Abu Reyan Ahmed

Professor Carlos Scheidegger

Computational Data Science

28th March 2017

Proposal Matching

We are developing an information retrieval tool which given a call for proposal, returns a list of relevant people. We are representing every people by a document which is generated by concatenating the abstracts of the papers written by that people. We collect these abstracts from Google scholar and UAVitae. Each document can be considered as a vector and the collection of all the documents provides a vector space. We are using Lucene which is a java based library for text processing. We now want to search a CFP to find the corresponding peoples for this CFP. We represent a CFP by a list of 100 words that have the highest TF-IDF value. We calculate the similarity with respect to every people and find out the top 20 similarities.

For the project of this course we will develop a visualization of the proposals and people. Visualization helps to evaluate a system. For example if we can generate a 2d plot where each proposal will be represented by a region in the plane and a person will be represented by a dot, then it will help us to understand how people and proposal are related with each other. In order to get this visualization we first get all the research topic like “Machine learning”, “Natural language processing”, “Data science”, “Graph theory” etc. One good source to collect all these research topics is google scholar. Now we determine a similarity matrix of these topics. Again

there are many ways to do this and we can again use google scholar. For example suppose a person mentioned both “Machine learning” and “Data science” in his google scholar profile. This indicates that these topics are correlated. Hence we add one for this pair of topics in similarity matrix. We can fill up the whole similarity matrix table using this strategy.

Now from this similarity matrix we run PCA and project every topic in a 2d plane. Now we embed the proposals into the plane. In order to do this we first represent a proposal with a set of topic that the proposal is about. Now we draw a blob that contains the topics in the plane. Similarly we represent a person by a blob in the plane. Or we can use a dot to represent a proposal of a person. We have a set of topics for a proposal. We take the coordinates of the topics and determine the barycentric coordinate. We represent the proposal at that point. It we be easy to visualize if we represent the proposals with a blobs and people with points.

Our objective of this project is to try different ways to compute the similarity matrix of the research topics and different representations of the proposals and persons. There are many analysis like SVM, PCA, MDS, t-SNE that we can use, also many of them have different parameters. We can compare among these different approaches and figure out which one works better in practice.