CS156 Final Project

Introduction

The dataset that we explored consist of a list of poll answers, their associated HCs/ LOs, and the score, for all the classes in Minerva in the past year. We used this dataset to explore two models: k-means clustering, and topic modeling using Latent Dirichlet Allocation (LDA) with either bags of words or term frequency (TF-IDF) technique, to explore potential hidden semantic spaces in the dataset that we explored.

Data Processing

The data is first downloaded and explored for assessment scores, poll responses, their means and counts. We discard the poll answers that are not scored, and when there are no LOs/ HCs associated with the poll responses. We then extract the text from the poll responses and tokenize and stem the words. Essentially, we split the response text

into individual words, then convert them to their root form, (e.g. ‘discuss’, ‘discussing’, ‘discussed’ and ‘discussion’ all stem from the same word discuss). We also filter the unnecessary words that does not matter in a clustering context (e.g. “I”, “we”, “a”, “the”).

Additionally, the dataset doesn’t provide the associated colleges and classes for the poll response, so we need to map the poll response to the colleges ourselves by matching the LOs and HCs to the associated colleges. To do this, we make use of two additional dataframes provided in our Capstone Seminar classes - the LOs Master List and the HCs spreadsheet. We filter the name and the name, perform basic cleaning, then merge the HCs and LOs dataset together, following by an inner merge with the original dataset. Additionally, the mapping is incomplete because some of the HCs/ LOs change name. For instance, `#objectivemorality` becomes `#objmorality` while `#multiplecauses` become `#complexcausation`. To minimize the changes, we use fuzzy string mapping to find the best match for all of the unmatched HCs/ LOs (provided the match is at least 70%).

K-means clustering

We first do k-means clustering, which assigns each poll to a cluster based on their similarities, which is the Euclidean distance measured in the feature space. Before running the model, we transform the text to a TF-IDF matrix. TF-IDF (term frequency - inverse document frequency) is a numerical statistic that reflects how important a word is to a poll in all answers we get. The value increases as the number of time the word appears in the document but is offset by the number of time it appears in all documents, thus adjusting for the fact that some words just appear more frequently in general. This gives more accurate measurement for K-means clustering because it’s not just dependent on words frequency but word frequency specific to the document itself. This can be seen from the word cloud visualization betwen bags of words and TF-IDF technique.

We see that both have similar common words (because, also, etc.), however TF-IDF information is more useful overall. Bags of words only show the most common words e.g. think, also, like, one, example, etc. which is used across major, whereas TF-IDF highlights things such as active, company, bias, argument, etc. which seem to be more specific to majors.

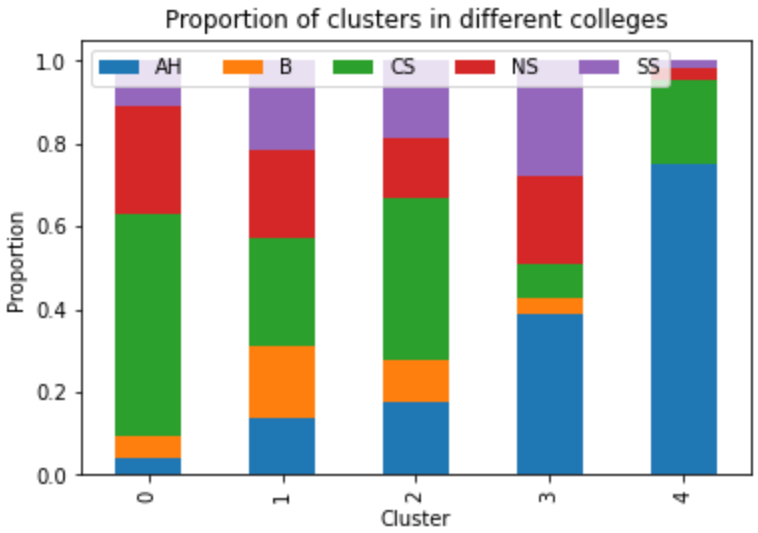
5 clusters

We first run a 5-cluster model, which is most appropriate because there are 5 colleges in Minerva, so the clusters would probably center around those words. We also get the 15 most common words from each clusters.

```
Top terms per cluster:

Cluster 0 words: data, variabl, model, sampl, use, distribut, would, test, hypothesi, prob
abl, observ, valu, mean, studi, differ,
Cluster 1 words: would, becaus, compani, market, use, time, one, valu, chang, product, inc
reas, countri, also, need, could,
Cluster 2 words: poll, complet, student, present, fac, facil, facia, facial, facialex, fac
ialexpress, facien, faciilit, facilat, facevook, facili,
Cluster 3 words: use, think, differ, one, peopl, problem, would, understand, becaus, make,
exampl, level, way, help, system,
Cluster 4 words: argument, thesi, sentenc, evid, induct, logic, deduct, premis, conclus, t
rue, statement, use, truth, claim, clone,
```

