

POI vadászat szolgáltatói honlapokról

#POI #webscraping #adatok #Python

OpenStreetMap Meetup
2018. június 25.

Szala Kálmán
@kami911kami

HA LENNE 10.000 FORINTOM,

Az előző rész tartalmából

**MEGVEHETNÉM A GÉPHEZ EGT AZ ÚJ,
EREDETI MICROSOFT WINDOWS 95 PROGRAMOT**

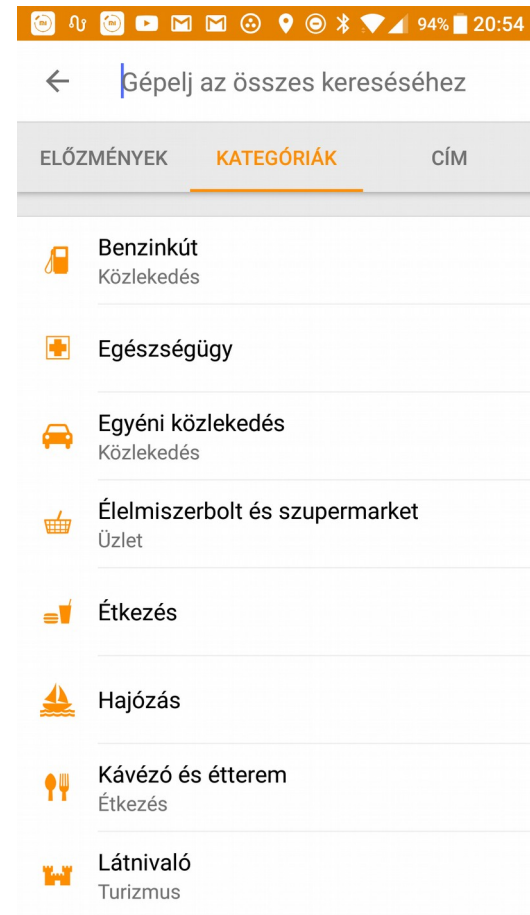
Ezeket terveztem



- Új adatok gyűjtése (POI)
 - Boltok, áruházláncok
 - Bankok
 - Éttermek, és étteremláncok
- Meglévő POI-k egységesítése
 - Azonos írásmód
 - Azonos kategória
 - Wiki oldal → közösen összeállítani
- Azonosítási mechanizmus, szabály
 - Elegendő-e az OSM ID, vagy valami ref:* szükséges

A POI-k fontosak

- Sokszor van rá szükség
- Keresés
- Meglévő POI-k
 - Sokszor kevés adat
 - Nincs fent minden POI
 - Nem egységesek a nevek





Minden tökéletes, lehet nyugodtan aludni :)

Sokszor kevés adat I

shop=convenience

A convenience store

[Áttekintés](#)[Kombinációk](#)[Térkép](#)[Wiki](#)[Projects](#)

Kombinációk

Ez a táblázat csak a leggyakoribb címkék leggyakoribb kombinációit mutatja.

◀◀	Oldal 1 / 31	▶▶	↻	JSON	Elemek megjelenítve 1 - 14, összesen 432	🔍
Darab →		Más címkék			→ Darab	
7 397	86.75% ▬	name=*			7 397	1.85%
5 195	60.92% ▬	source=*			5 195	1.33%
3 243	38.03% ▬	wheelchair=*			3 243	14.89% ▬
3 160	37.06% ▬	fixme=*			3 160	7.90% ▬
3 093	36.27% ▬	source=route4u			3 093	33.70% ▬
2 978	34.92% ▬	fixme=verify			2 978	33.09% ▬
1 959	22.97% ▬	url=*			1 959	3.13%
1 941	22.76% ▬	source=turistautak.hu			1 941	2.06%
1 838	21.56% ▬	source:date=*			1 838	3.60%
1 360	15.95% ▬	wheelchair=no			1 360	17.54% ▬
1 180	13.84% ▬	wheelchair=yes			1 180	12.43% ▬
1 137	13.33% ▬	toilets:wheelchair=*			1 137	18.09% ▬
879	10.31% ▬	addr:street=*			879	0.42%
861	10.10% ▬	addr:postcode=*			861	0.38%

Sokszor kevés adat II

amenity=pharmacy

A shop where a pharmacist sells medications

[Áttekintés](#)[Kombinációk](#)[Térkép](#)[Wiki](#)[Projects](#)


Kombinációk

Ez a táblázat csak a leggyakoribb címkék leggyakoribb kombinációit mutatja.

