Datasets. Our performance study is on four datasets:

- DBLP title: A set of titles of recently published papers in DBLP[1]. The set has 1.9M titles, 152K unique words, and 11M tokens.
- CS abstract: A dataset of computer science paper abstracts from Arnetminer[2]. The set has 529K papers, 186K unique words, and 39M tokens.
- TREC AP news: A TREC news dataset (1998). It contains 106K full articles, 170K unique words, and 19M tokens.
- Pubmed abstract: A dataset of life sciences and biomedical topic. We crawled 1.5M abstracts[3] from Jan. 2012 to Sep. 2013. The dataset has 98K unique words after stemming and 169M tokens.