



Software / SW

Kategorie výzkumu, experimentálního vývoje a inovací EV – experimentální vývoj

Druh publikačního výstupu VaVal: R

Vývoj SW SW byl vyvíjen jako hlavní výstup projektu podpořeného Technologickou agenturou ČR Výzkum a vývoj podpůrných sítí a informačních systémů pro neformální pečující o osoby po cévních mozkových příhodách (TAČR–VISNEP), registrační číslo projektu TL02000050, identifikační číslo výsledku TL02000050-V1.

Název SW **CMP-Rádce-klasifikátor textů**

Autorský tým SW Petr Šaloun, David Andrešič, Lenka Krhutová, Miroslav Paulíček

Distributor SW Vysoká škola báňská – Technická univerzita Ostrava / VŠB – TUO a Ostravská univerzita / OU

Datum vydání SW 30. 11. 2021

Datum veřejného zpřístupnění SW 3. 1. 2022 v licenci Creative Commons CC-BY-NC-SA



Dostupné: <https://github.com/tacrcmp/cmp-radce>

Požadavky na systém Linux Debian server, webový server (Nginx, Apache) s podporou PHP minimálně ve verzi 7.3. CMS Wordpress + plugins jako jQCloud... Poštovní server Postfix, Docker ve kterém běží: vlastní web CMP Rádce, Nextcloud projektu, databázový server MariaDB. Klasifikátor vyžaduje Java JRE ve verzi min. 8, raději však 11 (testováno na OpenJDK 11 64bit) a MySQL (testováno na verzi 8).

Popis SW

Funkční běžící informační systém vytvořený pro potřeby aplikačního garanta TAČR–VISNEP a hlavního příjemce výsledku SW, tj. Sdružení pro rehabilitaci osob po cévních mozkových příhodách z. s. komplexně podporující neformální pečující o blízké po CMP v kontextu jejich nízkoprahové informační podpory.

Softwarem je vyvinutý klasifikátor textů, který analyzuje dokumenty dostupné v rámci serveru CMP Rádce (dále „CMP-R“) a klasifikuje je podle kategorií tamtéž, pro potřeby navigace nad obsahem serveru i obsažené dokumentace. Jeho výstupy se transformují do databáze a vytváří/aktualizují slovní mrak (tag-cloud).

Tento původní software je začleněn do serveru, na kterém běží informační systém CMP-R a pod licencí Creative Commons typu CC-BY-NC-SA a je k dispozici zájemcům pro další využití a rozvoj. Klasifikátor je založen na původních algoritmech spojujících zpracování textů v přirozeném jazyce s umělou inteligencí, kategorizuje dokumenty dostupné na serveru CMP-R

a naviguje nad jejich obsahem pomocí tag-cloud. Jeho hlavním cílem je identifikovat klíčová slova každého dokumentu a sestavit z nich model jednotlivých kategorií obsahující základní morfologické prvky nejvýstižnějších klíčových slov seřazených dle důležitosti. Z těchto slov pak generuje tag-cloud sloužící k navigaci nad obsahem CMP-R. Nové dokumenty pak klasifikuje na základě podobnosti s tímto modelem. Klasifikátor je také schopen se průběžně "doučovat" z nově přidávaných dokumentů formou přepočtu modelu tříd a je tak rovněž schopen si poradit s rozdílnou větnou skladbou uživatelů CMP-R, jejich překlady a "stylem psaní". Samotný algoritmus je také díky svému principu schopen si v budoucnu poradit i s jinými jazyky, než je čeština, což otevírá cestu dalším jazykovým mutacím.