Text Categorization (by Topic)

Acăr Bogdan

Facultatea de Matematică și Informatică

May 11, 2019

Overview

1. Problem Statement
2. Dataset
3. Features
4. Learning Methods
5. Results

Text Categorization

- Classify document context according to a set of pre-defined topics.
- One document might contain more than one topic.

Dataset

- Dataset Reuters-21578 news articles containing 118 unevenly distributed topics.
- Modified Apte Split: 9063 training documents and 3299 test documents.
- Using top 10 categories.

class	train	text
earn	2877	1087
acq	1650	719
money-fx	538	179
grain	433	149
crude	389	189

class	train	text
trade	369	119
interest	347	131
ship	197	89
wheat	212	71
corn	182	56

Features

- Clean document text, remove stop-words, stemming.
- tf-idf document representation of monograms and bigrams.

Learning Methods

- Multinomial Naive Bayes, K-Nearest Neighbours and Linear Support Vector Machine
- Use grid search to look for best parameters over a set of pre-defined values.

Results

F1 top 10 categories scores per class and micro-averaged.

	NB[1]	NB[3]	kNN[2]	kNN[3]	SVM[1]	SVM[3]
earn	96	98	97	93	98	99
acq	88	97	92	82	94	97
money-fx	57	84	78	84	75	85
grain	79	90	82	91	95	95
crude	80	87	86	87	89	91
trade	64	80	77	85	76	87
interest	65	76	74	79	78	80
ship	85	58	79	75	86	78
wheat	70	67	77	74	92	81
corn	65	55	78	71	90	82
micro-avg	82	91	82	87	92	94

Dumais et al. (1998) [1], Joachims (1998) [2], Presented work [3]

Thank You!