Text Technology

Van Hoang, Florin Rheinwald, Pavlos Musenidis

December 28, 2020

1 Project Description

Motivation: To prepare a computational linguistics database in which users can do query on to find authors, certain topics and related papers.

1.1 Collect

Get the full ACL Anthology (with paper abstracts) in BibTeX format from its website[1].

1.2 Prepare

Write a grammar and use an XML Parser (i.e. ElementTree XML API on Python[2]) to encode the data into XML.

1.3 Access

Insert the grammar to a SQL or NoSQL database for querying. Ideally, users should be able to do query on (1) authors to find their papers, and (2) topics/terms (BERT, dependency parsing, LTSM, etc.) to find related papers sorted in decreasing year order. The terms are ideally included in the papers' titles.

SQL, NoSQL difference: https://www.xplenty.com/blog/the-sql-vs-nosql-difference/

1.4 Extension

Import the database to Neo4j[3] for querying and visualization.

Motivation: As Neo4j is a popular graph database, we want to learn about it and its advantages compared to SQL/NoSQL.

[1] https://www.aclweb.org/anthology/

[2] https://docs.python.org/3.8/library/xml.etree.elementtree.html

[3] https://neo4j.com/