Text search informed by word frequencies and topic modeling:

A human-machine collaborative approach to analyzing English text data from multilingual, multicultural students

Consultant:

Seth D Temple

Statistics PhD student

University of Washington - Seattle

Clients:

Salwa Al-Noori

Associate Teaching Professor in Biology

University of Washington – Bothell

Sami Al-Hasnawi

Assistant Professor in English

University of Al-Qadisiyah

December 17, 2021

Abstract: I present a procedure for the systematic review of student posts to discussion boards which considers the cultural, linguistic, and academic diversity of the class. This approach promotes a human-machine collaboration to complement the strengths of human and artificial intelligence in critical thinking and computation, respectively. Word frequency methods and topic modeling are employed to identify interesting terms in the documents corpus. These terms are then searched for in sentences and laid in juxtaposition based on the two cultural groups for human review. Some intriguing case examples of this approach are discussed.

1 Introduction

Professors Salwa Al-Noori and Sami Al-Hasnawi reached out to the University of Washington (UW) Statistical Consulting Service in Autumn 2021 for help in the analysis of survey and text data from a unique university course. They used the pedagogical approach Collaborative Online International Learning (COIL) to promote multicultural learning in their classes at UW Bothell in the United States and University of Al-Qadisiyah in Iraq. Independently, Professor Al-Noori led a biology course and Al-Hasnawi led a group of linguistics and English students; jointly, the classes met synchronously and asynchronously with online platforms to complete shared assignments. They collected two datasets in the process: (1) students completed pre- and post-course surveys related to the COIL learning outcomes of intercultural learning, and (2) students posted to Padlet discussion boards in response to various prompts. Their primary interest is in a text data analysis with the survey responses complementing the findings.

After multiple meetings with my clients, I have refined their research questions into the following list:

1. Is there evidence in the surveys that the students achieved COIL learning outcomes?
2. Can we find instances of intercultural learning in posts on the Padlet discussion boards?
3. How did the American and Iraqi student experiences differ and/or align?
4. How did English use positively and/or negatively impact intercultural learning, especially given the power dynamic between native and ESL speakers?

While these questions require qualitative analysis, we can leverage natural language processing (NLP) to assist in a systematic review of the text data. For instance, Zhang, Ren, and de Rijke (2021) developed a human-machine collaborative framework that assigned easy to classify dialogues to a machine learning (ML) model and hard to classify dialogues to a human.[1] Similarly, Jhaver, Birman, Gilbert, and Bruckman (2019) studied the use of human and automated content moderation on the Reddit website.[2] In our setting, we can use statistical and data visualization methods in NLP to efficiently scan the text corpus for dialogue of research interest. On the other hand, the clients have considerable biases as the instructors of the COIL course and participants in the Padlet discussion boards. In deploying NLP, I can act as an outsider and provide insights into the text data analysis that may have been missed by the clients. This is a common strategy to improve research projects in the social sciences.

This report consists of the following sections. Section 2 summarizes the findings from the survey data analysis and offers suggestions for future surveys. Section 3 describes the NLP methods considered in the text data analysis. Section 4 concludes with discussion about the ongoing research project and its limitations. Appendix A presents various tables from the survey data analysis. Appendices B and C include reports sent to the clients earlier in the quarter.

2 Survey data

Based on unique anonymous identifiers, I matched responses between the pre- and post-course surveys. Since some students did not complete the post-course survey, I ended up with twenty-six paired survey responses. Moreover, only some prompts appeared in both surveys. (The pre-course survey included various questions about demographics and may be summarized in a paragraph format.) These prompts were organized into three groups: International Interests (II), Intercultural Awareness and Intercultural Citizenship (IAIC), and English Language Use and Learning (ELUL). Responses were on five-point Likert scales, e.g., "Strongly disagree", "Disagree", "Neutral", "Agree", "Strongly agree". Most prompts were phrased in such a way that the instructors expected "Agree" or "Strongly agree" responses in the post-course survey if learning outcomes were met.

Initially, I considered some hypothesis testing based on sign statistics to evaluate if post-course responses had higher Likert scores than pre-course responses.[3] These nonparametric tests assign plus or minus values and leverage the binomial distribution to determine if the probabilities of plus and minus differ. I decided against this approach given the small sample sizes and the challenge in presenting statistical hypothesis testing to a non-STEM audience. I instead created tables to display the raw data in terms of (1) increases and/or decreases on the Likert scale and (2) percentage of "Agree" or "Strongly agree" responses. The clients agreed with this decision, noting that the surveys were designed for observational purposes and that human subjects approval was not attained for experimentation. Appendices B and C contain these tables for both cohorts, for the Iraqi cohort, and for the American cohort as requested by the clients.

For the tables, I looked closely at prompts that had more increased or decreased Likert scores and higher or lower agreement percentages. These correspond to prompts that improved or did not in the post-course responses. Many of the IAIC prompts improved on the Likert scale and in terms of agreements, providing aggregate evidence that the COIL course achieved its learning outcomes. Interestingly, many Iraqi students scored higher on the II question "Because I speak English, I don't have to learn a new language to be able to speak with people from other countries" in the post-course survey. Four of the seven American students changed their minds on the II question "Knowledge of other cultures helps me better understand my own", reaching one hundred percent agreement in the post-course survey. The II question "I purchase songs or CDs with music that is sung in a language other than my own" had perplexing results, perhaps because compact discs are not in the mode among current youth. Iraqi students selected higher Likert scores for ELUL prompts about how the English language will help them in future study, work, and cross-cultural communication. American students did not respond as much to these prompts, maybe because they read those prompts from a different perspective. Both cohorts scored higher on the ELUL prompts about studying abroad.

