Xixian CHEN 1155036600 cxxnju@hotmail.com

**Which tentative dataset do you use in your project? Show the basic statistics of the dataset.**

The dataset is very large, almost consisting of 6.75GB training data and 2.20GB testing data.

For training data, it has 6034195 items (questions). Every item includes Id, Title, Body and Tags. For testing data, it has 2013337 items. Every item only contains Id, Title and Body.

Following are some other things about the dataset,

Id - Unique identifier for each question

Title - The question's title

Body - The body of the question

Tags - The tags associated with the question (all lowercase, should not contain tabs '\t' or ampersands '&')

The questions are randomized and contain a mix of verbose text sites as well as sites related to math and programming. The number of questions from each site may vary, and no filtering has been performed on the questions (such as closed questions).

**What techniques/algorithms will you apply?**

Word Frequency Analysis

Much early work concerned the frequency of term usage in the text, but most of this work focused on deﬁning keywords in relation to a single document. The idea of statistically analyzing the frequency of keyword usage within a document in relation to multiple other documents became more common. This technique, known as Term Frequency - Inverse Document Frequency or simply TF-IDF, weights a given term to determine how well the term describes an individual document within a corpus. It does this by weighting the term positively for the number of times the term occurs within the speciﬁc document, while also weighting the term negatively relative to the number of documents which contain the term. When the TF-IDF function is run against all terms in all documents in the document corpus, the words can be ranked by their scores. A higher TF-IDF score indicates that a word is both important to the document, as well as relatively uncommon across the document corpus. This is often interpreted to mean that the word is signiﬁcant to the document, and could be used to accurately summarize the document. In this case, we simply regard the questions as multiple document/text. However, the title of question may tell more useful information than the body. As a result, we may weight more on the title.

Word Co-Occurrence Relationships

Word co-occurrence aims to find similarity between words or similarities of meaning among word patterns. The sentences in the question text are considered as a set of words; So we extract keywords from text uses word co-occurrence to build a co-occurrence matrix. Words are important to the text if they co-occur with other words more often in the text than they would if every instance of the word were randomly distributed. For a certain word wi, this can be thought of as the ratio of the number of co-occurrences of words wi、wj to the number of all other co-occurrences involving wi. Under the given assumptions, a high ratio would mean that the word wi is a likely keyword for the question text.