A Case Study for Massive Text Mining: K Nearest Neighbor Algorithm on PubMed data

**Nhan Do**[1], Murat Dundar[2]
[1]Department of Electrical and Computer Engineering, Purdue School of Engineering and Technology;
[2]Department of Computer and Information Science, Purdue School of Science

US National Library of Medicine (NLM) has a huge collections of millions of books, journals, and other publications relating to medical domain. NLM creates the database called MEDLINE to store and link the citations to the publications. This database allows the researchers and students to access and find medical articles easily. The public can search on MEDLINE using a database called PubMed. When the new PubMed documents become available online, the curators have to manually decide the labels for them. The process is tedious and time-consuming because there are more than 27,149 descriptor (MeSH terms). Although the curators are already using a system called MTI for MeSH terms suggestion, the performance needs to be improved. This research explores the usage of text classification to annotate new PubMed document automatically, efficiently, and with reasonable accuracy. The data is gathered from BioASQ Contest, which contains 4 millions of abstracts. The research process includes preprocess the data, reduce the feature space, classify and evaluate the result. We focus on the K nearest neighbor algorithm in this case study.

Mentor: Murat Dundar, Department of Computer and Information Science, Purdue School of Science