Demo

WikiCSSH: Extracting CS Subject Headings from Wikipedia for scholarly data

Kanyao Han, Pingjing Yang, Shubhanshu Mishra and Jana Diesner

Workshop on Scientific Knowledge Graphs (SKG 2020)

This material is based upon work supported by the Korea Institute of Science and Technology Information under Grant No. C17031.

ILLINOIS
School of Information Sciences

Outline

- The WikiCSSH hierarchy
- Using WikiCSSH for tagging scholarly text
- Using Hierarchical Subject Headings for Computer Science
- Downloading WikiCSSH data and code

The WikiCSSH hierarchy

categories: 7,354
pages: 181,070
redirects: 580,312

Parents	Node	Children
• Computer_systems	• Embedded_systems	• Avionics_computers • Firmware • Onboard_computers • Real-time_computing • System_on_a_chip • Microcontrollers • Information_appliances • Avionics • Engine_control_systems • Ada_(programming_language) • Robots • Single-board_computers • Embedded_operating_systems • Automotive_software • Synchronous_programming_languages • Embedded_microprocessors • Graphing_calculators • Robotics

Pages

Ecomechatronics | Monokub | LiteOS | Point_of_sale | List_of_wireless_sensor_nodes | In-circuit_emulation | Digifant_engine_management_system | LwIP | JOVIAL | RMX_(operating_system) | Embedded_system | Embedded_Board_eXpandable | Power-on_reset | Simatic_S5_PLC | Sensor_node | Spy-Bi-Wire | Heartbeat_(computing) | Engine_control_unit | Paparazzi_Project | Tiva-C_LaunchPad | Sensing_floor | Micro_Bit | Cockpit_display_system | Continuous_Computing | OSIAN | OSGi | Background_debug_mode_interface | NicheStack_TCP/IPv4 | CMS-2_(programming_language) | DAvE_(Infineon) | Garmin | Unmanned_aerial_vehicle | Chassis_Management_Controller | CEN/XFS | Motronic | Assembly_language | Unified_Diagnostic_Services | RAM_image | Picotux | Slugs_(autopilot_system) | Intelligent_environment | Input_capture | Tetrix_Robotics_Kit | In-target_probe | PBASIC | DO-254 | Tiger-BASIC | Microcontroller | List_of_PowerPC-based_game_consoles | CompactRIO | Powertrain_control_module | Flash_memory_emulator | Hardware-in-the-loop_simulation | Oscillator_start-up_timer | Pacemaker_crosstalk | Context-aware_pervasive_systems | MPLAB | Hume_(programming_language) | Embedded_software | Hawkboard | Lego_Mindstorms | FITkit_(hardware) | Smart_camera | Cardiac_Pacemakers,_Inc. | Hardware_reset | Immunity-aware_programming | MyRIO | Robotic_spacecraft | SECU-3 | Embedded_hypervisor | InfinityDB | IAR_Systems | Radisys | MULTICUBE | Universal_Avionics | Microwave_Imaging_Radiometer_with_Aperture_Synthesis | Automatic_system_recovery | Telematic_control_unit | Jetronic | Falcon_(programming_language) | TRANZ_330 | JTAG | Embedded_C | Watchdog_timer | Direct_numerical_control | Front-end_processor | Standard_Test_and_Programming_Language | IGEPv2 | Systxc3xa8me_d'aide_xc3xa0_la_conduite,_xc3xa0_l'exploitation_et_xc3xa0_la_maintenance | Ceibo_emulator | Honeywell | Bond-out_processor | Computer-on-module | ISEE_(company) | Low-voltage_detect | CodeSynthesis_XSD/e | Autotech | West_Bridge | HAL/S | EEMBC | Flexible-fuel_vehicle | AVR_Butterfly | Diebold_10xx | PC/104 | MXCHIP | Shaheen-III | UAV-related_events | Barcode_reader | Open_Programming_Language | Handheld_game_console | Open_Kernel_Labs | Wahoo_Fitness | Ada_(programming_language) | Biological_pacemaker | Corelis | Execute_in_place | Peripheral_DMA_controller | CompactDAQ | Otis_Boykin | SREC_(file_format) | Ultra-low-voltage_processor | SUMIT | Automated_teller_machine | Interactive_kiosk | Emission-aware_programming | Graphing_calculator | MCU_8051_IDE | Board_support_package | Tektronix_extended_HEX | Xpeak | Passenger_drone | ROM_image | Mobile_phone | UniPro_protocol_stack | PX4_autopilot | ADvantage_Framework | Tessy_(software) | Logic_analyzer | ClearSpeed | TI_StarterWare | AC_20-115 | Artificial_cardiac_pacemaker | VersaLogic | C166_family | MISRA_C | System_on_module | Pacemaker_failure | LGM-30_Minuteman | Worst-case_execution_time | Atom_(programming_language) | Apache_Celix | Embedded_C++ | Tillie_the_All-Time_Teller | UIP_(micro_IP) | Udhcpc | Radio_science_subsystem | Biotronik | MicroPython | Bit_banging | Gas_flow_computer | Vortex86 | Trillium_Digital_Systems | SolidRun | Coremark | OLogic | Nano-RK | EPIA | BasicX | IC_programming | Embedded_Java | Priority_inversion | Communication_Access_Programming_Language | Mechatronics | DO-178C | M-Module | Teller_assist_unit | Mikroelektronika | Instant-on | ILAND_project | Electronic_control_unit | Debit_card | NesC | Firmware | Intel_HEX | List_of_Wi-Fi_microcontrollers | Output_compare | DO-160 | ZMDI | Venus_Express | Rockwell_Collins | FeaturePak | SiRFstarIII | ArduPilot | Lego_Mindstorms_EV3 | Lua_(programming_language) | Dynamic_simulation | SWAP_(instrument) | Anti-hijack_system | S-TEC_Corporation | DO-178B | UniPro

