Miskolci Egyetem
Általános Informatikai Intézeti Tanszék

Nappali tagozat
Tavaszi félév

GEIAL527-M
Szövegbányászat és dokumentum kezelés
Mérnökinformatikus mesterszak (MSc)

A tárgy előadója, leckekönyvi jegyzője: Dr. Baksáné Dr. Varga Erika, egyetemi docens
A tárgy lezárásának módja: aláírás és vizsga
Tárgytípus: választható
Kredit: 4
Kontakt órák száma / hét: 2 előadás, 2 gyakorlat

ÜTEMTERV

Hét	**Előadás**	**Gyakorlat**
1.	Dokumentum fajták és dokumentumkezelés alapfogalmai. Dokumentum-tárolási módok. E-learning keretrendszerek.	Dokumentumkezelés Latex környezetben.
2.	A szövegbányászat feladata, alkalmazási területei. Dokumentumok előfeldolgozása.	Dokumentumkezelés Latex környezetben.
3.	Dokumentum reprezentálása vektortér-modellben. A vektortér-modell dimenziójának csökkentése.	Szövegelemzés R-ben. A tidy text formátum. Szavak és szókapcsolatok gyakoriságának elemzése.
4.	Az információ-visszakeresés alapjai. Mintaillesztési módszerek. Az információvisszakereső-rendszerek értékelési módszerei.	Szövegelemzés Phyton-ban. Szöveg jellemzése, a szavak gyakoriságának meghatározása.
5.	Tartalomkeresés web dokumentumokban. Keresőrendszerek működése.	Szövegelemzés Phyton-ban. Szöveg témájának meghatározása.
6.	Információkinyerés feladatai és módszerei. Az információkinyerés során felmerülő nyelvészeti problémák.	Szövegelemzés Phyton-ban. Tulajdonnevek felismerése.
7.	Természetes nyelvek szintaktikai szintű feldolgozásának feladatai és módszerei. Morfológiai elemzés és szófajmeghatározás.	Magyar nyelvű szöveg szintaktikai elemzése.
8.	Természetes nyelvek szemantikai feldolgozásának feladatai és módszerei. Szövegelemzés, annotácó.	Véleményelemzés Phyton-ban.
9.	Szemantikus web technológiák: XML, ontológiák, RDF adatbázisok.	Ontológia szerkesztés Protege-ben.
10.	Válaszkereső rendszerek. Természetes nyelvű adatbázis-interfészek. Chat robotok működése.	Chat robot tervezése és megvalósítási lépései. Esettanulmány.

Hét	Előadás	Gyakorlat
11.	Dokumentumok osztályozásának algoritmusai. Osztályozók elemzése.	Dokumentum osztályozó készítése Phyton-ban.
12.	Dokumentumok csoportosítása. A csoportosító módszerek elemzése.	**Projekt prezentáció**
13.	Kivonatolás. Összegzéskészítő eljárások és hatékonyságuk mérése.	**Projekt prezentáció**
14.	**Elővizsga**	**Projekt prezentáció pótlása**

Tananyag:
www.iit.uni-miskolc.hu → Munkatársak: Baksáné V.E. → Oktatott tárgyak → Szövegbányászat

Ajánlott irodalom:

- Tikk Domonkos (szerk.): Szövegbányászat (Az Informatika alkalmazásai sorozat), Typotex, 2007.
- 2. D. Jurafsky and J.H. Martin: Speech and Language Processing (3rd Edition), 2017, web.stanford.edu/~jurafsky/slp3/
- 3. J. Silge and D. Robinson: Text Mining with R (A Tidy Approach), O'Reilly 2017, ISBN 978-1-491-98165-8, www.tidytextmining.com

Az aláírás megszerzésének feltételei:

Egy önálló szövegbányászati projekt feladat megoldása és nyilvános bemutatása.

A vizsga menete: írásbeli

GEIAL527-M
Szövegbányászat és dokumentum kezelés
Vizsga minta

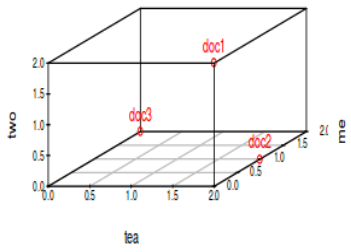
A vizsga írásbeli. A kérdések megválaszolására 60 perc áll rendelkezésre.
Értékelése:

0%-50% : elégtelen
51%-62% : elégséges
63%-75% : közepes
76%-88% : jó
89%-100%: jeles

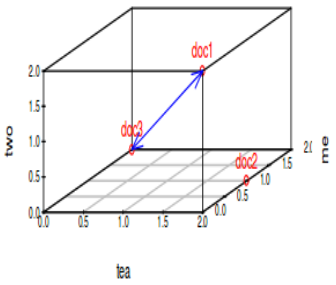
1. Hogyan mérjük a dokumentumok hasonlóságát? Mérőszámok és jellemzésük. 10 pont

A dokumentumokat a bennük található szavak halmazaként reprezentáljuk. Több dokumentum összehasonlításához a dokumentumot leíró vektorokat felvesszük egy term-document mátrixba (TDM). Dokumentumok hasonlóságának mérése:
- Két dokumentum akkor hasonló, ha az Euklideszi távolságuk kicsi. Vagyis a hasonlósági mérőszám értékét úgy kapjuk, hogy 1-ből kivonjuk a dokumentumok távolságát. Probléma: a rövid dokumentumok origótól vett távolsága kicsi, ezért ez a mérőszám a rövid dokumentumokat mindig egymáshoz hasonlónak méri.

