The management of unstructured text-based data (i.e., data not amenable for storage inside a relational database – emails, faxes, web pages, etc.) is a major problem pervading the information technology industry. It is widely accepted that at least 80% of the data held by companies is unstructured. Over recent years there has been a growing interest in creating automatic systems that assist users in managing documents such as emails. This project provides scope for students to learn about different facets of dealing with unstructured (text) data, and in particular, about how clustering and classification techniques can be successfully applied to it. To take this topic, students must have received a strong mark in CSE2AIF, and one or more of CSE3ALR and CSE3CI. Students should also have some background in both the Python and Matlab computer programming languages.