Project Proposal:
Latent Semantic Indexing

Cole Helgaas, Steven Jin

Oct 12, 2022

Instructions: Your proposal will consist of 3-4 paragraphs addressing the prompts below, followed by 4-5 formal references, entered into a bibtex file. These references should be reputable sources (no wikipedia pages, youtube videos, or blog pages). The template will automatically make these references appear at the end of this proposal (you can comment out the example references in the bibtex template file).

1. Provide a description of topic. Explain your topic, its broader impact, and the role that SVD plays in this application.

   Our project topic is Latent Semantic Indexing (LSI) which is a topic clustering algorithm. The algorithm count vectorizes documents to create a document-count matrix. It then runs SVD on the document-count matrix. From the output of SVD, we find:

   (a) topic-word associations

   (b) word-topic associations

   (c) low-rank representations of documents.

   SVD was originally created as a information retrieval algorithm, but has since been applied in ....

2. Describe what you intend to accomplish in your paper. Look back on the SVD project page at the components every project must include. Explain what modifications/extensions you will do in this project that is different that what you are seeing in resources (how will this project be different/unique to your team?).

   In this project, we run LSI on course descriptions for courses offered at Middlebury for the 2022-23 academic year. We then will analyze the results to see if low-rank representations of course descriptions capture high-level attribute of courses. For example, we will look at whether low-rank representations discriminate STEM vs non-STEM classes. Further, we will look for algorithmic bias by looking at word-topic associations for words associated with historically marginalized groups. Finally, something about word clouds. **TODO**

3. Give some ideas for what your team might do for the live script portion of this project.

   For the live script portion of this project, we will create a live script projects course descriptions into a 2-dimensional space. This will allow users to see how different low-rank representations cluster. Further, we will allow users to draft their own course description and use a nearest-neighbor algorithms to predict the department of the class. Both these examples will show that LSI can (hopefully) capture semantic ideas.

4. **DEI + J:** Think about the work we have done learning about diversity, equity, inclusion, and justice in math. What are some proactive steps you might take in this project to create windows, mirrors, and sliding glass doors for your audience? Promote inclusion and/or equity?

Honor Code Language on Duplicate Use of Work:
Any work submitted to meet the requirements of a particular course is expected to be original work completed for that course. Students who wish to incorporate any portion of their own previously developed work into a new assignment must consult with the involved faculty members to establish appropriate expectations and parameters.

By signing below, you are indicating that the proposed project will be original work for this course. If the proposed project overlaps with any previous work/experiences, you are required to discuss plans with Professor Kubacki before continuing.

Collaborative Work Pledge and Permission:
I was an active collaborator on this assignment. I approve of its final content.

Include signatures below to indicate agreement with the above statements.

References

[1] Kendall Atkinson. *Introduction to Numerical Analysis*. Wiley, 2 edition, 1989.

[2] William Layton and Myron Sussman. *Numerical Linear Algebra*. Lulu, 2014. `http://www.lulu.com/spotlight/Layton_Sussman`.

[3] MathWorks. MATLAB. Version R2017a. `https://www.mathworks.com`.

[4] Lewis Fry Richardson. The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Philosophical Transactions of the Royal Society*, 210:307–357, 1910.

[5] Alex Townsend and Lloyd Trefethen. Gaussian elimination as an iterative algorithm. *SIAM News*, 46(2), March 2013.

[6] Lloyd Trefethen. *The Princeton Companion to Mathematics*, chapter IV.21, pages 604–615. Princeton University Press, 2008.