While there is aggregate evidence that most students achieved learning outcomes related to intercultural awareness and global citizenship and left the course feeling better prepared for being abroad, there were issues with the surveys. I have discussed these issues with the clients. They intend to refine the survey over time. Some resources from SurveyMonkey and Pew

Researcher Center may be helpful in designing the next iteration of surveys. Below are some additional suggestions on survey design.

1. Design a shorter survey, reducing redundant prompts or prompts of less interest
2. Phrase the prompts in generic language as opposed to using course-specific terms, e.g., global citizenship
3. Be careful with prompts and language that are interpreted differently by cohort, e.g., Iraqi students and American students interpret "foreign language" differently
4. Require response, allow only one response (not select all feature) for important prompts
5. Use the same Likert scale, say five-point balanced scale from "Strongly disagree" to "Strongly agree", for all prompts
6. Put demographic questions at the end of the pre-course survey
7. Ask some volunteers to take the survey beforehand and provide feedback
8. Use consistent labeling and prompt language for the pre- and post-course surveys

3 Text data

Following group discussions in the synchronous online meetings, students posted asynchronously to discussion boards by Padlet. Prompts focused on (1) learning in a pandemic, (2) gender-issues, (3) English as a lingua franca, and (4) food and music. The clients expressed the most interest in prompts (2) and (3). Posts used informal language and had mistyping/misspellings. They also included emojis, pictures, Arabic characters, and hyperlinks, especially in the fourth prompt. Student names were anonymized per FERPA guidelines in US education. Given the many challenges associated with analyzing such text data, I focused on characters in the English alphabet and words in the English dictionary. Periods, question marks, and exclamation marks were interpreted to end sentences.

I developed a Python package to automate the text data analysis. This package combines pre-processing, wrangling, and NLP methods into user-friendly, documented functions. It depends on the ML and NLP libraries Gensim[4], NLTK[5], Numpy[6], Pandas[7], and Scikit-learn[8]. Users can explore printed outputs and data visualizations to identify anchor words of research interest. Next, search on the text corpus can find sentences containing the anchor words and flanking sentences for context. Finally, humans can interpret these anchored sentences, especially in comparing posts by cohort.

3.1 Word frequencies

Many methods in NLP rely on word frequencies. In this setting, these statistics summarize the content of the student posts. I computed various common statistics on word frequencies for the entire corpus, for the Iraqi student posts, and for the American student posts to identify potential anchor words. My current implementations count every instance of a word, versus the clients have requested that I also consider post counts for the presence/absence of a word.

3.1.1 Word clouds

Word clouds showcase the most frequent words in a visually appealing way. Larger fonts correspond to higher frequency. This data visualization can be useful as an exploratory tool, especially when juxtaposing word clouds from Iraqi and American student posts. For example, Figure 1 displays similarities and differences between the Iraqi and American posts on gender-issues. The American biology students used the terms “STEM”, “field”, “career”, and “pay” to discuss gender inequality in their discipline and workplaces. Iraqi students used the term “men” with higher relative frequency than the American students. Interestingly, there were many occurrences of “Iraq” in the American student posts. This is one example of how we may select anchor words for an informed text search.

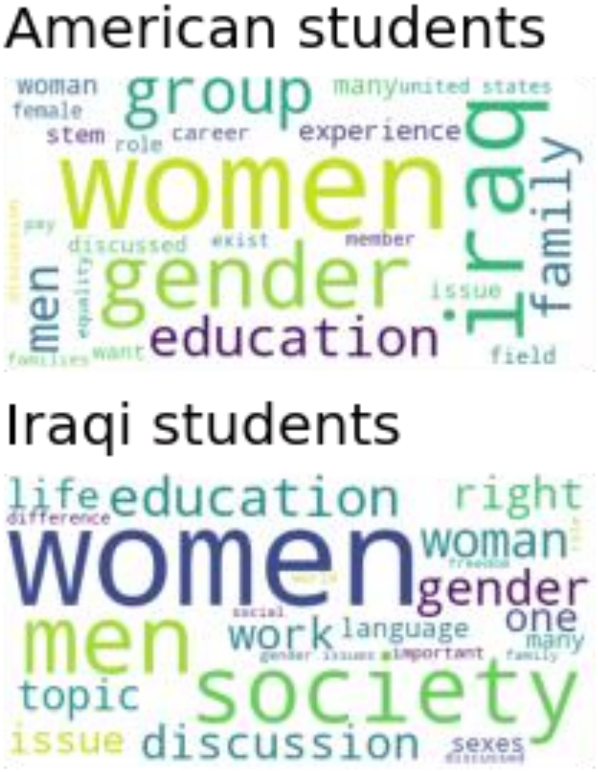


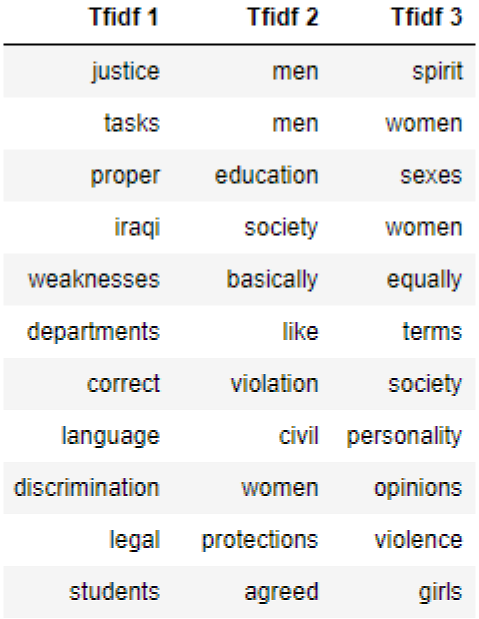
Figure 1 Gender-issues word cloud for American and Iraqi students

