Tanácsi Roland
Micsik András
HUN-REN SZTAKI DSD

2024.11.28

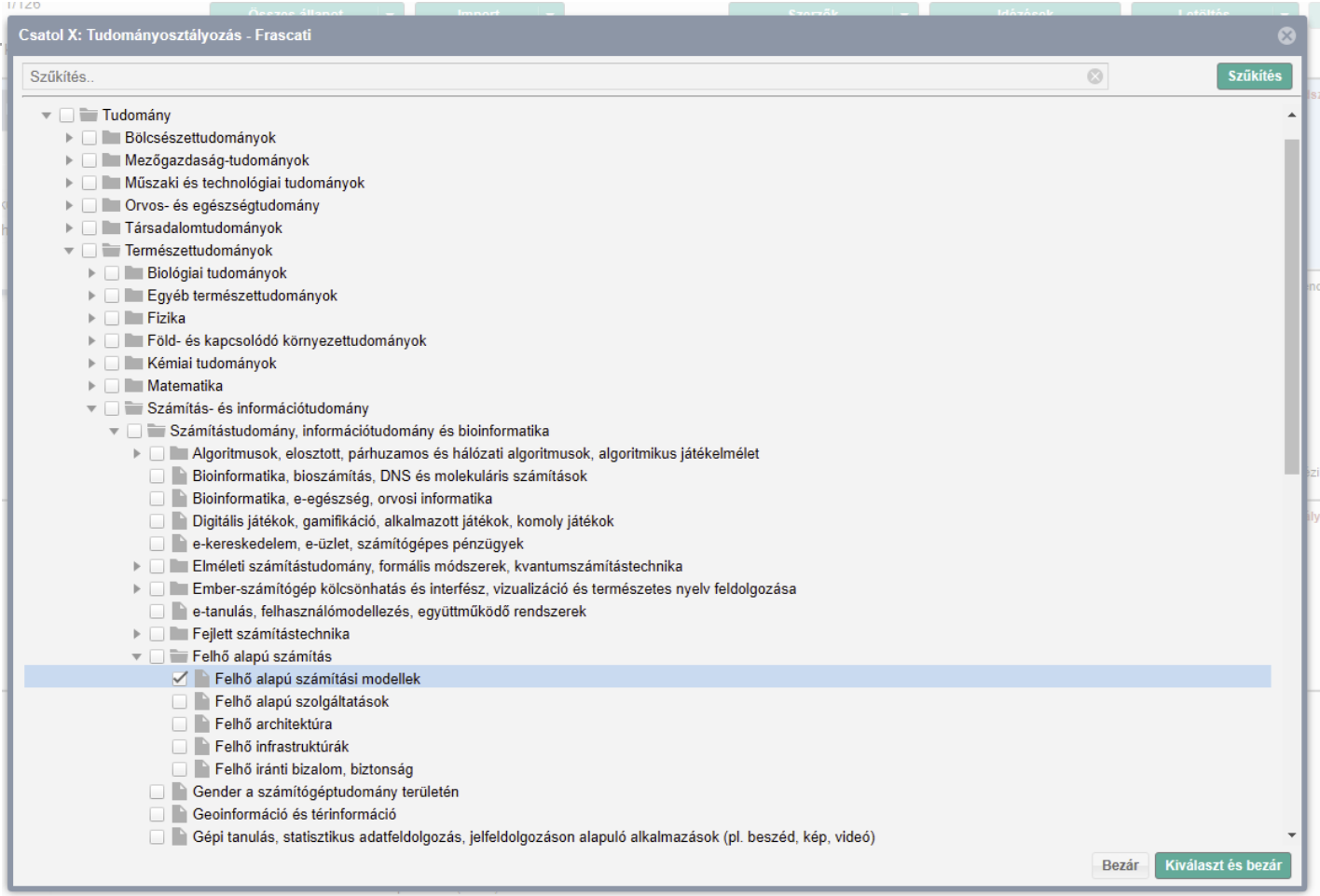
Kutatás.
Innováció.
Hatás.