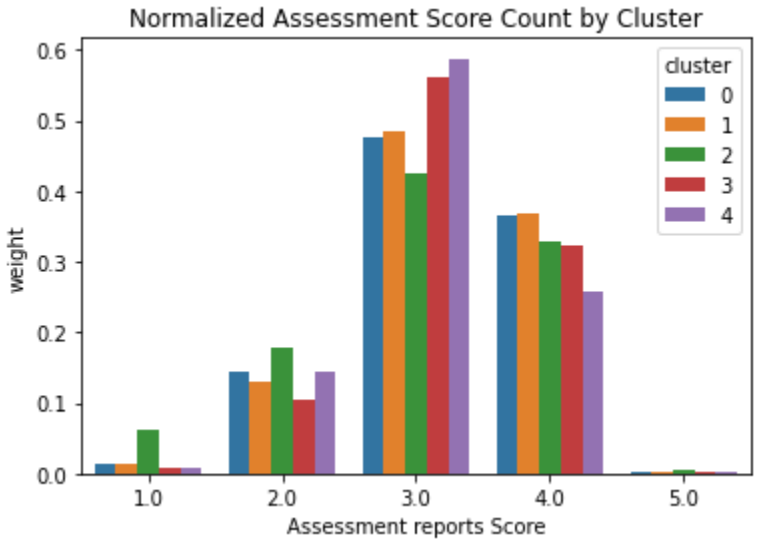
Based on the top term, it seems that cluster 0 maps best to CS major (with word like data, variable, distribution, etc.), cluster 1 maps best Business major (Company, market, product, country, etc.), cluster 3 seems to be the Social sicence cluster (people, system) and cluster 4 seems to be related to Computer Science formal analysis or Arts and Humanities major (premise, conclusion, logic, etc.).

Because we can map each poll response to a cluster, we can then map the proportion of colleges in each cluster.



We see that the mapping is pretty accurate for cluster 0 (majority CS) and cluster 4 (majority AH), but a bit more mixed for the other clusters. Note that Business is a small portion for all clusters because we don't have cornerstone course in Business, thus all the HCs scores will not have the Business college tag. As a result, the overall counts in Business is smaller.

We also use that to check the assessment score count by cluster, to see whether different clusters leads to different grade distribution.



We can see that clusters 0 and 1 is associated with higher ratio of 4s to 3s, while cluster 4 has the highest proportion for 3. It seems like Arts and Humanities professors are more likely to grade students 3s than CS professors!

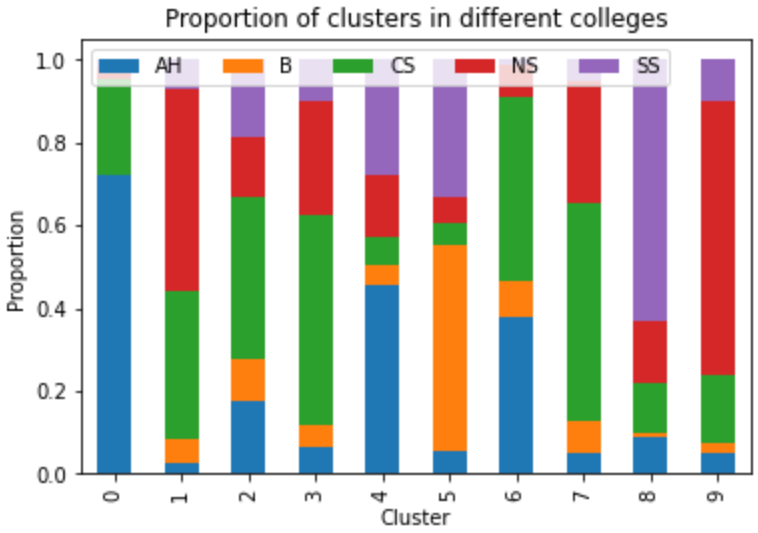
There's also an interesting peak for score 1 for cluster 2, which is probably because the phrase "The student is present but did not complete the poll" is in this cluster (and are the most frequent words), which probably is associated with more 1.

10 clusters

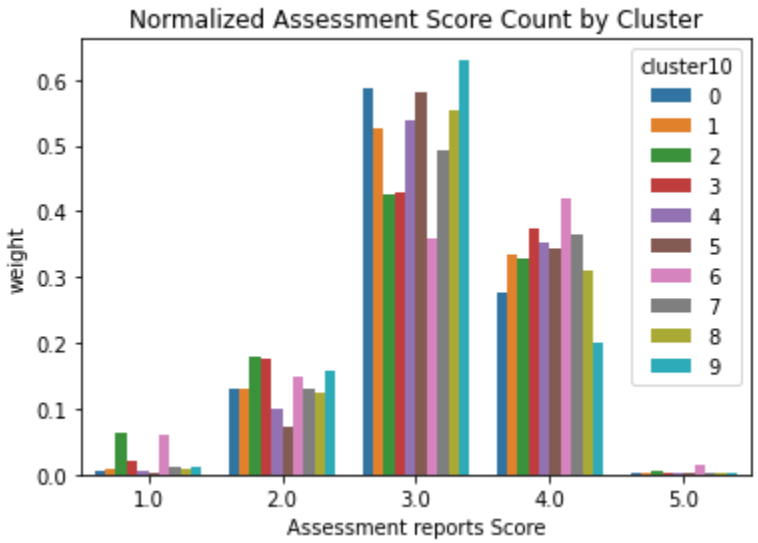
We also tried running the algorithm with 10 clusters instead of 5 and got the following results:

```
Top terms per cluster:

Cluster 0 words: argument, thesi, evid, induct, deduct, premis, conclus, use, statement, c
laim, support, clone, reason, make, valid,
```

```
Cluster 1 words: variabl, studi, hypothesi, test, treatment, control, observ, group, effec
t, experi, would, differ, confound, use, result,
Cluster 2 words: poll, complet, student, present, fac, facil, facia, facial, facialex, fac
ialexpress, facien, faciilit, facilat, facevook, facili,
Cluster 3 words: would, valu, use, probabl, number, time, becaus, function, distribut, mea
n, one, model, sampl, first, get,
Cluster 4 words: peopl, use, think, would, one, becaus, differ, make, like, way, also, exa
mpl, understand, could, work,
Cluster 5 words: compani, market, product, custom, risk, countri, busi, invest, cost, econ
omi, price, would, growth, increas, financi,
Cluster 6 words: doc, https, com, googl, edit, document, usp, share, kgi, edu, minerva, sp
readsheet, drive, colab, gid,
Cluster 7 words: data, model, use, line, would, observ, graph, hypothesi, set, predict, co
llect, differ, regress, variabl, point,
Cluster 8 words: system, level, emerg, agent, interact, individu, properti, network, compl
ex, differ, predict, behavior, analysi, model, social,
Cluster 9 words: problem, solut, solv, constraint, water, use, rightproblem, identifi, sta
te, differ, goal, breakitdown, subproblem, step, one,
```



