1. Motivation

Online information is wildly increasing since the last decade. Providing meaningful information timely has become urgent needs for companies like Microsoft, Google, Facebook, etc. Naively searching all the data assuredly won’t meet the demand. A basic solution would come into mind: category the information by topics or keywords. This encounters a problem: not every item has been given topics or keywords. Manually extracting the keywords may work in small datasets, but it will never be applicable for a search engine company. Thus, we need to design an algorithm which would work on disparate stack exchange sites to predict the tags (a.k.a. keywords, topics, summaries).

2. Most related topics

MapReduce and Frequent Itemsets;

Locality Sensitive Hashing;

Mining Data Streams;

Dimensionality Reduction;

Online Learning.

3. Deliverables

We plan to participate in competition on the website: <https://www.kaggle.com/c/facebook-recruiting-iii-keyword-extraction>. So we would deliver a runnable system and get a rank in top 20.