3.1.2 *n*-grams

Adjacent words may indicate compound ideas or common phrases. *n*-grams refer to *n* terms that occur next to each other with high frequency. For instance, a bigram is two adjacent words and a trigram is three adjacent words. I implemented the *n*-grams method in Gensim[4] to find such words. As an example, text from the gender-issues prompt contained the bigrams “gender (in)equality”, “united states”, “religious freedom”, “medical care”, and “higher education”. At the request of the clients, I intend to generalize my word-anchored search to find sentences with such *n*-grams. This functionality will enable text search for compound words and word associations.

3.1.3 Term frequency inverse document frequency

The numerical statistic tf-idf measures the importance of a word to a given document. It divides the word frequency in the document by the number of words in the document.[9] We may find other candidate anchors by looking at words with high tf-idf values for each document. For example, Figure 2 reports the tf-idf matrix for the Padlet on gender-issues. Words like “legal”, “violence”, “civil”, and so on that appear in the tf-idf matrix but not in the word cloud may be promising candidate anchors.



Tfidf 1	Tfidf 2	Tfidf 3
justice	men	spirit
tasks	men	women
proper	education	sexes
iraqi	society	women
weaknesses	basically	equally
departments	like	terms
correct	violation	society
language	civil	personality
discrimination	women	opinions
legal	protections	violence
students	agreed	girls

Figure 2 Three most frequent terms for 10 posts in the Padlet on gender-issues

3.2 Topic modeling

Finding groups/clusters is a common statistical task in which the goal is to uncover unknown structure in a dataset. Topic modeling refers to grouping words into categories based on their co-occurrence in a text corpus. There is an assumed organization to written and spoken text. Documents concern a finite set of topics, and those topics contain a finite vocabulary of related terms. Topic models try to learn both the distribution of topics in a documents corpus and the distribution of terms that describe the topics.

In our setting, we know that the discussion prompts concern specific topics. I instead use topic modeling as an exploratory tool to identify anchor words. That is, uncovering meaningful topics (groups/clusters) is not the aim of this analysis. I focus more on the most frequent terms in the topics as possible candidate anchors. Because I am less interested in the topics themselves, I look for more topics and accept that the topics may not be well-separated. This technique is akin to the tf-idf matrix in that I may find topics that mostly belong to a few documents.

3.2.1 Latent Dirichlet allocation

A popular way to find topics in a corpus is to consider a probabilistic model for the distribution of topics and for the distributions of terms. The jargon for this probabilistic model in the statistical literature is latent Dirichlet allocation (LDA). Pritchard, Stephens, and Donnelly (2000) developed this method originally to infer ancestral populations from the genetic data of people.[10] Their observation was that the nucleotides adenine (A), cytosine (C), guanine (G), and thymine (T) at each genetic marker/basepair appear with different multinomial probabilities depending on the local chromosomal background of individuals. That is, in a chromosomal region where an African-American (admixed) individual has local African ancestry, the nucleotide frequencies for A, C, T, and G ought to be those of some ancestral African population. This genetics example is a simple topic model where the terms are the nucleotides, the topics are the ancestral populations, and the documents are the genomes (nucleotide sequences). Blei, Ng, and Jordan (2003) extended the model of Pritchard, Stephens, and Donnelly to study text from human languages.[11]

LDA is a Bayesian method where make inferences by drawing samples from the probability model after the sampling procedure has reached some convergence definition. This is a hierarchical model that can be sampled in a Gibbs way. Matthew Stephens provides tutorials on this and related models on his GitHub website. Let i and j index positions and documents such that $z_{i,j}$ and $w_{i,j}$ refer to the topic and term for the i^{th} position in the j^{th} document. Parameters θ_i and $\psi_{z_{i,j}}$ follow (sparse) Dirichlet distributions with hyperpriors α and β with (probability) support the set of topics and terms. Namely,

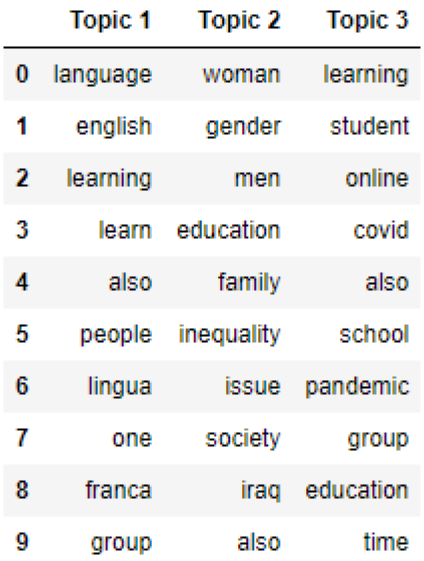
$$z_{i,j} \sim \text{Multinomial}(\theta_i)$$

$$w_{i,j} \sim \text{Multinomial}\left(\psi_{z_{i,j}}\right)$$

$$\theta_i \sim \text{Dirichlet}(\alpha)$$

$$\psi_{z_{i,j}} \sim \text{Dirichlet}(\beta)$$

We collect many samples of the topics and their terms/word, and then use the most frequent words assigned to a topic to describe it qualitatively. Figure 3 demonstrates this inference for a three topics topic model on Padlet text data from the first three discussion prompts. I found that LDA uncovers the three topics known *a priori*. I considered a topic model with more potential topics, but the inferred topics did not appear well-separated nor well-defined.



