#KARATEKIID

Picipolo

Zarys projektu

Projekt ma na celu automatyczne pobieranie dokumentów KIID oraz procesowanie ich

Nasze rozwiązanie składa się z niezależnych modułów: info_csv, bag_of_words, data_extracter, expression_checker

Taka forma zapewnia, że kod jest przejrzysty oraz pozwala na łatwy rozwój aplikacji w przyszłości poprzez dodawanie nowych modułów

Taki rodzaj rozwiązań zapewnia wysoką skalowalność

1. Webscrapping

W celu pobranie dokumentów ze stron zaimplementowaliśmy webscraper w oparciu o pakiet BeautifulSoup4

Nasza implementacja wykorzystuje rekurencyjne przeglądanie podstron oraz automatycznie pobiera pliki pdf, których nazwa wskazuje na to, że jest to dokument KIID

Aby uniknąć zapętlenia się rekurencji, nasz algorytm zapamiętuje wcześniej odwiedzone strony, a także umożliwia ustawienie głębokości poszukiwań

W przypadku napotkania znacznej ilości szukanych dokumentów, algorytm przerywa dalsze szukanie podstron z dokumentami, aby przyspieszyć działanie programu

2. Data extraction

Ekstrakcja danych została oparta na wyrażeniach regularnych (Regex).

W tak krótkim czasie jaki przeznaczony jest na hackaton nie byliśmy w stanie zaimplementować rozwiązania, które byłoby w stanie rozumieć kontekst dokumentu.

3. BagOfWords

Pobraliśmy listę polskich „stopwords”, a następnie usunęliśmy je z pliku wraz ze znakami interpunkcyjnymi oraz specjalnymi

W celu normalizacji danych użyliśmy lematyzacji i usunęliśmy z tekstu wszystkie cyfry

Końcowo zaimplementowaliśmy zliczanie zarówno dla danych surowych, jak i znormalizowanych

4. Podstawowe wyrażenia

W celu weryfikacji występowania podstawowych wyrażeń, zamieniliśmy dokument pdf na string

Korzystając z Fuzzy Matchingu sprawdziliśmy, czy w naszym stringu występują szukane przez nas wyrażenia lub wyrażenia podobne do nich w 95%

Otrzymane wyniki reprezentowane są w postaci plików .csv

Dalszy potencjał projektu

Do wyszukiwania informacji w tekście można by użyć algorytmów Deep Learningowych (bazujących na architekturze transformerów i mechanizmie atencji) Multimodalne transformery mogłyby pomóc w analizie podanych wyrażeń.

Dzięki takiemu rozwiązaniu nie trzeba byłoby się ograniczać do samych dokumentów pdf, ale można by także szukać potrzebnych informacji w np. skanach dokumentów

Dodatkową zaletą była by możliwość szukania informacji i innych rodzajów dokumentów, nie tylko KIID. Wystarczyłoby poświęcić trochę czasu na wytrenowanie modelu dla odpowiedniego typu dokumentu