Term project – default task: Text categorization

What is a Text Categorization Task?

Text categorization is the task of deciding whether a piece of text belongs to any of a set of prespecified categories.

It is a generic text processing task useful in indexing documents for later retrieval, as a stage in natural language processing systems, for content analysis, and in many other roles.

Reuters-21578 Text Categorization Collection

Classic Reuters-21578 collection has been widely used as a benchmark for text classification evaluation.

A collection of 21,578 newswire articles, manually labelled

Freely available
`http://archive.ics.uci.edu/ml/machine-learning-databases/reuters21578-mld/reuters21578.html`

Read the information file `readme`

The articles are assigned classes from a set of 118 topic categories. A document may be assigned several classes or none, but the commonest case is single assignment

Reuters-21578 – example classes

Example topic categories – 10 largest classes

	class	# train	# test		class	# train	# test	
	earn	2877	1087		trade	369	119	
	acquisitions	1650	179		interest	347	131	
	money-fx	538	179		ship	197	89	
	grain	433	149		wheat	212	71	
	crude	389	189		corn	182	56	