Text Enrichment

Mesterséges intelligencia alapú szöveginformáció gazdagító szoftvermegoldás

Követelményspecifikáció

Szoftverarchitektúrák tárgy házi feladat

Benda Krisztián és Szántó Tamás

krisztianbenda@gmail.com
tmas.szanto@gmail.com

FELADATKIÍRÁS

Az elkészítendő megoldás egy szövegfeldolgozó szolgáltatást megvalósító szerveralkalmazás és a szolgáltatásra épülő kliensalkalmazás. A szolgáltatás szöveges tartalmat képes feldolgozni, és az eredményét visszajelezni.

A szöveg feldolgozása során az algoritmus elemzi a bemeneti szavak jelentését, szófaját és információ tartalmát. Az elemzés eredménye a szavak jelentéséhez kapcsolodó információk feltárása és új információkkal való ellátása. A felhasználó a kliensalkalmazás segítségével interakcióba léphet a szolgáltatással és felhasználhatja annak eredményeit. Bővebb leírás a részletes feladatleírás részben olvasható.

A FEJLESZTŐI CSAPAT

A csapat tagjai:

Csapattag neve	Neptun-kód	E-mail cím
Szántó Tamás	ET7D8H	tmas.szanto@gmail.com
Benda Krisztián	J1CEl3	krisztianbenda@gmail.com

A csapatban dedikált szerepek kiosztását a csapat kis mérete miatt nem tartottuk fontosnak. A csapat összes tagja felelős az elkészítendő munka minden szegmensének minőségéért. A specifikáció, dokumentáció, implementáció és prezentáció elkészítése közös munka során kerül kidolgozásra.

RÉSZLETES FELADATLEÍRÁS

A projekt során célunk egy olyan szolgáltatás megvalósítása és használatának bemutatása egy kliens alkalmazáson keresztül, amely képes szöveges tartalmat elemezni a szövegben található szavak jelentése alapján.

A feldolgozandó szöveg kiegészítésre kerül kapcsolodó információkkal. A szolgáltatás előre megadott típusú adatok kinyerésére alkalmas. Ezek a következő típusok lehetnek: földrajzi, geopolitikai, esemény, szervezet, személy, műtárgy, idő. A felhasználónak lehetősége van szöveg bevitelére, és az elemzési funkció konfigurálására. Megadhatja, hogy az említett típusok közül melyek keresése valósuljon meg. A feldolgozott szövegben kiemelésre kerülnek a megadott típusú információk, ezek kiegészülhetnek hozzájuk kapcsolódó linkekkel. Lehetőség van továbbá egy összesített kimutatás készítésére, ami csak a felismert fogalmakhoz

kapcsolódó adatokat tartalmazza. A szolgáltatást a felhasználók webalkalmazás formájában érik el, az üzleti logika szerveroldalon kerül megvalósításra.

KÖVETELMÉNY SPECIFIKÁCIÓ

Kliensalkalmazás követelményei:

- Szöveg bevitelének támogatása másolás vagy gépelés formájában
- Feldolgozás során szükséges fogalmak beállítása
- Bevitt szöveg elküldése a szövegfeldolgozó szolgáltatás felé
- A feldolgozás eredményének megjelenítése, megtalált kulcsszavak kiemelése
- A feldolgozás során megtalált hivatkozások, linkek elérhetővé tétele
- Összesített adatok megjelenítése

Szövegfeldolgozó szolgáltatás követelményei:

- Megadott szöveg feldolgozása
- Szavakhoz kapcsolódó információk felismerése és entitásokhoz társítása: névelemek felismerése (Name Entity Recognition, NER)
- Felismerendő entitások: földrajzi, geopolitikai, esemény, szervezet, személy, műtárgy, idő
- A felismert kulcsszavakhoz információk társítása: “szöveg gazdagítása”
- Az egyes entitásokhoz megtalált kulcsszavak csoportosítása és lekérdezhetővé tétele
- A kulcsszavak és hozzájuk talált metainformációk alapján hivatkozások keresése
- Az egyes entitásokhoz a következő típusú hivatkozások keresése:
  - Földrajzi entitás: földrajzi hely azonosítására szolgáló Google Maps link
  - Geopolitikai entitás: kapcsolódó földrajzi hely azonosítására szolgáló Google Maps link
  - Esemény: Wikipedia oldal
  - Szervezet: Wikipedia oldal
  - Műtárgy: képre történő hivatkozás
  - Idő: adott időpontban esemény létrehozásra hivatkozás Google Calendarban
- A megtalált entitások szerint csoportosított kulcsszavak összegyűjtése és elérhetővé tétele

TECHNIKAI PARAMÉTEREK

A definiált szolgáltatás webapplikáció formájában kerül megvalósításra, melyet böngészőből tudunk elérni a kliens alkalmazáson keresztül. A böngészőben futó kliens JavaScript alapú, melyet elsősorban Safari és Google Chrome böngészőkre optimalizálunk. A kliens egy Python webszerverrel kommunikál REST interfészen keresztül. A webszerver két részre oszlik: NER algoritmust futtató motorra és NER eredményét feldolgozó, információ összegyűjtő, kliens alkalmazással kommunikáló szolgáltatásra. A NER motor végzi el az entitások felismerését és kategorizálását. A feldolgozó pedig a kategóriák alapján a megfelelő linkek előállítását.

SZÓTÁR

Esemény entitás: egy köztudatban élő, megtörtént esemény. Pl.: sport rendezvények, hurrikánok, háborúk

Földrajzi entitás: földrajzi helyre utaló tulajdonnév pl.: Mount Everest

Geopolitikai entitás: országok, városok, települések absztrakciója

Idő entitás: dátum, vagy időpont megnevezése

Műtárgy entitás: képzőművészeti alkotások

NER: Named Entity Recognition, névelem (és tulajdonnevek) felismerése egy adott korpuszon belül

Szervezet entitás: egyedi, köztudott névvel rendelkező emberi csoportosulás. cégek, ügynökségek

USE-CASE DIAGRAM