Online sajtócikkek adatbázisba rendezése webaratás segítségével

Indig Balázs

Eötvös Loránd Tudományegyetem, Digitális Bölcsészet Központ
Nyelvtudományi Intézet, Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály

indig.balazs@btk.elte.hu

Vázlat

1. Motiváció
2. Módszerek
3. A klasszikus webaratás folyamata
4. Próbáljuk meg máshogy!
5. A puding próbája...
6. Következtetések

Motiváció

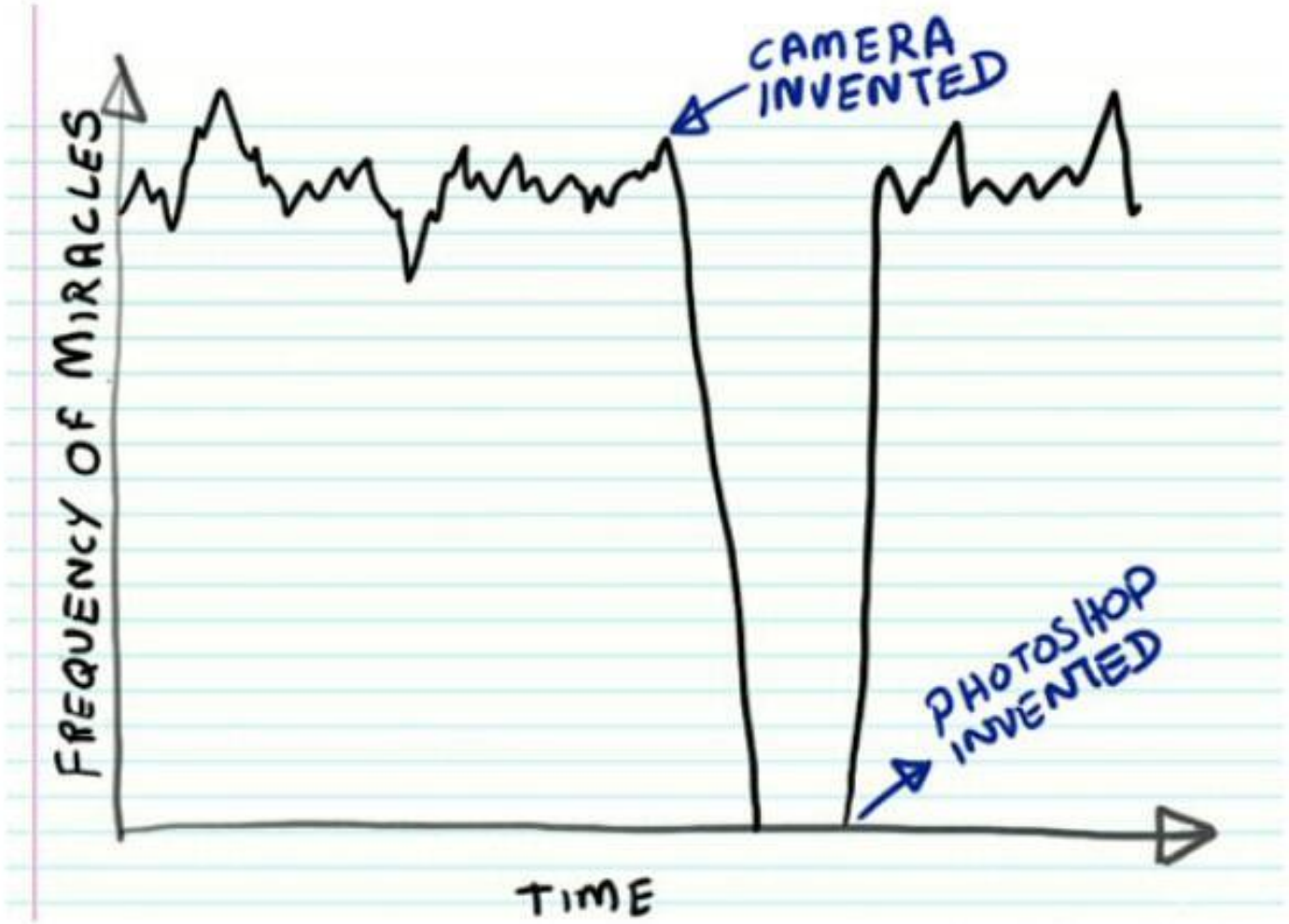
Hangzatos cikkek

- Egy gép elolvasott 3,5 millió könyvet. Ezt tudta meg a nőkről és a férfiakról
