



HARVARD

Office of the Vice Provost for Advances in Learning
RESEARCH

Harvard Syllabus Search and Exploration

Andrew Ang, Glenn Lopez, Devin McCabe, Ilia Rushkin, Dustin Tingley

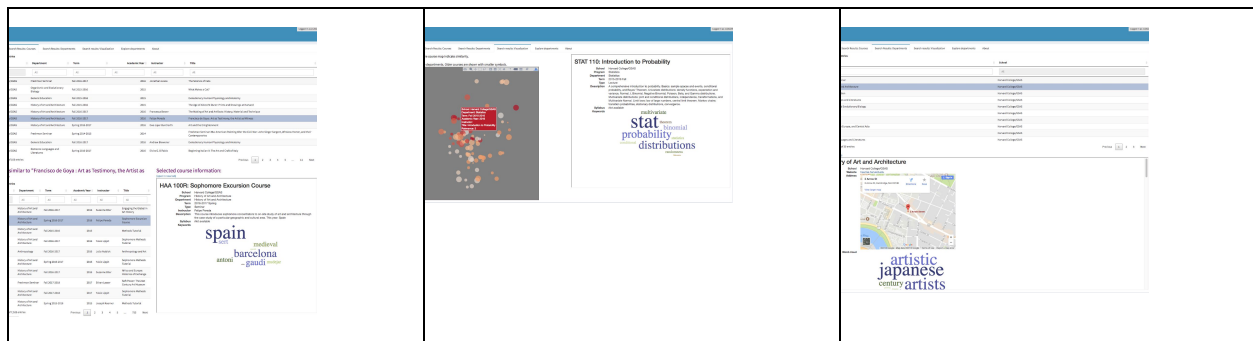
Goals

Allow members of the FAS community to easily search and explore the syllabi for FAS courses. In the future, allow any Harvard school to join. Collect syllabi without bothering busy people.

Product: vpal.hmdc.harvard.edu/syllabi (currently requires HUID be provided to VPAL-Research for HarvardKey login)

Syllabi for most courses are available on their respective Canvas sites, either created via a native Canvas templating system or uploaded as a file (e.g., PDF) by an instructor. These syllabi are already available to the public during each term's shopping period.

VPAL-Research built a data pipeline that automatically collects syllabi (from Canvas) and course descriptions (from the Registrar's office). We are able to collect syllabi for the vast majority of courses. The syllabi and course description text data is extracted and pre-analyzed using modern natural language processing tools. Results are presented in a HarvardKey-protected web dashboard that lets individuals search and explore syllabi, see courses "related" to other courses based on text in the syllabi, and explore departments that contain the courses.



How we did it

While there is no built-in platform feature to denote a specific uploaded file in a Canvas course as the syllabus file, instructors regularly name the file with "syllabus" as part of the name (or misspellings thereof), which we use to infer which files correspond to syllabi. With Canvas API access permission granted by the FAS Dean and HUIT, we programmatically download those syllabi files across all Harvard FAS courses. Additionally we download all course metadata (supplemental information about course departments, instructors, and terms) from the iCommons API, provided by HUIT Academic Technology.

Text from syllabi and associated metadata are cleaned, filtered, and reformatted for the purpose of modeling the text sources using an ensemble of different algorithms. These models are used to extract course and department keywords, compute similarity between courses, and enable searching of the documents and their metadata. A final data set is loaded in an [RShiny application](#), where the data can be searched and explored using a web browser.