◀ ◀ Oldal 1 / 20 ▶ ▶ ↻ JSON Elemek megjelenítve 1 - 14, összesen 273

Darab →		Más címkék	→ Darab	
1 225	70.28%	name=*	1 225	0.31%
602	34.54%	addr:postcode=*	602	0.26%
601	34.48%	addr:street=*	601	0.28%
598	34.31%	addr:city=*	598	0.27%
566	32.47%	addr:housenumber=*	566	0.31%
506	29.03%	opening_hours=*	506	4.22%
379	21.74%	wheelchair=*	379	1.74%
371	21.29%	source=*	371	0.10%
312	17.90%	url=*	312	0.50%
310	17.79%	source=turistautak.hu	310	0.33%
304	17.44%	source:date=*	304	0.60%
243	13.94%	wheelchair=yes	243	2.56%
227	13.02%	phone=*	227	2.42%
200	11.47%	building=*	200	0.02%

Megszűnt vagy átalakult POI-k



2016. augusztus 1-én az Eni Hungaria Zrt.-t - a kenőanyag üzletág kivételével - megvásárolta a MOL Nyrt., így többé nem része az Eni csoportnak.

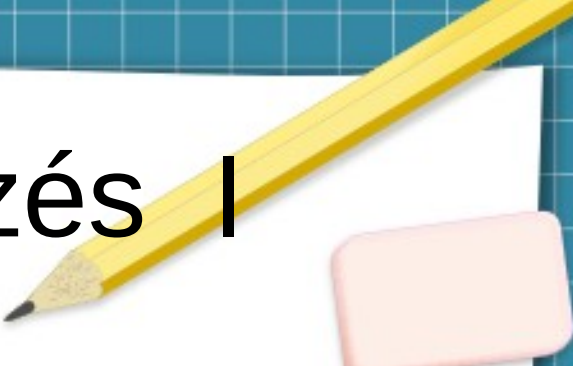
- 80 darab
 - <https://taginfo.openstreetmap.hu/tags/brand=Agip>

Nincs minden POI az OSM-ben



- Nemzeti Dohányboltok
 - 5799 vs. 1597
- K&H bankok és ATM automaták
 - 614 vs. 172
- Nem is lehet :(

Nem egységes elnevezés I



Name tag	Darabszám	Százalék
MOL	319	0.08%
Mol	35	0.01%

Nem egységes elnevezés II

Name tag	Darabszám	Százalék
OTP_Bank	304	0.08%
OTP	195	0.05%
OTP_ATM	25	0.01%
OTP_bankautomata	8	0.00%
OTP_Bank_+_ATM	6	0.00%
OTP_Bankomat_ATM	5	0.00%
OTP_Bankautomata	5	0.00%
OTP_Ingatlanpont	5	0.00%
OTP_bank	4	0.00%
OTP_Bank_és_ATM	3	0.00%
otpbank_ATM	3	0.00%
OTP_Bank_ATM	2	0.00%
OTP_-_ATM	2	0.00%

Szolgáltatói adatok



- Benzinkutak
 - Avia, Shell, OMV, MOL
- Gyógyszertárak:
 - Benu, Kulcs
- Posták, Posta Pontok, csomagautomaták, csekkautomaták
- Boltok
 - CBA, Príma, dm, Rossmann, Spar, Tesco
- Bankok és ATM-ek
 - K&H
- Egyebek
 - Foxpost, MOL Bubi