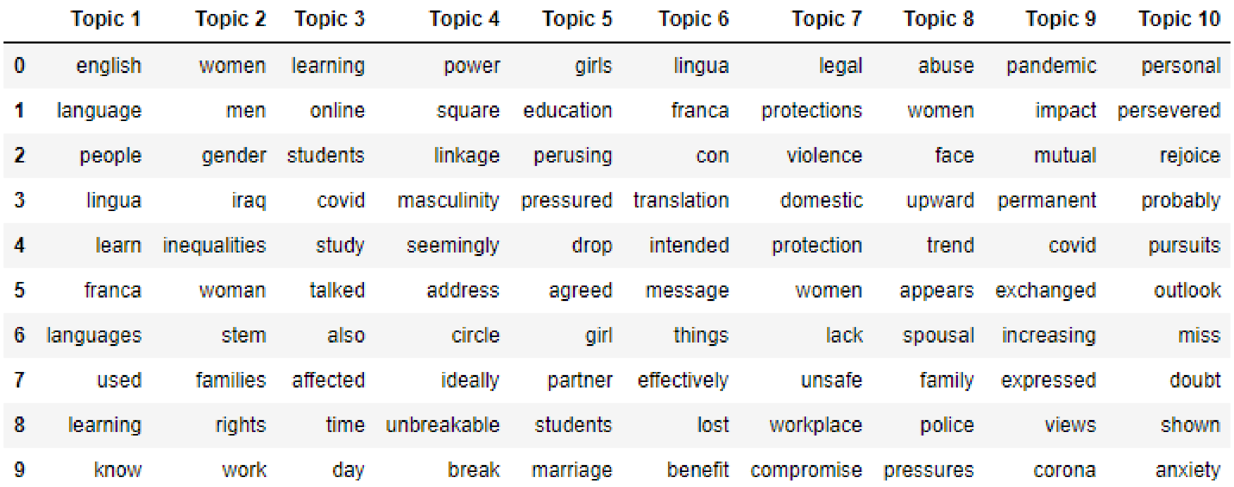
	Topic 1	Topic 2	Topic 3
0	language	woman	learning
1	english	gender	student
2	learning	men	online
3	learn	education	covid
4	also	family	also
5	people	inequality	school
6	lingua	issue	pandemic
7	one	society	group
8	franca	iraq	education
9	group	also	time

Figure 3 Most frequent terms in three topics LDA model

3.2.2 Non-negative matrix factorization

Non-negative matrix factorization (NMF) is another popular method for topic modeling. This technique comes from the research area of linear algebra whereby we decompose a high-dimensional matrix into the product of two matrices. Let V be an $m \times n$ matrix for the m documents and the n words in the corpus. The goal is to attain a factorization $V = WH$ where W is an $m \times p$ matrix, H is a $p \times n$ matrix, and p is much smaller than m, n. The dimension p is the number of topics. The matrix V contains the word counts for each document in the corpus. Like principal component analysis, the matrix W is loadings that determine how much topic weight to give each document. On the other hand, the matrix H gives weights to each word based on their relevance to the topic. That is, the highest values in each row of H are the most important words to that topic. NMF solves a constrained optimization problem: minimize the Frobenius norm of $V - WH$ given that the matrices are all non-negative. These mechanics are out of scope for this writeup.

Figure 4 demonstrates NMF applied to the same inference problem as in LDA. I found that NMF uncovers more interesting and nuanced words in its inferred topics. For more topics, NMF performed better as well. Both methods result in poorly defined topics outside of the known *a priori* discussion topics. Nevertheless, with NMF, I noticed multiple words like “power”, “masculinity”, “translation”, “compromise”, and “abuse”, to name a few, that may serve as good anchor words for text search.



	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
0	english	women	learning	power	girls	lingua	legal	abuse	pandemic	personal
1	language	men	online	square	education	franca	protections	women	impact	persevered
2	people	gender	students	linkage	perusing	con	violence	face	mutual	rejoice
3	lingua	iraq	covid	masculinity	pressured	translation	domestic	upward	permanent	probably
4	learn	inequalities	study	seemingly	drop	intended	protection	trend	covid	pursuits
5	franca	woman	talked	address	agreed	message	women	appears	exchanged	outlook
6	languages	stem	also	circle	girl	things	lack	spousal	increasing	miss
7	used	families	affected	ideally	partner	effectively	unsafe	family	expressed	doubt
8	learning	rights	time	unbreakable	students	lost	workplace	police	views	shown
9	know	work	day	break	marriage	benefit	compromise	pressures	corona	anxiety

Figure 4 Highest value words in ten topic NMF model

4 Conclusion

In this report, I have outlined a procedure for human-machine collaboration whereby an outsider identifies anchor words informed by word frequency and topic modeling in NLP to search for important sentences in a text corpus. Explicitly, this procedure is:

1. Setup and install the GitHub repository
2. Pre-process text corpus
    a. Anonymize identifying information
    b. Remove emojis, images, non-ASCII characters, and hyperlinks
    c. Tag each document if there are groups (American vs Iraq)
3. Identify anchors based on word frequency
    a. Word cloud
        i. Most frequent words
        ii. Different word frequencies in different groups
    b. *n*-grams
        i. Compound words
    c. tf-idf matrix
        i. Most frequent words per document
        ii. Evocative words (qualitative assessment)
4. Identify anchors based on topic modeling
    a. NMF
        i. Word counts matrix
        ii. Larger dimension *p* to look for more nuanced topics
        iii. Evocative words (qualitative assessment)
5. Search for sentences based on anchors
    a. Truncate words to word parts based on expert knowledge (e.g. “equal” to catch “equality”, “inequality”, etc.)
    b. (Context) 1 or more flanking sentences before and after the anchored sentence
6. Systematically review sentences, with emphasis on group differences

This procedure is designed for a small corpus in which the systematic review of sentences is not too laborious for human researchers. Its advantage is to intentionally and more objectively approach the qualitative assessment of text data rather than reading tens to hundreds of pages without organization and order.

