Text Enrichment

Szántó Tamás
Benda Krisztián

Ötlet

- A népszerűbb keresőmotorok csak kifejezések feldolgozását támogatják
- Nem alkalmasak egy hosszabb szöveg értelmezésére
- Jó lenne egy olyan program, ami segítségével hosszabb szövegeket tudunk értelmezni.
- Pl.: Hírek, újságcikkek olvasásakor ne kelljen a kifejezésekre külön-külön rákeresni

Megvalósítás

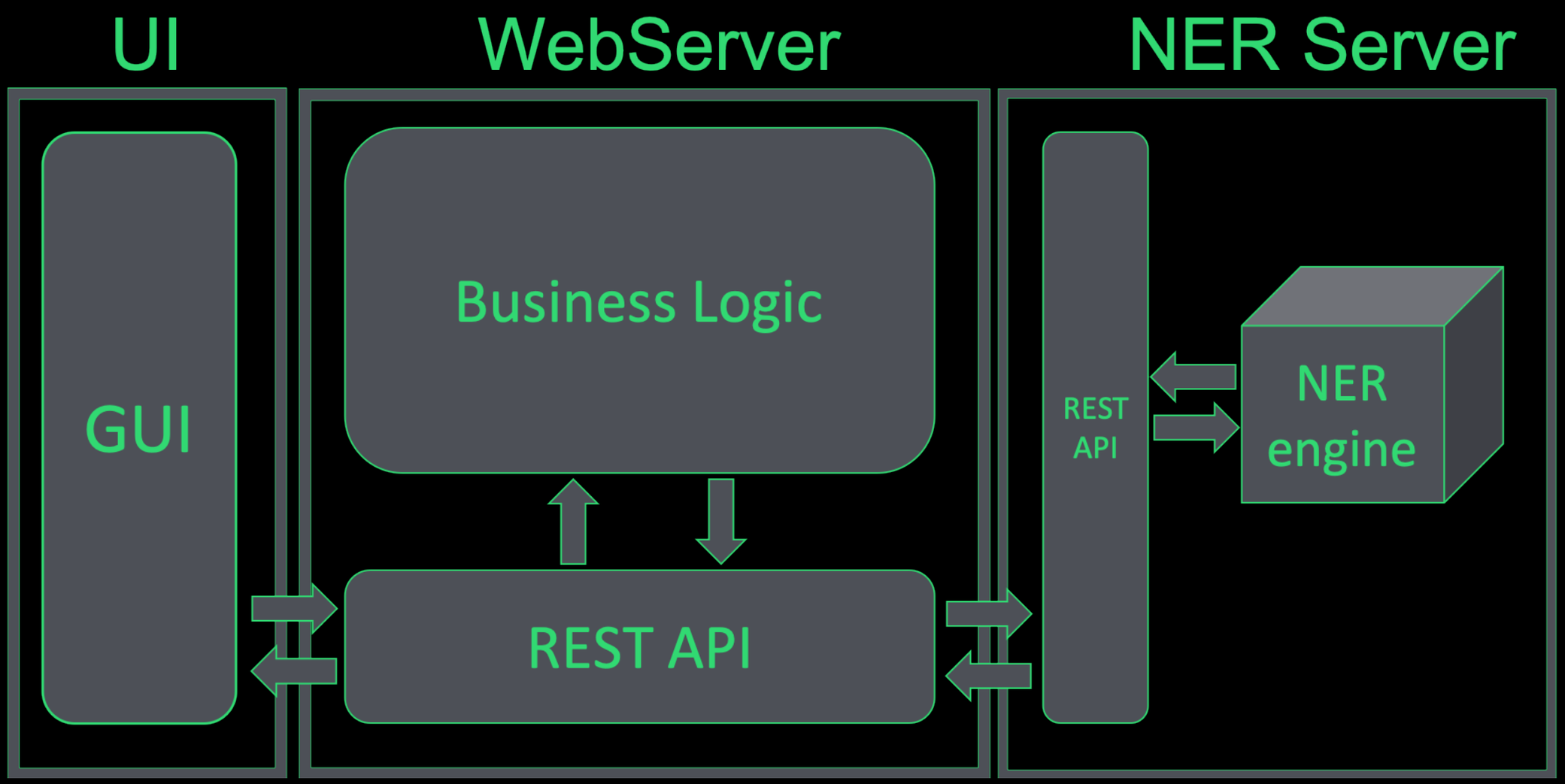
- Névelem detektálása szövegben (NER)
  - SpaCy
- Egy ráépülő szolgáltatás, ami az eredmények alapján plusz információkat keres
- Megfelelő grafikus felület a szolgáltatások használatához