POI aratás

Bemeneti adatok I



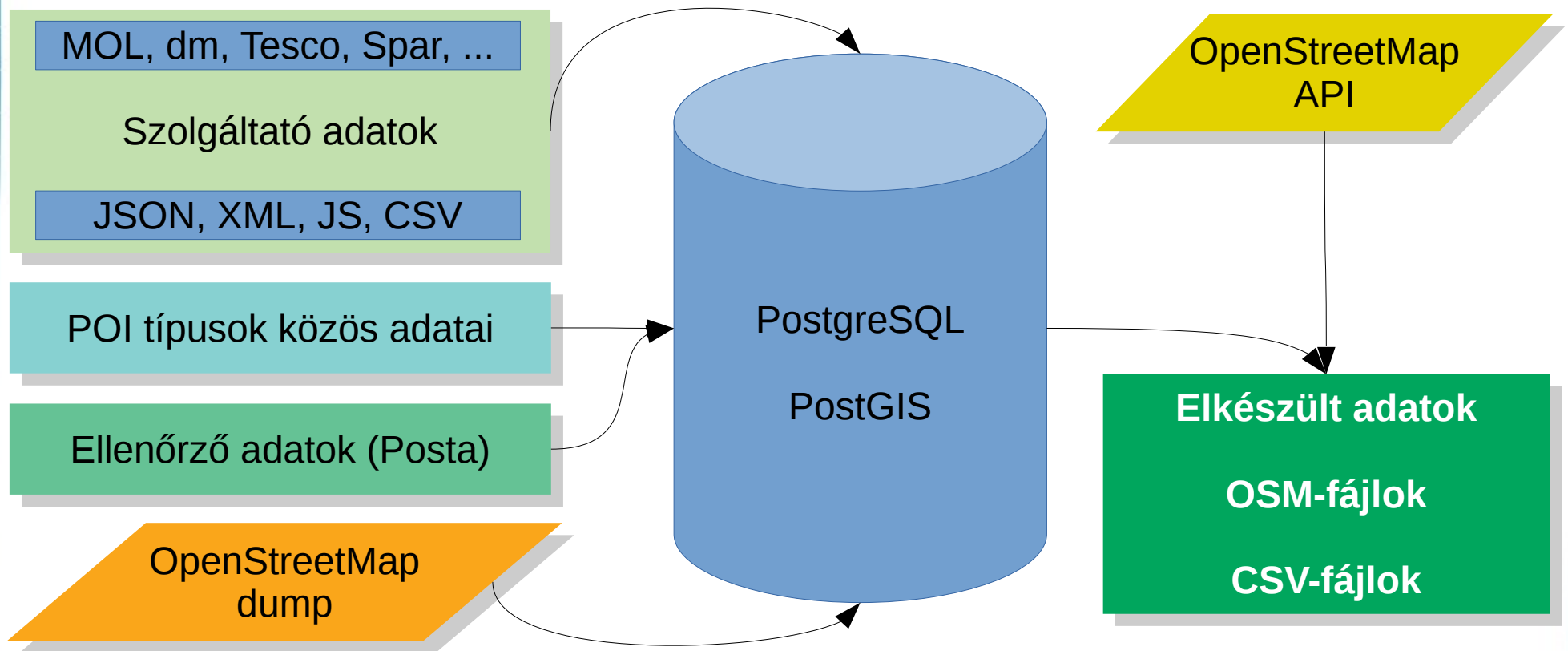
- Weben elérhető adatok
 - Posta Partner Extra
 - XML
 - Településnevek és irányítószámok
 - Közterület nevek
 - <https://www.posta.hu/partnerextra>
 - Szolgáltatói adatok
 - JavaScript var
 - JSON
 - CSV
 - XML

Bemeneti adatok II

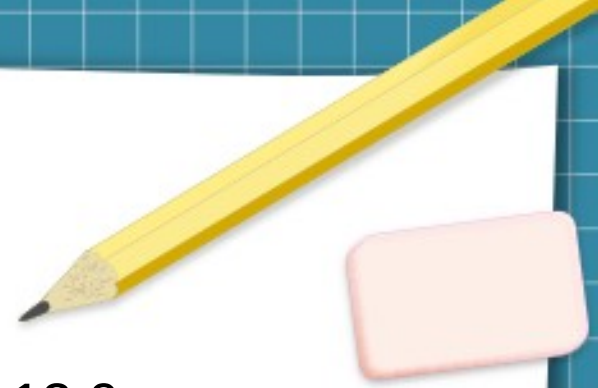


- Weben elérhető adatok
 - OpenStreetMap API
 - OsmApi Python modul
 - OpenStreetMap dump
 - <http://download.geofabrik.de/europe/hungary.html>
 - osm2pgsql → PostgreSQL, PostGIS betöltés
 - Egyedi stílusfájl alapján

POI kereső



Mi kell hozzá?