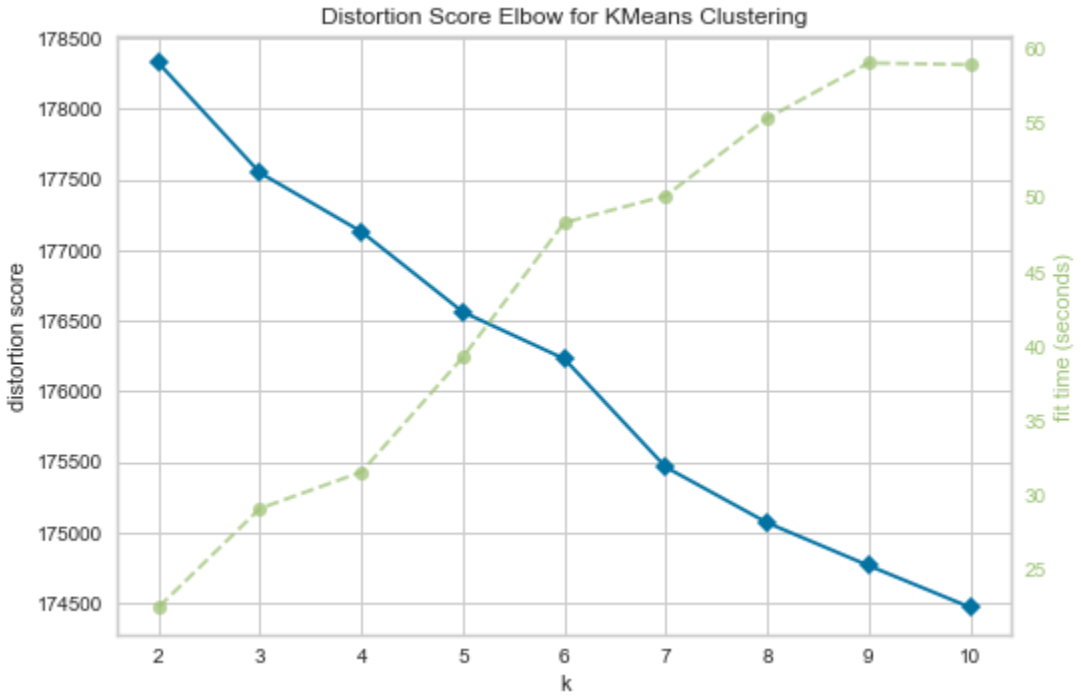
Interestingly, while the clusters contain more noise overall, some clusters do get better at classifying different colleges. We can more clearly identify cluster 0 as AH, cluster 9 as NS, cluster 5 as Business, cluster 8 as SS. We also get cluster 1 and 7 as a combination of NS and CS majors.



Once again, the normalized assessment count tells the same story: cluster 9 has the most 3s in proportion, followed by cluster 0, then 5, corresponding to NS, AH and Business major accordingly. Meanwhile, if the poll response is in cluster 6, it's more likely to get 4 and 5, but also more likely to get 1! The explanation is probably that the associated strings `doc, https, com, googl, edit, document` is probably associated with when students have to paste the link of their pre-class work, which gives more time (and words) to get high scores, but can also get low scores if students are not well prepared.

Selecting the best number of clusters

We also run a K-Ellow Visualizer to find the best number of clusters. It runs the algorithm for all values of k for 2 to 10, then compute an average distortion score for all clusters, which is the sum of square distances from each point to its assigned center. We are looking for an inflexion point beyond which the distortion score stays constant.



While the elbow point is not clear, it seems that 7 is the best number of clusters. We run K-means again with 7 clusters to see if we get better results.