HUN-REN
Magyar Kutatási Hálózat

Többcímkés osztályozási feladat optimális megoldásának keresése transformer modellek finomhangolásával

Tudományosztályozás az MTMT-ben

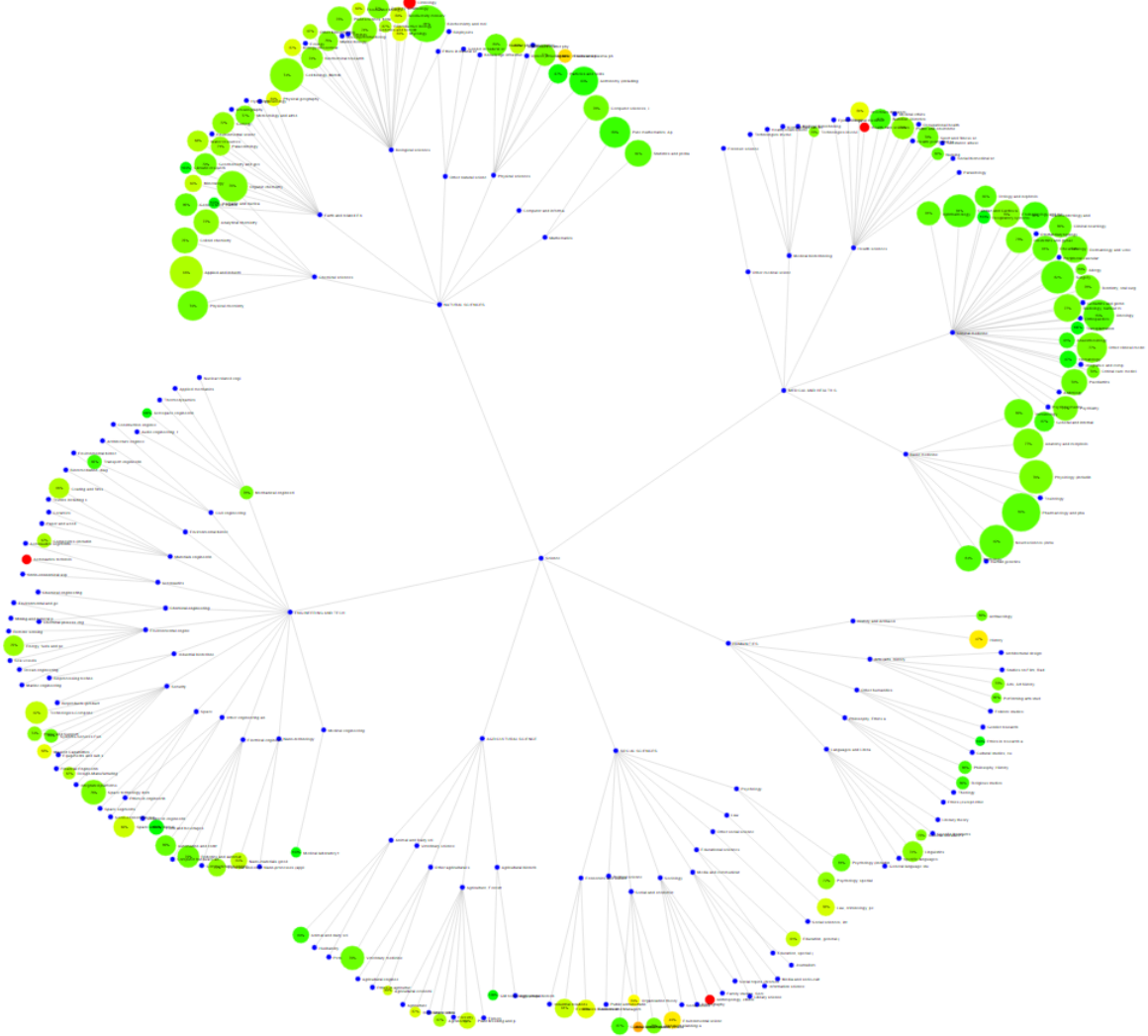
HUN-REN
Magyar Kutatási Hálózat



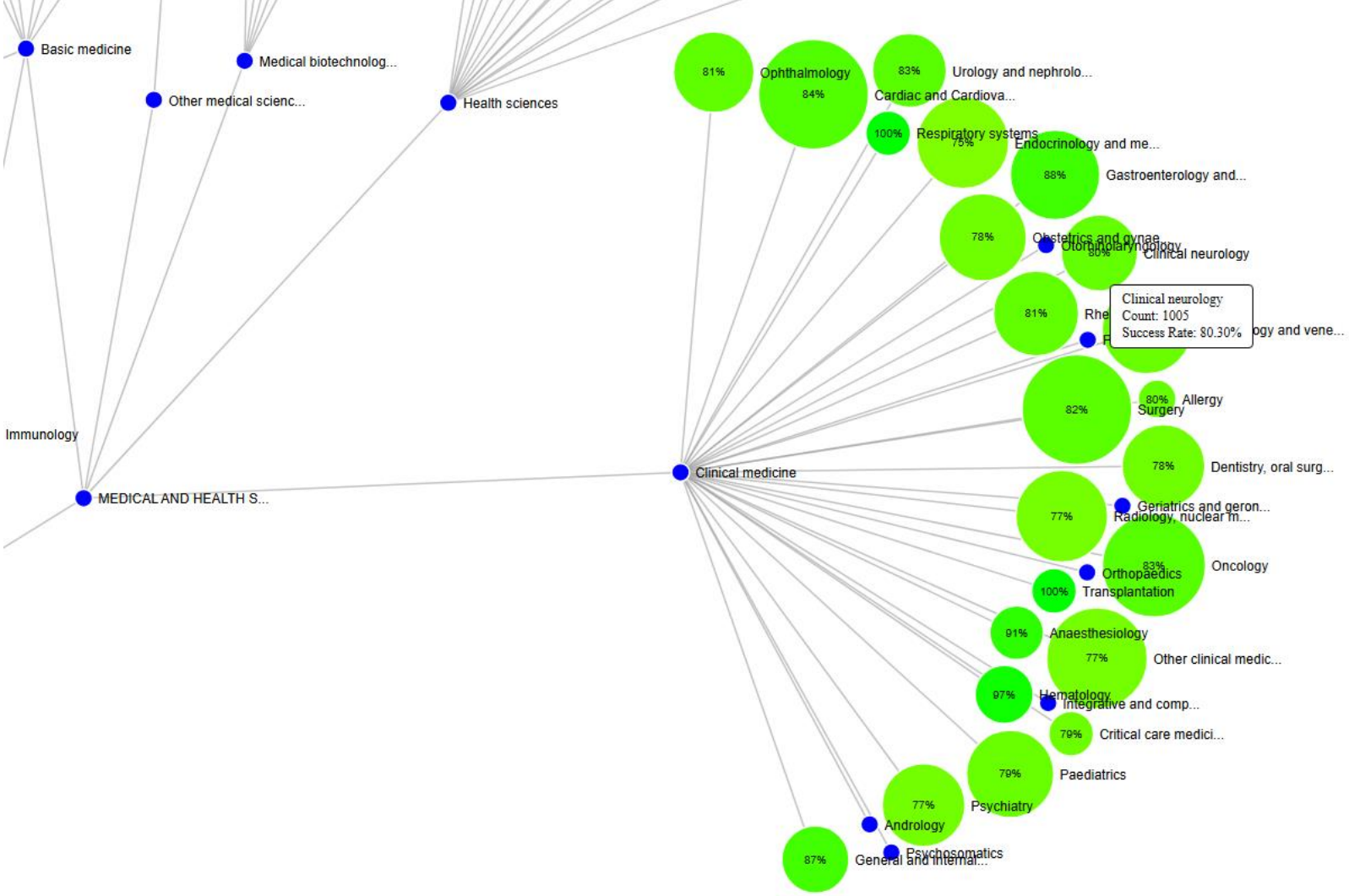
A projekt célja

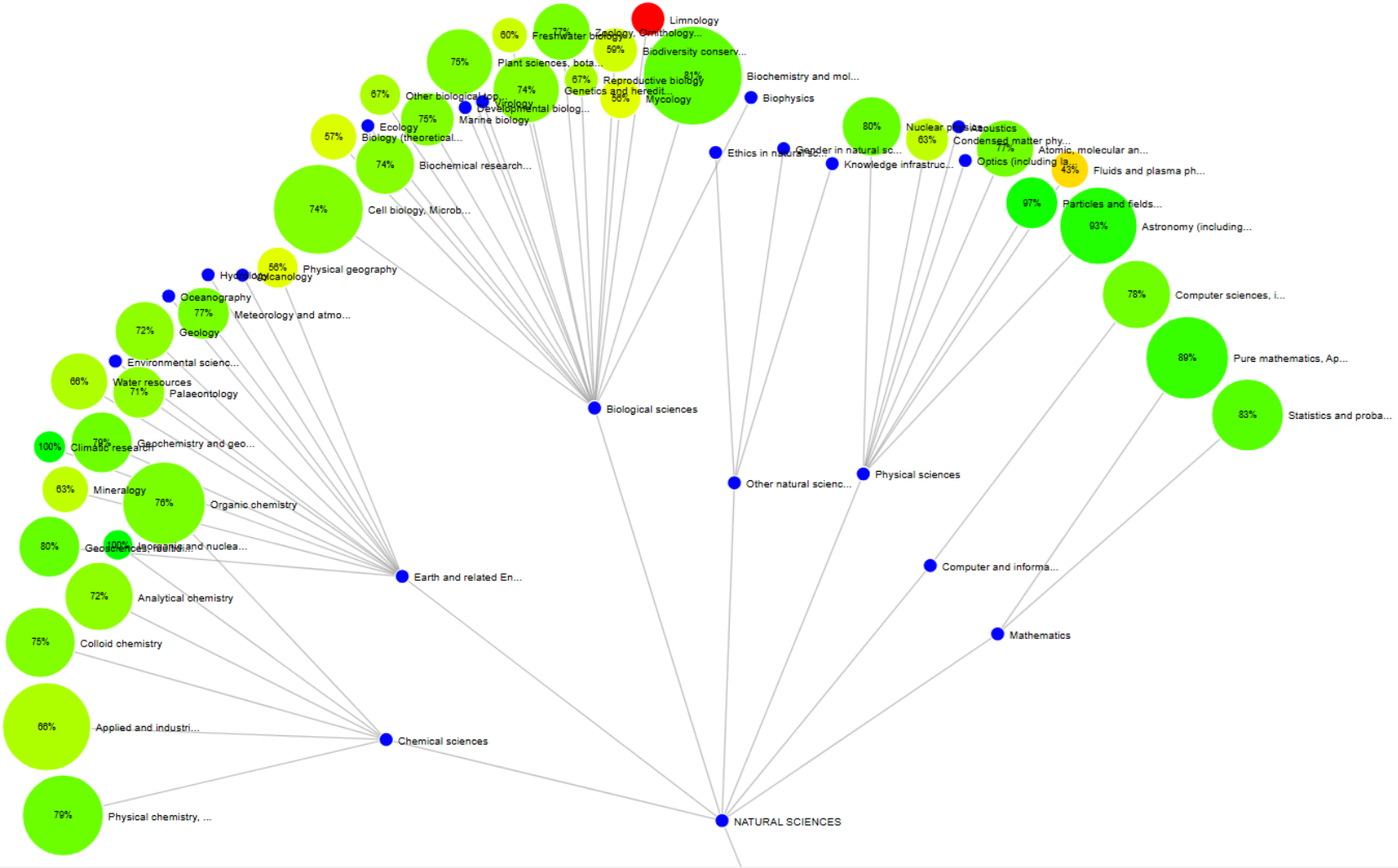
HUN-REN
Magyar Kutatási Hálózat

- Az MTMT-be felvitt közlemények osztályozásához az OECD által fejlesztett ún. Frascati hierarchiát alkalmazzák
  - 6 szinten több, mint **3000** elem
  - 4. szinten az elemek száma: **309** (ebből 160, ami szerepel az MTMT-ben)
- A forrásközlemények 80 százaléka nincs besorolva: cél, hogy MI segítségével ezeket pótoljuk
- Tanításra felhasználható: a maradék, ami 4-es szintű Frascati tárgyszóval rendelkezik (160 különböző tárgyszó)
- Nehézségek:
  - Kevés a tanításra felhasználható minőségi adat
  - Az egyes tudományágakhoz tartozó példák száma nagyon változó (0-20000)
  - A hierarchia elég bonyolult, sokszor szakértőnek is nehéz választani



