# Tracing Trends in Academic Interests

*Uncovering the Structure of Stanford Course Offerings*

**By Kevin Chavez**

## Introduction

Every year there are fluctuations in the number of students enrolling in the large spectrum of courses offered at Stanford University. Some courses, such as CS 106A and CS 229, grow immensely in popularity. Others remain steady or have slow growth or gradual decline in enrollment. These trends presumably reflect the shifting academic interests of the institution--both on a student level (elective courses) and an administrative level (major and university requirements). However, enrollment on a course-by-course basis is not particularly insightful since any individual course may abark many independent topics or focus too narrowly on a subset of an interest. On the other hand, bundling enrollment by department obscures the diversity of academic interests within departments and inter-department correlations. If we are to examine the trends in academic interests, we need a better representation. The goal of this project is to:

Stage 1. Use course descriptions to generate a more useful representation of academic interests--one that captures the variance in course offerings, but is significantly smaller than the set of individual classes.

Stage 2. Use course enrollment data as a marker of how much weight the institution places on these academic interests, and trace the dynamics of these interests over the past several years.

This report will describe results from the first stage of this project--representing academic interests--and delineate the steps needed to complete the second stage.

## Representing Academic Interests

### Feature Selection

Course descriptions were initially encoded using “bag of words” encoding, using terms that appeared less than 80% of the time and more than 0.2% Further, the vectors were weighted and normalized using term frequency - inverse document frequency convention. Some hand selecting of stop words was necessary because of common phrases appearing in ExploreCourses that contribute little to the overall semantic meaning. Examples of these are “Instructor consent required,” “Topics include:” and such. The final set of words was of size 1352, rather small for text analysis, but suitable for our purposes.

### Co-clustering with Non-negative Matrix Factorization

The problem of finding a useful representation of academic interests can be viewed as an instance of a co-clustering problem. In other words, a useful representation will capture clusters of terms (words or n-grams) as well as clusters of courses. The method of non-negative matrix factorization [1] can yield these clusters simultaneously. This method was chosen over latent semantic analysis followed by clustering because of its more intuitive results and claimed improved peformance [2]. Let the matrix $A \in \mathbb{R}^{nxd}$ be the matrix of course description encodings, where n is the number of words used as features and d is the number of course descriptions. We then seek non-negative matrices $U \in \mathbb{R}^{nxk}$ and $V \in \mathbb{R}^{dxk}$ such that we minimize

$$J = || A - UV^T ||_2 \quad (1)$$

That is, the matrix product $UV^T$ is an approximate factorization of the course description matrix. In general, $k < n$ and $k < d$. Matrices U and V lend themselves to natural interpretations. In particular, the columns of U are vectors in word-space that represent academic interests and provide the basis for a lower dimensional vector space, and the $i^{th}$ row of V is the projection of the $i^{th}$ course description onto that vector space of academic interests. Further, the membership of any particular course in a cluster is determined by its largest coordinate along the axes of the academic interest basis.

## Example Projection

To make this more concrete, let's consider the factorization for $k$ = 200, and look at the projection of a specific course onto the academic interests vector space.

Here's a course description:
*Topics: statistical pattern recognition, linear and non-linear regression, non-parametric methods, exponential family, GLMs, support vector machines, kernel methods, model/feature selection, learning theory, VC dimension, clustering, density estimation, EM, dimensionality reduction, ICA, PCA, reinforcement learning and adaptive control, Markov decision processes, approximate dynamic programming, and policy search. Prerequisites: linear algebra, and basic probability and statistics.*

Its projection onto the 200 dimensional space is visualized by the bar chart below (Fig. 1)

