Odpovídání na otázky nad českou Wikipedií

Autor: Jonáš Sasín
Vedoucí: doc. RNDr. Pavel Smrž, Ph.D.

2020/2021
jonasssn@gmail.com

Popis řešení

Cílem bakalářské práce je prozkoumat dostupné možnosti a vytvořit systém odpovídající na otázky nad otevřenou doménou pro češtinu.

Retriever

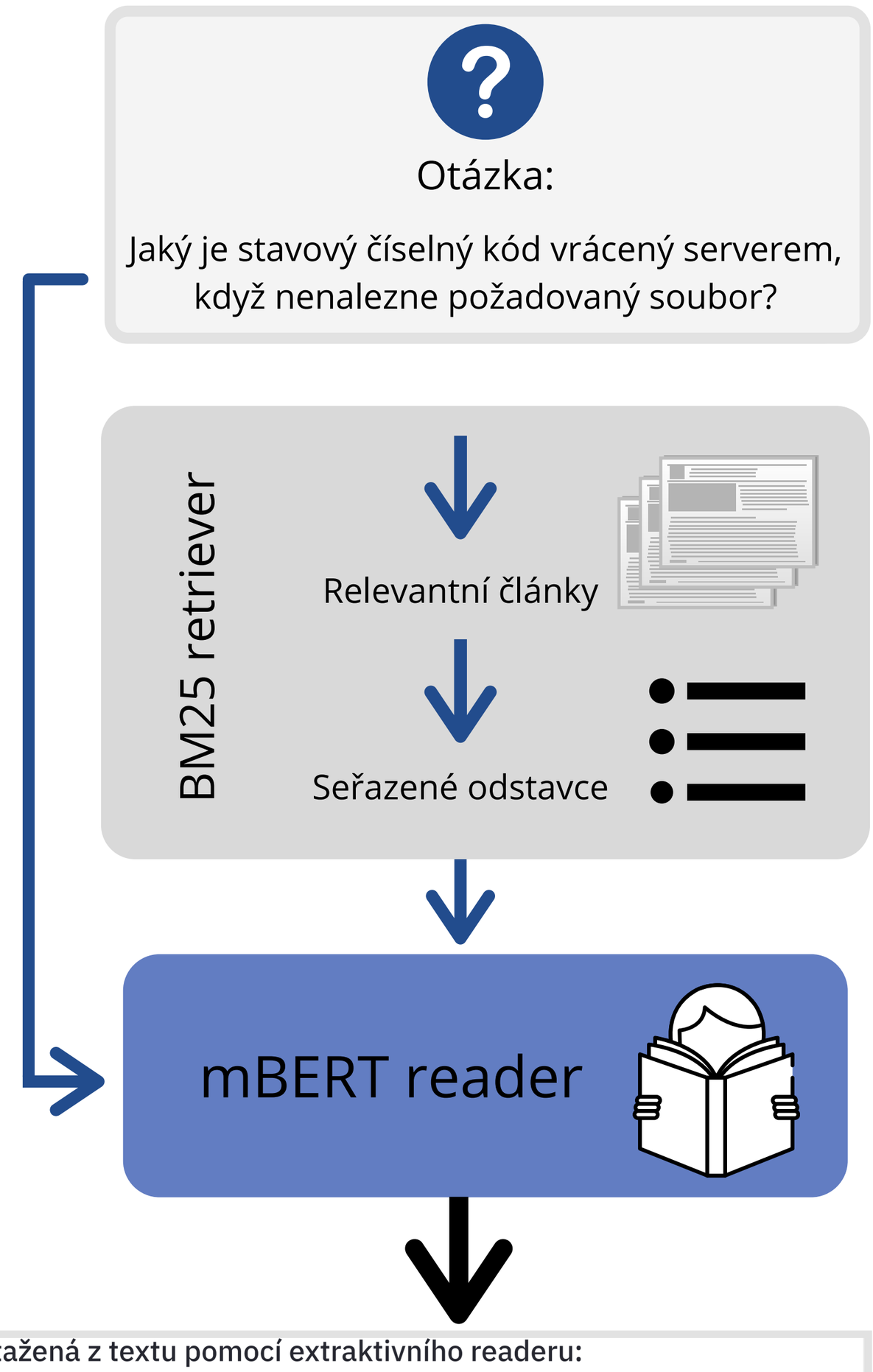
Systém nejdříve získá relevantní články, které zpracuje a rozdělí na odstavce. Ty jsou pak pomocí funkce BM25 ohodnoceny dle relevance k otázce.

