We only have access to a big data set via boolean search.

Sub-sets of the data set returned by a single key word on a single day are large.For example for the single search term 'election' for the single day of 1 August 1868 returns 1982 articles. The search term 'riot' for the same day (close to the 1868 General Election) returns 360 articles, 228 of which do not also contain the word election.

With even a small number of search terms it quicly becomes impractical to examine all the documents even for a single day.

Following King et al we define S - the search set of all documents in the British Newspaper Archive T - the target set of all documents in the British Newspaper Archive which are about election violence R - a reference set of documents which are about election violence

The task is to identify T from S in a form where T can be

It is trivial to define an algorithm which obtains a subset of S which contains T, because $S \subseteq S$ and $T \subset S$. Algorithms which aim to maximise the chances of obtaining all of T will tend to return S.

Our task is to find a good method for returning T from S in a form which we can analyse. By a *good* method we mean a method which returns a greater proportion of T, and a greater ratio of T to $\neg T$ than alternative methods. The main alternative method is manual searching by historians.

1 Keyword Identification

Note it is important to make keyword identification somewhat selective of T from S otherwise even very good stage 2 & 3 selection processes the false positives will overwhelm the true positives.

Algorithm:

1. use classifier on R and S to identify two lists of keywords
2. generate probability from classifier parameters
3. add to keyword list based on probability