Önálló labor feladatkiírásaim – 2015. tavasz

(ezekhez kapcsolódó saját témával is megkereshetnek)

Mészáros Tamás
http://www.mit.bme.hu/~meszaros/

Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs Rendszerek Tanszék

Információkeresés és -szolgáltatás, intelligens ágensek

- Információkeresés (information retrieval)
  - tárolt szöveghalmazban az igényelt információ megtalálása
  - szövegek elemzése és keresések futtatása
  - strukturált szövegek (XML) létrehozása, feldolgozása és felhasználása

    pl.: webes keresőrendszerek, könyvtári keresők, üzleti adatelemző és -tisztító rendszerek, spam szűrés, desktop kereső, tudáskinyerők, stb.

- Információszolgáltatás (kapcsolat a felhasználóval)
  - felhasználó modellezés (az igényelt információ kontextusa)
  - információbevitel, keresés és lekérdezés természetes nyelven

- Intelligens ágensek alkalmazása
  - észleli a környezetét és beavatkozói segítségével önállóan cselekszik
  - kommunikál más entitásokkal (ágensekkel, emberekkel, stb.)
  - közösségi entitás (több-ágens rendszerekben kooperál és verseng)
  - modellezési és alkalmazásfejlesztési eszköz (Java-alapú)

    pl.: web indexelő robot, szövegelemző hálózat, felhasználói interfész ágens, ...

Szövegfeldolgozás, szövegbányászat

- Szövegek...
  - strukturálatlan, szabad szöveg, természetes nyelvű
  - strukturált (pl. XML), meghatározott nyelvtani és értelmezési rendszerrel
  - jellemzően a kettő keveréke

- ... gépi feldolgozása, ...
  - statisztikai feldolgozás, index építés, kulcsszó kiemelés
  - kivonatolás, lényeges részek felismerése
  - nyelvi elemzés (szavak, entitások, mondatok, ...)
  - strukturált formára alakítás (többféle elemzési módszerrel)
  - XML technológiák (Schema, XML, XPath, XQuery) alkalmazása

- ... és az eredmények felhasználása
  - információkeresés és lekérdezés
  - különböző szövegkorpuszok jellemzése, összevetése és klaszterezése
  - tudásbányászat, összefüggések, állítások kinyerése
  - ...

Irodalmi szövegek számítógépes elemzése

(Az MTA Irodalomtudományi Intézettel közös feladatkiírás)

- Vajon Shakespeare írta az összes művét?
- Ki kivel állt kapcsolatban, kinek mely más szerzőre volt hatása
- Hogyan változott egyes szavak gyakorisága az idő folyamán?

  1800-1970 között a „nő” szó a „férfi”-hez képest elenyésző gyakoriságú volt, 1980 óta nagyjából egyforma arányban fordulnak elő az angol irodalomban.

  Az „1880” szó használatának gyakorisága 1912-re feleződött meg, míg az „1973” már nagyjából 1983-ra elérte ezt a szintet. Egyre gyorsabban felejtünk?

  Az 1800-1840 között találmányok nevei kb. 66 év után terjedtek el írásban, míg az 1880-1920 közöttieknek ez csak 27 évig tartott

- 2010 óta jelentek meg jelentősebb (angol) publikációk a témakörben

- A magyar írásbeliség ilyen jellegű vizsgálata úttörő munkának számít
  - Példaként Mikes Kelemen igen terjedelmes életművének vizsgálata a cél
  - Digitalizált változatban már rendelkezésre áll, most zajlik a szótárkészítés
  - A szövegnek létezik mai átírása, de az eredeti nyelvezet is vizsgálható
  - A szótár a szavak értelmezésében, eltérő szóalakok felismerésében segít
  - A feladat nyitott, egyéni ötletekkel is elő szabad, sőt kell állni

A Villanyspenót bővítése

(Az Országos Széchenyi Könyvtárral közös feladatkiírás)

- Az OSZK nagyszabású digitalizálási projektbe kezd
  - cél a teljes állomány digitalizálása és OCR-erezése
  - a szöveggyűjtemény mellé „szakértői rendszereket” és létrehoznak, amelyek összefogják egy-egy tudományterület anyagait
  - ezek más online tartalom és tudásforrásokat is integrálni kívánnak

- A Villanyspenót egy online irodalomtörténeti rendszer
  - ezt más adatbázisokkal és a digitalizált szövegekkel kapcsolnák össze
  - a kapcsolódási pontokat lehetőleg automatikusan kell felderíteni
  - ehhez tisztítani és elemezni kell a szöveget (programozottan)
  - szövegelemzési és szövegbányászati módszerek alkalmazásával
  - fogalmak, nevek, évszámok, korszakok, címek, bibliográfiai adatok, stb.