Using WikiCSSH for tagging scholarly text

Currently we just tag based on exact match of a keyword in WikiCSSH

Tagged document:

Methods for extracting entities (methods, research topics, technologies, tasks, materials, metrics, research contributions) and relationships from research publications
Methods for extracting metadata about authors, documents, datasets, grants, affiliations and others.
Data models (e.g., ontologies, vocabularies, schemas) for the description of scholarly data and the linking between scholarly data/software and academic papers that report or cite them
Description of citations for scholarly articles, data and software and their interrelationships
Applications for the (semi-)automatic annotation of scholarly papers
Theoretical models describing the rhetorical and argumentative structure of scholarly papers and their application in practice
Methods for quality assessment of scientific knowledge graphs
Description and use of provenance information of scholarly data
Methods for the exploration, retrieval and visualization of scientific knowledge graphs
Pattern discovery of scholarly data
Scientific claims identification from textual contents
Automatic or semi-automatic approaches to making sense of research dynamics
Content- and data-based analysis on scholarly papers
Automatic semantic enhancement of existing scholarly libraries and papers
Reconstruction, forecasting and monitoring of scholarly data
Novel user interfaces for interaction with paper, metadata, content, software and data
Visualisation of related papers or data according to multiple dimensions (semantic similarity of abstracts, keywords, etc.)
Applications for making sense of scholarly

Predicted categories:

Academia (12) | Methodology (11) | Knowledge (6) | Research_methods (4) | Research (4) | Meaning_(philosophy_of_language) (3) | Computer_science (3) | Software (3) | Metadata (2) | Ontology (2) | Graphs (2) | Data_modeling_diagrams (1) | Metrics (1) | Information_science (1) | Data_modeling (1) | Lexicography (1) | Vocabulary (1) | Numeral_systems (1) | Inductive_reasoning (1) | Abstraction (1) | Theories (1) | Critical_thinking_skills (1) | Quality_assurance (1) | Information (1) | Identification (1) | Structuralism (1) | Analysis (1) | Library_science (1) | Forecasting (1) | Human-machine_interaction (1) | Virtual_reality (1) | User_interfaces (1) | User_interface_techniques (1) | Papermaking (1) | Packaging_materials (1) | Printing_materials (1) | Mathematical_concepts (1) | Abstract_algebra (1) | Geometric_measurement (1) | Dimension (1) | Mathematical_notation (1) | Punctuation (1)

