Molnár Anna Enikő és Tamási-Mészáros Evelin

Sok a szöveg?!
Olvass inkább a sorok közt!

2023.05.18.

Clementine

VERSENY

EUROSTAT

- Több, mint 112 ezer webes álláshirdetés 400 weboldalról
- Feladat:
  - egymást átfedő álláshirdetések keresése többféle szempont szerint
- Nehézségek:
  - többnyelvű szövegek közötti hasonlóságok megtalálása (32 nyelv)
  - sablonok kiszűrése

We are looking for a data analyst. Full-time, 40 hours per week.

Nous recherchons un analyste de données. Temps plein, 40 heures par semaine.

Offerte di lavoro per analisti di dati.

We are looking for an accountant. Full-time, 40 hours per week.

Clementine

MÓDSZERTAN

- Adatelőkészítés kérdések
  - adatelőkészítés szokványos módszere megkérdőjeleződött
    - bizonyos adattisztításnak jelentésmódosító szerepe volt (pl. rövidítések, tulajdonnevek)

- Nyelvi egységesítés
  - Nincs mindegyik nyelvhez elérhető, jól működő szövegtisztító csomag
    - SpaCy-nek jelenleg 21 európai nyelvhez (pl. cseh, szlovén, szlovák nincs)
  - Egyesével mind a 32 nyelvre nagyon időigényes a tisztítás

- 6 milliárd feletti pár -> nagyon számításigényes

SENTENCE-TRANSFORMERS

- Mondatok, bekezdések, képek közötti hasonlóság mérésére használható modellek gyűjteménye

- Két felhasznált modell:
  - all-MiniLM-L6-v2
    - Több mint 1 milliárd angol nyelvű tanító páron tanítva
  - paraphrase-multilingual-MiniLM-L12-v2
    - 50+ nyelven előtanítva, köztük magyaron is

- Eredmény: több, mint 575 ezer potenciális pár az előre definiált kategóriák szerint
- Kategóriák:
  - Teljesen egyező
  - Részben egyező
  - Szemantikus duplikáció

DUPLIKÁCIÓK KEZELÉSE

- Egyező tartalom
  - torzíthatja a statisztikát
  - összezavarhatja a keresőmotorokat
- Nem csak a szószerinti egyezés számít duplikációnak!

- Duplikáció vs átvett tartalom
  - Duplikált tartalom: olyan kontent, ami nagyrészt vagy teljesen megegyezik egy másik domain tartalmával
  - Átvett tartalom: egy harmadik oldal jeleníti meg a cikket minél többször, akár eredeti formájában, akár egy jelentősen rövidített verzióban
    - Nem sajátjaként állítja be, hivatkozik a szerzőre