SpaCy

- Egyik legjobb természetes nyelvfeldolgozó megoldás
- Python nyelven elérhető
- 3 féle modellt is támogat:
  - en_core_web_sm
    - 35 MB
  - en_core_web_md
    - 115 MB
  - en_core_web_lg
    - 812 MB

SYSTEM	YEAR	LANGUAGE	ACCURACY	SPEED (WPS)
spaCy v2.x	2017	Python / Cython	**92.6**	*n/a* ⓘ
spaCy v1.x	2015	Python / Cython	91.8	13,963
ClearNLP	2015	Java	91.7	10,271
CoreNLP	2015	Java	89.6	8,602
MATE	2015	Java	92.5	550
Turbo	2015	C++	92.4	349

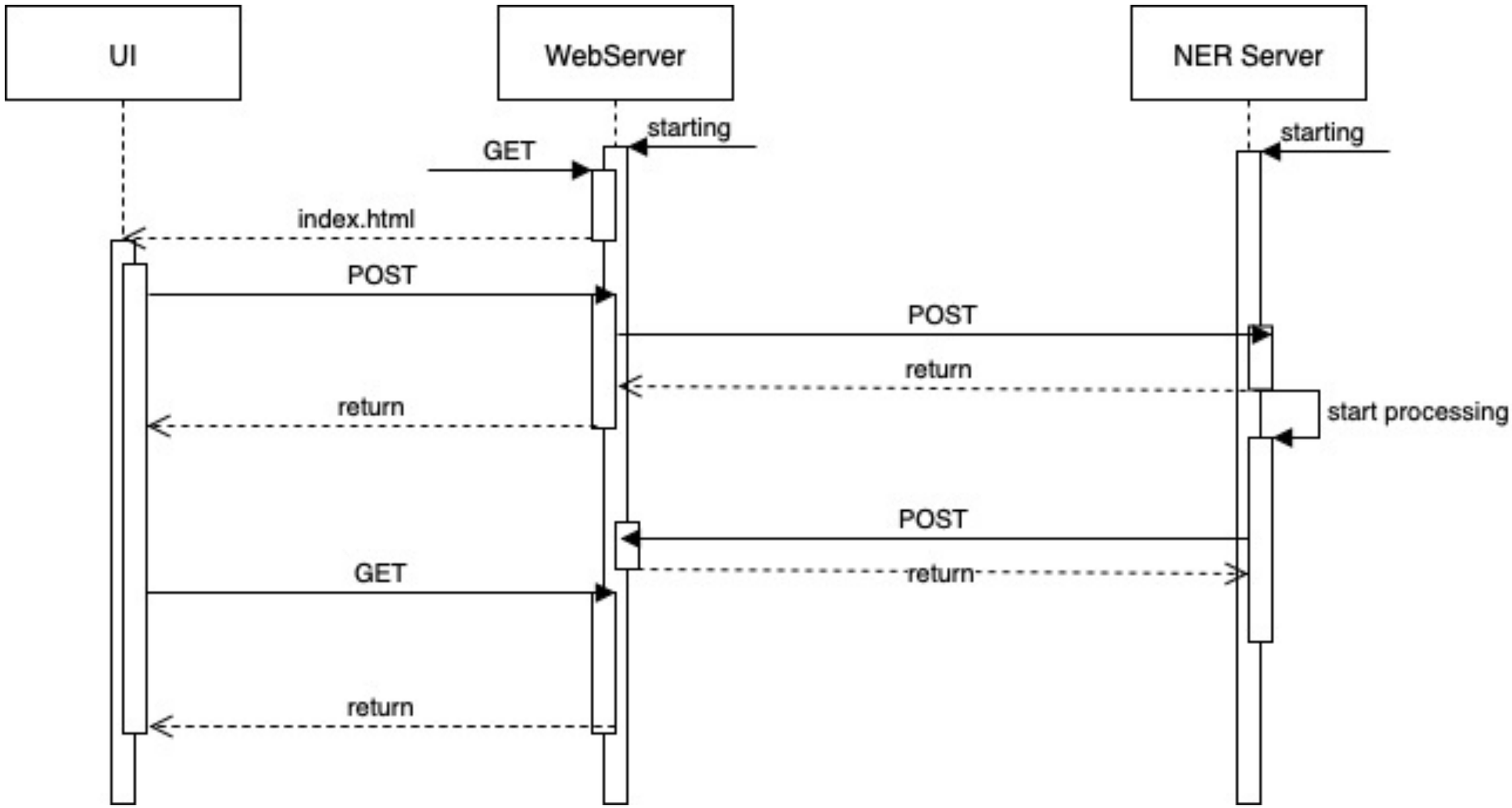
Architecture



Hatáskörök

- UI - grafikus megjelenítés
- NER Engine - Névelem felismerés
- NER Server Rest API - Kommunikációs réteg
- Buisness Logic - Szöveginformáció gazdagítás
- WebServer REST API - Kommunikációs réteg