- Egyre több a gyűlöletbeszéd...
- Kétszáz év alatt háromszor volt boldog a nyugati ember
- Írj 10 mondatot, megmondom, ki vagy! Nyelvészet a bűnüldözés szolgálatában
- Kulturális távolságok a nemzeti sztereotípiák alapján

De vajon hihetünk nekik? Reprodukálható, tudományos eredmények vagy bulvár? **Melyik a kakukktojás?**

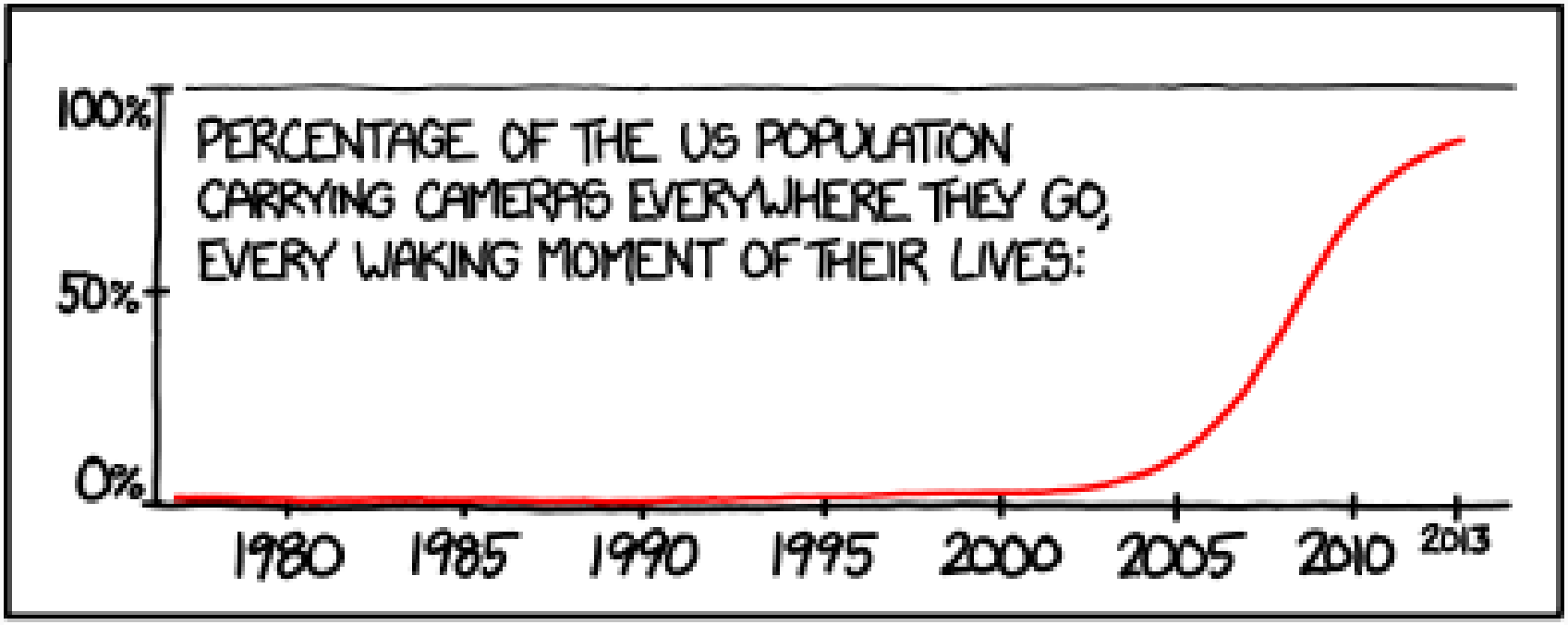
Ha mások meg tudják csinálni, akkor a tudományos álláspontra is szükség lesz előbb-utóbb...

