KNN - Automatické hodnocení písemek ze SUI (Check-Point odevzdání)

Autoři: xkaska01, xhemza05, xmagda03

1. Odkaz na repozitář:
https://github.com/atepos/knn.git

2. Definice úlohy:

Cílem je poloautomatizovat hodnocení písemných odpovědí studentů, které byly získány ze zkoušek SUI pomocí OCR.

Vstupem do našeho systému jsou digitálně naskenované zkoušky, u nichž je text převeden do digitální podoby pomocí OCR, a zároveň máme k dispozici jednotlivé hodnocení úloh. Víme, že každá otázka je ohodnocena maximálně 4 body na základě manuálního hodnocení. Na rozdíl od některých přístupů, kde se využívají referenční odpovědi, v našem případě přímo využíváme konečné známky jako referenční měřítko.

Výstupem našeho systému bude automaticky generované hodnocení každé odpovědi, které se snaží co nejvěrněji napodobit manuální hodnocení. Tímto způsobem budeme experimentovat s různými modeli dotrénovaných k automatickému hodnocení, přičemž manuálně udělené skóre slouží jako standard pro kalibraci a validaci výsledků.

3. Krátký přehled existujících řešení:

- **Automated Assignment Grading with Large Language Models: Insights From a Bioinformatics Course** - Tento článek jsme si vybrali i k prezentaci a tedy ho podrobněji rozebereme ve videu - Studie hodnotí využití velkých jazykových modelů (LLMs) pro automatické hodnocení písemných úkolů v kurzu bioinformatiky, kde více než 100 studentů odpovědělo na 36 textových otázek. Výsledky ukazují, že LLMs mohou poskytovat zpětnou vazbu srovnatelnou s lidskými hodnotiteli, přičemž open-source modely dosahují podobné přesnosti jako komerční řešení, což umožňuje školám zachovat soukromí studentů.
- **e-rater (ETS):**
  Využívají lingvistické a statistické metody pro hodnocení esejí. Přestože je primárně určen pro hodnocení esejí.
- **ASAP Short Answer Scoring:**
  Porovnává studentovy odpovědi s lidskými anotacemi pomocí statistických metod. Dosahuje výsledků srovnatelných s lidským hodnocením.
- **Can Large Language Models Be an Alternative to Human Evaluations?**
  Autoři ukazují, že hodnocení provedené LLM koreluje s výsledky expertů, což potvrzuje potenciál těchto modelů jako alternativy k manuálnímu hodnocení.
- **Check-Eval: A Checklist-based Approach for Evaluating Text Quality**
  Autoři v tomto článku představují nový rámec CHECK-EVAL určený k hodnocení textů. Vycházejí z předpokladu, že tradiční metriky (BLEU, ROUGE, METEOR)

často nedostatečně korelují s lidským hodnocením, zejména u kreativních či nuančních úloh.

4. Trénovací Dataset:

Dataset jsme přejali od studentů z minulého roku, kteří pomocí OCR převedli ručně psané písemky na jejich .json podoby. My jsme následně pro přehlednost jejich dataset upravili a nahráli do repozitáře zde. Provedli jsme následující úpravy:

Odstranili jsme z něj nepodstatné informace pro naše řešení jako (zahashovaný login studenta, rok vykonání zkoušky a zahashované ID otázky). Dále jsme upravili celkovou strukturu .jsonu, kdy v původním datasetu byly známky ke každé otázce v hlavičce a následovali jednotlivé pomíchané otázky s ukazatel k dané známce v hlavičce - my jsme známky z hlavičky vyjmuli a dali je ke každé otázce zvlášť. Dále bylo potřeba vyřešit chybějící otázky, tedy předchozí studenti špatně naparsovali data a např. v odpovědi se nacházel text k otázce i odpovědi zároveň - tyto případy jsme korektně separovali do `hodnot` náležitých `klíčů` v jsonu. Také bylo nutné opravit špatně namapované číslá otázek k náležitým otázkám. A tedy výsledná struktura datasetu je následující:

```
[
...
    {
        ...
    },
    {
        "questionNumber": číslo otázky,
        "questionText": Text zadané otázky,
        "answerTest": Odpověď studenta,
        "score": Získané skóre za opověď
    },
    {
        ...
    },
...
]
```

5. Způsob vyhodnocení:

V našem řešení hodnotíme kvalitu písemných odpovědí porovnáním výstupu modelu s manuálním hodnocením vyučujícího. Používáme metodu, která spočívá v tréninku modelu, aby na základě zadané otázky a odpovědi předpovídal bodové hodnocení, jež se co nejvíce blíží referenčnímu hodnocení udělenému lidským hodnotitelem.g

- **Definice a hodnotový rozsah:**
  Model předpovídá skóre, které je na škále například od 0 do 4 (0 – nejhorší, 4 – nejlepší).
- **Interpretace výsledků:**
  Pro vyhodnocení kvality modelu využijeme Mean Absolute Error. Pokud model například vrátí skóre 2 a ground truth bude 2 znamená to, že odpověď byla predikována

správně. Pokud model vrátí 1 a ground truth bude 4 znamená to, že model se spletl. Tímto způsobem projdeme všechny vzorky z testovacího datasetu, vypočítáme absolutní hodnotu rozdílu, poté výsledky sečteme a podělíme celkovým počtem vzorků z testovacího datasetu. Tedy v ideálním případě, kdy model všechny studentské odpovědi ohodnotí stejně jako učitel by byl výsledek 0.

Takže budeme opakovatelně hodnotit písemné odpovědi studentů a postupně zlepšovat přesnost modelu tak, aby co nejlépe napodoboval lidské hodnocení.

6. Baseline řešení

Jako Baseline je možné použít řešení našich předchůdců odkaz na jejich repozitář s jejich výsledky. Jedná se o finetunované modely `Ada`, `Babbage`, `Curie` a `Davinci`.

Kromě toho jsme natrénovali vlastní model – finetunovaný **DistilBERT** na nově vytvořeném datasetu. Tento model používáme ke klasifikaci, přičemž dataset jsme upravili do vhodného formátu. Skóre jsme rozdělili do kategorií:

`[0, 0.25, 0.5, ..., 4.0] → [0,1,2,...,16]`

Otázky s odpověďmi jsme transformovali do formátu:

`"{question} {answer}"`

Prvotní model dosáhl *MAE ~1.64*.

Použité trénovací parametry:

- **Dataset:** `2021_1_A.json`
- **Learning rate:** `2e-5`
- **Počet epoch:** `12`
- **Weight decay:** `0.01`

7. Plán útoku:

V dalším kroku plánujeme otestovat využití LLM modelů, u kterých předpokládáme, že dosáhnou lepších výsledků než naše současné řešení. Tento předpoklad také vychází z dosavadních výsledků našich předchůdců, které naznačují, že by tyto modely mohli být vhodnější.

Reference:

Automated Assignment Grading with Large Language Models: Insights From a Bioinformatics Course: https://arxiv.org/abs/2501.14499 ETS e-rater: https://www.ets.org/erater.html ASAP Short Answer Scoring: https://www.kaggle.com/competitions/asap-sas
Can Large Language Models Be an Alternative to Human Evaluations?: https://doi.org/10.48550/arXiv.2305.01937
Check-Eval: A Checklist-based Approach for Evaluating Text Quality: https://doi.org/10.48550/arXiv.2407.14467