Large Scale Hierarchical Text Classification

Classify Wikipedia documents into one of 325,056 categories

119 teams · 3 years ago

Overview

We are pleased to announce the 4th edition of the Large Scale Hierarchical Text Classification (LSHTC) Challenge. The LSHTC Challenge is a hierarchical text classification competition, using very large datasets.

Hierarchies are becoming ever more popular for the organization of text documents, particularly on the Web. Web directories and Wikipedia are two examples of such hierarchies. Along with their widespread use comes the need for automated classification of new documents to the categories in the hierarchy. As the size of the hierarchy grows and the number of documents to be classified increases, a number of interesting machine learning problems arise. In particular, it is one of the rare situations where data sparsity remains an issue, despite the vastness of available data: as more documents become available, more classes are also added to the hierarchy, and there is a very high imbalance between the classes at different levels of the hierarchy. Additionally, the statistical dependence of the classes poses challenges and opportunities for new learning methods.

The challenge is based on a large dataset created from Wikipedia. The dataset is multi-class, multi-label and hierarchical. The number of categories is roughly 325,000 and number of the documents is 2,400,000.

This challenge builds upon a series of successful challenges on large-scale hierarchical text classification. More information can be found at http://lshtc.iit.demokritos.gr/

This track concerns multi-label classification based on the Wikipedia dataset. The hierarchy is a graph that can have cycles. The number of categories is roughly 325,000 and the number of documents is 2,400,000. A document can appear in multiple classes.

Organizers

Ioannis Partalas, LIG, Grenoble, France

Massih-Reza Amini, LIG, Grenoble, France

Ion Androutsopoulos, AUEB, Athens, Greece

Thierry Artières, LIP6, Paris, France

Nicolas Baskiotis, LIP6, Paris, France

Patrick Gallinari, LIP6, Paris, France

Eric Gaussier, LIG, Grenoble, France

Aris Kosmopoulos, NCSR "Demokritos" & AUEB, Athens, Greece

George Paliouras, NCSR "Demokritos", Athens, Greece

Acknowledgements

Class-Y ANR project, University of Grenoble, University of Pierre and Marie Curie, NCSR "Demokritos", and Athens University of Economics and Business. We would also like to thank the Kaggle team for their support.

Leaderboard

1 anttip
2 Alexander D'yakonov (MSU, Moscow, Russia)
3 nagadomi
4 Martin Martin
5 paithan
6 student2012
7 Dmitriy Anisimov
8 Harvard Stat 183

29 discussion topics

Is there a way to know to which word a feature bel...
4 replies · an hour ago

Do Category IDs correspond to Wikipedia IDs?
0 replies · 2 months ago

Winning Solution Description
6 replies · 2 years ago

Hierarchy is not on labels
3 replies · 2 years ago

MaF-score: Theoretical performance of 0.5?
0 replies · 3 years ago

Launch Close

119 Teams
167 Competitors

Points This competition awarded standard ranking points
Tiers This competition counted towards tiers