Clustering tweets about Machine Learning using self-organizing maps

1 Description

Internet can be used as one important source of information for machine learning algorithms. In particular, twitter has become a valuable tool of informations for companies and social agents. Clustering algorithms have been extensively applied to the analysis of tweets [1, 2, 3]. They can serve to identify tweets with a common topic, select exemplar tweets, or study the similarities between groups of tweets.

2 Objectives

The goal of the project is to extract the most recent tweets (e.g., from the last hour) addressing a given topic (e.g., "Machine Learning") and to cluster these tweets using self-organizing maps algorithm, and any predefined criterion of similarity between tweets (e.g., similarity of their content, how close is the ML area of application they refer to, etc.). The student should: 1) Capture the tweets (see suggestions below), 2) Design any required preprocessing. 3) Use available implementations to visualize the clusters. 4) Discuss the results. 5) Answer to the following questions in the report:

- What class of problems can be solved with the NN? (e.g., supervised vs unsupervised problems)
- What is the network architecture? (e.g., type and number of layers, parameters, connectivity, etc.).
- What is the rationale behind the conception of the NN?
- How is inference implemented? (e.g., How is the information extracted from the network?). Type of prediction or type of inference process.
- What are the learning methods used to learn the network ? Algorithms used for learning the network.

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

- Use the `pattern` Python package `https://github.com/clips/pattern` to retrieve recent tweets from twitter. See example `https://github.com/clips/pattern/blob/master/examples/ 01-web/04-twitter.py` to see tools for tweet parsing in `pattern`.
- Extract any relevant information from the tweets to measure the similarity between any two tweets.
- Use this measure of similarity to cluster the tweets in a 2-dimensional lattice. Use Python packages `Somoclu` (`https://somoclu.readthedocs.io/en/stable/`) or `NeuPy` (`http://neupy.com/pages/home.html`).

- Implementations can use any other Python library.
- Visualize the clusters.

References

[1] Daniel Godfrey, Caley Johns, Carl Meyer, Shaina Race, and Carol Sadek. A case study in text mining: Interpreting twitter data from world cup tweets. *CoRR*, abs/1408.5427, 2014.

[2] Nattiya Kanhabua and Wolfgang Nejdl. Understanding the diversity of tweets in the time of outbreaks. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1335–1342. ACM, 2013.

[3] Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. Topical clustering of tweets. *Proceedings of the ACM SIGIR: SWSM*, 2011.