

d={
      'dewey':comm.columns.tolist(),
      'labels':["Computer Science, Information,\n& General
Works","Philosophy & Psychology","Religion","Social
```

```
Sciences","Language","Science","Technology","Arts &
Recreation","Literature","History & Geography"],
      'colors':["#ac92eb","#4fc1e8","#a0d568","#ffce54","#ed5564","black
","pink","green","blue","purple"],
      'means':comm.mean().tolist()
}

# define x axis as years
x = comm.index.tolist()

# plot
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

plt.figure(figsize=(25,15), dpi=200)
for i in range(9):
      plt.plot(x, comm[d['dewey'][i]], d['colors'][i],
label=d['labels'][i], alpha=1)
      plt.axhline(d['means'][i], color=d['colors'][i], linestyle = '--')

plt.legend(loc='upper right', fontsize='xx-large', ncol=2)
plt.xlabel('Year', fontsize=25)
plt.ylabel('Checkout Amount', fontsize=25)
plt.yticks(fontsize=20)
plt.xticks(np.arange(2006, 2024),fontsize=20)
sns.set_style('darkgrid')
```

In addition to the checkout counts over time, I also plotted the average checkout count across time for all Dewey categories. This can be observed in Figure 1 below as the colored horizontal lines crossing the y-axis.

Figure 1 (checkout counts of communication books across Dewey categories over time):



Overall, it appears that the highest checkout counts of communication-related titles are from the categories of technology (average = 337), followed by philosophy and psychology (average = 278), social sciences (average = 162), and language (average = 79).

To examine this further, I wanted to identify subtopics of communication that were being checked out within each of these broader Dewey classifications. To do so, I ran a topic model analysis using latent dirichlet allocation (LDA) to identify topics from a group of titles. Using the returned data from SQL Queries 2-5, I set the LDA model to identify 3 subtopics within each group of titles from technology, philosophy and psychology, social sciences, and language. Topic modeling was done using *gensim* in Python 3.7.

Python 2 (topic model):

```python
# load packages
import re
import numpy as np
import pandas as pd
from pprint import pprint
```

```
# Gensim
import gensim
import gensim.corpora as corpora
from gensim.utils import simple_preprocess
from gensim.models import CoherenceModel

# spacy for lemmatization
import spacy

# Enable logging for gensim
import logging
logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s',
level=logging.ERROR)

import warnings
warnings.filterwarnings("ignore",category=DeprecationWarning)

# prepare stopwords
import nltk
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
stop_words.extend(['communication'])

# set data
import pandas as pd
psych = pd.read_csv(path + '/psych.csv')
social = pd.read_csv(path + '/social.csv')
language = pd.read_csv(path + '/language.csv')
technology = pd.read_csv(path + '/technology.csv')

data = [i for i in social.title.tolist()] # swap out with each category

# tokenize words
def sent_to_words(sentences):
      for sentence in sentences:
      yield(gensim.utils.simple_preprocess(str(sentence), deacc=True))
data_words = list(sent_to_words(data))

# Build the bigram and trigram models
bigram = gensim.models.Phrases(data_words, min_count=5, threshold=100)
```

```
trigram = gensim.models.Phrases(bigram[data_words], threshold=100)
bigram_mod = gensim.models.phrases.Phraser(bigram)
trigram_mod = gensim.models.phrases.Phraser(trigram)

# Define functions for stopwords, bigrams, trigrams and lemmatization
def remove_stopwords(texts):
      return [[word for word in simple_preprocess(str(doc)) if word not
in stop_words] for doc in texts]

def make_bigrams(texts):
      return [bigram_mod[doc] for doc in texts]

def make_trigrams(texts):
      return [trigram_mod[bigram_mod[doc]] for doc in texts]

def lemmatization(texts, allowed_postags=['NOUN', 'ADJ', 'VERB',
'ADV']):
      """https://spacy.io/api/annotation"""
      texts_out = []
      for sent in texts:
      doc = nlp(" ".join(sent))
      texts_out.append([token.lemma_ for token in doc if token.pos_ in
allowed_postags])
      return texts_out

data_words_nostops = remove_stopwords(data_words)
data_words_bigrams = make_bigrams(data_words_nostops)
nlp = spacy.load('en', disable=['parser', 'ner'])
data_lemmatized = lemmatization(data_words_bigrams,
allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV'])

# Create dictionary and corpus
id2word = corpora.Dictionary(data_lemmatized)
texts = data_lemmatized
corpus = [id2word.doc2bow(text) for text in texts]

# Build LDA model
lda_model = gensim.models.ldamodel.LdaModel(corpus=corpus,
                                            id2word=id2word,
                                            num_topics=3,
```

```
                                        random_state=100,
                                        update_every=1,
                                        chunksize=100,
                                        passes=10,
                                        alpha='auto',
                                        per_word_topics=True)

