CS671: Introduction to NLP
Assignment #1: Tokenization, segmentation, morphological analysis

Due on: 14-8-2016, 23.59 5-8-2016
MM: 200

HFST (https://sourceforge.net/projects/hfst/files/) is a mature platform for FSTs for natural languages. However, you are free to use any other open source or free platform or library. Avoid proprietory software.

1. The 20 news group data set contains news group data from 20 news groups. It has approx. 19000 documents. The data set can be downloaded from the ftp site. Some more information about the data set is at: http://qwone.com/ jason/20Newsgroups/.
   (a) Design a tokenizer, sentence segmenter and morphological analyser using HFST (or other FST toolkit). Run it on the news group data set mentioned above. You can test your tokenizer and segmenter using some of the other resources available on the webpage above.
   (b) Build a machine learning based sentence terminator classifier using your tokenizer and morphological analyser to produce feature tags for the context words around the period sign. You can use any binary classification algorithm available in any ML library. If you are new to ML ask your TA or fellow student who has done ML for help. You should be able to use the ML libraries without a deep knowledge of all the algorithms.