Need for a Subject Headings for Computer Science

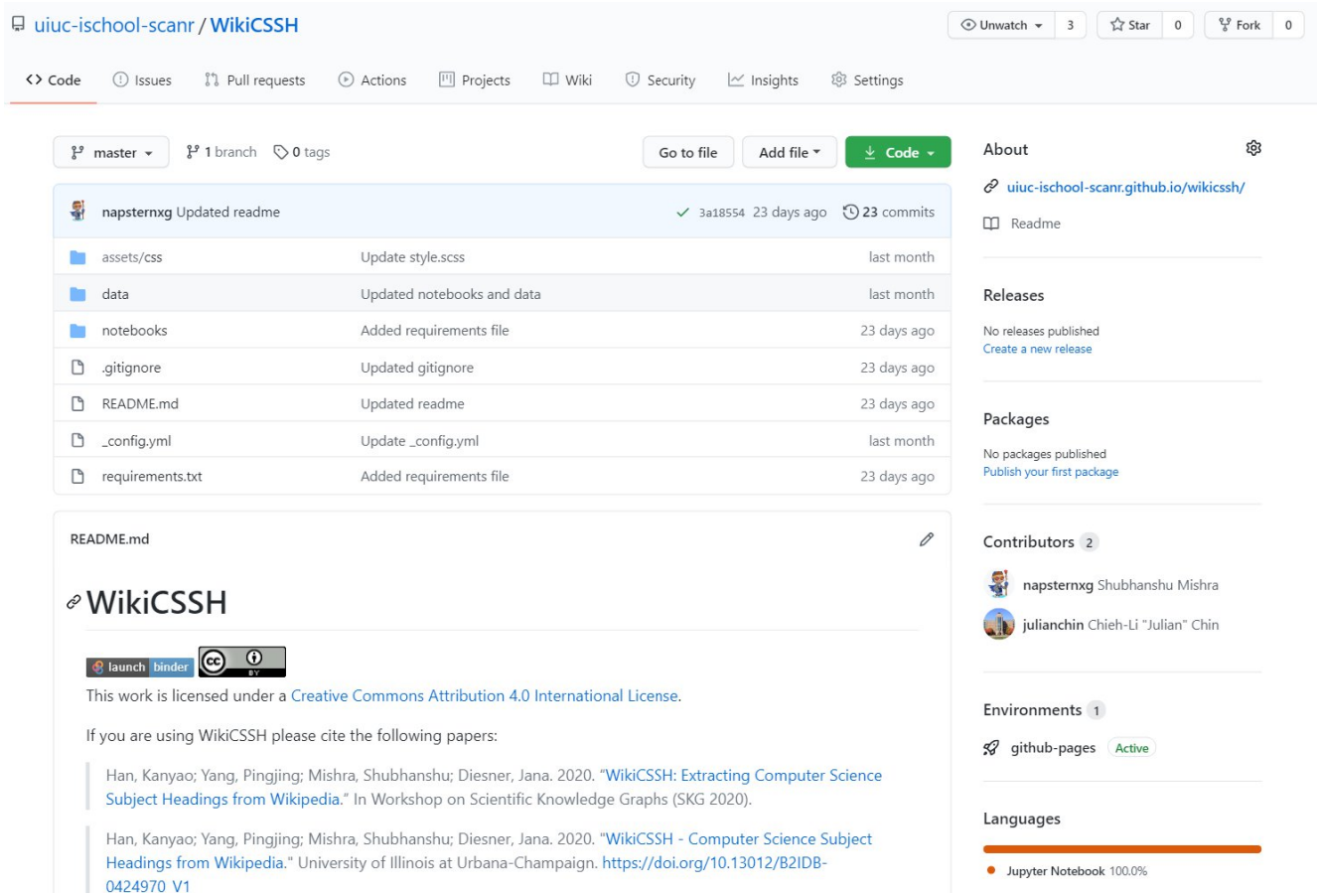
Hierarchical Subject Headings allow us:

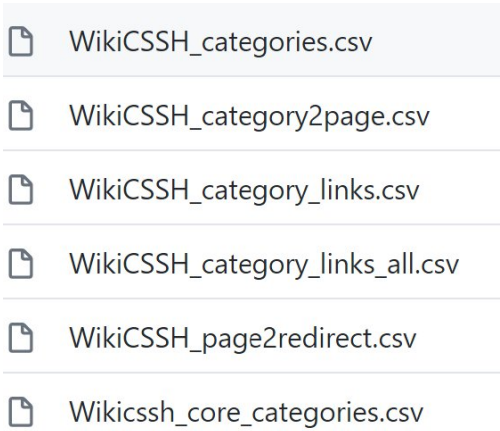
- Understand temporal evolution of concepts in scholarly data [1]
- Track novelty of authors over time [1]
- Quantify relative conceptual expertise of authors on a paper [2]
- Compute expertise of authors on concepts over time [2]

[1] Mishra, Shubhanshu, and Vetle I. Torvik. 2016. “Quantifying Conceptual Novelty in the Biomedical Literature.” *D-Lib Magazine : The Magazine of the Digital Library Forum* 22 (9–10). https://doi.org/10.1045/september2016-mishra.

[2] Mishra, Shubhanshu, Brent D. Fegley, Jana Diesner, and Vetle I. Torvik. 2018. “Expertise as an Aspect of Author

Downloading WikiCSSH





Code: https://github.com/uiuc-ischool-scanr/WikiCSSH
Data: https://databank.illinois.edu/datasets/IDB-0424970
Website: https://uiuc-ischool-scanr.github.io/WikiCSSH/

Thank You

- Code: https://github.com/uiuc-ischool-scanr/WikiCSSH
- Data: https://databank.illinois.edu/datasets/IDB-0424970
- Project website: https://uiuc-ischool-scanr.github.io/WikiCSSH/
- Citations:
  - Han, Kanyao, Pingjing Yang, Shubhanshu Mishra, and Jana Diesner. 2020. "WikiCSSH: Extracting Computer Science Subject Headings from Wikipedia." In Workshop on Scientific Knowledge Graphs (SKG 2020).
  - Han, Kanyao; Yang, Pingjin; Mishra, Shubhanshu; Diesner, Jana (2020) WikiCSSH - Computer Science Subject Headings from Wikipedia. University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-0424970_V1
- This material is based upon work supported by the Korea Institute of Science and Technology Information under Grant No. C17031.