TEAM D

i. *Reference:*

i1. Swapna Gottipati, David Lo, Jing Jiang. 2011. Finding Relevant Answers in Software Forums. In Proceedings of the 2011 26th IEEE/ACM International Conference on Automated Software Engineering (ASE).

ii. *Keywords:*

***ii1*: Software Forums**: Software forums are web applications where the programmers look for solutions to their coding problems. A very widely used example of a software forum is www.stackoverflow.com where some programmers post their problems and others provide solutions to those.

***ii2*: Semantic Search**: Provides ability to search text in more meaningful way by considering various points like context and intent of search, location, etc. For example, Google can refine and personalize search results for a user based on such points.

***ii3*: Data Mining**: The practice of examining large databases in order to generate new information. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

***ii4*: Feature Extraction**: **It is** a type of dimensionality reduction that efficiently represents interesting parts of the data. **Feature Extraction** starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations.

iii. *Notes on 4 of 19:*

iii1: **Motivational statements**: Traditional information retrieval techniques perform poorly on software forums as the complexity of software systems leads to use of several jargons in posts and duplicate content. Manually filtering relevant answers in these long threads is tedious and confusing. To address these issues, this paper proposes a semantic search engine framework which infers semantic tags of posts and uses them for effective search results.

iii2: **Related Work:** In this particular article, the authors are inclined towards extracting information from the tags of the posts of different software forums to utilize them to retrieve useful information. Similar work has been mentioned by *S. Thummalapenta and T. Xie in their article "Spotweb: Detecting framework hotspots and coldspots via mining open source code on the web."* But, in their work they are extracting data from Google code rather than software forums. Apart from this, Wang et al. used natural language and execution trace information in bug reports to detect duplicate bug reports. While the process is similar, but in this current article the authors are focusing on retrieval of relevant answers on software forums rather than finding bug reports.

iii3: **Data**: The dataset in this study is constructed by crawling webpages corresponding to several posts in different software forums. The authors have analyzed three different forums (softwaretipsandtricks.com, DZone and sun.com) and retrieved information like posts' message content and author names, to aid them in their research.

iii4: Future Work:` There is a lot of scope of future work in this field. The extracted tags can be further analyzed to extract answers with positive feedback and same can be sent to experts along with the questions with no answers. There is also a scope in future to detect noise in human written communication by looking into spelling variations. An automated approach to cluster the forum posts in hierarchical fashion is also a possible future extension.

iv. *Needs Improvement*

iv1: We can make the search more personalized according to user's preferences (for e.g. language [C++, Java]).

iv2: The data pre-processing phase needs to incorporate the fact that informal language is commonly used in Software Forums. This leads to lots of spelling and grammatical errors in the posts. By correcting those with sophisticated Natural Language Processing techniques can improve the quality of data and hence the search results.

iv3: There is a lack of subject-completed study instruments in the paper. There is hardly any mention of surveys or questionnaires which should have been performed on the user base so as to determine the overall need for this research and the user inclination towards their method.