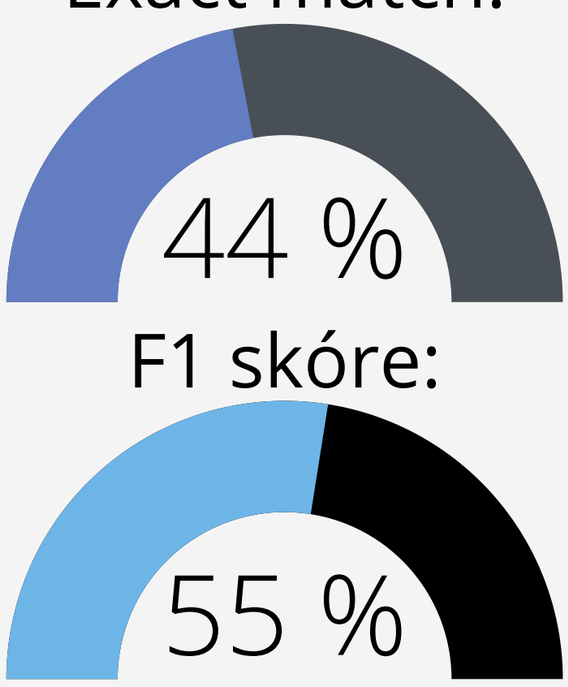
Reader

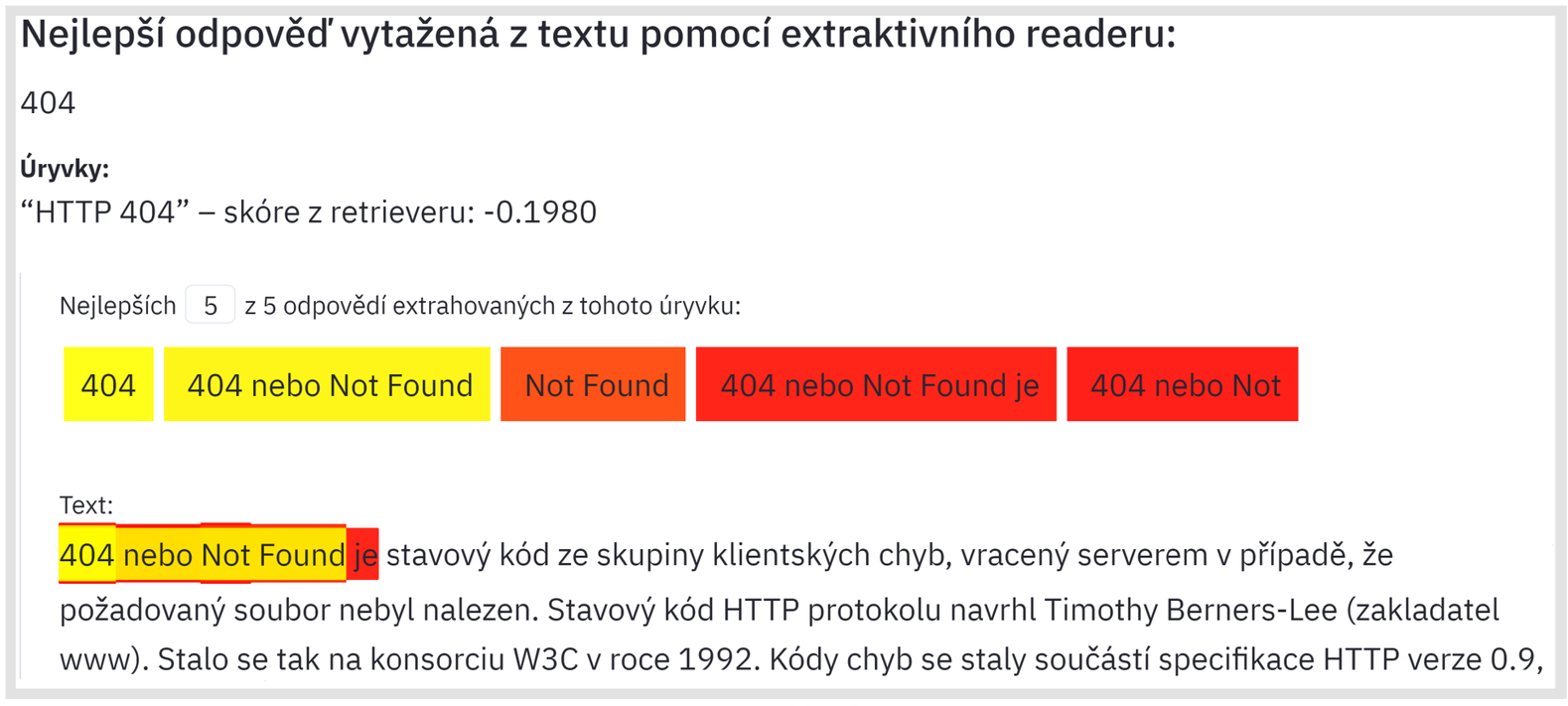
Pro extrakci odpovědi je použit vícejazyčný model BERT trénovaný na českém překladu datasetu SQuAD 2.0.

Dosažené výsledky

Systém byl vyhodnocen na datové sadě SQAD v3.0, čítající 11273 otázek (bez ano/ne). Dosahuje úspěšnosti **0,44 EM** a **0,55 F1** skóre. Při ručním vyhodnocení je úspěšnost až 69 %. Výsledné řešení převyšuje výsledky systému AQA z roku 2018, který je nejrelevantnějším porovnáním pro výstupy této práce. Vytvořena byla také demonstrační aplikace prezentující funkčnost řešení.







Demonstrační aplikace: r2d2.fit.vutbr.cz/.cs