A csodák gyakorisága



Forrás: https://twitter.com/pickover/status/528708130520461312

Megnyugtatóan rendezett kérdések (xkcd.org/1235)



A szöveges tartalmak hiteles (!) megőrzése

- Szöveget a szövegboltból… *Nemzeti könyvtárak*
  - Ha egy régi könyvet akarunk elolvasni, akkor könyvtárba megyünk
  - Van katalógus, amiben lehet keresni (metaadat)
  - Sárgák a lapok, tehát régi a könyv (hitelesség)
- A Web 2.0 óta, rengeteg szöveg **eleve digitális (born digital)**
  - Van a *Common Crawl* és az *Internet Archive*), csak hiányosak
  - **Egyik napról a másikra megváltozhat vagy eltűnhet egy oldal**
  - Viszont könnyeben gyűjthetők, feldolgozhatók és hitelesíthetők (?)
  - A papír sárgul, a bitek rohadnak

Módszerek

A webaratás korának hajnala

Mit mond a **Főnök**?

- Az általános **nyelvtechnológiai** munkafolyamat:
  ‘`Szerezz VALAMILYEN szöveget, amivel dolgozhatunk! A pontos tartalom nem számít.`’
- Az átlagos **digitális bölcsész, társadalomtudós** munkafolyamat:
  ‘`Szerezd meg AZT A SPECIFIKUS szöveget, amivel dolgozhatunk! A pontos tartalom számít igazán.`’

A klasszikus webaratás folyamata

A klasszikus webaratás folyamata

1. Indíts el egy **tipegőt (crawler)** valamilyen kezdeti paraméterekkel
    - Induló oldalak, domain, mélység, szélesség
2. Automatikusan nyerjél ki metaadatokat
3. Rendezd adatbázisba és szolgáld ki

Mi van akkor ha egy történész egy specifikus hírportál összes cikkén szeretné vizsgálni az eredeti megnyilvánulásokat?

A klasszikus webaratás folyamata (problémák)

1. Miért bízzak meg az archívumban/archiválóban?
    - Feltörhették, módosíthatták
2. Mi van ha hiányos az archívum?
    - Nem elég a mélység
    - SEO csapdák
    - Dinamikus oldalak
3. Mi van ha sok a szemét, amiből ki kell bányászni a szöveget?
    - Túl sok a mélység
4. Minőségbiztosítás? Mire is?
    - Nem tudjuk, hogy 100 év múlva mire lesz szükség
    - De ha a mai igényekre nem jó, akkor biztos nem jó
    - A távoli olvasás (*distant reading*) a jövő útja

A klasszikus webaratás folyamata

Próbáljuk meg máshogy!

A webarchiválás „távoli olvasás” megközelítésben

0. **Tegyük fel a kutatási kérdéseinket tágan értelmezve**
1. Gondosan válasszuk ki a learatandó oldalakat
2. Vizsgáljuk meg őket, hogy kinyerjük a lényeges tulajdonságaikat
3. A megszerzett információval felvértezve indítsuk az aratást
4. Mentsük el az oldalakat – **ezek az elsődleges forrásdokumentumaink!**

5. Használjunk *portálra szabott* sablonszűrést és metaadatkinyerést, futtassuk az eszközöket (szótövezés, stb.)
7. Mentsük el a korpuszt máshova – **hiszen automatikusan reprodukálható**
8. Szolgáljuk ki a szöveget és válaszoljuk meg a kérdéseinket

9. Találjunk és javítsunk hibákat a használt rendszerben
10. Menjünk vissza az 5-ös lépéshez és kezdjük újra **UGYANAZZAL** a szöveggel

A fő ötlet [Indig et al., 2019]

„Ha egy **CIKK** nincs **A (PORTÁL) ARCHÍVUM(Á)BAN**, akkor nem is létezik!” (ferdítve a *Csillagok Háborújá*ból)

A technikai részletek [Indig et al., 2019]

Kétszintes webaratás és **portál-alapú sablonszűrés**:

