CLASSIFY BIBLIOGRAPHY DATA

TARUN ASHUTOSH

BIBLIOGRAPHY DATA

- Tags:

  author, booktitle, edition, editor, issue, journal, location, month, pages, proceeding, publisher, title, volume, year

Tag	Example		Tag	Example
author	A. L. Barabási and F. Slanina		month	Apr.
booktitle	in Evolutionary Design by Computers		pages	pp. 41–44.
edition	3rd ed.		proceeding	in IAAI '90:
editor	W. Banzhaf and C. Reeves, Eds.		publisher	Springer
issue	no. 1		title	"The Group Lasso for Logistic Regression,"
journal	Genetics		volume	vol. 24
location	Cambridge, MA		year	2007

FEATURE ENGINEERING

Possible Differences between phrases:

- Number of alphanumeric words
- Number of numeric entities
- Number of words belonging to each pos tag
- Position of each phrase within the reference

Available pos tags in NLTK library:

'PRP$', 'VBG', 'VBD', '``', 'VBN', ',' ,
'""", 'VBP', 'WDT', 'JJ', 'WP', 'VBZ', 'DT', 'RP', '$', 'NN', ')', '(', 'F
W', 'POS', '.', 'TO', 'LS', 'RB', ':', 'NNS', 'NNP', 'VB', 'WRB', 'CC
', 'PDT', 'RBS', 'RBR', 'CD', 'PRP', 'EX', 'IN',
'WP$', 'MD', 'NNPS', '--', 'JJS', 'JJR', 'SYM', 'UH'

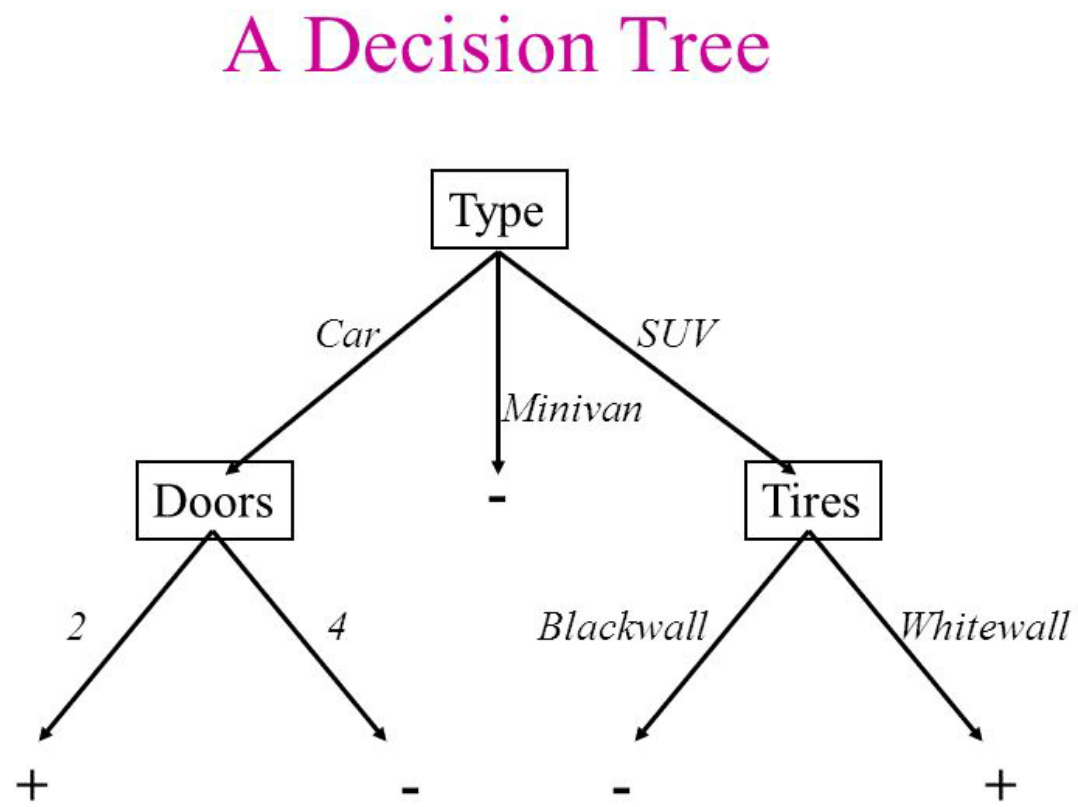
And these differences were made the features for learning

RANDOM-FOREST CLASSIFIER

It is a multiclass classification problem.

Reasons to choose Random Forest:

- Doesn't have the problem of class imbalance as may arise while doing one vs all classification.
- Is a little less sensitive to noise as compared to other classifiers like SVMs.
- The classification is faster.
- Has the functionality of showing feature importance based on impurity measure



TRAINING AND TESTING

1

Training and Testing accuracies with just the pos tags as features ---------- ~70%

2

Training and Testing accuracies with just the number of words as features ----- ~50%

3

Training and Testing accuracies with just the number of numeric words as features ----- ~40%

4

Training and Testing accuracies with just the position of tags as features ---------- ~50%

5

Training and Testing accuracies with all the above 4 as features ---------- ~90%

THANK YOU