# Extracting Documents

Please read the description and readme file of the dataset first.
http://archive.ics.uci.edu/ml/datasets/Bag+of+Words

The format of the docword.nips.txt file extracted from docword.nips.txt.gz is 3 header lines, followed by NNZ triples:

- - -
D
W
NNZ
docID wordID count
docID wordID count
docID wordID count
docID wordID count
. . .
docID wordID count
docID wordID count
docID wordID count
- - -

The format of the vocab.nips.txt file is line contains wordID=n.

Extract all documents containing the words given in vocab.nips.txt file as per the word count given in docword.nips.txt file. Each document can be named by docID.txt.