Emily Hillenbrand

Metis Data Science
Project 04 - NLP and Unsupervised Learning
18 Nov 2019

Math Topics: Summary

Overview

I originally wanted to find a way to explore any subject of interest in a semi-automated way, by using natural language processing (NLP) to create a list of relevant keywords or similar subjects instead of reading a Wikipedia page. To explore this idea, I used this Kaggle dataset made up of subtitles from youtube lectures on 11 different math/computer science subjects. I ended up with a model that can be used to compare the different subjects, explore relevant topics, or act as a starting point for creating a curriculum.

Pipeline

I explored multiple combinations of NLP tools and ultimately settled on:

1. **Standardization:** Lemmatize (SpaCy), plus a combination of custom and generic functions to remove punctuation, numbers, and stop words.
2. **Vectorization:** Tf-Idf (Scikit-Learn)
3. **Dimensionality Reduction:** NMF + Normalize (Scikit-Learn)
4. **Visualization/Comparison:** Heatmaps (Seaborn), Cosine Similarity (Scikit-Learn)

Results

I was able to create this list of topics for finding keywords associated with a subject, and this heatmap for finding similarities between subjects. Both are useful if you already have domain knowledge in the areas of math and computer science; you can use them to inform a mental map of where you've been in your educational journey, or you can build a curriculum leading up to studying NLP, or they can help you explain some of the important concepts in these areas.

I loved this project. My favorite part of this new type of analysis was seeing how I could break down the subjects I've spent so much time with, in a way that could help future learners. I also enjoyed figuring out how to create and import my own python modules to automate my pipeline.

Appendix

1. Final list of topics

   Topic 0: lambda, eigenvalue, matrix, eigenvector, transpose, inverse, zero, diagonal, symmetric, orthogonal

   Topic 1: probability, event, outcome, conditional, toss, occur, coin, sample, axiom, head

   Topic 2: list, log, element, string, hash, code, algorithm, loop, array, sort

   Topic 3: integral, theta, delta, curve, dx, area, formula, field, pi, region

   Topic 4: vector, transform, dot, plane, length, product, scalar, member, direction, component

   Topic 5: laplace_transform, infinity, st, integral, sine, pre, inverse, zero, convolution, dt

   Topic 6: equation, solution, differential, derivative, pre, negative, slope, constant, initial, zero

   Topic 7: node, tree, address, link, subtree, insert, child, pointer, list, search

   Topic 8: edge, vertex, graph, path, algorithm, weight, cycle, short, tree, node

   Topic 9: determinant, row, matrix, submatrix, column, cofactor, entry, diagonal, triangular, product

   Topic 10: random, variance, variable, expectation, distribution, pdf, probability, cdf, pmf, conditional

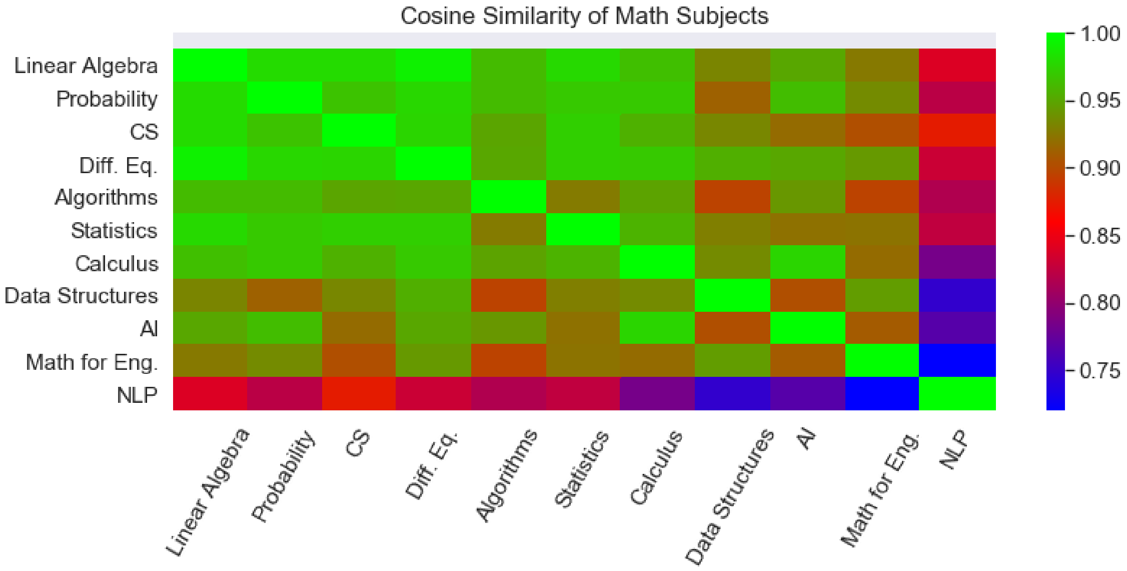
   Topic 11: particle, temperature, beta, energy, log, alpha, gas, density, volume, heat

   Topic 12: state, sub, arrival, markov, probability, chain, renewal, process, random, transition

Topic 13: matrix, column, row, transpose, space, null, pivot, basis, subspace, echelon

Topic 14: model, neural, word, datum, error, network, hypothesis, learning, input, sort

2. Comparison Heatmap


Cosine Similarity of Math Subjects
Linear Algebra
Probability
CS
Diff. Eq.
Algorithms
Statistics
Calculus
Data Structures
AI
Math for Eng.
NLP