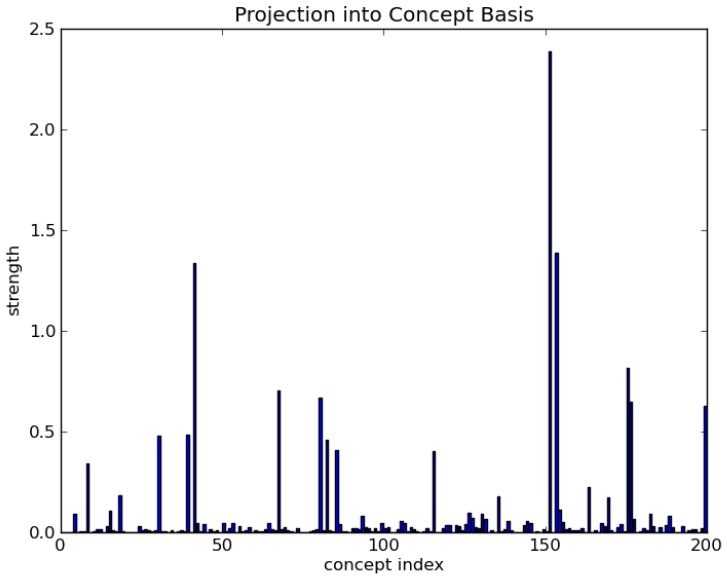


Figure 1. Projection of CS 229 course description onto 200 dimensional academic interests space

Notice that this course has largest components along axes 151, 153, and 41, which correspond to the "academic interests" represented by the following collection of terms:

| Index | Top Words |
|---|---|
| 41 | method, numerical, statistical, quantitative |
| 151 | probability, random, statistics, regression |
| 153 | learning, service |

Since the strongest component is along axis 151, this class would be assigned to that cluster, but we can see how its really a linear combination of these other areas as well.

## Selecting Dimensionality of Factorization

There is freedom in selecting the inner dimension, k, of the factorization. To find an optimal value of k, we examine two metrics: the similarity between cluster directions in the space of academic interests and the stability of the distribution of courses over the separate clusters.

To capture and summarize the structure of Stanford academics, it is desirable for the cluster directions to be dissimilar. By nature of non-negative matrix factorization, these directions will not be orthogonal, but we can plot the maximum cosine similarity between directions as a function of $k$ as well as the average similarity.

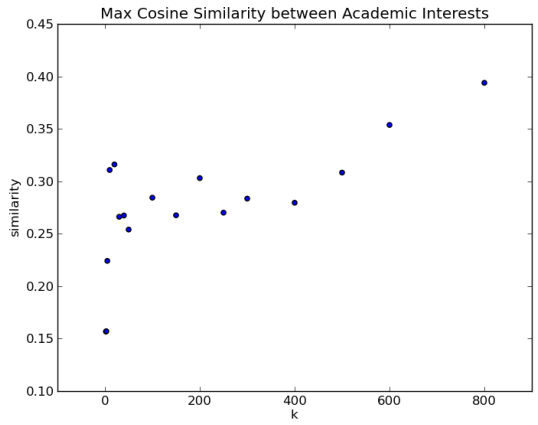


Figure 2. Maximum cosine similarity between any academic interest vectors.

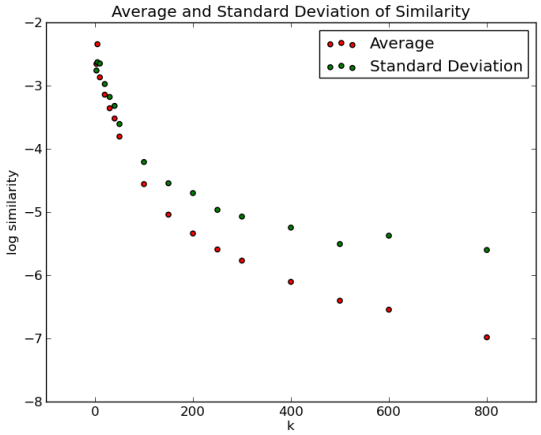


Figure 3. Average and standard deviation of cosine similarity.

Decomposing the courses into more than 400 academic interests starts increasing the maximum cosine similarity, which can loosely be interpreted as beginning to do fine-splitting of clusters. Further, the average similarity decays rapidly until about 150 separate interests. Thus, the proposed useful range for $k$ is between 150 and 400.