- Python (3.6.5)
- PostgreSQL (10.4) / PostGIS (2.4)
- Python modulok
- Virtualenv
- Egy adatbázis
 - PostGIS extension
 - osm2pgsql
 - osm2pgsql/default.style
- numpy \geq 1.13.0
- osmapi
- pandas \geq 0.23.1
- geopandas \geq 0.3.0
- scipy \geq 0.19.0
- requests
- bs4
- sqlalchemy
- geoalchemy2
- psycopg2-binary
- xlrd



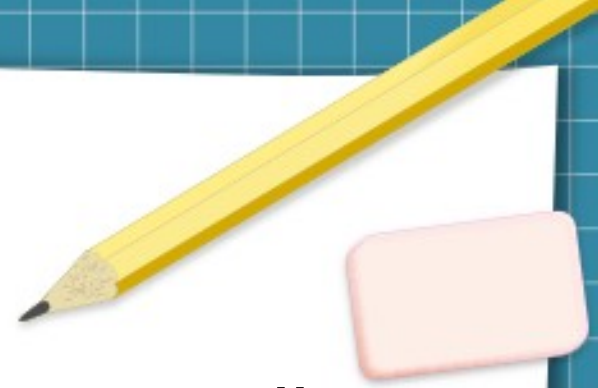
**osm2pgsql -c -m -s -d poi -C 8000 -U poi -W -style
osm2pgsql/default.style --extra-attributes -H
localhost ~/hungary-latest.osm.pbf**

READ THE SOURCE LUKE



https://github.com/KAMI911/osm_poi_matchmaker

Futtatás



- app.conf
 - # Database settings
 - db.type=postgresql
 - db.write.host=localhost
 - db.write.port=5432
 - db.write.username=poi
 - db.write.password=*****
 - db.poi.database=poi
- python3 ./create_db.py

Statisztikák



- Jelenlegi POI-k száma
 - 6956 darab
- Ebből az OSM-ben
 - 30 méteres körzetben keresve
 - 2233 darab
- vagy
 - 70 méteres körzetben keresve
 - 3228 darab