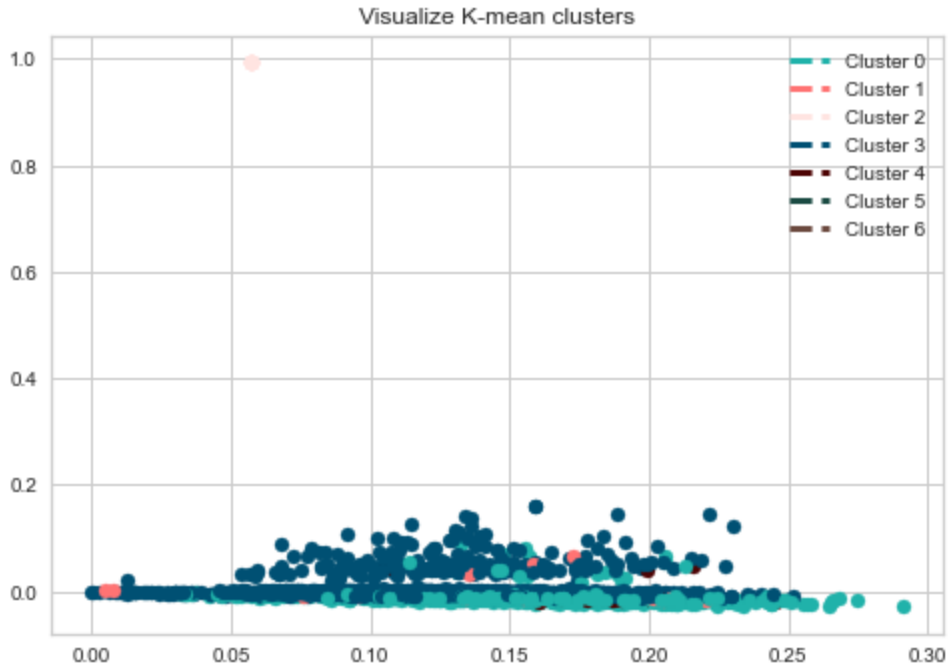
```
Top terms per cluster:

Cluster 0 words: problem, solut, solv, constraint, water, use, rightproblem, differ, ident
ifi, state, goal, breakitdown, subproblem, step, one,
Cluster 1 words: poll, complet, student, present, faazillexzvnfbceqmeprnivsrzaadmewqcnu, f
abian, fabianokafor, fabiola, fabl, fabric, fabul, fac, faabi, facad, facbook,
Cluster 2 words: peopl, use, one, think, would, becaus, differ, make, exampl, argument, un
derstand, system, also, way, like,
Cluster 3 words: data, variabl, model, hypothesi, studi, observ, use, test, would, predic
t, differ, control, treatment, one, regress,
Cluster 4 words: doc, https, com, googl, edit, document, usp, share, kgi, edu, minerva, sp
readsheet, drive, colab, gid,
Cluster 5 words: compani, market, product, custom, risk, countri, busi, economi, cost, inv
est, would, price, growth, financi, increas,
Cluster 6 words: would, valu, use, probabl, time, becaus, number, function, mean, distribu
t, one, sampl, get, first, node,
```

There seems to be no significant improvement for 7 clusters compared to either the 5 or 10 clusters. It is also possible that the word clusters are not just dependent on majors, but also cultures, grades, English capacity, etc. But if we are judging based on the

proportion of clusters in different colleges alone, the 7 clusters seem to do a mediocre job.

Visualizing K-means result

We wanted to do Principal Component Analysis and plot the clusters, but as the TF-IDF matrix is a sparse matrix, we used an alternative technique, `TruncatedSVD`, which performs dimensionality reduction for sparse matrices. Even after using this, however, our matrix was still too heavy, so we slice to 3000 datapoints. However, the result doesn't seem to separate as well as we have hoped for. This might be due to the slicing, or a non-optimal the dimension axes used.

