УНИВЕРЗИТЕТ „Св. Кирил И Методиј“ – Скопје
Факултет за Информатички науки и компјутерско инженерство

-ПРОЕКТ-
по предметот
ВОВЕД ВО НАУКАТА ЗА ПОДАТОЦИ

Тема

Креирање и анализа на податочно множество со прашања и одговори од доменот на ПМ

Ментор:
Проф. Д-р Слободан Калајџиски
М-р Благоја Јанкоски

Изработила:
Сара Манасиева (216054)

Скопје, јули 2024

Вовед

Во рамките на овој проект, главната цел е да се направи податочно множество со прашања и одговори од областа на Проектен менаџмент (ПМ). Конкретно, јас се фокусирам на екстракција и прочистување на прашања од PDF документи и подоцна ги структурирам во .jsonl формат кој го делам на тренинг, тест и валидациско множество со користење на квантитативна статистичка анализа.

Екстракција и прочистување на прашања од пдф документи

За екстракција на прашања и одговори од PDF документите, користам две основни библиотеки:

1. **pdfplumber**: Оваа библиотека ми овозможува да отворам и да работам со PDF документи во Python. Со неа можам да извлечам текст од секоја страница на PDF документот. Првичниот чекор е да ја отворам PDF датотеката и да ја извлечам содржината од секоја од страниците.
2. **re (регуларни изрази)**: За да го пронајдам и издвојам потребниот текст кој содржи прашања и одговори, користам регуларни изрази. Регуларните изрази ми овозможуваат да дефинирам шаблони за текст кои ги користам за да го локализирам и издвојам секојот дел од текстот кој ми е потребен за прашањата и одговорите.

Процесот на прочистување на текстот вклучува следните чекори:

- **Отстранување на header, footer и број на страна**: Претпоставувајќи дека во PDF документот има заглавја, подножја и броеви на страници кои не се дел од прашањата и одговорите, користам регуларни изрази за да ги отстранам од текстот.
- **Отстранување на веб линкови и референци**: Во многу PDF документи може да има веб линкови или референци кои не се дел од прашањата и одговорите. Со регуларни изрази ги отстранувам овие делови од текстот.
- **Отстранување на специјални знаци, математички формули, празни места итн.**: За да го подобрам квалитетот на текстот и да го оставам само со содржина што е потребна, користам регуларни изрази за да отстранам специјални знаци, математички формули, празни места и други непотребни карактери.

Овие чекори ми помагаат да го прочистам текстот од PDF документот и да го подготвам за креирање на JSONL фајл со прашања и одговори кои потоа може да се користат за квантитативна статистичка анализа и обучување на модели.

Квантитативна статистичка анализа

За квантитативната статистичка анализа на pmQA со визуелизација користев следни библиотеки и техники:

1. **NLTK**: За токенизација на текстот, идентификација на сврзувачки термини (како "и", "но") и термини за негација (како "не", "никогаш").

2. **Fuzzywuzzy**: За споредба на текстови и откривање на термини за негација со специфична граматичка структура.
3. **TextBlob**: За анализа на сентимент и идентификација на термини за негација со користење на регуларни изрази.
4. **Udpipe**: За морфосинтаксичка анализа на реченици и идентификација на пропозиционални фрази.

Оваа анализа вклучува детекција на негативни изрази, употреба на сврзувачки термини и размислувачки изрази во текстот на pmQA, како и визуелизација на овие податоци преку генерирање на квантитативни мерки како број на фрази, содржечки изрази и други податоци во табеларен формат за подетална анализа.

