What Project Should I Choose?

Andrew Poon
poon-andrew@stanfordalumni.org

Abstract

This work analyzes the distribution of past CS229 projects by applying hierarchical agglomerative clustering. The clusters reveal which topics are very popular and which topics are more unique. Tracking the clusters over time also provides insight into how student projects have shifted over time. This knowledge will help future students select interesting and unique projects.

1 Introduction

Choosing a project for CS229 is difficult. Students want to choose projects that are not only interesting, but are also unique from other projects chosen by current and past students. This work attempts to address this problem by using text analysis and clustering to organize past projects. The distribution of projects changes over time and provides insight into where student interests lie and how the field is evolving. As a result, this work is able to identify popular topics worked on by many students, and highlights the most unique projects that have been submitted over the years.

2 Data Set

The data set for this project comes from the archive of past CS229 projects. These can be accessed at the course website[1]. The papers for each year have been collected, converted to text[2], and converted into word frequency vectors (i.e. a histogram of words).

2.1 Document Processing

Converting project papers into useful word frequency vectors requires processing. Standard processing techniques such as stop word removal [4, p. 27], lower casing all words, and word stemming [4, p. 32] were applied. By far, the most tedious part of this process was building a filter list to remove "contentless" words. This problem was exacerbated by the fact that the texts were CS229 project papers with different vocabularies than standard English texts. For example, words such as "algorithm" and "learning" would be relevant for ordinary topic modeling. But in this case, every project applies some algorithm that attempts to learn something so these words do not indicate anything particular about the project. Words like this are considered "contentless" in this context and were subsequently added to the stop word list.

After painstakingly removing as many irrelevant words as possible, each paper is converted into its own word frequency vector. The resulting vector indicates the key topics of the paper. An example of the top entries in a word frequency vector is shown in Table 1. In this example, the paper [2] investigates using neural networks to identify handwritten digits and training a robot to write those digits. Thus, is it not surprising that terms such as "nn" and "digit" appear frequently.

Unfortunately, not every project paper was available for analysis. Some papers (and all papers in 2004) were not available for download. Additionally, some papers failed to convert from PDF to

[1] http://cs229.stanford.edu

[2] The UNIX utility `pdftotext` was used.

Table 1: Word frequency vector of a paper investigating digit recognition using neural networks.

nn	32	image	29	program	28	motor	28
reconstruction	22	recognition	17	digit	15	mnist	15

plain text. In the end, 1179 papers were converted to word frequency vectors to form a sufficient data set.

3 Clustering

Hierarchical Agglomerative Clustering (HAC) [3, p. 520] was used to cluster the projects. HAC starts by initializing each sample in its own cluster. On each iteration, the clusters that are most similar are merged together. As the clustering iterates, the threshold for merging clusters becomes more relaxed. The result is that the most similar clusters are merged first, less similar clusters are merged later, and dissimilar clusters are left unmerged. This continues until the similarity threshold drops below a certain value, or until sufficiently few clusters remain.

Similarity between clusters was measured using Cosine Similarity [5] given by the expression

$$\frac{A \cdot B}{||A||||B||}$$

where A and B are the two word frequency vectors. This essentially performs a vector dot product. This approach is straightforward and provides a simple way to measure the similarity of two papers.

After two clusters are merged, the word frequency vectors are averaged and only the twenty most frequent words are kept. This was done so that the new vector would only contain the most relevant words. Otherwise, the vocabulary of the word vector would grow after each merge, causing many clusters to merge together quickly.

The standard *k*-means clustering algorithm was also attempted, but did not perform well. The random initial clusters resulted in inconsistent final clusters. There were also cases with empty clusters and cases with all samples in one huge cluster. After several attempts, *k*-means clustering was abandoned in favor of HAC.

