Project Proposal – Team 43

- What is the problem or task addressed in the project?

  Avoiding users from asking previously asked questions on StackOverflow.

  Suppose a user wants to know -> 'how to find the least value in an array'.

  A search on the site will show only the questions that contain these specific texts.

  Questions with text 'minimum value' or 'smallest value' will not be shown. As a result the same question (semantic meaning) gets asked multiple times in different words. We plan to provide a method to reduce such repetition.

- Why is it interesting, and why is it challenging?

  It involves handling a real world problem of database growth with redundant information. Also, user query will be unstructured, searching a similarity between this text and other previously asked questions will involve applications of many NLP techniques.

- Who would benefit from this work, and how?

  a. The site user: Due to more semantic search, he will save time if the question was already asked in a different format.

  b. The site administrator/manager: Benefits in cost reduction, and less maintenance overhead.

- What data and knowledge sources will be used?

  For POS tagging the words in the question, we will be using Brown, Gutenberg, and Treebank corpuses. Stack overflow questions will be obtained using already existing StackOverflow web services. Word Net and Big Huge Thesaurus will be used to obtain the synonyms for the words.

- How will data be collected and annotated (if applicable)?

  Some sample set of questions will be manually labelled into correct clusters and used for evaluating the approach.

  Stack overflow questions will be obtained using the Stack overflow web services and then POS tagging will be done on this questions using NLTK postagger.

- What specific techniques do you expect to use? (Answer if possible. This is not a firm commitment, especially for projects on topics not yet covered in class, e.g. dialogue)

  a. Similar questions need to be grouped together. To achieve this we'll use a POS tagger with PCFG to identify the word list that will be used as inputs for clustering similar questions.
  b. We plan to try out k-means with cosine similarity and Euclidian distance, then evaluate which performs better.

- How will the approach be evaluated?

We will try to manually group some of the questions into different clusters and label them accordingly. We will try our algorithm and see how it performs, by evaluating how many groups we could form and how many were correctly tagged into that group versus how many were incorrectly tagged. We can further calculate the F-Score as the correctness measure.

- What does each member of the team plan to do? (Subject to change!)

  Vishal, Pradeep: Programmatic data collection from stack overflow using the provided APIs to access their question base and parse them.

  Darshan, Vishal: Generate set of manually labelled questions for evaluations

  All 3 members: Developing the core algorithm to pos-tag, parse, cluster the questions.

  Pradeep, Darshan: Testing the algorithm using different types of clustering like k-means, cosine similarity.