This text data analysis ignored many aspects of the original data out of convenience. For instance, emojis and punctuation were stripped from the text data and may be informative about student personality and intention. Using the hotkey Ctrl + F, the clients can search for sentences in the original document if desired. The small size of the corpus also limited the topic modeling. Typically, NLP methods are applied to huge corpuses, e.g., all articles on Wikipedia or all articles from a specific journal/conference. I therefore emphasize that NMF and LDA are employed in this application to discover interesting words and word associations, not to infer topics. Ultimately, the NLP methods are an intermediate step to find and organize anchored sentences. The qualitative research questions are answered by human researchers reading and interpreting the texts.

Professors Salwa Al-Noori and Sami Al-Hasnawi intend to publish their pedagogical findings in some research journal. I will implement a few additional package features at their request and conduct another text data analysis for a second COIL course. Additionally, I have offered to help create tables and figures for them for their publication. I will provide them with an official letter detailing my expected work and time to complete these tasks.

References

1. Zhang, Y., Ren, P. & de Rijke, M. A Human-machine Collaborative Framework for Evaluating Malevolence in Dialogues. in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* 5612–5623 (Association for Computational Linguistics, 2021). doi:10.18653/v1/2021.acl-long.436.
2. Jhaver, S., Birman, I., Gilbert, E. & Bruckman, A. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* **26**, 1–35 (2019).
3. Bian, G., McAleer, M. & Wong, W.-K. A trinomial test for paired data when there are many ties. *Math. Comput. Simul.* **81**, 1153–1160 (2011).
4. Radim Rehurek, P. S. Software Framework for Topic Modelling with Large Corpora. in *IN PROCEEDINGS OF THE LREC 2010 WORKSHOP ON NEW CHALLENGES FOR NLP FRAMEWORKS* (Citeseer, 2010).
5. Bird, S., Klein, E. & Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. ("O'Reilly Media, Inc.," 2009).
6. van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering* **13**, 22–30 (2011).
7. McKinney, W. & Others. pandas: a foundational Python library for data analysis and statistics. *Python for high performance and scientific computing* **14**, 1–9 (2011).
8. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).

9. Rajaraman, A. & Ullman, J. D. *Mining of Massive Datasets*. (Cambridge University Press, 2011).

10. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**, 945–959 (2000).

11. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *the Journal of machine Learning research* **3**, 993–1022 (2003).

Appendix A

Tables including aggregate statistics from the pre- and post-course surveys are displayed. I have made separate tables for the American and Iraqi cohorts to better explain the findings. Codes for the survey prompts are listed at the end of this section.

Table 1A Summary of paired surveys to International Interests (II) and Intercultural Awareness and Intercultural Citizenship (IAIC) prompts

Prompt	Count of Likert responses				Percent of Likert response >= 4		
	Decreased	Same	Increased	Incr. - Decr.	Pre-survey	Post-survey	Post - Pre
II.A	7	10	9	2	0.27	0.31	0.04
II.C	6	11	9	3	0.89	0.92	0.04
II.D	2	18	5	3	0.92	1.00	0.08
II.E	13	2	11	-2	0.15	0.27	0.12
II.F	3	15	8	5	0.12	0.19	0.08
II.G	3	18	4	1	0.96	0.92	-0.04
II.H	2	18	6	4	0.92	0.96	0.04
II.I	5	15	5	0	0.72	0.73	0.01
IAIC.1	4	16	6	2	0.62	0.73	0.12
IAIC.2	2	17	7	5	0.96	1.00	0.04
IAIC.3	4	11	11	7	0.69	0.88	0.19
IAIC.4	4	17	5	1	0.89	0.89	0.00
IAIC.5	4	14	8	4	0.92	0.92	0.00
IAIC.6	2	12	12	10	0.92	0.96	0.04
IAIC.7	3	20	3	0	0.96	0.96	0.00
IAIC.9	4	17	5	1	0.81	0.81	0.00
IAIC.10	2	13	11	9	0.81	0.92	0.12
IAIC.11	5	13	8	3	0.62	0.73	0.12
IAIC.12	4	12	9	5	0.62	0.80	0.18
IAIC.13	4	15	6	2	0.80	0.85	0.05
IAIC.14	3	11	12	9	0.73	0.81	0.08

Table 1B Summary of paired surveys to International Interests (II) and Intercultural Awareness and Intercultural Citizenship (IAIC) prompts for Iraqi students