Point	tea	me	two
doc1	2	0	2
doc2	2	1	0
doc3	0	2	0

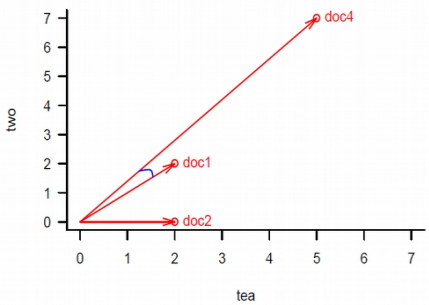


$$sim(doc1, doc2) = -\sqrt{(2^2 + 2^2 + 2^2)} = -\sqrt{12}$$



- Jaccard hasonlóság: dokumentumok esetén a közös szavak és az eltérő szavak aránya. Duplikátum ellenőrzéshez használjuk.
- A két dokumentum vektor által közrezárt szög koszinusza. Ez mindig 0 és 1 közé eső érték.

Point	tea	two
doc1	2	2
doc2	2	0
doc4	5	7



2. Mit jelent a lemmatizálás? Írjon rá angol és magyar példát! Milyen eszközzel végezzük? Mi a különbség a lemmatizálás és a közöstő keresés között? 10 pont

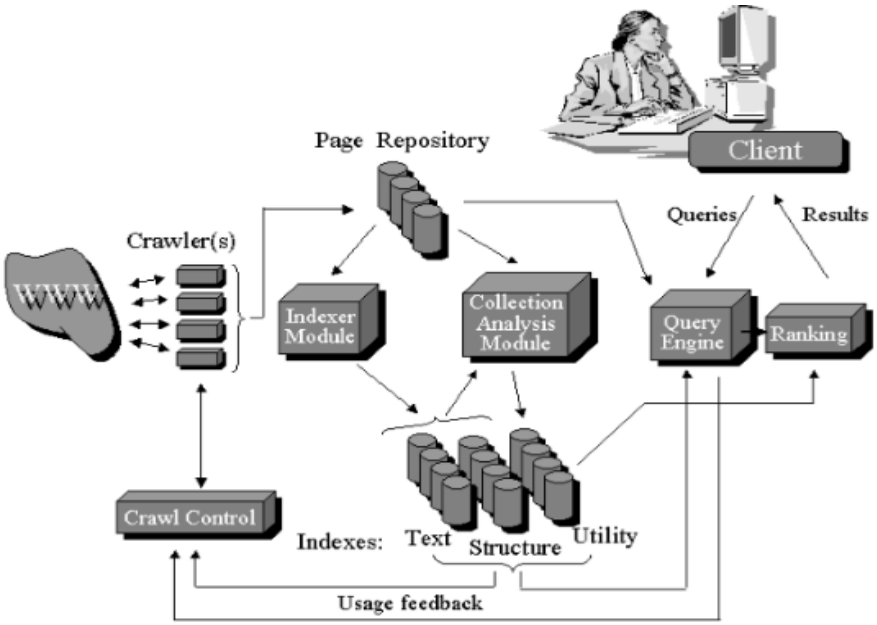
Lemmatizálás = szótövesítés. A morfológiai elemző feladata: a szóalakot tőre (lemmára) és toldalékokra bontja.

Magyar példa:
elemzendő szó: fémkapunk
szabályok:
fém (N) + kapu (N) + nk (S, toldalék) → összetettszó
fém (N) + kap (V) + unk (S, toldalék)

Angol példa:
elemzendő szó: unfaithfully
un (prefix) + faith (stem) + ful (suffix) + ly (suffix)

A közöstő keresés esetén több szónak keressük a közös tövét, ami lehet hogy nem szótári alak (lemma). Például:
magyar szavak: kamat, kamatozik, kamásni, kamra → közös tő: „kam”
angol szavak: sing, singing, single → közös tő: „sing”

3. Írja le és grafikusan is ábrázolja egy webes kereső motor belső felépítését, működését. 10 pont



4. Milyen dokumentum osztályozó eljárásokat ismer? 10 pont

Rocchio-osztályozó
Neurális hálózat alapú módszerek
Statisztika alapú osztályozás (naiv Bayes-módszer)
Döntési fa alapú osztályozás
Legközelebbi szomszédokon alapuló osztályozás (k-NN)
Szupportvektor-gépek (SVM)
Regressziós modellek

Lásd Tikk Domonkos (szerk.): Szövegbányászat (Az Informatika alkalmazásai sorozat), Typotex, 2007. - 5.4 fejezet Osztályozó algoritmusok