TANSZÉKVEZETŐ

SZAKDOLGOZAT FELADAT

Kacsó Zoltán
Mérnökinformatikus hallgató részére

Természetes nyelvű szövegek kategorizálása adatbányászati eszközökkel

A természetes nyelvek kategorizálása napjainkban széles körben felhasznált technológia. Leggyakoribb felhasználása a szövegek témakörének megismerése, viszont ezen adatbányászati eszközök felhasználhatóak az adatok más szempontbeli csoportosítására is. Esetünkben a szöveg írójának kiderítésére szeretnénk alkalmazni. Komoly feladatot jelent a felhasznált algoritmusoknak a szövegeket olyan formában ábrázolása, ami alapján az író szóhasználata hangsúlyozódik ki.

A hallgató feladata különböző adatbányászati algoritmusok keresése és megvizsgálása, valamint kiválasztani, a megvizsgált algoritmusok közül melyek alkalmazhatóak jelen feladatához. A kiválasztott algoritmusokhoz az adatok elő feldolgozása, az algoritmus belső paramétereinek (ha létezik) konfigurálása minél jobb pontosság elérésének érdekében szintén a hallgató feladata. Végül a kapott eredmények statisztikai elemzése, az algoritmusok összehasonlítása és konklúziók levonása a cél.

A hallgató feladatának a következőkre kell kiterjednie:

- Ismertesse a kategorizálás alapeszközeit.
- Mutassa be a természetes nyelvű szövegek kategorizálásának nehézségeit.
- Mutassa be a felhasznált algoritmusok működését.
- Készítsen implementációt a választott algoritmusokhoz Java nyelven.
- Elemezze a kategorizálás eredményeit. Hasonlítsa össze a különböző algoritmusok pontosságát.

Tanszéki konzulens: Dr. Dudás Ákos

Budapest, 2016. szeptember 19.

Dr. Charaf Hassan
egyetemi tanár
tanszékvezető