Prompt	**Count of Likert responses**				**Percent of Likert response >= 4**		
	Decreased	Same	Increased	Incr. - Decr.	Pre-survey	Post-survey	Post - Pre
II.A	4	7	8	4	0.26	0.26	0.00
II.C	5	9	5	0	1.00	0.90	-0.11
II.D	1	15	2	1	0.95	1.00	0.05
II.E	10	0	9	-1	0.21	0.37	0.16
II.F	2	10	7	5	0.16	0.26	0.11
II.G	1	15	2	1	0.95	0.89	-0.06
II.H	2	12	5	3	0.90	0.95	0.05
II.I	2	12	4	2	0.72	0.79	0.07
IAIC.1	1	14	4	3	0.79	0.84	0.05
IAIC.2	1	13	5	4	0.95	1.00	0.05
IAIC.3	2	9	8	6	0.69	0.95	0.26
IAIC.4	2	14	3	1	0.95	0.95	0.00
IAIC.5	3	12	4	1	0.95	0.89	-0.05
IAIC.6	2	9	8	6	0.90	0.95	0.05
IAIC.7	2	16	1	-1	0.95	0.95	0.00
IAIC.9	3	12	4	1	0.79	0.74	-0.05
IAIC.10	1	11	7	6	0.74	0.90	0.16
IAIC.11	5	11	3	-2	0.74	0.74	0.00
IAIC.12	2	9	7	5	0.63	0.89	0.26
IAIC.13	2	11	5	3	0.72	0.84	0.12
IAIC.14	3	8	8	5	0.74	0.74	0.00

Table 1C Summary of paired surveys to International Interests (II) and Intercultural Awareness and Intercultural Citizenship (IAIC) prompts for American students

Prompt	Count of Likert responses				Percent of Likert response >= 4		
	Decreased	Same	Increased	Incr. - Decr.	Pre-survey	Post-survey	Post - Pre
II.A	3	3	1	-2	0.29	0.43	0.14
II.C	1	2	4	3	0.57	1.00	0.43
II.D	1	3	3	2	0.86	1.00	0.14
II.E	3	2	2	-1	0.00	0.00	0.00
II.F	1	5	1	0	0.00	0.00	0.00
II.G	2	3	2	0	1.00	1.00	0.00
II.H	0	6	1	1	1.00	1.00	0.00
II.I	3	3	1	-2	0.72	0.57	-0.14
IAIC.1	3	2	2	-1	0.14	0.43	0.29
IAIC.2	1	4	2	1	1.00	1.00	0.00
IAIC.3	2	2	3	1	0.71	0.72	0.00
IAIC.4	2	3	2	0	0.72	0.72	0.00
IAIC.5	1	2	4	3	0.86	1.00	0.14
IAIC.6	0	3	4	4	1.00	1.00	0.00
IAIC.7	1	4	2	1	1.00	1.00	0.00
IAIC.9	1	5	1	0	0.86	1.00	0.14
IAIC.10	1	2	4	3	1.00	1.00	0.00
IAIC.11	0	2	5	5	0.29	0.72	0.43
IAIC.12	2	3	2	0	0.57	0.57	0.00
IAIC.13	2	4	1	-1	1.00	0.86	-0.14
IAIC.14	0	3	4	4	0.71	1.00	0.29

Table 2A Summary of paired surveys to English Language Use and Learning (ELUL) prompts

Prompt	Count of Likert responses				Percent of Likert response >= 4		
	Decreased	Same	Increased	Incr. - Decr.	Pre-survey	Post-survey	Post - Pre
ELUL.2A	3	17	6	3	0.77	0.89	0.12
ELUL.2B	4	12	8	4	0.67	0.73	0.06
ELUL.2C	3	14	9	6	0.85	0.92	0.08
ELUL.2D	1	18	7	6	0.81	0.96	0.15
ELUL.2E	2	16	8	6	0.89	0.92	0.04
ELUL.2F	6	13	7	1	0.81	0.77	-0.04
ELUL.2G	3	16	6	3	0.85	0.96	0.11
ELUL.2H	6	14	6	0	0.81	0.77	-0.04
ELUL.2I	4	13	8	4	0.46	0.48	0.02
ELUL.3A	2	9	9	7	0.56	0.91	0.34
ELUL.3B	1	13	5	4	0.63	0.86	0.23
ELUL.3C	2	10	10	8	0.73	0.82	0.09

Table 2B Summary of paired surveys to English Language Use and Learning (ELUL) prompts for Iraqi students

Prompt	Count of Likert responses				Percent of Likert response >= 4		
	Decreased	Same	Increased	Incr. - Decr.	Pre-survey	Post-survey	Post - Pre
ELUL.2A	3	11	5	2	0.74	0.89	0.16
ELUL.2B	4	7	6	2	0.59	0.68	0.10
ELUL.2C	1	9	9	8	0.79	1.00	0.21
ELUL.2D	1	11	7	6	0.79	1.00	0.21
ELUL.2E	0	11	8	8	0.89	1.00	0.11
ELUL.2F	4	8	7	3	0.84	0.84	0.00
ELUL.2G	3	10	5	2	0.84	1.00	0.16
ELUL.2H	5	10	4	-1	0.90	0.79	-0.11
ELUL.2I	2	10	6	4	0.42	0.50	0.08
ELUL.3A	2	6	8	6	0.53	0.89	0.36
ELUL.3B	1	12	2	1	0.71	0.88	0.18
ELUL.3C	2	10	6	4	0.74	0.78	0.04

Table 2C Summary of paired surveys to English Language Use and Learning (ELUL) prompts for American students

Prompt	Count of Likert responses				Percent of Likert response >= 4		
	Decreased	Same	Increased	Incr. - Decr.	Pre-survey	Post-survey	Post - Pre
ELUL.2A	0	6	1	1	0.86	0.86	0.00
ELUL.2B	0	5	2	2	0.86	0.86	0.00
ELUL.2C	2	5	0	-2	1.00	0.71	-0.29
ELUL.2D	0	7	0	0	0.86	0.86	0.00
ELUL.2E	2	5	0	-2	0.86	0.72	-0.14
ELUL.2F	2	5	0	-2	0.72	0.57	-0.14
ELUL.2G	0	6	1	1	0.86	0.86	0.00
ELUL.2H	1	4	2	1	0.57	0.72	0.14
ELUL.2I	2	3	2	0	0.57	0.43	-0.14
ELUL.3A	0	3	1	1	0.67	1.00	0.33
ELUL.3B	0	1	3	3	0.43	0.75	0.32
ELUL.3C	0	0	4	4	0.71	1.00	0.29

