InfoSztrájk!!

- https://www.facebook.com/infosztrajk
- Szólhatna erről az egész óra, mert (a) a téma rettentően fontos és (b) a professzor (ma még) arról beszél amiről akar
- Mi történik? Szerintem masszív, sokszázmilliárd forintos lopás. Mikor lesz ez a "szerintem" elhagyható? Ha a bíróság kimondja. ajánlott olv ajánlott 2
- És még? Szerintem szabályos nemzetárulás, a jövendő generációk oktatása ugyanis nemzetállami feladat, merthogy egy nemzet ereje a kiművelt emberfők sokaságában rejlik.
- tl;dr A magam részéről 100%-ban támogatom az InfoSztrájk céljait és eszközeit, mindenki más azt csinál amit akar

Mit tehet a számítógépes nyelvész?

"And just how do you arrive at that remarkable conclusion, Mr. Mayor?"
"In a rather simple way. It merely required the use of that much-neglected commodity – common sense. You see, there is a branch of human knowledge known as **symbolic logic**, which can be used to prune away all sorts of clogging deadwood that clutters up human language."
"What about it?" said Fulham.
"I applied it. Among other things, I applied it to this document here. I think I can explain it more easily to five physical scientists by symbols rather than by words." Hardin removed a few sheets of paper from the pad under his arm and spread them out. "I didn't do this myself, by the way," he said. "Muller Holk of the Division of Logic has his name signed to the analyses, as you can see."
"As you see, gentlemen, something like ninety percent of the treaty boiled right out of the analysis as being meaningless, and what we end up with can be described in the following interesting manner:
"Obligations of Anacreon to the Empire: None!
"Powers of the Empire over Anacreon: None!"

A számítógépes nyelvészet hosszútávú céljai

- Nem egyszerűen beszélő gépeket akarunk, hanem olyan gépeket amik okosabban beszélnek az embereknél!
- Számos részterületen már most jobban gondolkodnak (nemcsak sakk, go, számolás)
- Bizonyos dolgokat már egész jól csinálunk, pl. viszonylag hatékonyan tudunk *ki, hol, mikor, kit, mivel* jellegű kérdésekre válaszolni (de nem *hogyan* és *miért* jellegű kérdésekre)
- Valamennyire el tudjuk dönteni, hogy egy kijelentés egy dologról azt pozitív vagy negatív szinben tünteti-e fel (sentiment analysis)
- Nagyon jól tudjuk klaszterezni a hírforrásokat bias szerint
- Troll-detekció, szerző-beazonosítás

Ki nyeri az értékes díjakat, és mire?

- Baseline: Kálmán László: “ 90% tanítóadattal, 10% tesztadattal, súlyponttal és euklidészi távolsággal számolva olyan 47% és 58% között találja el a hangokat”
- A “shared task” paradigma
- Eredetileg “bakeoff”, ma ez már nem PC
- Megteremtője a számítógépes nyelvészet valaha élt legfontosabb alakja, Allen Sears vadászpilóta

Mire lehetett tavaly díjat kapni?

- Gradient Boosted Tree (0.89,0.89) AHW1LF
- Decision Trees (0.84)
- Logistic Regression (0.84, 0.83) AHW1BB
- Linear Regression (0.83, 0.82) AHW1SY
- Random Forest (0.914, 0.926) AHW1RO
- feedforward NN (left as exercise to the reader!!)
- k-Nearest Neighbor Manhattan: (0.83, 0.90) AHW1LF

Mik a teendők

- Mindenki csináljon magának Slack és GitHub identitást, és ezeket írja be a Csoportbeosztás Google Sheet-re
- Akit érdekel a magyar NLP, az írjon nekem, és felveszem a HuNLP slack-re
- Kezdjen el mindkét (mindhárom?) csoport írni vagy lopni egy jupyter notebookot (saját gép vagy colab)
- Cél: rendesen kivesézzük a prototipus-alapu módszereket
- Euclidean (0.52, 0.58); scaled (0.68, 0.75); Canberra (0.74,0,81); log (0.79,0.81)
- Látszik. hogy a távolság-fogalom fontos
- http://yunzhishi.github.io/voronoi.html

Módszertan

1. Itt nincs se idiolektus, se “intuíció”
2. Idiolektus éppenséggel lehetne, csak adat nincs hozzá
3. Rengeteg múlik a hiperparamétereken
4. És a rejtett trükkökön (sajnos)
5. Szabályos kisérletes módszertan (ezt köszönhetjük Sears-nek)

A lineáris eset

1. Az elméleti minimum: átlag, szórás, várható érték, feltételes valószínűség
2. Elsőnek `highleyman_1962.pdf`
3. Terminológiai változások az elmult 60 évben: *receptor* $\rightarrow$ *feature extractor*; *measurement space* $\rightarrow$ *feature space*; etc.

Régi idők

Even if the optimal decision function were known, its implementation would require, in general, the use of a digital computer or other complex equipment. The cost of such equipiment may, in many cases, outweigh the advantages of mechanized categorization.