Text Classification by NB

P. Rullo

The training set consists of 5 documents associated with 2 classes - sport and politics. We use a binary representation, that is, for each word, the presence/absence within each document is reported.

The following table represents the document/word matrix, after lemmatization and stopword removal have been performed.

	parliament	italian	loose	win	soccer	championship	government	Class
d1	0	1	0	1	1	1	0	sport
d2	0	0	1	0	1	1	1	sport
d3	0	0	0	0	0	1	0	sport
d4	1	1	1	0	0	0	1	politcs
d5	1	1	0	1	0	0	1	politcs

Now, perform the following tasks:

1. Feature selection by using entropy – select the first k=5 best features
2. Prior probabilities estimates, over the (reduced) training set, by using both fractions and m-estimates
3. Classification of the following docs
    - d1 = “The team of Juventus will likely not win the Italian soccer championship”.
    - d2 = “the Italian parliament, based on the initiative of the government, has approved a law for Italian soccer teams”