UNSUPERVISED ACADEMIC CURRICULA EVALUATION THROUGH TOPIC MODELING

Jean Michel Rouly

April 16, 2015

George Mason University

What to expect from this talk

- Overview

What to expect from this talk

- Overview
- Background

WHAT TO EXPECT FROM THIS TALK

- Overview
- Background
- Data

WHAT TO EXPECT FROM THIS TALK

- Overview
- Background
- Data
- Trajectory

OVERVIEW

MOTIVATION

Evaluating a department on its conceptual coverage involves...

MOTIVATION

Evaluating a department on its conceptual coverage involves...

Relative Standing comparing against other, similar departments

MOTIVATION

Evaluating a department on its conceptual coverage involves...

Relative Standing comparing against other, similar departments

Absolute Performance benchmark against standardized expectations (published by the ACM)

MOTIVATION

Evaluating a department on its conceptual coverage involves...

Relative Standing comparing against other, similar departments

Absolute Performance benchmark against standardized expectations (published by the ACM)

Both of these require data about the topics covered in a course.

MOTIVATION

What we've got:

What we've got:

1. Widely available university course description data.

What we've got:

1. Widely available university course description data.
2. Descriptions detail what concepts are taught in a course.

What we've got:

1. Widely available university course description data.
2. Descriptions detail what concepts are taught in a course.
3. **Human-readable** descriptions require manual inspection.

What we've got:

1. Widely available university course description data.
2. Descriptions detail what concepts are taught in a course.
3. **Human-readable** descriptions require manual inspection.

So what to do?

RESEARCH GOAL

Through the application of probabilistic machine learning methods, specifically LDA topic modeling, a corpus of unstructured course descriptions 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 data from the Internet, digest into a common internal format, pipeline into an LDA topic model, and ultimately visualize in an interactive manner.

Ultimately the automatically discovered topics can be used in end-user university evaluation processes.

In other words

Build a tool that automatically...

RESEARCH GOAL

In other words

Build a tool that automatically...

- infers topics from a collection of course descriptions

RESEARCH GOAL

In other words

Build a tool that automatically...

- infers topics from a collection of course descriptions
- computes comparisons between departments

RESEARCH GOAL

In other words

Build a tool that automatically...

- infers topics from a collection of course descriptions
- computes comparisons between departments
- evaluates departments on their concepts

BACKGROUND

Attempts to discover the abstract topics of a dataset.

Topics

A topic is a frequency distribution over terms, roughly representing a concept taught in a course.

Overview

Latent Dirichlet Allocation (LDA) is a form of *Latent Variable Modeling* that can infer topics from within a document.

LDA takes a generative approach to latent variable modeling, assuming the topics occur in some proportion within each document.

DATA

UNIVERSITY DATASET

University	Course Count	Web
American University	32	american.edu
George Mason University	145	gmu.edu
Kansas State University	83	ksu.edu
Louisiana State University	59	lsu.edu
Portland State University	190	pdx.edu
Rensselaer Polytechnic Institute	61	rpi.edu
University of South Carolina	64	sc.edu
Stanford University	69	stanford.edu
University of Utah	142	utah.edu
University of Tennessee, Knoxville	29	utk.edu
ACM Exemplar Courses	68	—

Table 1: University course descriptions

SAMPLE COURSE DESCRIPTIONS

Raw Course Description

Capstone course focusing on design and successful implementation of major software project, encompassing broad spectrum of knowledge and skills, developed by team of students. Requires final exhibition to faculty-industry panel.

Cleaned course description

capston focus design success implement major softwar project encompass broad spectrum knowledg skill develop team student requir final exhibit faculti industri panel

TRAJECTORY

Trajectory is a tool that automatically ingests course description data from the Internet and presents an accessible interface for departmental evaluation.

Lines of Python	3193
Lines of Java	631
Lines of HTML/CSS/JS	1828
Lines of JSON	3219
Lines of Bash	165
Size on disk	6.7M

Table 2: Code statistics

TRAJECTORY

Four primary modules:

Scrape	web-scrape online university catalogs
Import/Export	pass structured data between Learn and Scrape
Learn	estimate LDA topic model on data
Web	visualization tool

Underneath the entire system is a structured relational database layer.

TRAJECTORY / WEB

Browse	collected data by university or department
Understand	courses through inferred topics
Analyze	conceptual overlap in prerequisite chains
Compare	departments based on conceptual composition
Evaluate	departments against ACM benchmarks

Live Demo

wish me luck!

SUMMARY

Try out **Trajectory** online at

`trajectory.staging.rouly.net`

Get the source of this presentation and the **Trajectory** project at

`github.com/jrouly/trajectory`

FUTURE GOALS

- Learning Outcomes meta-analysis
- Alternative methods
- Topic summarization

CREDITS

Co-authors:

- Huzefa Rangwala
- Aditya Johri

Presentation theme:

- `github.com/matze/mtheme`

Questions?