## Visualization

Despite the reduction in dimensionality, it is nonetheless still difficult to visualize the structure of the academic interest space. In particular, the basis vectors are over a thousand dimensional. However, they tend to only have significant components in a few of those directions. An intuitive visual can be created through word clouds with the size of a words determined by the square of the component of the basis vector in that word's direction. Figures 4 and 5 show sample basis vectors in this represenation.


Figure 4. Three sample word clouds, representing basis vectors of academic interest.

It's also helpful to visualize the distribution of courses over these 150 interest clusters. (Note: The first cluster is left out of this figure because it consisted entirely of courses without a description. Apparently, the default cluster--for courses having no description--is most strongly correlated with the term 'research').

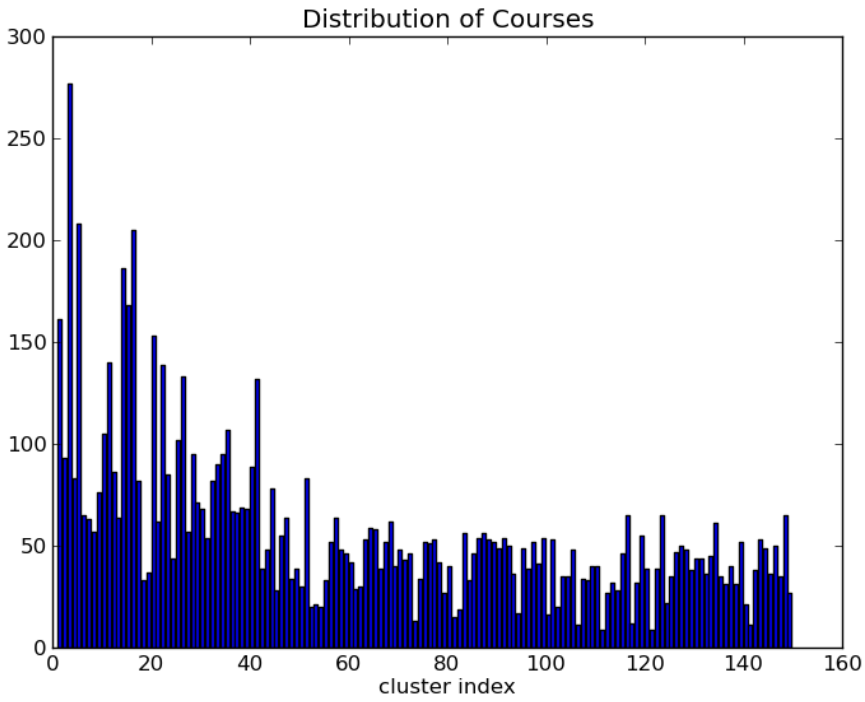


Figure 5. Distribution of 2013-2014 course descriptions over the 150 topics of academic interest

For a full set of the clusters of courses and academic interests derived by the $k = 150$ factorization, see http://www.stanford.edu/~kjchavez/tracing_trends/structure/.

# Next Steps: Tracking Shifting Interests

With this framework in place, the next step is to use enrollment data to examine the shifting focus on these interests. In this section, the steps to do so are explicitly stated. This phase of the project is not yet completed, but will be in the near future. Results will be available at http://www.stanford.edu/~kjchavez/tracing_trends/dynamics/

## Enrollment per Academic Interest

Let $E \in \mathbb{R}^{dxm}$ be the course enrollment matrix, where $d$ is the number of courses offered and $m$ is the number of years over which this is tracked. The row of E corresponding to CS 229 for the past 5 years looks like:

$$E_i^T = [\ 266 \quad 318 \quad 378 \quad 574 \quad 720\ ] \tag{2}$$

Clearly, we can see a growing interest in this course. However, such a growing interest can reflect a growing interest in any or all of different aspects of the course. By projecting these enrollment numbers onto the academic interest basis vectors, we can obtain this course enrollment's contribution to overall interest in the various academic topics. Let $N$ be a diagonal matrix where $N_{jj}$ is the total number of Stanford students in year $j$. Then we have the academic interest fractional enrollment matrix $\tilde{F} \in \mathbb{R}^{kxm}$.

