Wikidata

David Arroyo Menéndez

March 8, 2020

Presentation

Wikidata: A Free Collaborative Knowledge Base Denny Vrandečić and
Markus Krötzsch
URL: https://ai.google/research/pubs/pub42240 Citations: 824
Published on: Association for Computational Machinery

Abstract

Unnoticed by most of its readers, Wikipedia is currently undergoing dramatic changes, as its sister project Wikidata introduces a new multilingual 'Wikipedia for data' to manage the factual information of the popular online encyclopedia. With Wikipedia's data becoming cleaned and integrated in a single location, opportunities arise for many new applications. In this article, we provide an extended overview of Wikidata, including its essential design choices and data model. Based on up-to-date statistics, we discuss the project's development so far and outline interesting application areas for this new resource.

We need good tools written from scratch by the teachers.

Choice of Programming Language

- A shallow learning curve, so that novice programmers get immediate rewards for their efforts
- The language must support rapid prototyping and a short develop/test cycle; an obligatory compilation step is a serious detraction
- The code should be self-documenting, with a transparent syntax and semantics
- It should be easy to write structured programs, ideally object-oriented but without the burden associated with languages like C++.
- The language must have an easy-to-use graphics library to support the development of graphical user interfaces

Design Criteria

Requirements

- Easy to use
- Consistency
- Extensibility
- Documentation
- Simplicity
- Modularity

Non Requirements

- Comprehensiveness
- Efficiency
- Cleverness

Modules

- Parsing Modules
- Tagging Modules
- Finite State Automata
- Type Checking
- Visualization
- Text Clasification

Documentation

- Tutorials
- Reference Documentation
- Technical Reports

Uses of NLTK

- Assingments (Example: Chunk Parsing)
- Class demonstrations (Example: Chart Parsing Tool)
- Advanced Projects (Example: Probabilistic Parsing)

Evaluation

- A possitive experience for students and teachers
- A problem was find corpora.

Other approaches

- Linguistic Students
- Grammar Developers
- Other Researchers and Developers

Conclusions and Future Work

The NLTK original idea was its combination of three factors:

1. It was deliberately designed as courseware and gives pedagogical goals primary status.
2. Its target audience consists of both linguists and computer scientists, and it is accessible and challenging at many levels of prior computational skill.
3. Finally, it is based on an object-oriented scripting language supporting rapid prototyping and literate programming.