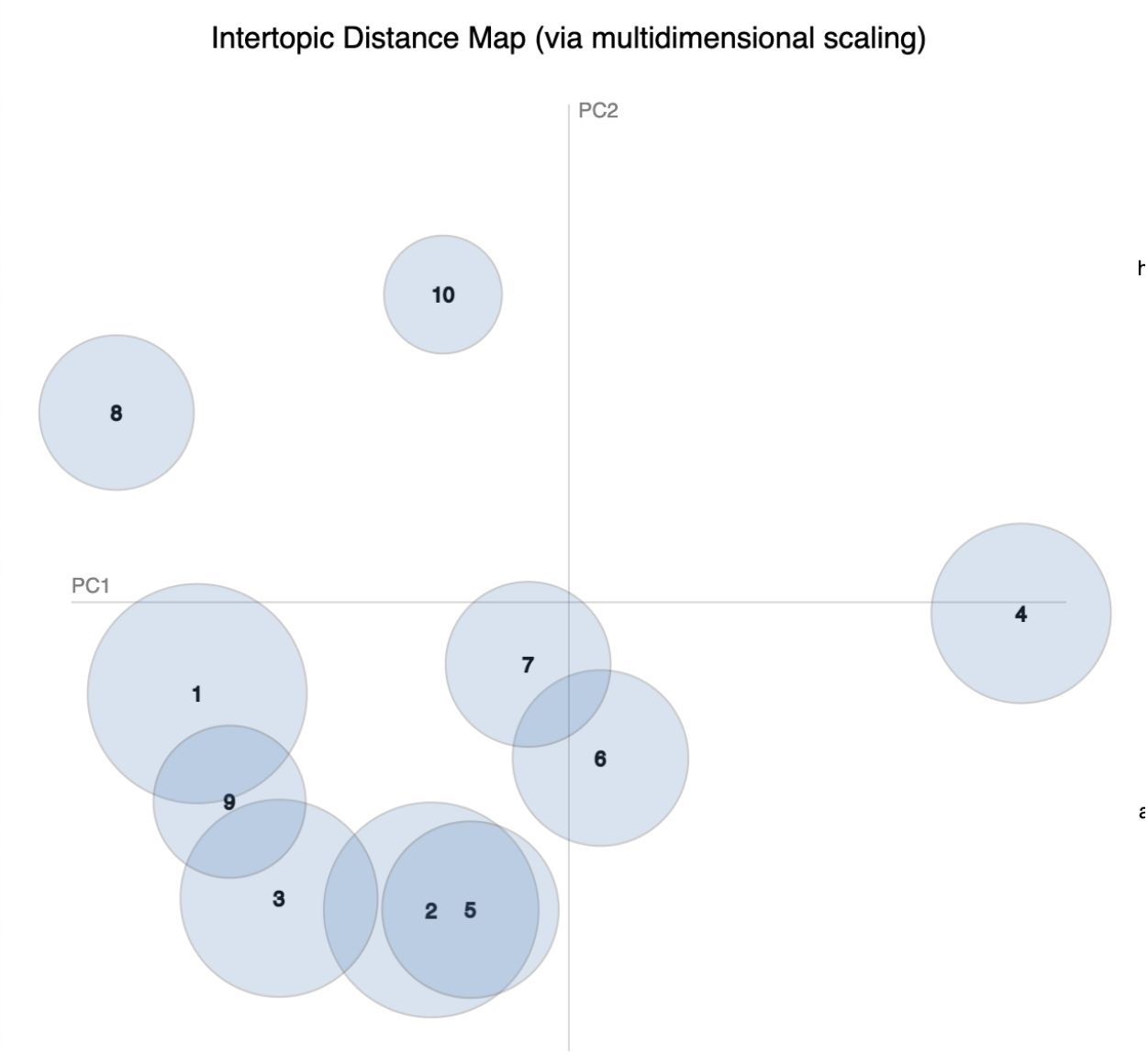


Topic modelling

K-means clustering problem is that it only assign each poll response to one cluster. Topic modeling allows for more flexibility by assigning each document to a probability of belonging to each topic. We use Latent Dirichlet Allocation (LDA) with both bags of words or term frequency (TF-IDF) technique.

Bags of word technique

Bags of word is essentially using the cleaned response directly and mapping it to the LDA model. As a result, our clusters are not very separated.

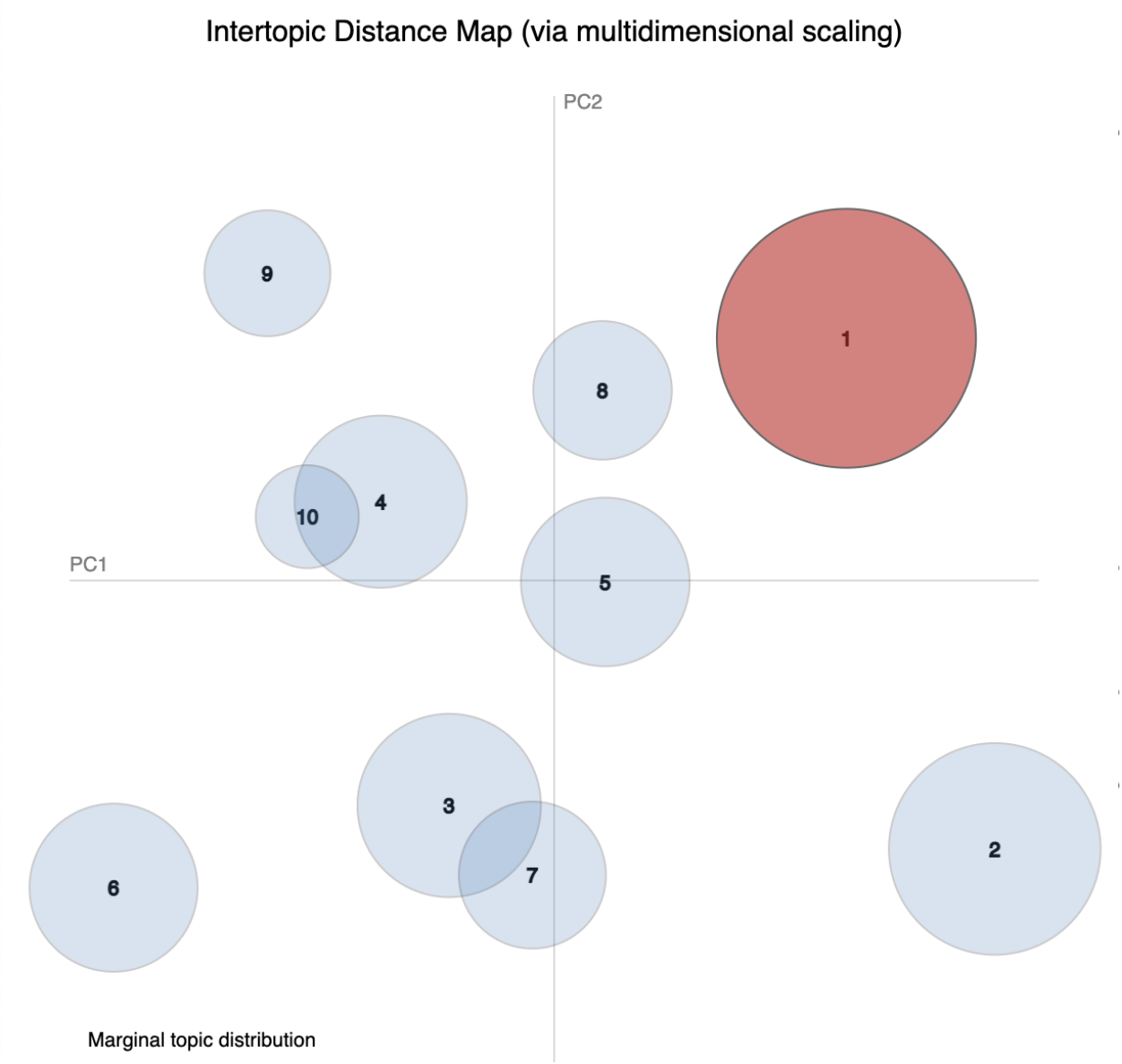


We can also see the most frequent word within each cluster. This is determined by a parameter λ, which is between 0 and 1. The closer λ is to 1, the values will be ranked according to their frequency in the overall paper (without being exclusive to only the cluster itself), whereas the closer λ is to 0, the values will be ranked according to how exclusive they are to the cluster. That means the higher the λ, the more representative the sample (since it's determined by the frequency), but the less descriptive it is. Meanwhile, smaller λ can highlight interesting connections, but also can be noisier. Thus, we choose λ to be equal to 0.5 to balance between these two considerations.