- Technológiák (gyakorlatban megismerhetők, előismeret előny)
  - web (Javascript, node.js, Angular, Java)
  - adatbázis (nosql, Mongodb)
  - szövegelemzés és -bányászat (reguláris kifejezések, XML, stb.)

Közlekedési hálózatok ágensalapú modellezése

- Úthálózatok és az ott közlekedő autósok modellezése
  - Bizonyos szempontok szerint reális modell legyen
  - Az ágensalapú megközelítés érdekes lehet

- Próbáljuk meg az autósok viselkedését is modellezni
  - Agresszív, szabálykövető, tutyi-mutyi, stb.
  - A különböző viselkedésű autósok megfigyelése a szimulációban

- Mire használható?
  - A hálózat megfigyelése a szimulált rendszerben (a valóságban nehéz)
  - Kísérletezésre (viselkedések, hálózati konfiguráció, stb.)
  - Közlekedési lámpák jobb hangolására
  - Úthálózatok átalakítására (lásd Wekerle és az egyirányú utak)
  - Nagy ágenspopuláció hatékony futtatása (pl. A GPGPU érdekes lehet)

- Sokan dolgoztak már a témában, az ő eredményeik felhasználhatók.

Kontrollált természetes nyelvek alkalmazása

- Természetes nyelvű kommunikáció a számítógéppel – ideális lenne

- Kontrollált természetes nyelvek
  - Mintha természetes lenne (jól érthető, megtanulható)
  - Alkalmazási területre szabható (mesterséges)
  - Jól elemezhető (igen egyszerű elemző is elég lehet)
  - Jól értelmezhető (egyértelmű szemantika, automatikus fordítás)

- Mire használható?
  - Bonyolult interfészek egyszerűvé válnak:
    - a nyelv biztosítja az összetettségüket, nem a felületi elemek
    - Olyanok is használhatják, akik egy számítógépes felületen elvesznek
    - Akkor is működik, ha nem áll rendelkezésre megfelelő interfész eszköz
    - Hangalapú kommunikációt is lehetővé tesz (pl. Google Speech Input)
  - Tudásbevitel
    - Elemezhető, a számítógép által megérthető szöveg
    - Formális (pl. logikai) reprezentációra alakítható

Nyelvalkotás

- A természetesre hasonlító, mesterséges nyelvek létrehozása
  - Komolyabb példák: Basic English (850 szóból álló angol), légiirányítás, technikai dokumentációk nyelvezete, ontológiák leírása, stb.
  - Könnyedebb példák: Klingon, Dothraki, Valyrian, Castithan, Irathient, stb.

    (Ezek mögött is komoly elmélet húzódik meg.)

- Mi a célunk?
  - ~~Általános kommunikáció megvalósítása~~ (ezzel nem foglalkozunk)
  - Valamilyen problémakörhöz kapcsolódó feladat megoldása (résznyelv) azaz alkalmazási területhez és célhoz kötött nyelv megalkotása

- Mire és hogyan használhatjuk?
  - Adat- és tudásbázisok (természetes nyelvű) bővítése és lekérdezése

    Konkrét példa: tudományos közlemények állításainak formálisabb leírása ún. szemantikus absztrakt segítségével. A szókincset a tárgyterület adja, a feladat a bevitelhez és a lekérdezéshez használható nyelv létrehozása

Általános tudnivalók – mit várok és mit nyújtok

- „Mindent szabad, ami örömet okoz”
  - a motiváció érdekében a feladatkírást a hallgatóval közösen véglegesítem
  - szabad saját (akár irreálisnak tűnő) ötletekkel változtatni a feladaton
  - a feladat méretének helyes meghatározása a konzulens feladata

- Előzetes jelentkezés a konzulensnél
  - rövid bemutatkozás (előismeretek)
  - miért érdekli a téma (mi a motiváció)
  - milyen elképzelései vannak a feladatkiírás módosításával kapcsolatban

- Az önálló labor menete
  - közösen megalkotott specifikáció és vázlatos munkaterv
  - heti rendszeres konzultációk (aki jól halad, annál ritkábban is lehet)
  - fontos a terület önálló felfedezése, de irodalmat bőségesen adok
  - a terveket megbeszéljük, az implementáció önálló munka
  - nem a tökéletes termék a cél, hanem a terület megismerése, megértése