- Színkódok (teljesítmény):
  - Zöld - jó
  - Sárga - közepes
  - Piros - rossz
  - Kék - nincs adat
- Körméret - előfordulás gyakorisága


Basic medicine
Medical biotechnolog...
Other medical scienc...
Health sciences
Immunology
MEDICAL AND HEALTH S...
Clinical medicine
81%
Ophthalmology
84%
Cardiac and Cardiova...
83%
Urology and nephrolo...
100%
Respiratory systems
75%
Endocrinology and me...
88%
Gastroenterology and...
78%
Obstetrics and gynae...
Otorhinolaryngology
86%
Clinical neurology
81%
Rhe...
Clinical neurology
Count: 1005
Success Rate: 80.30%
ogy and vene...
82%
80%
Allergy
Surgery
78%
Dentistry, oral surg...
77%
Geriatrics and geron...
Radiology, nuclear m...
83%
Oncology
Orthopaedics
100%
Transplantation
91%
Anaesthesiology
77%
Other clinical medic...
97%
Hematology
Integrative and comp...
79%
Critical care medici...
79%
Paediatrics
77%
Psychiatry
Andrology
87%
Psychosomatics
General and internal...


NATURAL SCIENCES
Biological sciences
Physical sciences
Chemical sciences
Earth and related En...
Other natural scienc...
Computer and informa...
Mathematics
Limnology
Freshwater biology
Zoology
Ornithology...
Biodiversity conserv...
Plant sciences, bota...
Biochemistry and mol...
Biophysics
Genetics and heredit...
Reproductive biology
Mycology
Other biological top...
Virology
Developmental biolog...
Marine biology
Ecology
Biology (theoretical...
Biochemical research...
Cell biology, Microb...
Ethics in natural sc...
Gender in natural sc...
Knowledge infrastruc...
Nuclear physics
Acoustics
Condensed matter phy...
Optics (including la...
Atomic, molecular an...
Fluids and plasma ph...
Particles and fields...
Astronomy (including...
Computer sciences, i...
Pure mathematics, Ap...
Statistics and proba...
Physical geography
Hydrology
Volcanology
Oceanography
Meteorology and atmo...
Geology
Environmental scienc...
Water resources
Palaeontology
Geochemistry and geo...
Climatic research
Mineralogy
Organic chemistry
Geophysics
Inorganic and nuclea...
Analytical chemistry
Colloid chemistry
Applied and industri...
Physical chemistry, ...

Erőforrások

HUN-REN
Magyar Kutatási Hálózat

- A projekthez igényelt erőforrások:

Név	Flavor	CPU cores	RAM (GB)	GPU	VRAM (GB)
milab1	g2.2xlarge	16	62,8	V100	32
milab2	m2.4xlarge	32	62,8	-	-
milab3	g2.xlarge	8	31,3	V100	16

- Adatok, modellek kezelése: **HuggingFace**
- Szoftverek
  - Pl. transformers, pytorch, huggingface_hub stb.
  - A legtöbb könnyen telepíthető egy **“pip install”** segítségével

Fejlesztési irányzatok

HUN-REN
Magyar Kutatási Hálózat

- Főbb kipróbált megoldások:
  - **Annif**
  - **Mistral-7B**
  - **SciBERT**
- Egyéb kipróbált megoldások:
  - Spacy
  - GraphSage
  - Scikit-learn modellek:
    - MLP
    - PassiveAgressiveClassifier
    - DecisionTreeClassifier

