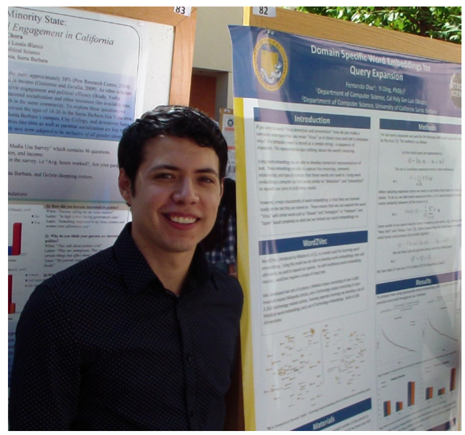
2017 Summer Undergraduate Research Experience Project Showcase

University of California, Santa Barbara
August 17, 2017

Funded by the National Science Foundation via an IGERT in Network Science (DGE-1258507)

Domain Specific Word Embeddings for Query Expansion

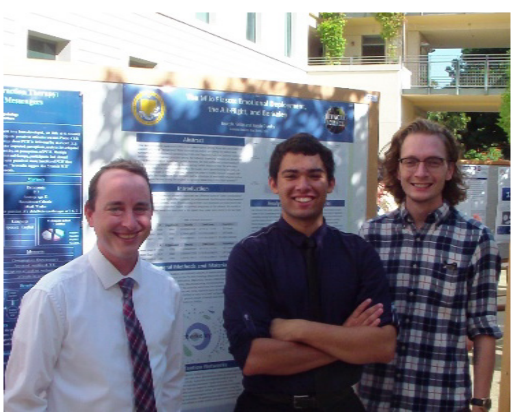
Fernando Diaz, Computer Science, California Polytechnic State University (Cal Poly)
Yi Ding, Ambuj K. Singh, Computer Science



Using the continuous space word embedding model word2vec we study the use of term relatedness in the context of query expansion. Moreover, with our models trained on domain specific corpora we examine the performance of these embeddings against models which have been trained over arbitrary corpora. Our query expansion methods are used for standard ad-hoc information retrieval on standard TREC data (Disks 4,5 with query sets 301-450, 601-700). We demonstrate that when classifying queries under a specific domain and using the appropriately trained word2vec embeddings, our retrieval methods achieve results which are negligibly better than using embeddings trained over arbitrary corpora. However, we notice the less ambiguous queries, which can be easily categorized into one specific domain, allow our embeddings to show significant improvement. Our results hint that if trained over larger domain-specific corpora, using these embeddings may yield more accurate results.

The Milo Fiasco: Emotional Deployment, the Alt-Right, and Berkeley

Jacob Gursky, Communications, University of Pennsylvania
Joseph Heide, Sociology, University of California, Santa Barbara
Alex T. Kulick, Sociology



How does emotional deployment differ in the conversations among those who used Twitter to support right-wing “provocateur” Milo Yiannopoulos before and during the protest event surrounding his controversial speaking tour? We used PageRank, hub, and authority centrality measures to identify principal actors among the networks formed by retweets and @mentions within the conversation about Milo and UC Berkeley. We then performed a close reading of the tweets concerning emotional claims. By comparing Twitter discourse from the same users conversing about the event – in general before the event, and in general during the event – we can analyze changes in both content and network structure of Twitter accounts engaged with the Alt-Right. We hypothesize that centrality will be a function of emotionally negative content in the network formed during the protest to a greater extent than in the other networks.

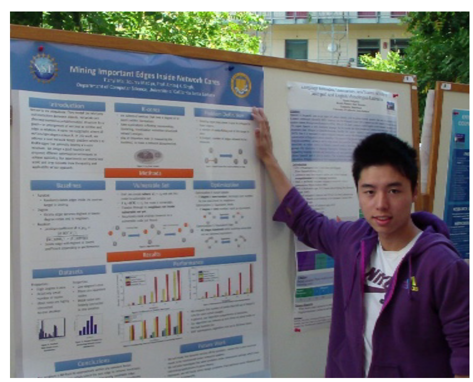
Mapping New York Times Coverage of the Department of Education with LDA Topic Modeling

Emma Kerr, Sociology, University of California Santa Barbara
Devin Cornell, Sociology

Secretary of Education Betsy DeVos is responsible for setting the political tone for the U.S. Department of Education for the duration of the Trump administration. However, given her lack of experience in public education, she has been an extremely controversial choice for the position. This project's motives were twofold: to explore potential applications of topic modeling for social science topics, and to examine the popular perception of Betsy DeVos as a deviation from previous Secretaries of Education. We pulled more than 650 articles from the New York Times website and analyzed them using LDA topic modeling. Those topic models were then interpreted and categorized into Policy and Personal categories, omitting irrelevant topics. We anticipated finding that Betsy DeVos has generated more discussion than previous Secretaries of Education, but that this exposure would mainly be related to her personal characteristics, history, and actions. By contrast, other secretaries would generate smaller volumes of work with discussion centered on specific educational policy. Results indicate that this hypothesis is holding true, with Betsy DeVos generating 15 Policy topics, 16 Personal topics, and 9 Neither topics out of 261 articles. Arne Duncan generated 32 Policy, 0 Personal, and 8 Neither out of 105 articles; Margaret Spellings generated 33 Policy, 5 Personal, and 2 Neither out of 268 articles.

Important Edge Mining Inside Cores

Tianyi Ma, Computer Science, University of California Santa Barbara
Sourav Medya, Ambuj K. Singh, Computer Science



K-cores of a network are defined as subsets of vertices that have a degree of at least k within themselves. These small dense subgraphs are relevant in many applications, ranging from community detection to network visualization. Moreover, k-cores play an important role in the robustness of a network, which is a measure of how hard it is to make the network disconnected. In this work, we investigate optimal strategies for breaking k-cores apart by removing a fixed number of connections. These small modifications in the graph structure often have a cascading effect, as the removal of a single link might cause vertices to leave the k-core, one by one, until its collapse. First, we show that there is no efficient optimal or even approximation algorithm for breaking k-cores with size larger than 2. We then resort to the study of heuristics for breaking k-cores that are efficient and work well in practice. Our experiments show that our approach achieves effective results in large real-world datasets, significantly outperforming the baseline methods.