For instance, for cluster 4, the most common values seem to be variables in math equations, suggesting this is a cluster for math responses (as calculation, regression, matrix, equation, vector are all present). Cluster 8, meanwhile, have terms like emotion, music, poem, etc. which all tend to the Arts and Humanities college. Other clusters, however, are not as clear. For instance, cluster 1 has words like use, make, context, argument, moral, truth, which suggests that it might be Arts and Humanities, but the words are too general to be useful overall. Similarly, cluster 9 has social and group, culture, which makes us think it might be AH, but also has contrarian (a social science HC) and narwhal (a CS reference).

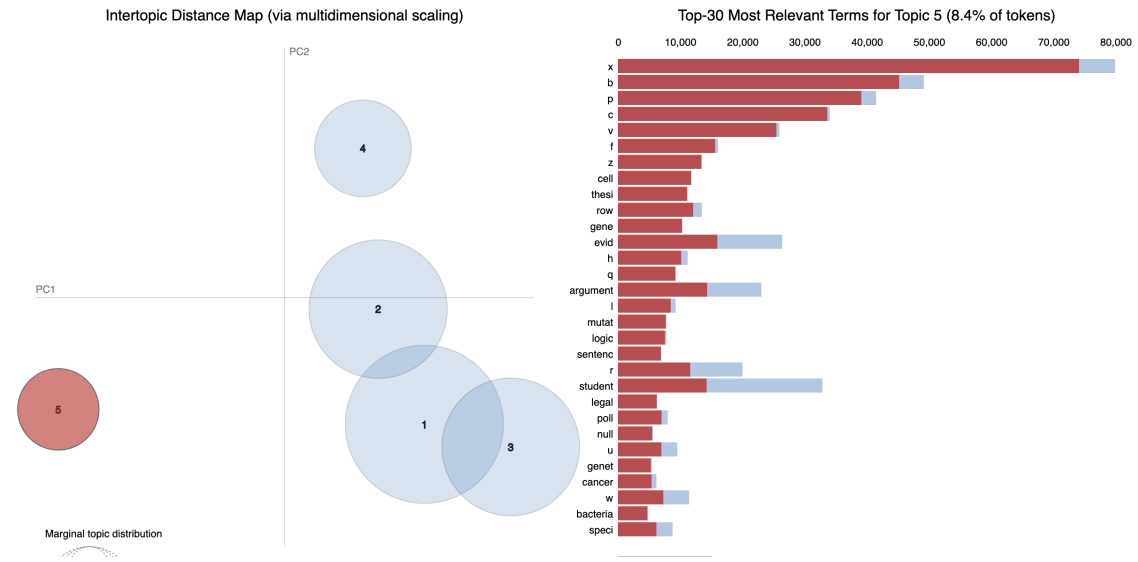
TF-IDF technique

We try the topic modeling again with TF-IDF technique, which takes into account the frequency of the word in each response offset by overall frequency in all the polls. We were able to achieve much better separation, partly because the technique itself makes for a better measurement.



Here, we see a few clear clusters. For instance, 1 contains ‘company’, ‘market’, ‘product’, economy, business, growth, price, brand, all of which suggest it’s a business-related cluster. 2 meanwhile has audience, music, story, narrative, thesis, art, etc. which is a humanity cluster. Cluster 6 contains p, n, treatment, hypothesis, sampling, which means this is probably either statistics or natural science. There’s also gist and github which suggests it’s more likely to be a statistics/ CS cluster. 9 has mineral, earth, atmosphere, ant, carbon, etc. suggesting it’s a NS cluster, while 4 is likely to be a mathematics-related cluster (x, vector, equation, etc.). The other clusters are to be harder to predict, as they contain a mix bag of results.