Annif

HUN-REN
Magyar Kutatási Hálózat

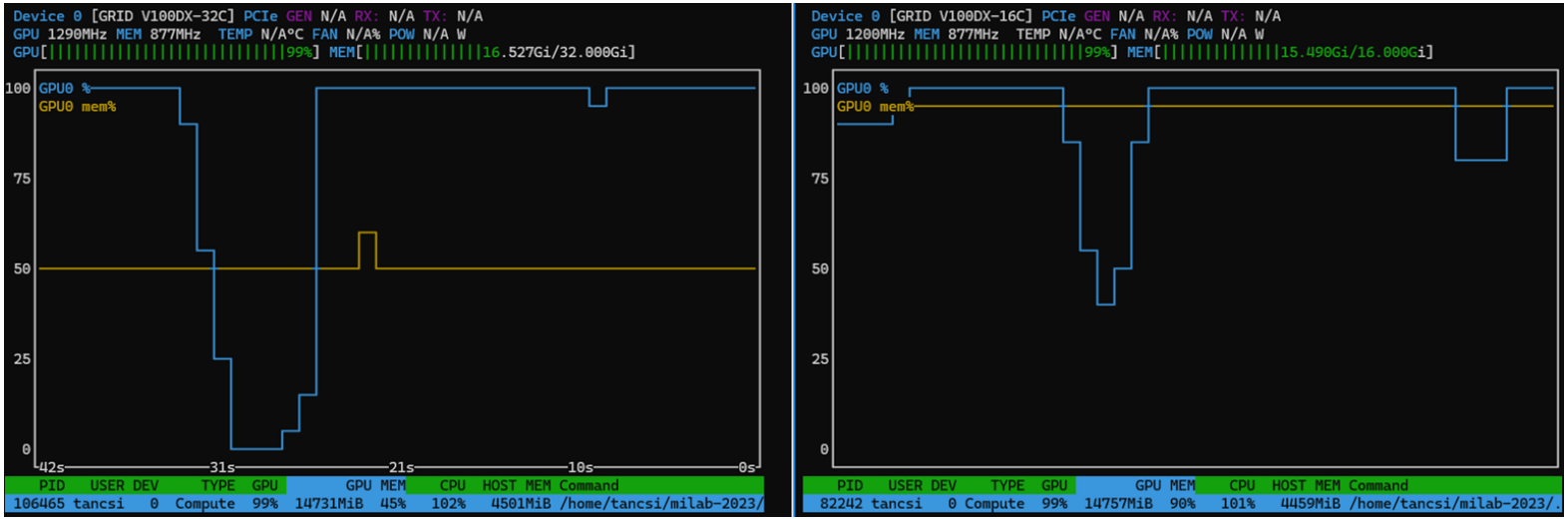
- Professzionális szövegosztályozó program - finn fejlesztés
- A tanítható modellek (backend-ek) és algoritmusok mind beépítettek, csak CPU-s tanítás lehetséges
- **Jelenlegi legjobb:** NN-ensemble modell, amely kombinálja a tfidf és omikuji-bonsai modelleket.
- A tanítási idő függ:
  - tanító adatok mennyiségétől
  - A konfigban beállított “epoch” számától
- Általános tanítási idő (10 epoch, 325 ezer adatsor, 32 CPU mag): **~70 perc**
- HPO keresés esetén (pl. 20 darab próbálkozás esetén): 20*70 = **~24 óra**

Mistral-7B

HUN-REN
Magyar Kutatási Hálózat

- A **transformers** és **pytorch**, “Casual Language Model” feladatként
  - A **system prompt** az utasítással és lehetséges tárgyszavak listájával, felügyelt tanítás (SFT) és Q4-es kvantálással (aka. **QLoRA**, a **bitsandbytes** könyvtárral)
- A tanítás hossza függ az adatok mennyiségétől, hosszától (tokenben mérve), az epoch-tól és a batch_size paramétertől
  - **4560 tanító adatsor**, **10 epoch** és **4-es batch_size** esetében a tanítási idő: **~25óra**
  - 4-es batch_size esetén a VRAM használat **~25-27 GB** volt
- Itt már lehetséges volt elosztott módon tanítani (2-es batch_size-al és **accelerate** segítségével)

Mistral-7B “data parallelism” példa



- **nvtop** linux parancs, GPU monitorozó felület
- **Kék:** végrehajtás, **Sárga:** memória használat
- Bal oldal **master node**, jobb oldal **worker node**
- A tanító adatok ketté lettek osztva, két teljes Mistral be lett töltve a GPU-ba
- A letörések a két gép közötti súly szinkronizáció miatt