- A legtöbb (hír) portál **permalinkeket** használ a cikkek azonosítására és van egy **cikkarchívuma** amiben a cikkek kereshetők
  - A cikkarchívum egyszerű felépítéséből fakadóan könnyen kinyerhetők a cikkek linkjei (**dilemma**: szabályok vagy gépi tanulás?)
- Csak ezeket a linkeket járjuk végig
  - Gyakorlatilag nincs duplum vagy szemét!
  - **Kevesebb zaj, kisebb terhelés, gyorsabb aratás**
- Az adott portálnak van egy sajátos designja, ami azonos vagy nagyon hasonló minden cikkre
  - Egyszerű, hatékony szabályokkal vagy célzott gépi tanulással kezelhető (újra egy **dilemma**)

A technikai részletek [Indig et al., 2019] (folyt.)

- Az **ISO szabvány WARC archívum formátumot** használjuk
  - **Innentől minden reprodukálható, de még nem hiteles!**
- A kiválasztott oldalakhoz igazítottuk a webaratás és sablonkinyerés folyamatát
  - Mivel egy oldal sablonja ritkán változik, **minden nap learatható**
  - *Egy könnyen ellenőrizhető keretrendszerben*
- Szükség szerint felülvizsgálhatók és javíthatók a szabályok

A puding próbája…

A feladat és az erőforrásaink

A feladat:

- Hat (struktúrálisan) eléggé különböző magyar hírportálról
- Nyerjünk ki metaadatokat: *szerző*, *megjelenés dátuma*, *cím*, *lead*, *kulcsszavak*, *szöveg*
- Legyen az egész **precíz és fenntartható**, a futásidő másodlagos
- Hasznosítsunk újra mindent, ha csak lehet!

Az erőforrásaink:

- Egy „olcsó” irodai gép (4 GB RAM, Intel i3, 4 szál)
- 100 Mb/s kapcsolat

Összehasonlítás

A tipegő:

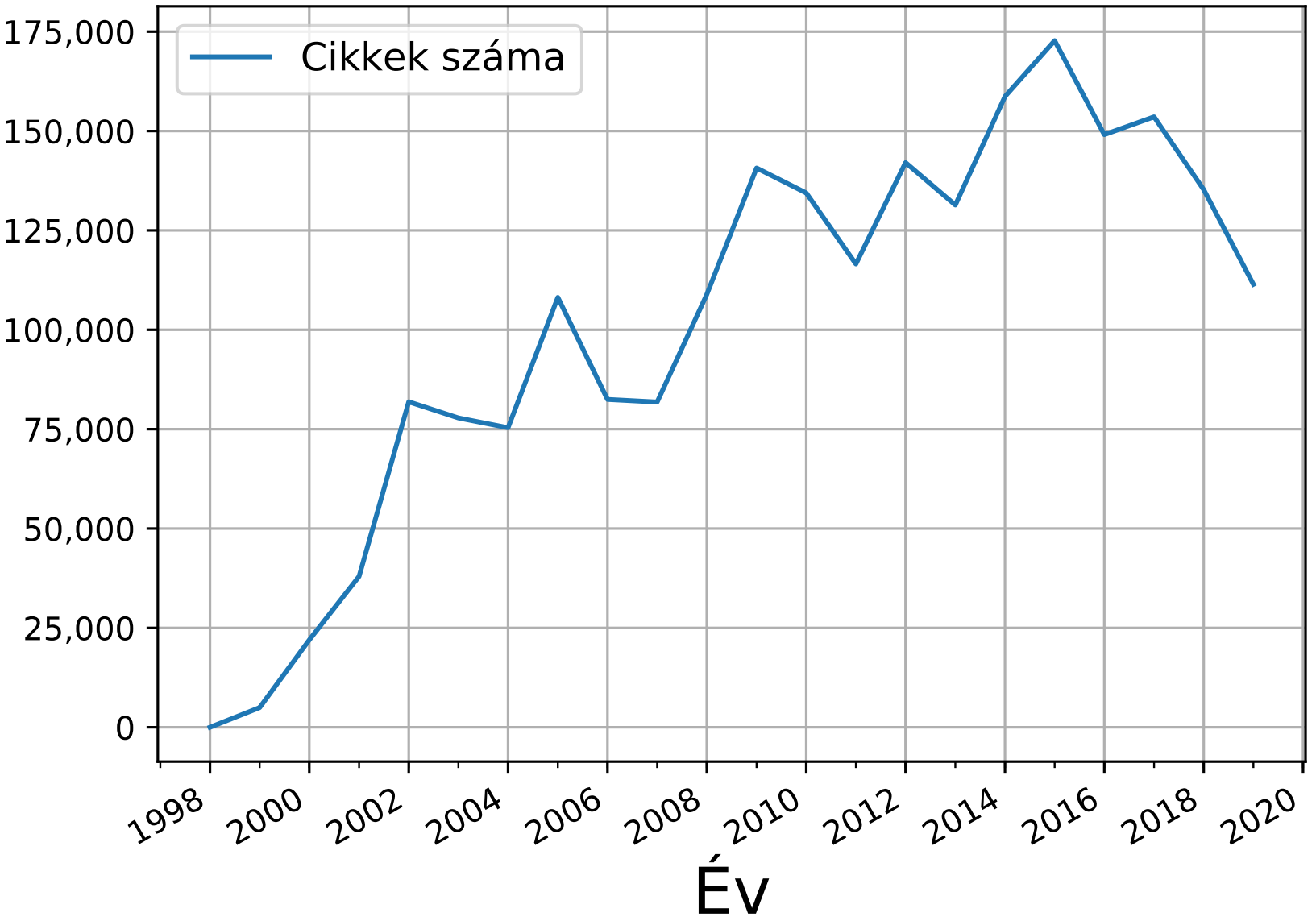
- A program működése nem összevethető a meglévőkkel
- Csak az eredmény!