Houston, we have a problem

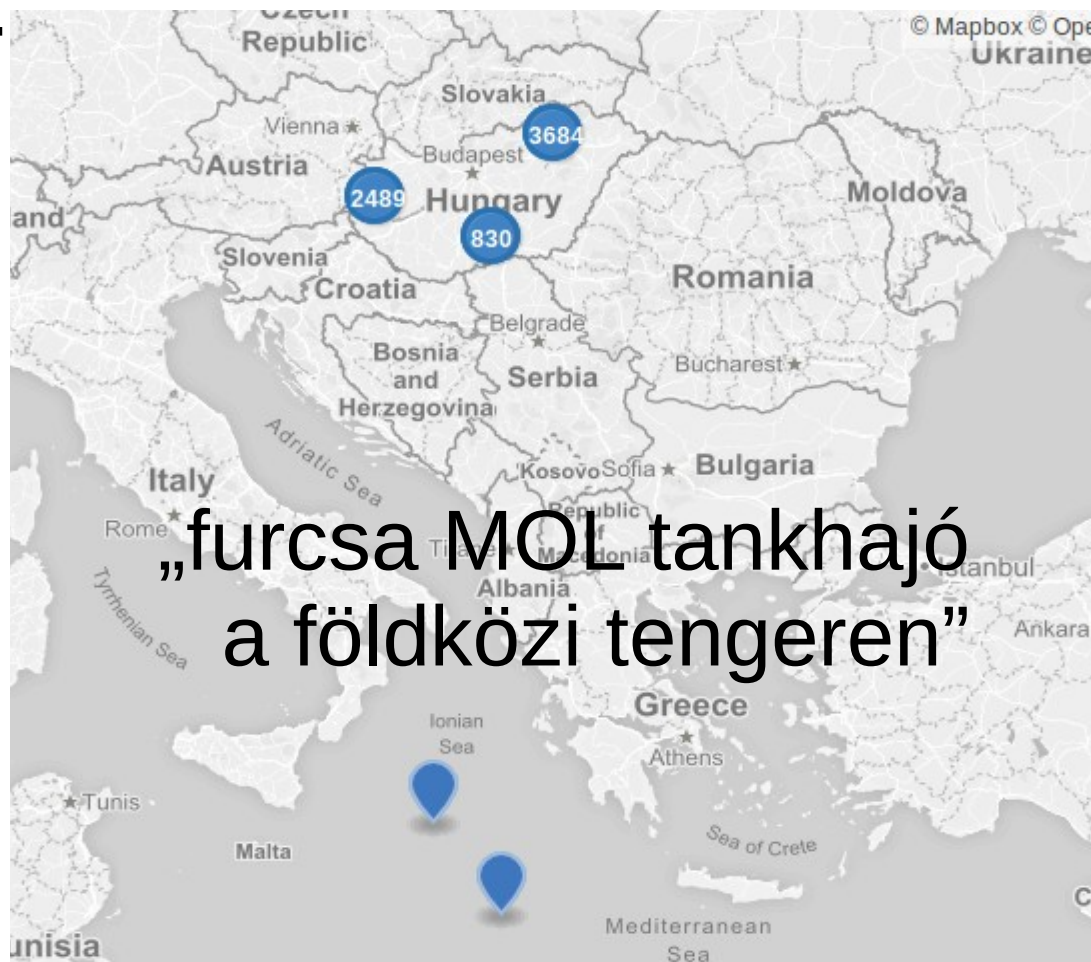
címEK



- BERCSÉNYI U.1 2934/5 HRSZ
- BATTHYÁNY TÉR 5-6, Bp. Batthyány tér 5-6
- 9737 Bük, 0131/3 hrsz.
- Kaszás u. 2.-Dózsa György út 87.
- Rétköz u.10/b. (volt Nagyszeben tér)
- 9653 Répcelak, 076/15. hrsz 86. számú főút mellett
- CSIPKEGYÁRI UTCA 11., Győr,Csipkegyári utca 11.
- Zalai u.2. (Szarvas Gábor u.33)
- Pécs, belterület

Koordináták

- Felcserélt koordináták
- Rossz koordináták:



Visszajelzések I



„Köszönettel vettük egyes MOL-töltőállomások nem megfelelő koordinátáira vonatkozó észrevételét.

A hibás paramétereket illetékes kollégáink javították rendszerünkben.

Jelzését köszönjük, és bármely további kérdés, észrevétel esetén készséggel állunk rendelkezésére ügyfélszolgálatunk elérhetőségein!”

Visszajelzések II



- „Jelzése alapján, az állományok javításra kerültek. Amennyiben további hibát, pontatlanságot talál, akkor azt kérjük Ügyfélszolgálatunkon keresztül jelezze, és javítjuk.”
- „Észrevételét illetékes munkatársak részére továbbítottuk.”

Bejövő adatok



- `osm_poi_matchmaker/dataproviders`
 - `hu_aldi.py`
 - `hu_avia.py`
 - `hu_benu.py`
 - `hu_cba.py`
 - `hu_cib_bank.py`
 - `hu_dm.py`
 - `hu_foxpost.py`
 - `hu_generic.py`
 - `hu_kh_bank.py`
 - `hu_kulcs_patika.py`
 - `hu_mol.py`
 - `hu_mol_bubi.py`
 - `hu_omv.py`
 - `hu_penny_market.py`
 - `hu_posta.py`
 - `hu_rossmann.py`
 - `hu_shell.py`
 - `hu_spar.py`
 - `hu_tesco.py`
 - `hu_tom_market.py`

Adattárolás



- poi_common
 - Közös adatok
 - Név
 - Közös tag-ek
 - Kategóriák
 - Amenity
 - Shop
 - Kapcsolati címek
 - Alap URL
- poi_address
 - A POI adatai
 - Cím → city tábla
 - Branch
 - Nyitvatartás
 - URL kiegészítés



POI-k felkutatása

POI-k felkutatása



- Koordináta alapján
- Azonos kategória
- 70 méteres körzetben
 - Konfigurációs fájlból konfigurálható
- Node, ha még nincs
- Ha létezik, akkor a már meglévő frissítése
 - Node
 - Way



hutescosup?

Do You Speak This Language?



huaviafu hubenupha hububibir hucbacon
hudmche hufoxpocso hukhatm hukhbank
hukulcspha humobpefu humolfu huomvfu
hupostacse hupostacso hupostapo hupostapp
huprimacon hurossmche hushellfu husparexp
husparint husparsup hutescoexp hutescoext
hutescosup

Azonosítók



huaviafu hubenupha hububibir hucbacon hudmche
hufoxpocso hukhatm hukhbank hukulcspha
humobpefu humolfu huomvfu hupostacse
hupostacso hupostapo hupostapp huprimacon
hurossmche hushellfu husparexp husparint
husparsup hutescoexp hutescoext hutescosup

hu tesco sup

Hungary Tesco Supermarket

A close-up photograph of a terracotta pot filled with dark soil. Numerous small, bright green seedlings with rounded leaves are growing densely together. The seedlings in the foreground are in sharp focus, while those in the background are blurred. A semi-transparent dark horizontal band is positioned across the lower third of the image, containing white text.