4 Results

The clustering sequences for 2005 and 2013 are shown in Figure 1. Initially, each project is placed in its own cluster. As the algorithm runs, the most similar clusters are merged first. These early clusters are marked and keywords from those clusters are shown in Table 2. As the algorithm continues to run, the threshold for merging clusters relaxes and clusters with lower similarity are merged. This results in a very large cluster which can be seen at the bottom of the graphs. This large cluster does not have very distinctive keywords and is not meaningful. However, there are a handful of clusters that remain separate from the large cluster. These remaining individual clusters can be interpreted as the most unique papers.

The clustering sequences for each year follow the same pattern: similar clusters merge early on, but eventually most clusters merge into a huge cluster while only a few unique clusters remain separate. By analyzing the early clusters and the remaining unique clusters for each year, we can see how popular topics and unique topics change over time.

Clustering reveals several popular project areas such as: finance, robotics, music, image classification, text analysis, biology, sports, and movies. The cluster sizes over the years are shown in Figure 2. Tracking these clusters over time shows some trends in student interest. For example, all clusters, except for robotics, are generally growing over time. This is expected as the class size has grown significantly over time.

In the case of robotics, student interest has shifted as more projects are attempting to control other machines such as cars and rockets rather than the typical robotic arm, causing the keyword “robot”

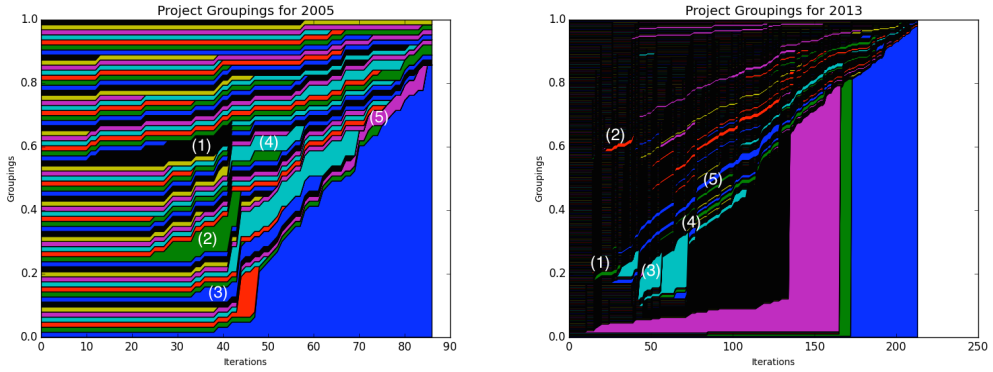

Figure 1: Evolution of clusters.

Table 2: Keywords from early clusters.

2005	
Grouping	Keywords
(1)	clustering, mean, word, partition, node
(2)	image, feature, classify, skin, object
(3)	note, music, component, instrument, microtiming
(4)	query, document, classify, precision, default
(5)	reinforcement_learning, light, player, traffic, game

2013	
Grouping	Keywords
(1)	feature, josquin, note, music, classify
(2)	tag, word, question, feature, classify
(3)	stock, return, feature, price, trading
(4)	game, team, season, prediction, play
(5)	click, query, search, rank, url

to decline in frequency. The cluster for movies fluctuates wildly during the early years. Apparently, the surge in popularity was due to the introduction of the Netflix Prize [6]. In the years following a glut of Netflix projects, students were less interested in applying machine learning to movies. Also of interest is the finance cluster. Surprisingly, there were no stock trading projects in 2005. However, there were two large peaks in 2009 and 2011. The first peak came after the housing bubble bust, while the second peak is due to a sponsored project[3] investigating stock trading based on Twitter messages. Finally, we can also see that image classification (plotted separately) is the most common topic because of wide applicability in areas such as computer vision and medical imaging. Oddly, there was a drop in 2012 in image classification projects. Closer inspection shows that many projects that year investigated astronomical applications such as detecting dark matter, causing the keyword "image" to not appear as often.

At the other end of the spectrum, we also find many examples of unique projects by examining the clusters that remain separate. Some examples of unique projects include: ionospheric corruption of radio waves, tracking vehicles using an autonomous helicopter, optimizing wind farms, and detecting arguments in online forums. Ironically, this algorithm found a unique project [1] from 2012 that also analyzed past CS229 projects in a similar fashion. This algorithm, while attempting to find unique projects, discovered that it itself was not unique. This was truly a surprising result!

