Gépi Tanulás 3

Kornai András

KálmánCL 4/27

Hogy állunk a jupyterrel?

- Készen van a Slack/HuNLP-ben a kalmancl csoport, eddig 3 nem-tanár tag (Ittzés, Jánosy, Sánta).
- A github/kornai/aml2020-ra meghívva, de még nem reagált: Meszáros, Hegyi, Kálmán) jelen van Ittzés, Jánosy, Sánta
- A legjobban az `aml2020/first_assignment` és `hw4` alatti cuccokból lehet lopni
- Aki szereti, pandas-szal dolgozhat, de nem kell.
- Sok kész python könyvtár van: NumPy, SciPy, Scikit-learn

Mit csináljanak a nem-programozók?

- A “B” csoport (Pöltl, Prótár, Mészáros, Goda) eddig nem iratkozott fel se slackre se githubra, pedig ezek nem programozói teendők
- Lehet, hogy ők az egész életükben nem fognak programozni (ezt kétlem) de annyi bizonyos, hogy fog kelleni tudniuk programozókkal együtt dolgozni, bárhová vesse is őket a sors
- Az egyik lehetséges megoldás az, hogy az “A” vagy a “C” csoportból áttelepítünk egy programozni tudót a “B”-be
- A másik az, hogy őket egy önkéntes majd levizsgáztatja
- Nyilvánvaló, hogy többek fel se vették volna a kurzust ha tudták volna, ezért lesz alternatív megoldás

Alternatív megoldások

- Weka megoldás Klautau (2002) alapján. Aki nem tud programozni, az azért egy kész programcsomag használatát megtanulhatja
- Prototipus-alapú osztályozók összehasonlítása (amit kaptam az nem ilyen)
- Beszámoló cikkekből (pl. Markov (1913), Makhoul and Schwartz (1986), Valiant (1984), Levy and Goldberg (2014) ...)
- Bármilyen empirikus munka (konkrét javaslatok később)
- Program-visszafejtés (kész, jól működő program elmagyarázása)
- Amit még most itt megbeszélünk

Mi a számítógépes nyelvészet?

- "A jelenlegi paradigma, aminek több köze van a paraméter-billegetéses versenyfeladatokhoz, mint az elméleti nyelvészethez, talán nem az utolsó szó a számítógépes nyelvészetben"
- (A) Jóval több köze van (B) biztos nem az utolsó szó
- (A') A paraméter-optimizálásnak *nincs alternatívája* nemhogy a látóhatáron, hanem el se tudjuk képzelni mi lehetne az
- A memorizálásról kb. Chomsky (1959) óta tudjuk, hogy nem megoldás
- Minden más per def paraméteres modellek beállítása, ebből a paradigmából nincs kiút (spéci a szabályalapú rendszerek szabálymegtalálása is valamilyen modellre való rátanulást jelent)
- (B') hogy is alakult ez így? Úgy, hogy a szakértők megbuktak

A fonetikusok összeomlása

Taháfut-al-fonetika

- Kb. 40 éven át hazudták a fonetikusok azt, hogy ők értik a folyamatokat (akusztikus és auditorikus oldalról egyaránt): Potter (1945) kezdte és Makhoul and Schwartz (1986) lezárta (bár van aki máig hazudik)
- A szakértők nem voltak szakértők. Amit mondtak az nagyjából igaz volt, de a teljes igazságnak kevesebb mint a harmada
- M&S 1986: tudomásul kell vennünk, hogy pont a részletek tekintetében (pedig az ördög ott lakik) tudatlanok vagyunk
- Ha ez így van, akkor csak a tudatlanságon alapuló modellezés lehet hatékony (addigra a statisztikai modellek már péppé verték a tudás-alapúakat)
- A fonetikusoknak volt 35 évük arra hogy visszavágjanak
- Fool me once, shame on you, fool me twice, shame on me. Ma nincs az a majom, aki elhinné, hogy a fonológusok, szintakták, szemanták bármivel is küönbek lennének a fonetikusoknál

Van-e hagyományos (nem-C) nyelvészet?

- Mint tudományszociológiai jelenség kétségkívül létezik (vannak diákok, kutatók, tanszékek, folyóiratok, konferenciák. . . .)
- Mint tudás-kupac szintén létezik (ahogy létezik népi orvoslás is)
- De kicsit se felel meg önnön elvárásainak: megmagyarázni az emberek nyelvi képességeit, a nyelvek csoportos (tipológiai) ill. egyedi sajátosságait
- Ahogy a gépi modellek egyre jobbak, egyre világosabb, hogy a "fővonal" egy hazugságon alapul (humán genóma-specificitásnak semmi nyoma)
- A strukturalista elméletek símán beilleszthetőek voltak a számítógépes fejlődésbe
- A modern generatív szintaxis még erre sem jó (pedig a cikkek 2/3 része ilyen)

Klasszikus Markov Modellek

- A kiindulás eleve nyelvészeti volt: Andrej Andrejevics Markov azt vette észre, hogy a következő hang (nála: betű) sokkal jobban megjósolható ha tudjuk mi volt az előző
- Véges automata, állapotai a betűk, élek x-ből y-ba valószínűségi súlyokkal (empirikus munka: megadni ilyen súlyokat pl. magyarra)
- Lehetnek az állapotok betüpárok (bigram), sőt betű-n-esek is (ngram) (empirikus munka: kimérni mennyivel jobb bigramra mint unigramra, trigramra mint bigramra, stb)
- Adatok: átmeneti mátrix ami megmondja hogy az i-edik állapotból a j-edikbe milyen valószínűséggel megyünk át
- Tétel: ebből a mátrixból egyszerűen kiszámolható a határeloszlás (ha sokat futtatjuk a folyamatot melyik állapotnak mi lesz a valószínűsége)

Fontos trükkök

- Nem kell n-állapotot visszakövetni, elég állapot-n-esekkel dolgozni *unigram modellben*
- Fel szoktunk venni (csendes vagy hangos) kezdő- és végállapotokat
- A hangos explicit kezdő- ill végjelet jelent (regexp ^, $)
- Simítunk! Nem hagyunk 0 valószínűségü élet akkor se, ha nem találkoztunk olyan adattal
- Matematikai oka: kikerüljük a szingularitásokat (nem fog soha kelleni zéróval osztani)
- Filozófiai oka: nem tudhatjuk milyen adatot fogunk még látni

Rejtett Markov Modellek

- Egy állapothoz többféle kimenő jel is tartozhat (diszkrét modell)
- Illetve jelek egy eloszlása is (folytonos modell)
- Példa diszkrét (szintaxis) modellre: az állapotok szófajok, a kimenő jelek szavak
- Mit tud ez? Szavak sorozatából visszafejteni a szófajokat
- Mennyire jól? Kb. 96%
- https://bit.ly/3aJcJT4 HMM tutorial