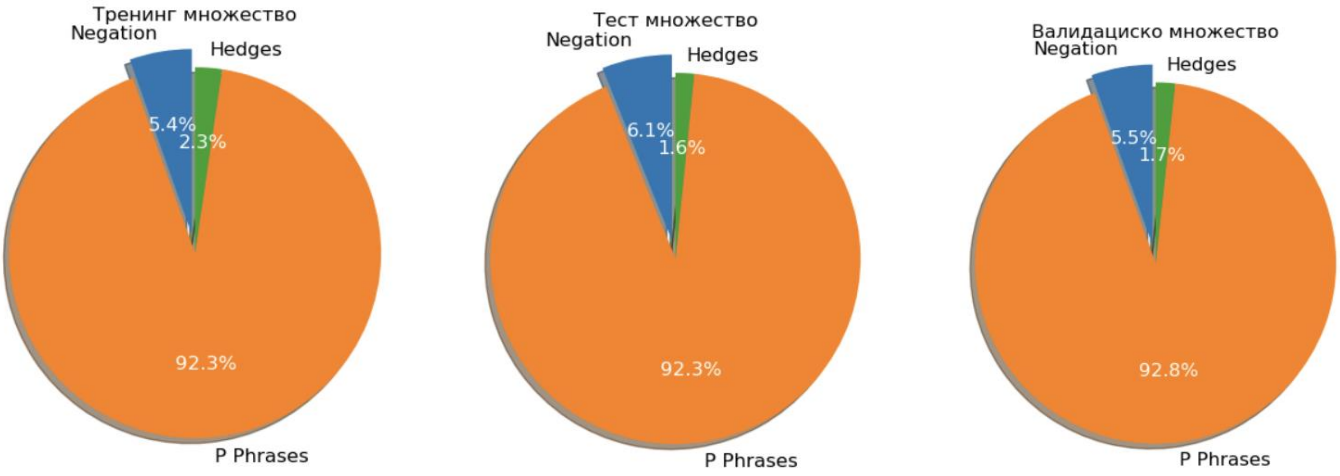
Рамномерно дистрибуирано податочно множество за прашања и одговори

Во овој чекор го делам податочното множество на тренинг, тест и валидациско множество во сооднос 80%, 10%, 10%. При податочното множество е рамномерно дистрибуирано во однос на :

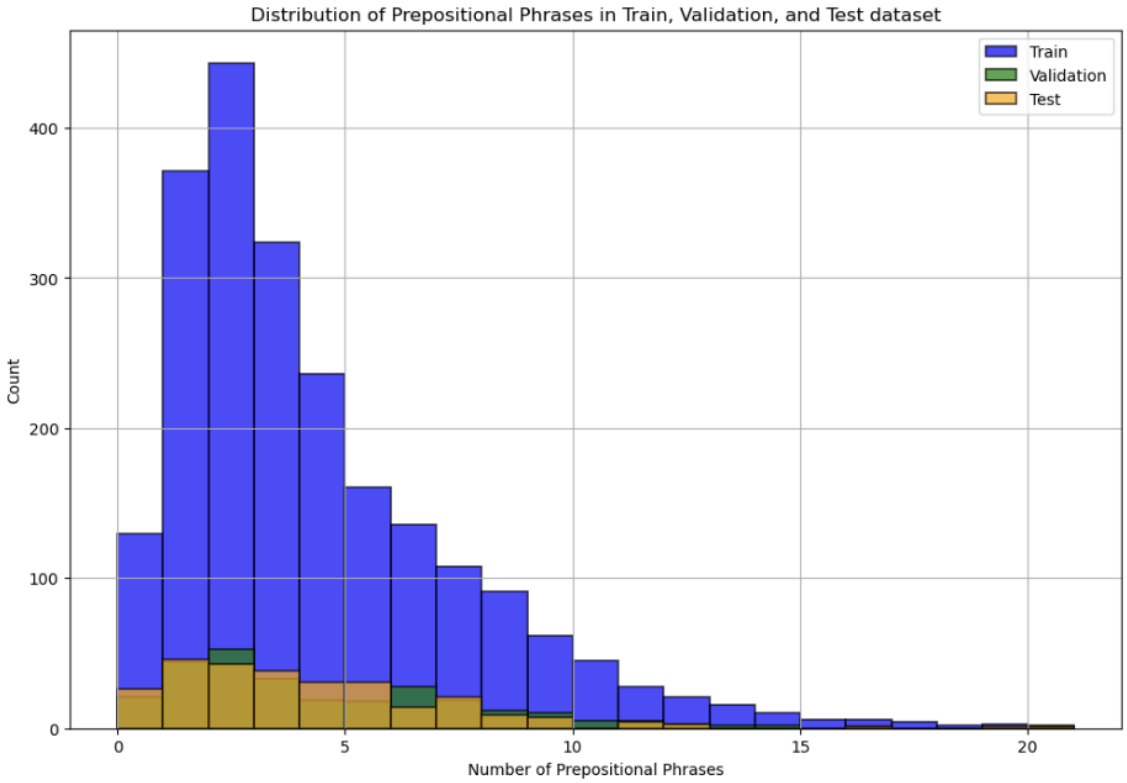
a. присуство на сврзувачки термини (на пр. и, но)
b. присуство на термини за негација (на пр. не, никогаш)
c. присуство на услови за заштита (на пример, понекогаш, можеби)

Визуелизација

Слика 1.



Слика 2.



Слика3.