Hogyan segíthetsz?

Hogyan segíthetsz?



- Kimeneti fájlok átnézése, ellenőrzése, javítása
 - <https://github.com/KAMI911/osm-import-data/tree/master/hu-poi-data>
- Hibás bemeneti adatok felkutatása
- Javaslatok további közös kulcs-érték párokra
- Hibák jelentése
- Újabb bemeneti adatok felkutatása
- Importálók készítése
- Javítások és tesztek készítése
- Továbbfejlesztés



Továbbfejlesztési lehetőségek

Továbbfejlesztési lehetőségek



- POI-k azonosítása
 - Közös ID
 - Cím alapján képzett SHA512 hash (?)
- Okosabb POI azonosítási eljárás
- POI koordináták gyorsítótárazása
- POI attribútumok gyorsítótározása
- További attribútumok tárolása és feldolgozása



Kérdések

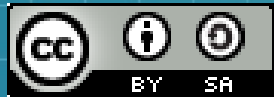


Köszönöm a figyelmet!

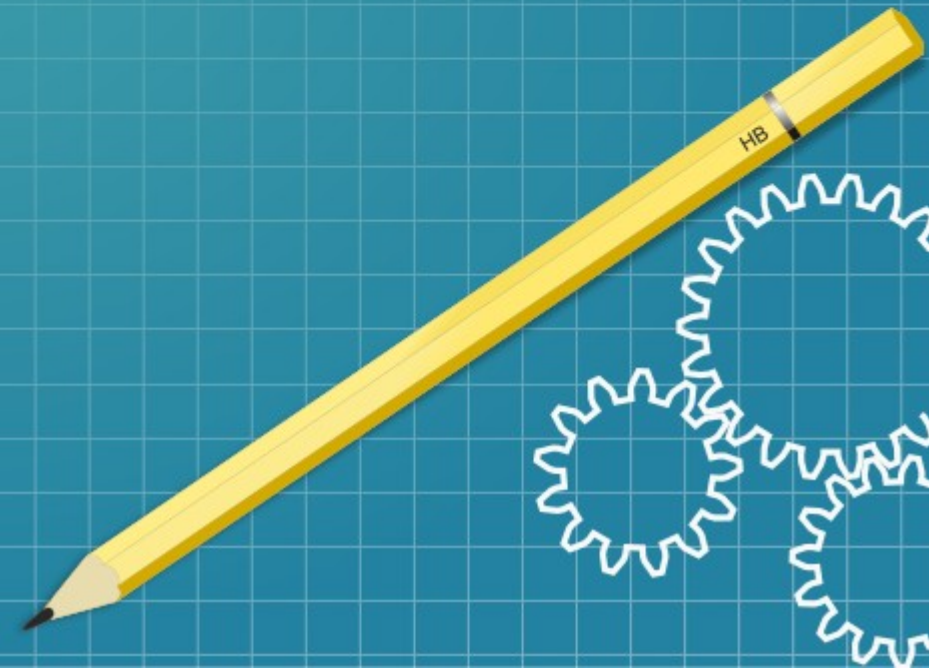
**Szalai Kálmán (KAMI)
@kami911kami**



This work is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License.



SlideShare:



Képek



- <http://images2.fanpop.com/image/photos/13900000/Cute-Dog-Wallpaper-dogs-13936344-1024-768.jpg>
- <http://barkpost.com/wp-content/uploads/2013/09/Tennis-Ball-for-Dogs-1.jpg>
- https://upload.wikimedia.org/wikipedia/commons/2/23/Rostov%27s_region_combain.jpg
- https://commons.wikimedia.org/wiki/File:Rostov%27s_region_combain.jpg
- https://cdn.pixabay.com/photo/2016/03/10/22/23/lettuce-1249414_960_720.jpg
- <https://hu.m.wikipedia.org/wiki/F%C3%A1jl:Dart>