# Compute Perplexity and coherence scores
print('\nPerplexity: ', lda_model.log_perplexity(corpus))

coherence_model_lda = CoherenceModel(model=lda_model,
texts=data_lemmatized, dictionary=id2word, coherence='c_v')
coherence_lda = coherence_model_lda.get_coherence()
print('\nCoherence Score: ', coherence_lda)
```

The keywords and values for each topic within each of the four Dewey categories are depicted in Tables 1-4 below.

Table 1 depicts the keywords and values for the three topics identified within the LDA model for technology (perplexity = -6.38, coherence = 0.57). We can infer that the first topic refers to books that guide readers in understanding communication technologies in the workplace. The second topic refers to books that describe communication technologies as a broadcasting system. The third topic refers to books that describe using communication technologies in parenting environments.

Table 1 (Technology):

Topic 1		Topic 2		Topic 3	
communication	0.055	communication	0.031	guide	0.024
guide	0.032	system	0.019	child	0.023
business	0.027	radio	0.018	skill	0.017
work	0.020	wireless	0.015	parent	0.016
practical	0.014	electronic	0.012	system	0.012
message	0.011	skill	0.011	design	0.012
world	0.010	organization	0.011	practical	0.011

secret	0.009	tool	0.010	improve	0.010
marketing	0.009	success	0.009	professional	0.009
satellite	0.008	power	0.009	family	0.009

Table 2 depicts the keywords and values for the three topics identified within the LDA model for philosophy and psychology (perplexity = -5.73, coherence = 0.56). We can infer that the first topic refers to communication more broadly across animals and people. The second topic refers to dealing with conflict communication. The third topic refers to defensive communication.

Table 2 (Philosophy and psychology):

Topic 1		Topic 2		Topic 3	
animal	0.028	conflict	0.022	communication	0.021
life	0.015	life	0.022	say	0.021
teen	0.015	mindful	0.022	defensive	0.021
voice	0.015	language	0.022	information	0.021
connect	0.015	face	0.022	medium	0.014
guide	0.015	love	0.015	relationship	0.014
contact	0.015	skill	0.015	word	0.014
side	0.015	relationship	0.015	critical	0.014
death	0.015	secret	0.015	war	0.014
personality	0.015	heart	0.015	powerful	0.014

Table 3 depicts the keywords and values for the three topics identified within the LDA model for social science (perplexity = -6.54, coherence = 0.60). We can infer that the first topic refers to political communication. The second topic refers to online communication and internet trends. The third topic refers to the history of the communication industry.

Table 3 (Social science):

Topic 1		Topic 2		Topic 3	
age	0.019	issue	0.014	communication	0.059
political	0.017	internet	0.014	industry	0.019
tech	0.014	graphic	0.012	history	0.016

high	0.014	voice	0.012	make	0.013
computer	0.014	traffic	0.012	cable	0.013
costume	0.014	essential	0.011	guide	0.012
campaign	0.014	social	0.010	connection	0.010
space	0.014	new	0.010	satellite	0.010
information	0.013	technical	0.008	strategy	0.010
guide	0.011	report	0.008	business	0.010

Table 4 depicts the keywords and values for the three topics identified within the LDA model for language (perplexity = -5.27, coherence = 0.52). We can infer that the first topic refers to early infant communication. The second topic refers to the practice of sign language. The third topic refers to guides for speaking ideas clearly and naturally.

Table 4 (Language):

Topic 1		Topic 2		Topic 3	
baby	0.048	sign	0.047	guide	0.030
language	0.047	language	0.036	clear	0.030
french	0.033	introduction	0.025	naturally	0.030
real	0.033	barron	0.025	story	0.030
basic	0.033	skill	0.014	grammar	0.017
early	0.033	practice	0.014	adolescent	0.017
sign	0.033	care	0.014	friendship	0.017
review	0.019	culture	0.014	talk	0.017
confident	0.019	read	0.014	authentic	0.017
ultimate	0.019	health	0.014	inter	0.017

Discussion/Analysis of results

My research aim was to identify which topics of communication are most interesting to the general public. My analysis was done in two steps: (1) identifying the Dewey topics of the most commonly checked out books relating to communication, and then (2) identifying the subtopics

of communication that were checked out within the most popular Dewey categories. I found that (in descending order) the general public checked out communication-related books in the Dewey categories of technology, philosophy and psychology, social science, and finally language. Upon further analysis, I identified that titles in the technology category focused on communication technologies in the workplace, in broadcasting contexts, and in parenting. Titles in the philosophy and psychology category focused on conflict communication, defensive communication, and communication more broadly across animals and people. Titles in the social science category focused on political communication, internet communication, and the history of the communication industry. Finally, titles in the language category focused on infant communication, sign language, and speaking clearly and naturally.