Bevezetés

Számítógépes nyelvészet – 2018 tavasz
1. óra

Simon Eszter – Mittelholcz Iván

MTA Nyelvtudományi Intézet

Tartalom

Bemutatkozás

Bemutatkozás

- mi
- ti

A félév bemutatása

Az órák

- összesen 13 óra
  - ebből 11-re van terv
  - egy lauf
  - egy úgyis elmarad...
- egy órán belül:
  - elméleti bevezetés slide-okkal
  - gyakorlatok gépen
  - házi feladat

Bevezetés a karakterkódolások rejtelmes világába

- Elmélet:
  - szöveges fájlok
  - karakterkódolás általában, karakterkódolás és fontkészlet
  - ASCII és kiegészítései
  - Unicode, Unicode kódolások (UTF és UCS)
  - karakterkódolás detektálása
  - konvertálás kódolások között
- Gyakorlat:
  - file és iconv parancsok
  - karakterkódolás python-ban
  - python2 és python3 közti különbségek

3. ÓRA

Bevezetés a héjak és szabályos kifejezések csodálatos világába

- Elmélet:
  - shell bevezetés
  - nyelvosztályok
  - regex elméleti alapok (reguláris nyelvek, automaták)
  - regex motorok működése, hatékonyság
- Gyakorlat:
  - sed, grep
  - python regex-ek
  - regexek és karakterkódolás
- Házi feladat:
  - -

Automaták, FST, kétszintű morfológia

- Elmélet:
  - mi az automata, hogyan kell csinálni
  - mire lehet használni: különböző morfofonológiai feladatokra
  - Kimmo és a kétszintű morfológia
  - automaták implementálásának alapjai táblázattal
- Gyakorlat:
  - hfst-nek van olyan parancsa, amivel szabályokból FST-t lehet építeni
- Házi feladat:
  - automata építése, ami egy nyelv minden elemét legenerálja, és csak azt
  - szorgalmi: játékautomata leprogramozása pythonban

Korpuszépítés

- Elmélet:
  - a forrás módja: hang, írott, multimodális → innentől csak írott
  - forrás: papír Vs elektronikus → kép Vs szöveg → txt
  - az annotáció formátuma: inline (XML) vagy standoff (tsv & BIE1)
  - annotációs séma → annotációs útmutató
  - kézi annotálás, annotációs eszközök, inter-annotator agreement
- Gyakorlat:
  - crawling: wget, scrapy
  - boilerplate removal: beautifulsoup4
  - odt –> xml-ből kinyerés
  - docx, pdf: tika
  - kézi annotáció segítése: excel, ana2html, GATE
- Házi feladat:
  - játékkorpusz annotálása (NER v. NP-chunk v. dependencia) ketten vagy hárman, inter-annotator agreement számolása
  - szorgalmi: NLTK-ban van rá eszköz, azzal kiszámolgatni

Korpuszannotáció 1.

- Elmélet:
  - kézi vs. automatikus annotáció, gold vs. silver standard
  - az automatikus korpuszannotáló eszközök kiértékelése (P, R, F)
  - mondatra bontás, tokenizálás
  - morfológiai elemzés
  - egyértelműsítés
- Gyakorlat:
  - GATE vagy NLTK, polyglot?
- Házi feladat:
  - egy szöveg végigtolása egy elemzőláncon

Korpuszannotáció 2.

- Elmélet:
  - NER
  - sekély szintaktikai elemzés
  - szintaktikai elemzés (konstituencia és dependencia)
- Gyakorlat:
  - ?
- Házi feladat:
  - az ötödik hét kézzel annotált játékkorpusza legyen a gold standard $\rightarrow$ az e-magyar teljesítményének a kiértékelése ezeken a korpuszokon (precision, recall, f-measure)

Korpuszlekérdezések, -statisztika

- Elmélet:
  - alapfogalmak: korpusz, korpuszlekérdező motor, nyelvek és felület
  - lekérdező nyelvek: CQL (MNSZ), MQL (Emdros)
  - token–type, gyakoriság, relatív gyakoriság, MLE…
- Gyakorlat:
  - MNSZ-en vagy ómagyar korpuszon parancssorból lekérdezgetni dolgokat, valami egyszerűbb statisztikát számolni
- Házi feladat:
  - ómagyar korpuszon egy nyelvi jelenség diakrón vizsgálatát elvégezni: pl. a főnevek száma az egyes kódexekben, relatív gyakoriság, diagram

Gépi tanulás 1.

- Elmélet:
  - történeti kitekintés: szabályalapú vs. statisztikai módszerek
  - supervised és unsupervised tanulás
  - gold standard adat
  - train-devel-test halmazok, keresztvalidáció
  - feature extraction, n-gramok
  - modellépítés
  - taggelés
  - kiértékelés
- Gyakorlat:
  - huntagen végigpróbáljuk az egyes lépéseket
- Házi feladat:
  - NLTK-ban egy korpuszon egy tanuló algoritmussal valamit kipróbálni

Gépi tanulás 2.

- Elmélet:
  - Bayes-tétel, noisy channel, HMM
  - supervised algoritmusok: döntési fa, maxent, CRF, neurális háló stb.
  - unsupervised: klaszterezés
- Gyakorlat:
  - Scikit-learn-ben megnézni egy-két dolgot
- Házi feladat:
  - ?

Kitekintés

- Elmélet:
  - ontológia, linked open data, RDF
  - információkinyerés, NER
  - információ-visszakeresés
  - kulcsszókinyerés
  - metaforák és metonímiák felismerése
  - automatikus szótárgenerálás
  - sentiment analysis
  - gépi fordítás
- Gyakorlat:
  - -
- Házi feladat:
  - játék kulcsszókinyerés

Adminisztráció

Házi feladatok és teljesítés

- összesen 8 házi feladat kerül kiadásra
- ebből legalább 3-at kell beadni a teljesítéshez
- a feladatokból 4 megoldható programozási tudás nélkül is
- a beadott házikra megajánlott jegyet lehet kapni
- akinek ez nem jó, írhat javító ZH-t

Technikai részletek

Technikai részletek

OS

- unix-like oprendszerek preferáltak (Linux, OS X)
- windows:
  - cygwin
  - Windows Subsystem for Linux
  - VirtualBox + Linux (Debian, Ubuntu, CentOS)

Shell

- Linux, OS X: ✓
- online: Unix Terminal Online

Python 3

- Linux, OS X: ✓
- Windows: python
- Anaconda
- online lehetőségek: PythonAnywhere, repl.it

Repó

https://github.com/m-ivan/compling

Git

- TryGit, The Simple Guide
- `git clone https://github.com/m-ivan/compling.git`

Jupyter Notebook

- tutorial
- `pip install jupyter` vagy `pip3 install jupyter`
- Anacondában elvileg benne van – ha mégsem: `conda install jupyter`

Fogalmi tisztázás

Fogalmi tisztázás

- számítógépes nyelvészet
- korpusznyelvészet
- NLP
- HLT
- stb.