Unsupervised academic curricula evaluation through Latent Dirichlet allocation

Jean Michel Rouly   Huzefa Rangwala

Updated January 25, 2015

Research goal

Through the application of probabilistic machine learning methods, specifically LDA topic modeling, a corpus of unstructured course syllabi can be digested and mined for topics. In this scenario, each topic represents a core concept covered by the courses.

A research framework will be constructed to read syllabus data from the Internet, digest into a common internal format, pipeline into an LDA topic model, and ultimately visualize in an interactive manner.

Background information

Program of study

A defined, ordered set of courses at a university with the goal of achieving a degree, certification, etc.

Background information

Program of study

A defined, ordered set of courses at a university with the goal of achieving a degree, certification, etc.

Course

A set of learning outcomes and techniques to achieve them, defined by a syllabus.

Background information

Program of study

A defined, ordered set of courses at a university with the goal of achieving a degree, certification, etc.

Course

A set of learning outcomes and techniques to achieve them, defined by a syllabus.

Syllabus

Unstructured collection of keywords and phrases that describes the core concepts and outcomes of a specific course.

Background information

Machine Learning

Interdisciplinary field combining elements of Artificial Intelligence and Statistics that allows programs to approximate unknown functions based on complex datasets.

Background information

Machine Learning

Interdisciplinary field combining elements of Artificial Intelligence and Statistics that allows programs to approximate unknown functions based on complex datasets.

Latent Variable Modeling

Subfield of Machine Learning that focuses on reconstructing "hidden" or unobservable variables that influence the structure of a dataset.

Background information

Machine Learning

Interdisciplinary field combining elements of Artificial Intelligence and Statistics that allows programs to approximate unknown functions based on complex datasets.

Latent Variable Modeling

Subfield of Machine Learning that focuses on reconstructing "hidden" or unobservable variables that influence the structure of a dataset.

Topic Modeling

Example of latent variable modeling that discovers topics that occur in a dataset.

Background information

Machine Learning

Interdisciplinary field combining elements of Artificial Intelligence and Statistics that allows programs to approximate unknown functions based on complex datasets.

Latent Variable Modeling

Subfield of Machine Learning that focuses on reconstructing "hidden" or unobservable variables that influence the structure of a dataset.

Topic Modeling

Example of latent variable modeling that discovers topics that occur in a dataset.

Latent Dirichlet allocation

Generative approach to topic modeling, starts with unknown variables and *generates* documents.

Latent Dirichlet allocation

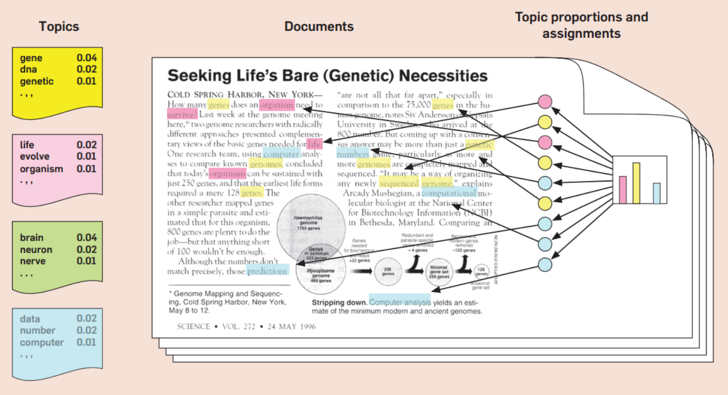

Figure: The LDA generative model.[1]

[1] David M. Blei. "Probabilistic Topic Models". In: *Commun. ACM* 55.4 (Apr. 2012), pp. 77–84. ISSN: 0001-0782. DOI: 10.1145/2133806.2133826. URL: http://doi.acm.org/10.1145/2133806.2133826.

Latent Dirichlet allocation

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}) = \frac{\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}}{w_{1:D}}$$

Gibbs sampling is used to approximate the probability of the denominator (evidence)[1].

- $\beta_{1:k} :=$ topic k
- $\theta_{d,k} :=$ topic proportion for topic k in document d
- $z_{d,n} :=$ topic assignment for word n in document d
- $w_{d,n} :=$ the n^{th} word in document d

[1] David M. Blei. "Probabilistic Topic Models". In: *Commun. ACM* 55.4 (Apr. 2012), pp. 77–84. ISSN: 0001-0782. DOI: 10.1145/2133806.2133826. URL: http://doi.acm.org/10.1145/2133806.2133826.

Project outline

- Collect preliminary syllabus dataset.
- Perform exploratory clustering.
- Expand initial prototype to include wider spread data sources, including multiple departments and universities.
- Expand initial prototype to include exploratory LDA computation.
- Complete exploratory results.
- Visualize exploratory results.
- Build formal syllabus data set from data collected online.
- Complete topic modeling analysis of big data set.
- Begin looking into analysis of topics to consider automatic labeling.

Project outline

- Collect preliminary syllabus dataset.
- Perform exploratory clustering.
- Expand initial prototype to include wider spread data sources, including multiple departments and universities.
- Expand initial prototype to include exploratory LDA computation.
- Complete exploratory results.
- Visualize exploratory results.
- Build formal syllabus data set from data collected online.
- Complete topic modeling analysis of big data set.
- Begin looking into analysis of topics to consider automatic labeling.