List A Codes for survey prompts

- II.A : If I were to go to a foreign country, I would go to one where most people speak and understand my language.
- II.C : Knowledge of other cultures helps me better understand my own.
- II.D : Learning about another culture or learning another language will better prepare me for the global workforce.
- II.E : I think there would be difficulties working on projects with someone from a different cultural background than my own.
- II.F : Because I speak English, I don’t have to learn a new language to be able to speak with people from other countries.
- II.G : Cultural differences can be understood if people are open minded about people from other cultures.
- II.H : I have a lot to learn from people with cultural backgrounds that are different than my own.
- II.I : I purchase songs or CDs with music that is sung in a language other than my own.
- IAIC.1 : To be able to communicate with someone in another language you have to understand their culture.
- IAIC.2 : Learning culture is part of learning another language.
- IAIC.3 : It is important to understand my own culture when learning another language.
- IAIC.4 : Learning another language means learning new kinds of behavior, beliefs and values.

- IAIC.5 : Languages can be linked to many different cultures (e.g. the English language can be used to express the cultures and countries in which it is used such as India, Singapore, China).
- IAIC.6 : Cultures may be defined and understood differently by different groups and individuals.
- IAIC.7 : It is important not to judge people from other cultures by the standards of my own culture.
- IAIC.9 : Experience of intercultural communication (at home or abroad) is important for becoming a global/ intercultural citizen.
- IAIC.10 : Learning about other cultures is important for becoming a global/ intercultural citizen.
- IAIC.11 : Speaking English is important for becoming a global/ intercultural citizen.
- IAIC.12 : Speaking other foreign languages is important for becoming a global/ intercultural citizen.
- IAIC.13 : An interest in global social issues (for example poverty, environmental protection, democracy, racism) is important for becoming a global/ intercultural citizen.
- IAIC.14 : Taking an active role in global social issues and trying to improve the world is important for becoming a global/ intercultural citizen.
- ELUL.2A : English will allow me to get good grades at university.
- ELUL.2B : English will allow me to pass exams.
- ELUL.2C : I will need English for further study.
- ELUL.2D : I will need English for my future career.
- ELUL.2E : English will allow me to meet and communicate with more and varied people from many different cultures.
- ELUL.2F : English will allow me to travel to many different countries and to learn about different cultures.
- ELUL.2G : English will allow me to meet and communicate with native speakers of English.
- ELUL.2H : English will allow me to have a fun and enjoyable experience.
- ELUL.2I : Other people will respect me more if I have knowledge of the English language.
- ELUL.3A : How well do you think your English classes have prepared you for studying abroad?
- ELUL.3B : Do you feel well prepared for intercultural communication when studying abroad?
- ELUL.3C : Do you hope to develop an identity or feeling of intercultural/global citizenship when you are abroad?

Appendix B

Client: Salwa Al-Noori and Sami Al-Hasnawi
Consultant: Seth D Temple
Date: 11/24/2021

Dear Salwa and Sami,

This report concerns pre- and post-course surveys from a COIL course in Winter 2021. It describes pre-processing of the survey data and recommendations for data analysis and presentation.

Pre-processing Survey Data

Survey data collected via the Google Forms API requires cleaning and data visualization outside the scope of the default plots. Cleaning the data manually in Excel or programmatically in R/Python can be achieved with a couple hours' time. I performed the following data cleaning steps:

- String preprocessing
    - Made all strings lowercase (e.g., "Agree" to "agree")
    - Created column for university
- Likert scale preprocessing
    - Kept highest scale response (e.g., "Agree; Strongly agree" to "Strongly agree")
- Removed duplicate entries
    - Kept most recent timestamp
- Paired pre- and post-course survey responses
    - Twenty-six paired responses
    - Sorted data frames by unique identifier
- Corrected mistyped identifiers
- Relabeled columns and generated a data dictionary

Not all students responded to both the pre- and post-course surveys. As a result, the paired response data is smaller in size than the original survey data. This observation is noteworthy because the original sample size is small. With only twenty-six paired responses, and considerable heterogeneity among the students, formal hypothesis testing is not advised.

Presenting Demographic Data

Most of the demographic data can be visualized as pie charts. However, such figures may occupy space in a publication while not adding significant value to the research. I recommend that the demographic data is summarized in a written paragraph to minimize its contribution to the paper length. For instance, the twenty-six paired responses come from 21 females and 5 males and 19 University of Al-Qadisiyah students and 7 University of Washington students. The University of Washington students were undergraduates mostly in the biological sciences and the University of Al-Qadisiyah students were mostly graduate students in English. Other sources of heterogeneity, say race or ethnicity, are strongly correlated with the university covariate; the University of Q students identified as Middle Eastern (Arabic; Turkish; Persian) whereas the University of Washington students identified as White or Asian/Pacific Islander. (There was a Hispanic UW student in the course, but that student did not complete the post-course survey.) These demographics challenge any formal hypothesis testing in that the covariates ought to be addressed to avoid confounding results. You can describe this data in an introductory paragraph where you set the scene for the COIL class and its objectives.