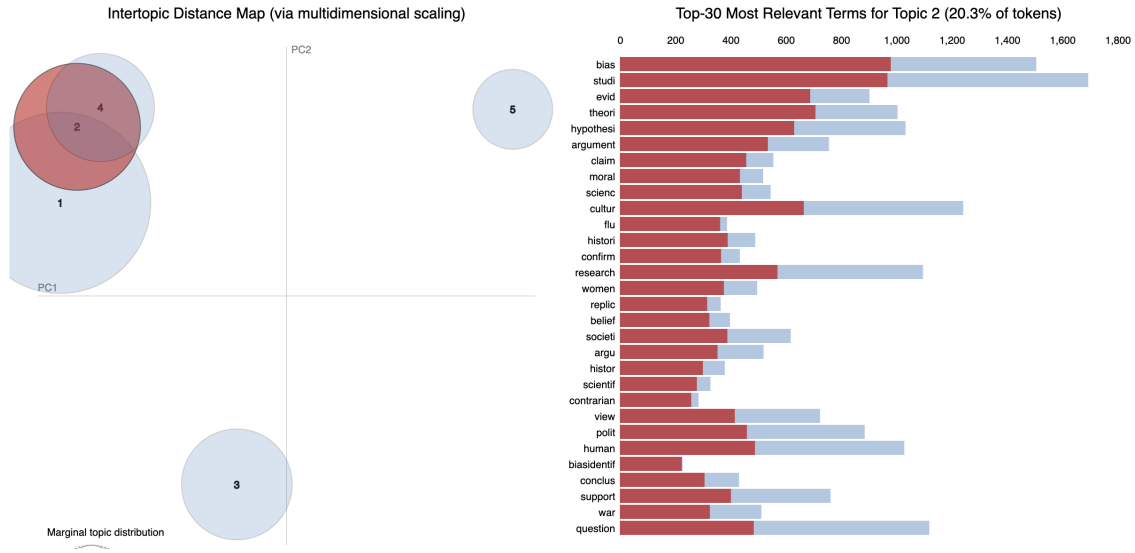
So maybe a problem might be that we have too many clusters, which make the results more diluted. We tried the topic modeling again with 5 clusters.



Once again, the result is not that clear with bag-of-word technique. The only differentiated cluster (5) is related to mathematics, but other clusters such as 4 and 2 also appear to be Computer science clusters. 3 is more likely to be business, but words like ‘policy’, ‘develop’, ‘resource’ and ‘economy’ suggests it can also be a mixture with the Economics social science classes.

5 clusters

The TF-IDF technique however, does a lot better with just 5 clusters.

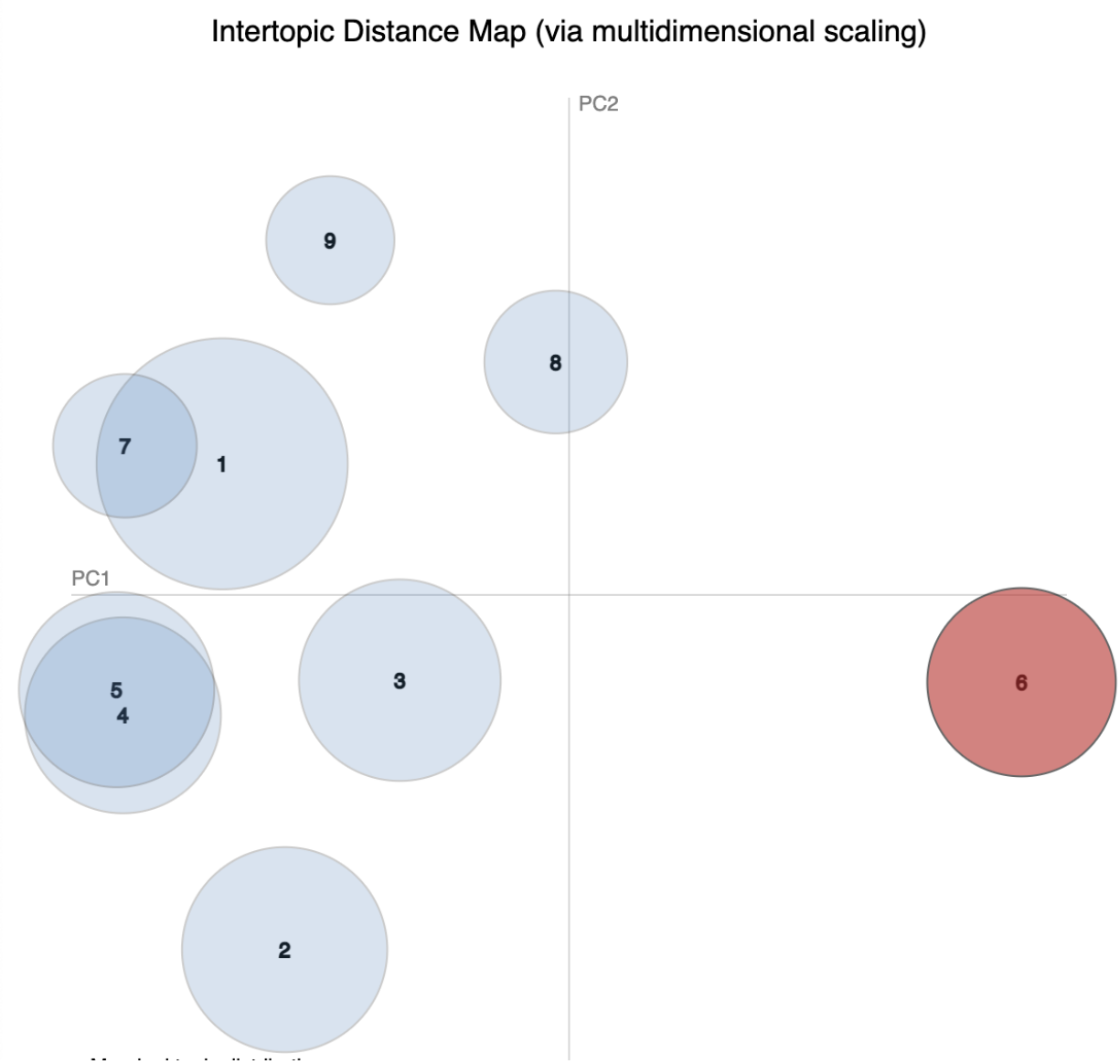


Even though the clusters do not look as separated, they actually map really well to the colleges. Cluster 5 contains “cells”, “gene”, “energy”, “entropy”, “mutation”, suggesting it’s a Natural science cluster, while cluster 3 contains the different variables for mathematics, as well as algorithm, google docs links, which is typical of a Computer science cluster. 1, 2 and 4 overlap a lot, but are actually pretty distinctive. 1 includes “company”, “market”, “product”, “strategy”, which suggests it’s clearly a Business cluster, while 4 has “music”, “audience”, “language”, “emotion”, “words” which makes it a Arts and Humanities cluster. 2 is the most unclear, with word such as “bias”, “study”, “evidence” “theory” which can belong to any major, but also there’s words like “culture”, “history”, “women”, “politics”, “war” that makes it somewhat indicative of Social science. Also, since social science seems to be the most multi-disciplinary, it’s probably harder to classify.