Resources

Software Toolkits

scikit-learn

Simple and efficient tools for data mining and data analysis. Built on NumPy, SciPy, and matplotlib.

http://scikit-learn.org

MALLET

"MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text."

http://mallet.cs.umass.edu

BeautifulSoup

Efficient and easy to use Web scraping and HTML manipulation library.

http://www.crummy.com/software/BeautifulSoup

Resources

Syllabus Data

Syllabus data collected from GMU Computer Science and Statistics departments, as well as Portland State University Computer Science and Chemistry departments.

Additional goal institutions:

- University of Colorado
- Rice University
- UNC, Greensboro
- Chaminade

(primarily because they offer easily-accessed public syllabus repositories).

Additionally, the Open Syllabus Project may prove a useful resource or collaborator in the future.

Framework

Scrape

Modular Python command line application that supports custom input data sources (syllabus archives) & multiple clustering tools. Pluggable backend scraping engines contribute to flexibility.

Learn

Java program that adaptively ingests data generated by the scrape module. Makes heavy use of MALLET to perform LDA.

Open source and available at `https://github.com/jrouly/trajectory`.

Preliminary results

Clustering

Data

- Scraped from `http://cs.gmu.edu/syllabus` archive.
- 1369 syllabus files, some empty.
- 1268 data rows (non-zero syllabi), 7189 features (terms).
- 292 categories (unique section numbers).

Execution time	0.144568s
Homogeneity	0.415
Completeness	0.877

Table: Preliminary clustering metrics

Preliminary results
Clustering

Cluster 1	Cluster 2	Cluster 3	Cluster 4
intelligence	chapter	software	operating
artificial	project	swe	systems
agents	sipser	testing	projects
learning	networks	web	aydin
tecuci	layer	interfaces	synchronization
expert	savitch	construction	scheduling
knowledge	data	design	homeworks
reasoning	dlc	constructing	processes
semantic	experimental	professor	group
intelligent	design	quality	friday

Table: 10 Most frequent terms in first four clusters

Preliminary results

Topic Modeling: Documents

Max Tokens:	12715
Total Tokens:	588131
Total Syllabi:	1570
Size on Disk:	9.1MB clean, 113MB raw

Doc	Topic	Proportion	Topic	Proportion
0	33	0.766641741676518	62	0.230550274742815
1	44	0.577637403706785 5	8	0.315250971602513 1
2	86	0.814329713432563 9	62	0.180157680477061 27
3	9	0.949110670067181 2	5	0.034876914241398 584
4	82	0.573669041236505 6	53	0.395394804342343 86

Table: First five documents and their top two topics

Preliminary results

Topic Modeling: Topics

Topic	Term	Term	Term
0	lisp (98)	june (57)	prolog (46)
1	systems (154)	operating (119)	system (101)
2	systems (304)	operating (252)	students (189)
3	randomization (66)	trials (57)	clinical (57)
4	database (389)	relational (151)	design (133)

Table: First five topics and their top four terms

Preliminary visualization tool

Simple combinatorially generated, cross-referenced HTML documents that display per-document topic breakdown (top n topics) as well as a definition of topics by frequent words (top n most frequent words).

Preliminary visualization tool

Document: 931 (raw) CS483.txt
(57) 0.8471050516082895: algorithms, design, algorithm, analysis, graph, credit, academic, techniques, assignment, discuss
(33) 0.08755366540491626: exam, final, office, class, homework, hours, midterm, students, assignments, grading
(70) 0.0613696456586234: week, october, september, november, december, group, lecture, analysis, article, review
(54) 0.0023807257563482347: data, trees, structures, binary, java, code, lists, linked, design, hashing
(50) 0.0001269923240291689: computer, science, mason, office, university, department, george, project, hours, description

---

Document: 932 (raw) CS390.txt
(87) 0.4333149747743147: design, user, software, interfaces, interface, human, development, students, project, computer
(42) 0.2970356100559934: research, dissertation, students, proposal, presentation, topic, project, degree, engineer, http
(30) 0.14859787680817188: class, line, blackboard, homework, lecture, questions, quizzes, work, exams, learn
(65) 0.09929787296385625: class, students, papers, paper, research, presentation, project, topics, team, instructor
(33) 0.013683214925854774: exam, final, office, class, homework, hours, midterm, students, assignments, grading

Figure: Per-document topic breakdown

Preliminary visualization tool

Topic: 51
Words: algorithms, software, testing, analysis, chapters, data, techniques, design, syllabus, structures
Known documents: 8 416 431 443 444 469 537 568 582 622 649 720 730 755 799 830 876 888 892 926 943 965 989 992 1018 1050 1070 1079 1087 1178 1187 1225 1237 1263 1284 1286 1305 1327 1348 1352

Topic: 52
Words: project, software, engineering, grade, class, work, email, writing, plagiarism, design
Known documents: 37 48 395 525 617 625 640 662 669 694 695 724 729 745 794 852 871 905 956 1016 1061 1062 1065 1118 1123 1151 1166 1228 1250 1297 1349

Topic: 53
Words: class, analysis, copy, tests, matrix, regression, back, test, page, work
Known documents: 48 63 83 179 182 187 188 189 204 226 260 267 275 297 301 341 370 803 879 1190 1314

Figure: Topic-word definitions

Continued development goals

Ultimately: visualization & comparison of university programs of study given unknown dataset of course descriptions.

Metadata awareness Track information like course number, semester, institutional information, etc.

Prerequisite chains Use metadata to track lists of prerequisite courses.

Rich visualizations Investigate use of D3.js[2] to develop rich, visually pleasing, interactive tools.

Evaluation suite Correlate results against existing third party evaluations, manual inspection, other instutitions.

[2] http://d3js.org

Questions?