URL categorization using machine learning

1 Description

Internet can be used as one important source of information for machine learning algorithms. Web pages store diverse information about multiple domains. One critical problem is how to categorize this information. Support vector machines [1] and other supervised classification algorithms [2, 3, 4] have been applied to text categorization.

2 Objectives

The goal of the project is to solve the *URL categorization* task using a supervised classification algorithm and using a database with a large number of categorized websites [1]. A supervised classified should be learned to predict the category a web page belongs to.

The student should: 1) Design any preprocessing of the web pages in the dataset (see suggestions below); 2) Define and learn the classifier using the training data. 3) Designthe validation method to evaluate the accuracy of the proposed classification approach.

As in other projects, a report should describe the characteristics of the design, implementation, and results. A Jupyter notebook should include calls to the implemented function that illustrate the way it works.

3 Suggestions

- Formalize the task as a classification problem where the classes of the problems are all possible categories included in the dataset.
- Access each web page and parse the information from the web page, mainly words, that could be used to predict how to categorize the content of the web page. The `urllib` Python package can be used to download the web information (see discussion here: `https://stackoverflow.com/questions/1825438/download-html-page-and-its-content`)
- Extract features from the text data in the web page. You may use the `nltk` or `gensim` Python packages that implement several useful functions for NLP.
- Implementations can use any other Python library.

References

[1] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features, 1998.

[2] Leah S Larkey and W Bruce Croft. Combining classifiers in text categorization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 289–297. ACM, 1996.

[1] The URL categorization dataset can be downloaded from `https://www.crowdflower.com/data-for-everyone/`.

[3] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[4] Bo Tang, Steven Kay, and Haibo He. Toward optimal feature selection in naive Bayes for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 28(9):2508–2521, 2016.