Finding the best number of clusters

To select the best number of cluster, we use coherence score to measure how interpretable the topics are to humans by measuring how similar the words are to each other. We use the CV coherence score, which is the default metric for valuation.

We first compare the coherence score between bags of words and TF-IDF and get a slightly better coherence score for TF-IDF (0.3885 vs 0.3381), which aligns with our observations when examining the data manually. Next we test the coherence score with different number of clusters, from 2 to 10, and get the highest score for 9 clusters (0.395).



It’s interesting that from PCA, it doesn’t seem better separated. Looking at the clusters indivudally, we do see the differences e.g. 6 is a Math/ CS cluster, 2 is Business cluster, 3 is a SS complex system cluster, 1 seems to be a mixture between links and hypothesis testing/ scientific method, 5 is a politics SS cluster, and 4 is more of a economics SS cluster. Others are more difficult to differentiate (e.g 8 is both CS/ SS brain major, 7 seems to be AH but also have Github inside, and 9 just seems completely random), and there’s no clear cluster for NS major. Once again, judging from observation, it seems that the 9 clusters do not do better than the 5 clusters.

Results

The results are shown in an interactive website here.

Conclusion

K-means clustering and LDA provide two interesting ways of looking at the same set of data and finding clusters of words surrounding the topic of interest. It's hard to judge their accuracy beyond human examination, but it's interesting that both seem to pick up on the same patterns of words especially for CS, Business and AH majors, but struggle more with NS and SS.

The next finding we have is that finding optimal number of clusters are not as accurate as we have expected, especially when comparing with manually chosen clusters. It's probably because in our cases, we already have a predisposition as to how many clusters the words should have, which is more accurate than the calculation of an approximation of the metrics we want. It's also possible that because we don't have a clear metric of success beyond "looking at the words and seeing if the clusters are interpretable", thus manually chosen clusters might feel more accurate despite performing objectively better in certain metrics if we had calculated. But since natural language processing is just an extension of human language processing, and given we are working with manageable datasets, we feel this is sufficient for exploration, but can potentially made more rigorous if we want to do further hypothesis testing.

Another interesting finding is that, judging by observation alone, k-means seems to get more concentrated clusters with higher number of clusters, while LDA seems to get better results with fewer clusters. A potential explanation is that since k-means only assign each response to 1 cluster, the more clusters there are, the better divided the space is, and thus some clusters get more accurate (even though you also get more noises in other clusters). Meanwhile, LDA get better response when there are fewer clusters due to the percentage assigned model - the more clusters, the more diluted the percentage, the more inaccurate the result will be.

In conclusion, this is an interesting exploration of topic modeling. In the process, we have learnt that 'coffee' and 'starbuck' is highly associated with CS majors, and 'vietnamese' is a word that frequently appears in the AH cluster. The things you find when you look at words closely...

References

BoW Model and TF-IDF For Creating Feature From Text. (2020, February 27). *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/

CS221. (n.d.). Retrieved April 21, 2022, from https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

Kapadia, S. (2020, December 29). *Evaluate Topic Models: Latent Dirichlet Allocation (LDA)*. Medium. https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

Kulshrestha, R. (2020, September 28). *Latent Dirichlet Allocation(LDA)*. Medium. https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2

LDAvis Demo. (n.d.). Retrieved April 21, 2022, from http://www.kennyshirley.com/LDAvis/#topic=0&lambda=0.01&term=

Zvornicanin, E. (2021, December 7). *When Coherence Score is Good or Bad in Topic Modeling?* | *Baeldung on Computer Science*. https://www.baeldung.com/cs/topic-modeling-coherence-score

Appendices

HCs Appendix

- #differences: One thing that makes the team efforts works really well is because we leverage our different abilities and skills to help each other complete the work. Esther is amazing at data science, telling stories and doing complex models, so she sets the direction for the project, thinks of different models to try and does a lot of the technical work. Ha is weaker in data science but stronger in writing and software engineering, so she helps with data cleaning and exploration, interprets the findings, writes the report and cleans up the website and Github. By focusing on each person's strength, we are able to create a coherent and clear report with a lot of technical analysis to back up.
- #responsibility: We were both very commited and proactive in the assignment. This is especially difficult since Esther is currently in Taiwan, and Ha is in San Francisco, so Esther goes to sleep when Ha starts working, and vice versa. To make this happen, we do check-in at the beginning and end of each person's work day, leave

many messages and updates along the day. We also check-in with prof to make sure we are on the right track and get feedback for our final project.

Code

All accompanied code is found on this Github repo. We have not added the dataset in the repo for security reason, but it will be in the zip files uploaded in the assignment submission.

We also downloaded the code and added it as PDF file in the next page.