Gyakorlati alkalmazás

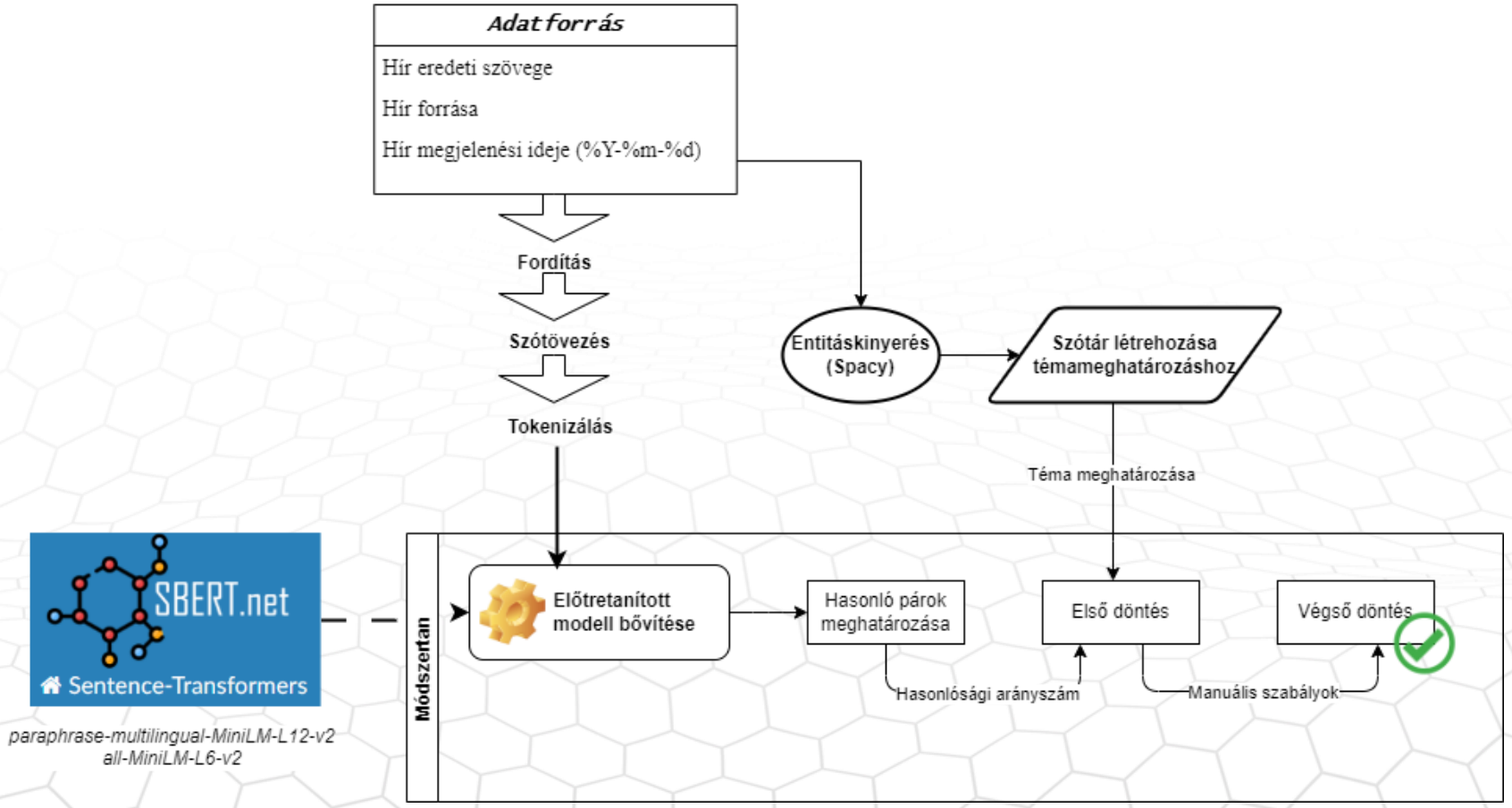
Clementine

GYAKORLATI ALKALMAZÁS

- Koronavírussal kapcsolatban a magyar elektronikus sajtóban megjelent cikkek
- 2020 és 2023 között
- Analóg kérdés: az egyes cikkek információtartalma mennyiben egyezik?
- A duplikációmentesítés finomítottabb kérdése

Clementine

FOLYAMAT



Clementine

KONKLÚZIÓ I.

- Adott nyelv sajátosságainak figyelembe vétele
- Megjelenés dátumának fontossága

Álláshirdetések	*Hírek*
Megjelenés dátuma	
Ország	
Nyelv	
Cég neve	*Forrás - hírportál neve*
Rövid leírás	*Hír címe/rövid leírás (ahol van)*

Szöveghez tartozó egyéb változók

A hasonlósági érték kapcsolatot határoz meg.

Clementine

KONKLÚZIÓ II.

Clementine

TAPASZTALAT

✓ „Média kapcsolati hálója”
⊗ Megjelenési dátum ~ hasonlóság mértéke

A Pfizer felgyorsítja az EU-nak szánt szállításait

Czinkóczi Sándor
JÁRVÁNY · 2021 március 16., 13:21

Az Európai Bizottság 10 millió adag vakcina gyorsított szállításáról állapodott meg a Pfizer - BioNTech konzorciummal a második negyedévre - közölte Ursula von der Leyen, az Európai Bizottság elnöke kedden a Twitteren.

Ursula von der Leyen hozzátette, hogy a gyorsított eljárás keretében az uniós tagállamoknak szánt 10 millió adag vakcina a második negyedévre korábban tervezett 200 millió dózison felül értendő. A friss megállapodás pótolhatja a vakcinaszállítások esetleges hiányosságait, és szélesebb mozgásteret biztosít a tagállamoknak oltási kampányuk teljesítésére.(DW/MTI)

200 millió darabot rendelt egy eddig nem használt vakcinából az Európai Unió

24.hu | 2021. 05. 03. 17:51

200 millió adag oltóanyagról állapodott meg Novavaxszal az Európai Unió- írja a Reuters. Az amerikai cég azt vállalta a hír szerint, hogy 2021 végéig szállítja le ezt a mennyiséget.

Az Európai Gyógyszerügynökség (EMA) már hónapok óta vizsgálja az oltóanyagot, amelyet eddig még sehol nem engedélyeztek. A megállapodás már korábban terítéken volt, de azt eddig elhalasztották.

Clementine

Köszönjük a figyelmet!

Clementine