A sablonszűrő eszközök összevetése (JusText [Pomikálek, 2011], Newspaper3k [Ou-Yang, 2013], mi szabályaink) [Indig et al., 2019]:

- **Mindegyik szabályalapú**, nehezen összevethetőek
- A miénk speciális és moduláris, a többi általtános és **monolitikus**
- A legtöbb eszköz, egyáltalán nem képes metaadat kinyerésre, vagy nem kezelik jól a magyar tipográfiát

Ahogy nőnek a számok, újabb problémák kerülnek felszínre

A learatott 2 227 180 darab cikk (6 hírportál) éves eloszlása



Következtetések

Következtetések

- **30 nap alatt** egy olcsó PC-vel (és sebességkorlátozással)
- Kevesebb mint 120 GB hely kellett (csak a HTML-ek warc.gz-ben)
- Nagyjából **egy milliárd** token körül lehet az archívum és nő
- Fenntartható, **alacsony terhelés mindkét oldalon**
- Reprodukálható, javítható, kiterjeszthető
- **Úttörő** munka számtalan későbbi kutatáshoz
  - Téma modellezés, stilometriai vizsgálatok (a rendelkezésre álló metaadattal)
  - Időbeli (socio-)lingvisztikai vizsgálatok (a megjelenés dátumával)
  - A munkafolyamat gépi tanulással való bővítéséhez tanuló adat
  - A célzott oldalak számának kiterjesztése
- Jövőbeli tervek:
  - Sztenerdizált munkafolyamat, TEI kimenet, több összehasonlítással
  - **A digitális dokumentumok hitelességének kérdése**
  - Szemantikus kereső szolgáltatás

Felhasznált irodalom i

Indig, B., Kákonyi, T., and Novák, A. (2019).
Crawling in reverse – lightweight targeted crawling of news portals.
In Kubis, M., editor, *Proceedings of the 9th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 81–87, Poznań, Poland. Wydawnictwo Nauka i Innowacje.

Ou-Yang, L. (2013).
Newspaper3k: Article scraping and curation.
`https://github.com/codelucas/newspaper`.

Felhasznált irodalom ii


Pomikálek, J. (2011).
Removing boilerplate and duplicate content from web corpora.
PhD thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic.