David LaCharite
Capstone Proposal

U.S.Agency for International Development Project Description Category Prediction :

Background:

The United States Agency for International Development (USAID) publishes project descriptions for USAID projects dating back to 1946. Projects are classified by categories (e.g. Conflict, Action Relating to Debt, Administration and Oversight). I propose to develop an NLP model which will predict the category classifications from the text in project descriptions.

1. What are your questions?

- Can NLP be used to evaluate consistency of text in USAID project descriptions?
- Does the text in project descriptions consistently reflect the project classifications?
- Can certain words or phrases be identified which can improve consistency of project descriptions?

2. Do you have the data?

Yes. The data is publicly available at the USAID data webpage (explorer.usaid.gov/query). There are over 40,000 observations with project descriptions and classifications.

3. Have you looked at the data?

I have familiarized myself with texts from each classification, but have not numerically processed the corpus texts.

4. What is your proposed approach to answering your question?

- To train Naive Bayes and SVM to predict project classifications their descriptions.
- Use K-means clustering to see if the classes converge in an unsupervised learning environment.
- Use bag of words to identify crucial words and phrases in different categories of project descriptions.

5. What is your minimum viable product?

Train/test a project description classification model and examine words and phrases which are crucial to distinguishing between categories.