Data extraction from a bibliographic data source and time series analysis.

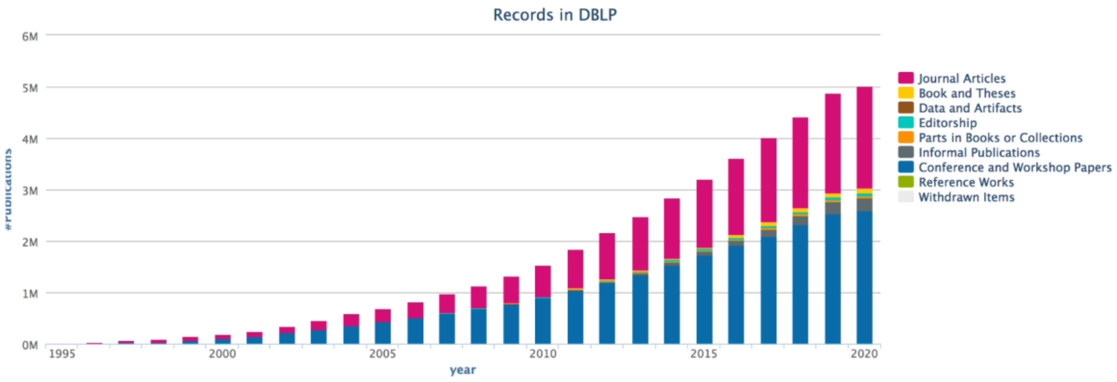
DESCRIPTION

DBLP (https://dblp.uni-trier.de/) is a website that maintains a bibliographic record of publications, mainly in the field of computer science. In addition to the ability to navigate and view the data of published papers through the site, this data is available as a compressed file in XML (Extensible Markup Language) format.

In the context of the project you are asked to extract from the above file the number of total publications per year. The extracted information will be in the form:

2020	543210
2019	432109
.....	....

and is illustrated graphically in the diagram below (illustrative representation). The extracted data correspond to a time series, which you are asked to study using time series analysis techniques.



Specifically, you are asked to build a prediction model that can be used to predict the total number of publications in the future. The prediction model should be evaluated on existing data for prediction accuracy. For example, you can use data up to 2010 to build the model, and test the prediction accuracy for years after 2010 for which you have the actual number of publications.

WORKING STEPS

The steps of the work are as follows:

- Step 1: Download from: http://dblp.org/xml/release/ the file: “dblp-2020-0401.xml.gz”. ATTENTION all work will be done with this file produced on 1/4/2020.
- Step 2 (**data preprocessing**): extract the number of publications made per year (it is suggested to use a programming language: C, Java, Python, ..., or Unix bark commands). Make a corresponding graph to the one above showing **the total number of publications per year** using a visualization tool (Gnuplot is suggested). Take care of the quality of the chart by following good practices.

- Step 3 (**data analysis**): apply time series analysis techniques to the extracted data, based on what you have been taught in the course and by reviewing the relevant literature. The techniques applied should be documented in the deliverables. The use and comparison of different techniques will be positively assessed.
- Step 4 (**evaluation of the analysis**): evaluate the accuracy of the prediction model, both quantitatively and by graphically visualizing the results (the use of Gnuplot is recommended).