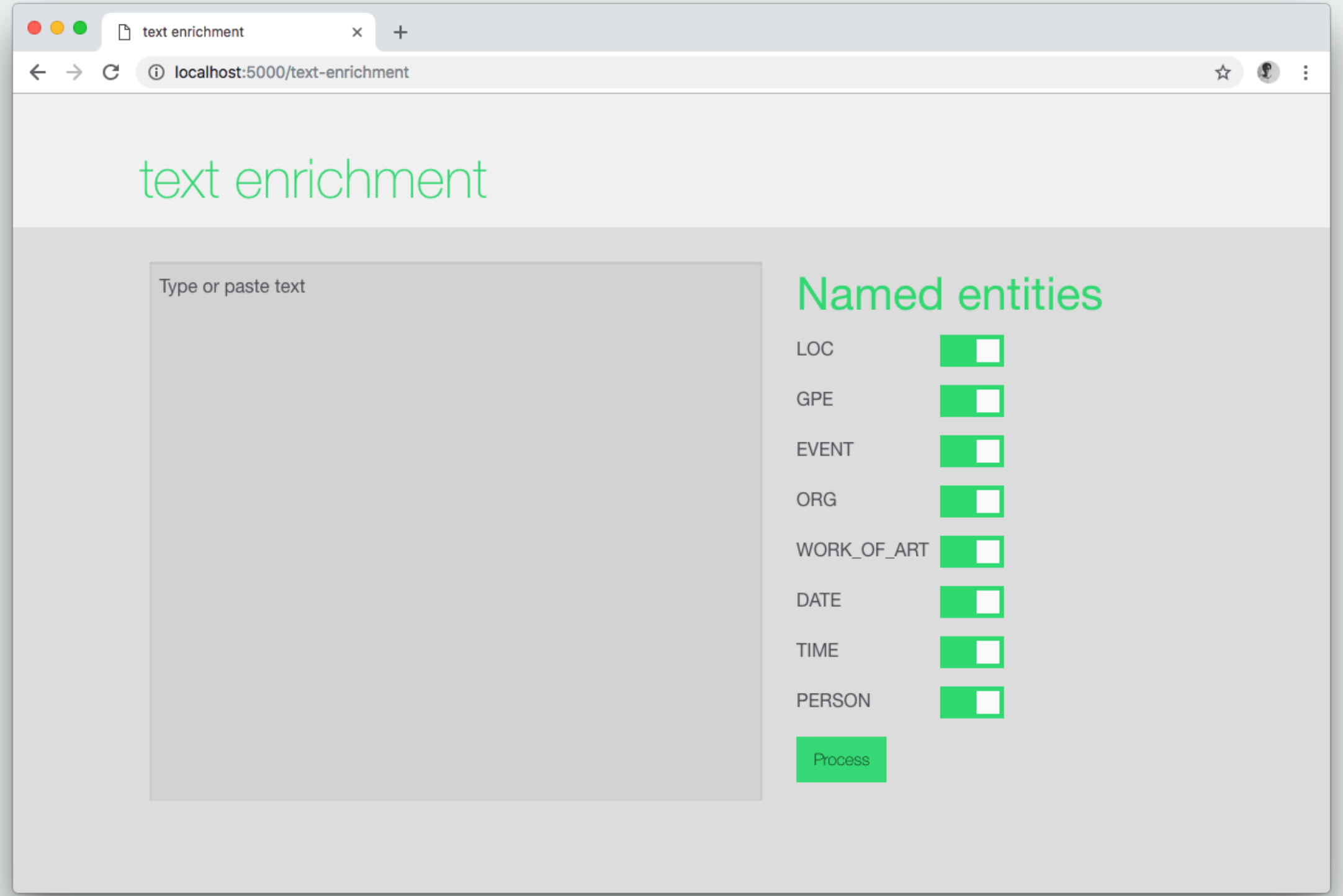
Kommunikáció

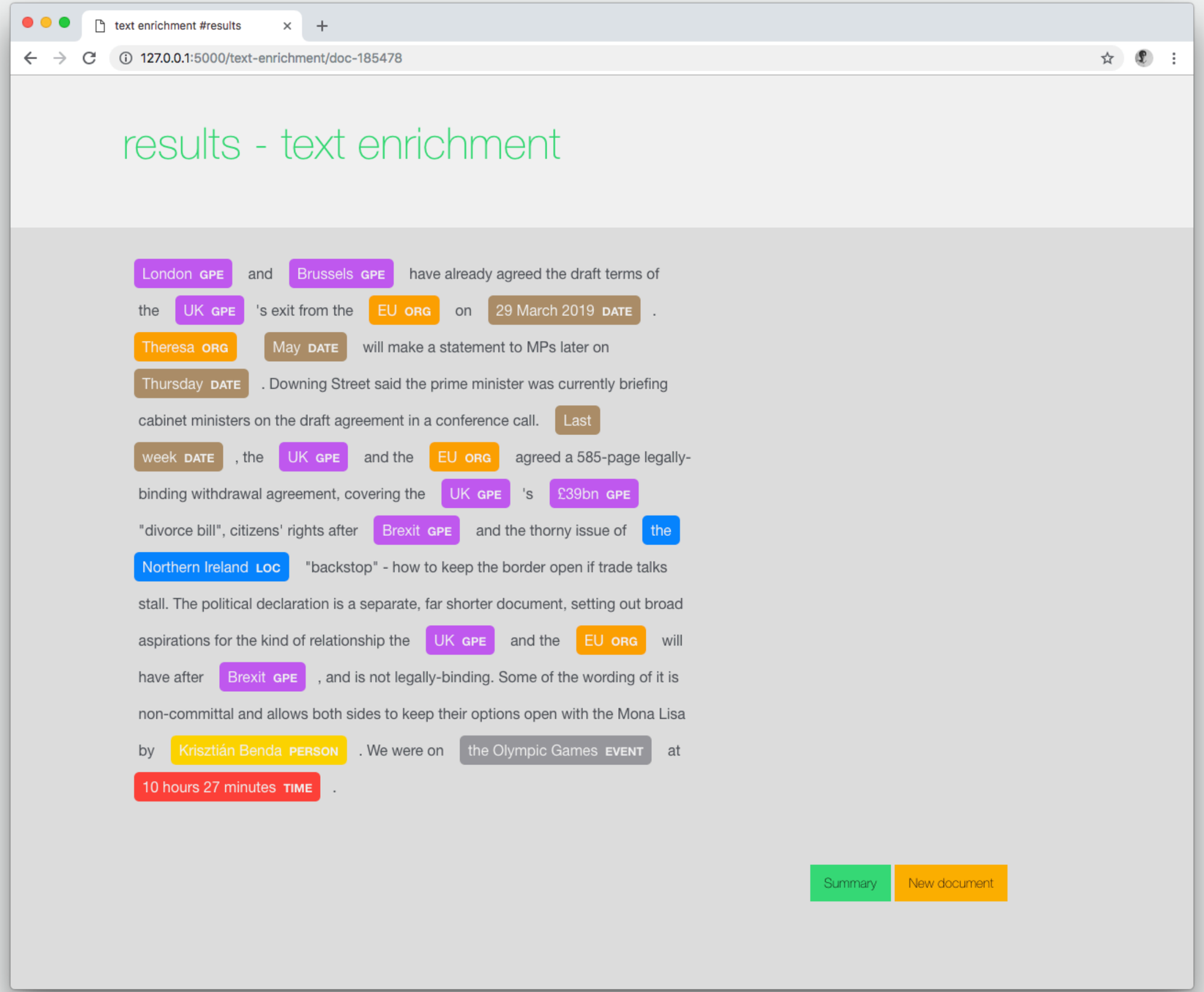


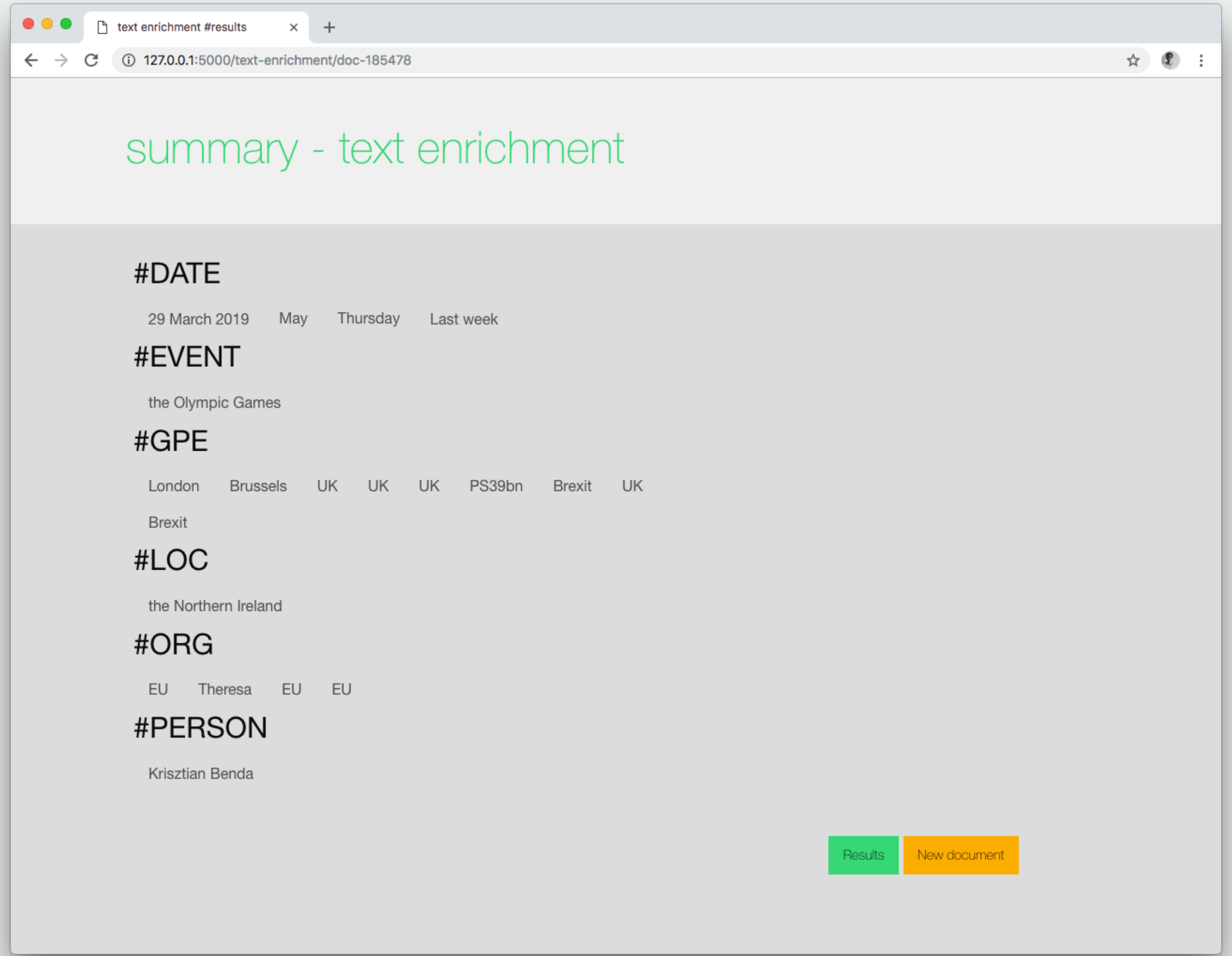
Miket ismerünk fel?

- Földrajzi helyszín -> Google Maps
- Geopolitikai entitások -> Google Maps
- Személyek -> Wikipédia, vagy Google Search
- Dátum, idő -> Google Calendar
- Művészeti alkotás -> Googla Image Search
- Szervezet -> Wikipédia, vagy Google Search
- Események -> Wikipédia, vagy Google Search

Megoldás


text enrichment
localhost:5000/text-enrichment
text enrichment
Type or paste text
Named entities
LOC
GPE
EVENT
ORG
WORK_OF_ART
DATE
TIME
PERSON
Process


text enrichment #results
127.0.0.1:5000/text-enrichment/doc-185478
results - text enrichment
London GPE and Brussels GPE have already agreed the draft terms of the UK GPE 's exit from the EU ORG on 29 March 2019 DATE . Theresa ORG May DATE will make a statement to MPs later on Thursday DATE . Downing Street said the prime minister was currently briefing cabinet ministers on the draft agreement in a conference call. Last week DATE , the UK GPE and the EU ORG agreed a 585-page legally-binding withdrawal agreement, covering the UK GPE 's £39bn GPE "divorce bill", citizens' rights after Brexit GPE and the thorny issue of the Northern Ireland LOC "backstop" - how to keep the border open if trade talks stall. The political declaration is a separate, far shorter document, setting out broad aspirations for the kind of relationship the UK GPE and the EU ORG will have after Brexit GPE , and is not legally-binding. Some of the wording of it is non-committal and allows both sides to keep their options open with the Mona Lisa by Krisztián Benda PERSON . We were on the Olympic Games EVENT at 10 hours 27 minutes TIME .
Summary
New document


text enrichment #results
127.0.0.1:5000/text-enrichment/doc-185478
summary - text enrichment
#DATE
29 March 2019
May
Thursday
Last week
#EVENT
the Olympic Games
#GPE
London
Brussels
UK
UK
UK
PS39bn
Brexit
UK
Brexit
#LOC
the Northern Ireland
#ORG
EU
Theresa
EU
EU
#PERSON
Krisztian Benda
Results
New document

Köszönjük a figyelmet!