Privacy Policy Analysis Report

Preliminary dataset for finding the best classifier:

- Total size: 1949 sentences
- 0.7 of the dataset assigned as training set
- Training set size: 1559 sentences
- Testing set size: 390 sentences

After labaling the preliminary dataset based on sensitive terms (e.g. share, use, consent, permiss, collect, etc.) we performed three classification algorithms for classifying sensitive/non-sensitive sentences. Below is the accuracy of each algorithm:

- Naïve Bayes: 0.63
- K-Nearest Neighbor (k=3): 0.76
- Support Vector Machine: 0.93

Primary dataset of privacy policy documents for SVM classification:

- Number of privacy policy links selected from Playdrone: 163
- Number of accessible links (March 16): 107
- Total dataset: 16822 sentence from 107 privacy policy webpages
- Training set size: 0.7
- Accuracy: 0.99
- Precision: 0.97
- Recall: 0.99
- F-measure: 0.98
- Sensitive sentences: 10545
- Non-sensitive: 6277
- 0.62 of sentences were sensitive
- 0.38 non-sensitive
- Average sentence per document: 157
- Average sensitive sentence per document: 97
- Average non-sensitive per document: 59