5 Discussion

The clustering algorithm presented here is able to discover some general patterns in past projects. However, the clusters are not very precise because they are based only on word frequency. The biggest factor that could improve this system is to use better natural language processing (NLP).

[3] Supervised by Mihai Surdeanu and John Bauer.

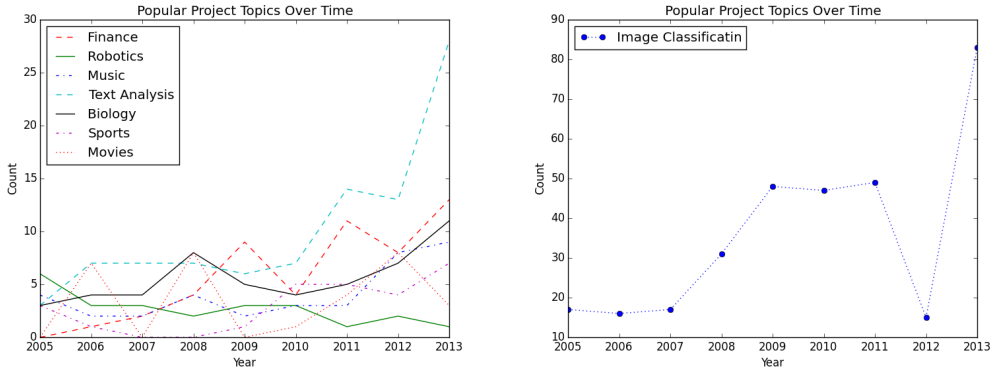

Figure 2: Topic trends over time.

The simple word frequency approach used here loses information that could be found in sequences of words (N-grams). Named entity recognition (NER) could be used to identify topic keywords and eliminate "contentless" words.

Another problem encountered in this work is that project topics are becoming more broad as well. For example, robot vision combines robotics and image classification. Trading stock based on Twitter messages combines finance and text analysis. These kinds of projects do not fit neatly into the common topics and introduce a lot of ambiguity in the results. Thus, a crude attempt at topic modeling was made by manually reading through past projects and labeling the paper with topic keywords. This provided more precise labels for each project, but was too tedious to be scalable. With better topic modeling, each topic would have a fingerprint of distinctive keywords which could be used for more accurate clustering.

6 Conclusion

This project performed unsupervised clustering on past CS229 projects. This revealed several common topic areas and well as some unique projects. Interest in the popular topic areas has also varied with time and has been influenced by external factors such as sponsored projects. These results provide guidance to future students by showing which topics have been very popular, and providing examples of unique projects. This information should help students choose more varied and unique projects in the future.

7 Future Work

This project has been described as "very meta" and ironically discovered itself to not be unique when searching for unique projects. But why stop there? We must go deeper. In the future, more projects can apply machine learning to past CS229 projects. Then, there could be a new project that analyzes the other project analyzers.

Regarding this system, the performance could be improved by incorporating better NLP and topic modeling. This would lead to better clustering and could reveal finer details than the general trends found in this work. The visualization of the clustering process also has much room for improvement.

References

[1] Michael Chang and Ethan Saeta. *Analyzing CS 229 Projects*, 2014. http://cs229.stanford.edu/proj2012/ChangSaeta-AnalyzingCS229Projects.pdf.

[2] John A. Conley and My Phuong Le. *Handwritten Digit Recognition: Investigation and Improvement of the Inferred Motor Program Algorithm*, 2005. http://cs229.stanford.edu/proj2005/ConleyLe-HandwrittenDigitRecognition.pdf.

[3] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer, 2009.

[4] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[5] Wikipedia. *Cosine Similarity*, 2014. http://en.wikipedia.org/wiki/Cosine_similarity.

[6] Wikipedia. *Netflix Prize*, 2014. http://en.wikipedia.org/wiki/Netflix_Prize.