

# 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

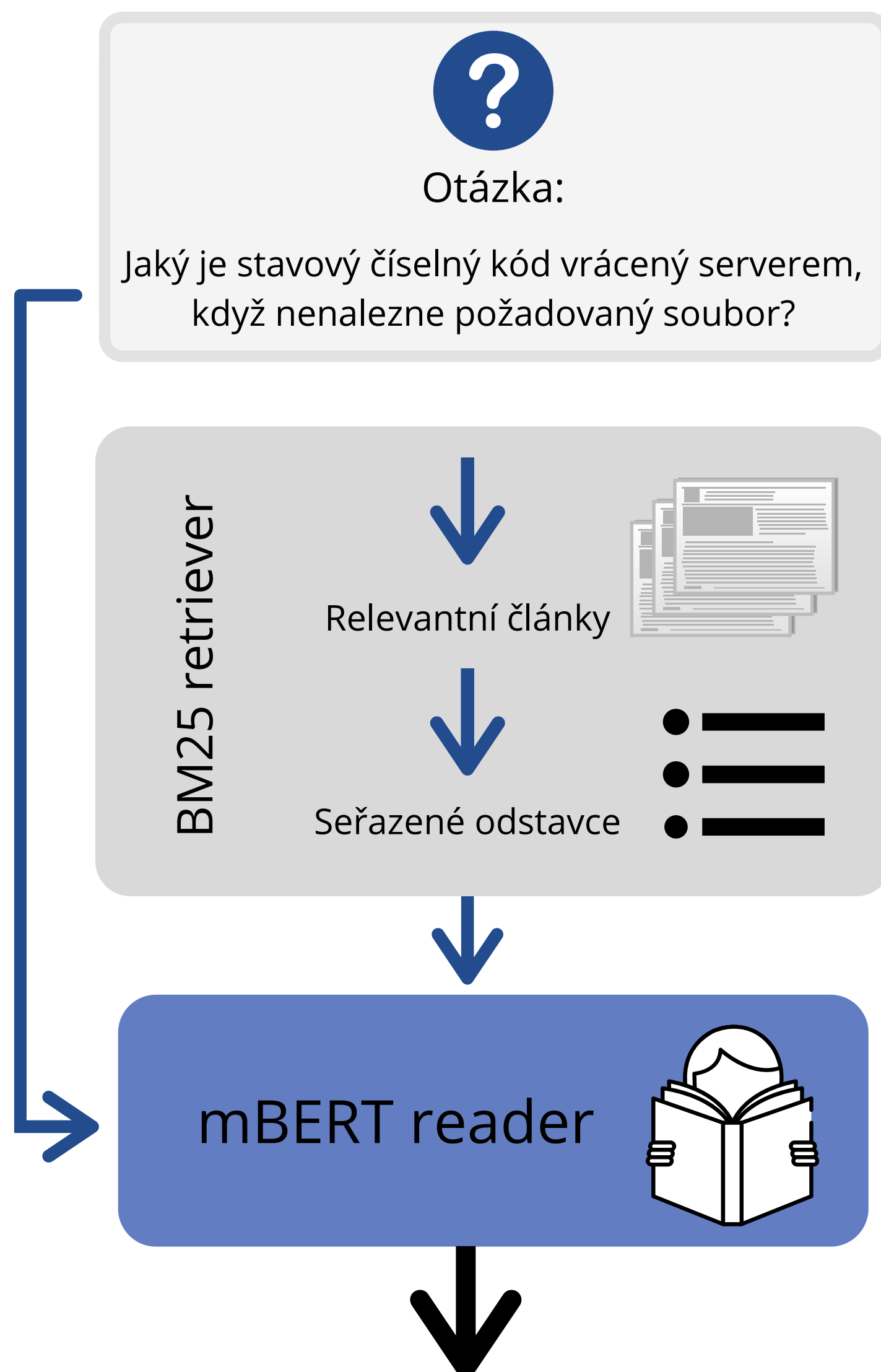
Systém nejdříve získá relevantní články, které zpracuje a rozdělí na odstavce. Ty jsou pak pomocí funkce BM25 ohodnoceny dle relevance k otázce.

### Reader

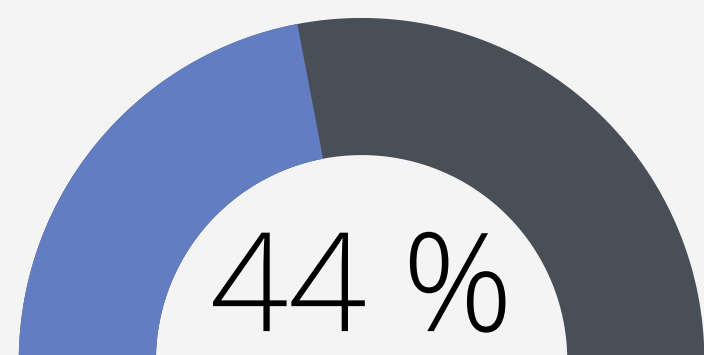
Pro extrakci odpovědi je použit vícejazyčný model BERT trénovaný na českém překladu datasetu SQuAD 2.0.

## Dosažené výsledky

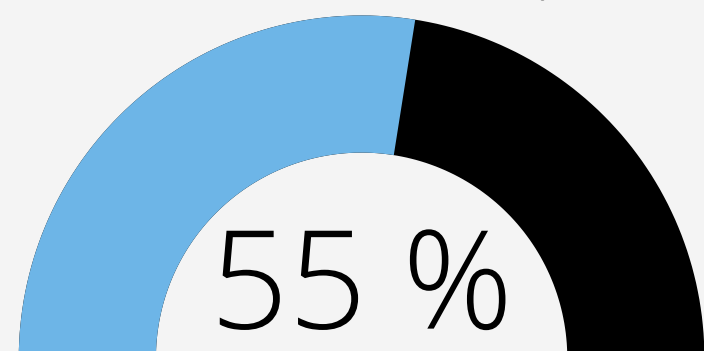
Systém byl vyhodnocen na datové sadě SQuAD v3.0, čítající 11273 otázek (bez ano/ne). Dosahuje úspěšnosti **0,44 EM** a **0,55 F1** skóre. Při ručním vyhodnocení je úspěšnost až 69 %. Výsledné řešení převyšuje výsledky systému AQA z roku 2018, který je nejrelevantnějším porovnáním pro výstupy této práce.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



Exact match:



F1 skóre:



Nejlepší odpověď vytažená z textu pomocí extraktivního readeru:

404

Úryvky:

“HTTP 404” – skóre z retrieveru: -0.1980

Nejlepších 5 z 5 odpovědí extrahovaných z tohoto úryvku:

404

404 nebo Not Found

Not Found

404 nebo Not Found je

404 nebo Not

Text:

404 nebo Not Found je stavový kód ze skupiny klientských chyb, vrácený serverem v případě, že požadovaný soubor nebyl nalezen. Stavový kód HTTP protokolu navrhl Timothy Berners-Lee (zakladatel www). Stalo se tak na konsorciu W3C v roce 1992. Kódy chyb se staly součástí specifikace HTTP verze 0.9,

Demonstrační aplikace: [r2d2.fit.vutbr.cz/cs](https://r2d2.fit.vutbr.cz/cs)