SciBERT

HUN-REN
Magyar Kutatási Hálózat

- BERT modell tudományos szövegeken előtanítva, 512 token maximális context size-al “Sequence classification” feladatként
- Kimenet, mint “One-hot vector” a lehetséges tárgyszavakkal
- V100-as GPU esetén, kb. 80%-os kihasználás és 4-5 GB-nyi VRAM használat
- Tanítási idő
  - 325 ezer tanító sor, 50 epoch esetén: **~97 óra**
  - 7310 tanító sor, 50 epoch esetén: **~4-5 óra**
- Elosztott tanítás lehetséges (pl. **Ray**, **accelerate**)
  - A modell mérete miatt nem tudja kihasználni kellőképpen (~12-15%-os GPU kihasználás)
  - Az elosztott tanítási idő több lett, mint egy GPU-s tanítás esetén

Eredmények

HUN-REN
Magyar Kutatási Hálózat

- Klasszifikációs feladat lévén a teljesítményt F1-score segítségével lehet jellemezni
  - Kétféleképpen kerül kiszámításra az F1-score
    - **Eredeti** formulával, amely a pontos megegyezést számolja a TP értékeként
    - **Megengedő** formulával, például a modell azt prediktálta, hogy **“Statistics and probability”**, de a valós válasz a **“Pure mathematics, Applied mathematics”**, de mindkét elem a **“Mathematics”** leszármazottja, ezért TP += 0,5

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Eddigi legjobb eredmények

HUN-REN
Magyar Kutatási Hálózat

	Eredeti			Megengedő		
Alapmodell	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Annif	0,6412	0,6364	0,6388	0,6908	0,6364	0,6625
Mistral-7B	0,7117	0,6320	0,6695	0,7838	0,6320	0,6998
SciBERT	0,7654	0,7496	**0,7576**	0,7955	0,7576	**0,7761**

HUN-REN Cloud használati tapasztalatok

HUN-REN
Magyar Kutatási Hálózat

- Felhasználói támogatás, ismeretek átadása
  - **Science-cloud** weboldalán lévő **“Referencia architektúrák”** megfelelő segítséget nyújtottak
  - Ellenkező esetben cloud-ban jártas szakemberek gyors segítsége (Pallinger Péter, Tenczer Szabolcs)
- A GPU-k mennyisége és minősége olykor nem elegendő
  - Nagyobb modellek tanításához **több VRAM** kell
  - Mindegyik flavor-ben 1 GPU található, csak **“multi-node”** elosztott tanítás lehetséges
  - Az alacsony **“CUDA Compute Capability”** miatt vannak könyvtárak, amik nem használhatók emiatt (pl. **Unsloth**)
  - **NVIDIA GRID driverek** csak speciális fiókkal érhető el, nincs lehetőség driver frissítésre (se linux kernel frissítésre)
- Hálózat minősége, rendelkezésre állás megfelelő