COIL Course as a Treatment Effect

The COIL program has defined learning objectives related to international interest, intercultural awareness and citizenship, and English language learning and use. These objectives inspired the Likert-scale items in the pre- and post-course surveys. Some Likert-scale items were asked in both surveys, in which case the course may serve as a treatment affecting the survey responses. I caution against any formal hypothesis testing for this treatment effect. For one thing, introducing statistical hypothesis testing to the publication could create confusion. Additionally, each hypothesis test is correlated in that the same students are responding to different survey items. Instead, I provide summary tables (end of document) that indicate the effect of the COIL course on the paired responses. These table focuses on (1) if post-course responses reach a greater value on the five-point ordinal Likert-scale and (2) if the percentage of “Agree” or “Strongly agree” responses changes between the surveys. In all but one case more post-course responses increment on the Likert-scale as opposed to decrementing. Similarly, the percentage of “Agree” or “Strongly agree” responses increases more often than not, although some items achieved high agreement from the onset. These findings in aggregate support the idea that the COIL course achieved its learning objectives.

Based on the paired responses, I conducted some tests for the null hypothesis that responses were as likely to increment as they were to decrement on the ordinal Likert scale. These were trinomial tests (article, blog), a special type of the sign test. There is evidence for some of the survey responses improving in the post-course survey, meaning the COIL course had some impact on some of the students. As said above, this statistical result is difficult to communicate; this is why I suggest presenting the summary table as a more compelling piece of evidence for the COIL program’s impact on the students.

Data Visualization

I created some R code to create pie charts for the demographic covariates. I have also explored some data visualization for the Likert-scale items in Excel. Some examples of how to present Likert-scale items can be found here: 4 ways to visualize Likert scales, Sharing course evaluation data. On request, I can generate figures to accompany your publication; however, unless requested, I do not recommend a data visualization for the survey data. Any summary figure for the surveys should only present a minimal, curated list of the Likert item responses so as to not overwhelm the audience.

Data Dictionary

I recorded the survey questions with unique abbreviations. This step is to have a data dictionary to compare survey responses between the different surveys. Attached to this correspondence is the data dictionary. For example,

ELUL.2A	It will allow me to get good grades at university.
ELUL.2B	It will allow me to pass exams.
ELUL.2C	I will need it for further study.

Sincerely, Seth D Temple

Appendix C

Client: Salwa Al-Noori and Sami Al-Hasnawi
Consultant: Seth D Temple
Date: 12/06/2021

Dear Salwa and Sami,

This report concerns preprocessing of the Padlet text data and some exploratory data analysis (EDA) for the COIL course in Winter 2021. The EDA is meant to present some early ideas about how to analyze and present text data. In our next meeting, we will want to define explicit research goals for the text data analysis. Here is an informative blog post summarizing natural language processing (NLP), the scientific field by which scientists try to make sense of text. I strongly encourage you to read the sections 'Common NLP Tasks & Techniques' and '11 Common Examples of NLP'. These sections will facilitate brainstorming about analysis directions and aims.

Preprocessing text data

Text data is messy! Before any text data analysis, it is common practice to take various preprocessing steps to clean the text data. The Padlet text data is especially challenging to work with because of:

1. Informal language
2. Varying levels of English proficiency
3. No editing for grammar and spelling
4. Images, emojis, hyperlinks, non-English words included.

(Published books or journal articles would be an example of an easier text dataset to work with.) Given these challenges, I took the following measures to clean the text data. These steps are informed by my prior research experience with NLP libraries in Python.

- Removed all emojis, images, URLs, and non-English words that involved a different alphabet (e.g. Farsi or Arabic words)
- Cut some generic pleasantries like 'Hello, …', 'Hi everyone, …', and so on
- Made all words lowercase (e.g. 'Iraq' is the same as 'iraq')
- Replaced ! and ? by . (periods denote sentences in NLP toolkits)
- Placed some periods in some text where they were obviously missing
- Replaced punctuation with whitespace (punctuation is hard to analyze programmatically)
- Replaced Dr., Prof., U.S. with Dr, Prof, and US to address the period issue above

These changes ensure that the text data can be interpreted as Unicode characters (ASCII or UTF) and that complete ideas can be encapsulated by periods.

I copied, pasted, and saved Padlet posts in the shared PDF files into *.txt files. This process required 2-4 hours, but I don't know of any easier way to read in PDFs into the correct formats. During this process, I noticed a few cases in which student names were not anonymized. I anonymized these names in my files when I uncovered them, but please be aware that not all student names are anonymized in the files you sent me. After creating these PDF files, I developed two scripts to store posts and sentences into *.csv files. Comma-separated value formatted files have a tabular design and play nice with common machine learning packages in Python like pandas. These two scripts and other code developments I have put in this online repository: https://github.com/sdtemple/coilnlp. You can access these *.csv files and *.txt files in this GDrive folder: text data.

Exploratory data analysis

Attached is a Jupyter Notebook where I demonstrate some exploratory data analysis of the text data. I conducted this EDA using functions I wrote in the 'functions.py' at the GitHub repository. One possible consulting aim could be that I provide a more thorough Python tutorial for this custom package and installations for its dependencies. This may be especially useful because many (unsupervised) methods in text data analysis require post-hoc interpretation, in which case your experiences in linguistics and in teaching would come in handy. Besides the main NLP use cases in the linked above, I could develop tools for:

- Word co-occurrences
- Topic modeling (latent Dirichlet allocation; non-negative matrix factorization)
- Comparing Padlets to an external text data set

I look forward to our research call this week. In clarifying goals and expectations, I can better provide a data analysis service that benefits you two as researchers. By December 17, I have to give a presentation and write-up for course credit; however, as a professional statistician, I do intend to continue to help you two for some time afterward as needed.

Sincerely,

Seth D Temple