$$\tilde{F} = \mathrm{VN}^{-1} E \tag{3}$$

## PCA to Extract Directions of Greatest Variance

Using PCA on the matrix $\tilde{F}$ will yield the directions in academic interest space along which enrollment varies the most. Using the top few principal components, we can generate a plot of the institution's trajectory. Consider the following table of pseudo enrollment data:

| | | | | | |
|---|---|---|---|---|---|
| CS229 | 266 | 318 | 378 | 574 | 720 |
| EE263 | 80 | 84 | 100 | 113 | 107 |
| CS109 | 104 | 103 | 105 | 97 | 67 |
| ECON102A | 34 | 27 | 29 | 37 | 38 |
| CS106A | 207 | 244 | 345 | 586 | 712 |

With the assumption that there were 1000, 1200, 1300, 1600, and 1800 total students in each of the years since 2009, respectively. We find that the greatest direction of fractional enrollment variance lies along the vector **v**, whose elements are displayed in Figure 6.

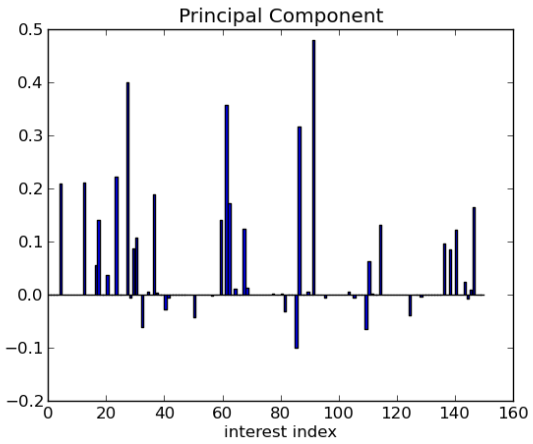


Figure 6. Vector elements of the first principal component.

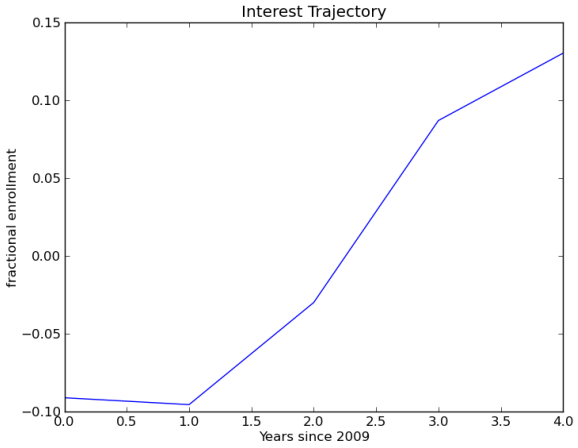


Figure 7. Trajectory of enrollmen in the first principal component

Thus one can understand the most important trends in academic interest, by examining one trajectory line and the components of a single vector. In particular, in this contrived case, we can see that interest in “computer, programming, software” (Element 91) and “language” (27), “engineering” (61) and “linear, optimization, algebra” (86) is increasing, while interest in “probability, random, variables” (85) and “math, fields” (109) is slowly decreasing. In this case, this principal component explains 98 % of the variance. With this infrastructure in place, we simply await true Stanford course enrollment data.

# Extensions

Non-negative matrix factorization gives us a natural way to find clusters of words and clusters of courses for any value of k. The structure of these clusters will differ by choice of k, but they differ smoothly. This sheds light on another opportunity for insight from this process. By examining how clusters merge as k decreases, we can see indications of which courses are semantically more similar than others, and possibly see suggestions for interdisciplinary cooperation. This extension may be pursued at http://www.stanford.edu/~kjchavez/tracing_trends/-course_search/.

# References

[1] Xu, Wei et. al. Document Clustering Based On Non-negative Matrix Factorization. In SIGIR ‘03. Toronto, Canada.

[2] Aggarwal, Charu and Checng Zhei. A Survey of Text Clustering Algorithms. In *Mining Text Data.* pp. 77 - 128.

[3] Landauer, T.K., Foltx, P.W., and Darrel Laham. An Introduction to Latent Semantic Analysis. In *Discourse Processes, 25.* pp. 259 - 284.