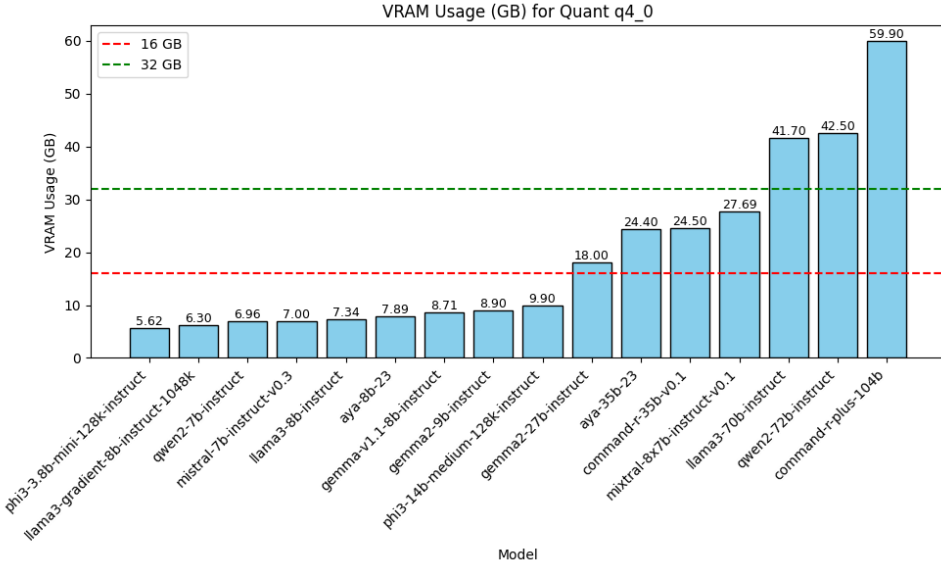
Egyéb opciók

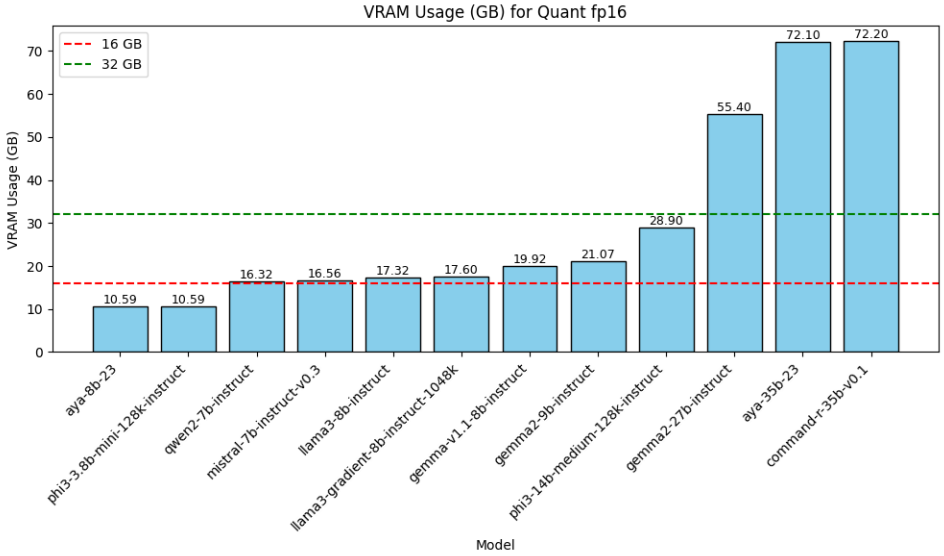
HUN-REN
Magyar Kutatási Hálózat

Szolgáltatás	GPU	Max. VRAM	CUDA cores	CUDA CC
HUN-REN Cloud (Wigner)	A100	40	6912	8.0
HUN-REN Cloud (SZTAKI)	V100	32	5120	7.0
Kaggle notebook (free tier)	P100	16	3584	6.0
Google Colab (free tier)	T4	16	2560	7.5

VRAM használat különböző modellek esetében (inference)

HUN-REN
Magyar Kutatási Hálózat





Összefoglalás

HUN-REN
Magyar Kutatási Hálózat

- A HUN-REN Cloud erőforrásai segítségével az MTMT használatát próbáljuk javítani, amely erőforrások nélkül jóval kevesebb kísérletet tudtunk volna elvégezni
- Viszont többféle kísérletet nem tudtunk végrehajtani, amelyek megoldhatók lennének a következő feltételekkel:

1. **Egy adott flavor-ben egynél több GPU**, lehetővé téve a **“tenzor alapú párhuzamosítást”** (gyorsabb tanítás/inference, esetlegesen nagyobb modellek betöltése)

1. **Gyorsabb GPU-k**, több kísérlet legyen egy egységnyi idő alatt (fejlesztési folyamat gyorsítása)

1. **Nagyobb VRAM-al** rendelkező GPU-k, nagyobb modellekkel való tanítás vagy inference érdekében

- Ígéretes, de még nem optimális eredmények a tudományos címkézésre (főleg a **SciBERT**)
  - További próbálkozások különböző **szintetikus** kiegészítő **adatokkal** és **más alapmodellekkel**

Köszönjük a figyelmet!

Kérdések és válaszok

HUN-REN
Magyar Kutatási Hálózat

https://hun-ren.hu