Využití moderních technologií pro práci s obsahem webového archivu

Jan Lehečka, 21.10. 2019

Workshop výzkumného projektu
“Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů”

KATEDRA KYBERNETIKY

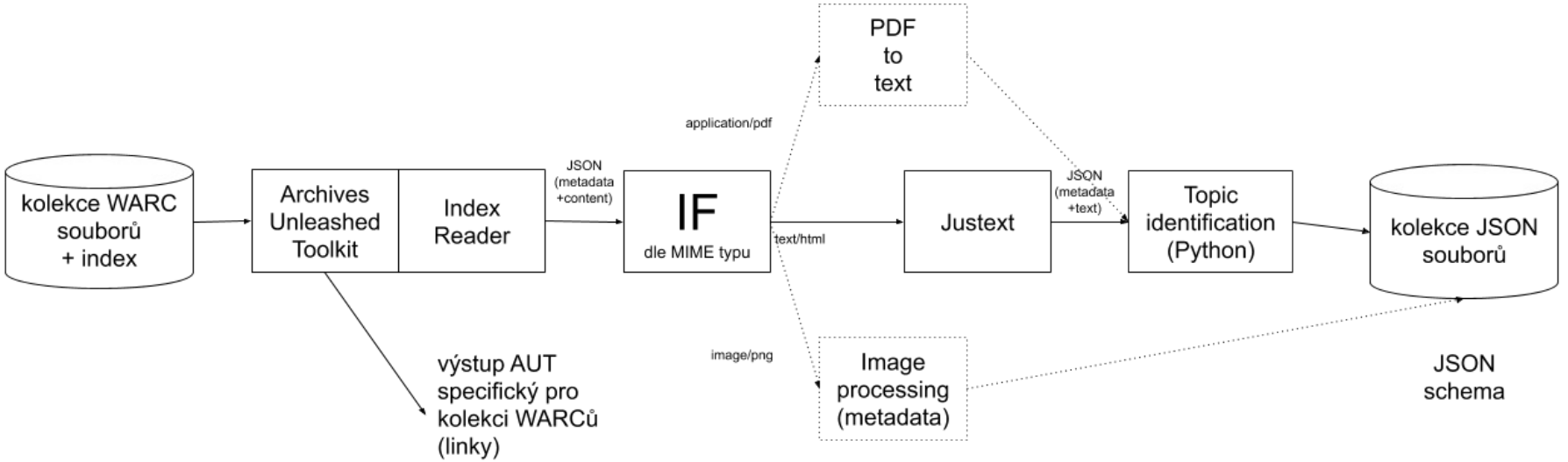
FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

LINDAT
CLARIAH-CZ

Osnova prezentace

- Schéma řešení a použité technologie
- Specifikace intermediary formátu
- Ukázka extrakce textu z webové stránky
- Ukázka identifikace tématu webové stránky

Schéma řešení a použité technologie

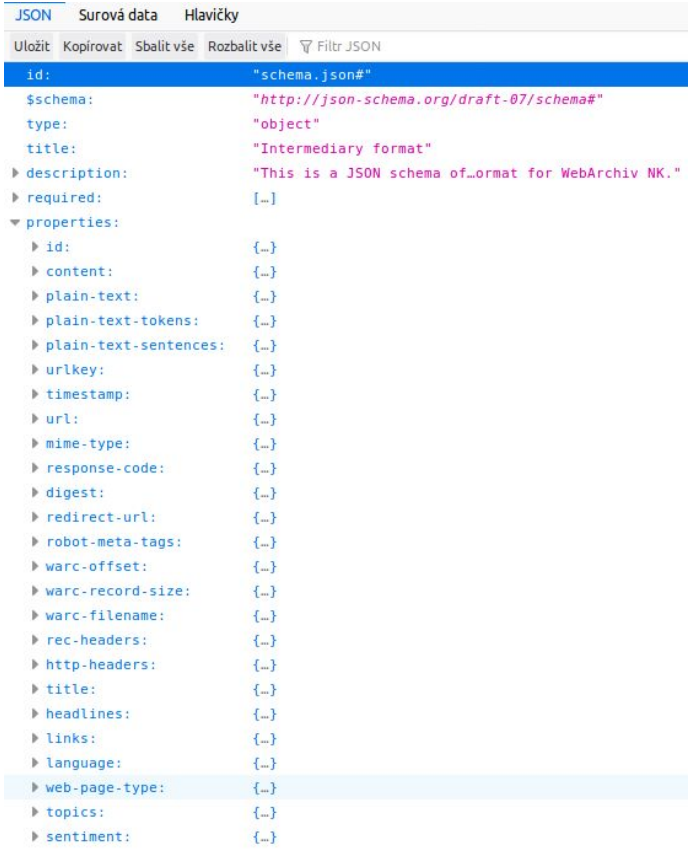


Webarchiv → Archives Unleashed → PySpark → python → scikit learn → APACHE HBASE

hadoop + APACHE Spark

Specifikace intermediary formátu

- *intermediary formát* = formát jednotlivých záznamů během zpracovávání webového archivu
- definuje položky pro surová stažená data a metadata
- specifikace definována pomocí JSON schéma
- každá webová stránka = 1 JSON objekt, do kterého postupně jednotlivé algoritmy ukládají získaná metadata



Ukázka extrakce textu z webové stránky

jusText

Neděle 20. října 2019, svátek má Vendelín
Vyzkoušet Premium zdarma
Premium
Přihlásit Můj účet
iDNES.cz / Sport

V semifinále MS se střetnou ragbisté Nového Zélandu a Anglie

Ostatní sporty
Dalších 8 fotografií v galerii
Australský ragbista Samu Kerevi (u míče) v duelu s Anglií. | foto: AP
říj 19 2019
Ragbisté Nového Zélandu deklasovali na mistrovství světa v Japonsku ve čtvrtfinále 46:14 Irsko a udělali další krok k třetímu titulu za sebou. Soupeřem All Blacks v semifinále bude příští sobotu Anglie, jež v sobotu porazila podobně jednoznačně 40:16 Austrálii.
Vstoupit do diskuse
Irsko sice z předchozích tří zápasů Nový Zéland dvakrát porazilo, v Tokiu ale vládli trojnásobní mistři světa. Položili sedm pětek, z toho první dvě Aaron Smith. Irové, kteří byli před turnajem světovými jedničkami a usilovali o první semifinále na MS v historii, vymazali nulu na kontě až v závěrečné desetiminutovce za rozhodnutého stavu.
Anglie se v souboji bývalých mistrů světa v Ótitě dostala do vedení 14:3 v prvním poločase díky dvěma pětkám Johnyho Maye během tří minut. Australané se sice krátce po pauze zásluhou pětky Mariky Koroibeteho přiblížili na rozdíl jediného bodu, na to ale okamžitě zareagoval pětkou Kyle Sinckler a Anglie už zbytku utkání dominovala.
Australané nepostoupili do semifinále teprve potřetí v devíti ročnících MS a ve všech případech podlehli Anglii. Ta přešla přes čtvrtfinále poprvé od finále v roce 2007.
V neděli určí druhou semifinálovou dvojici zápasy Wales - Francie a Japonsko - Jihoafrická republika.
Mistrovství světa v ragby v Japonsku
Čtvrtfinále
Anglie - Austrálie 40:16 Nový Zéland - Irsko 46:14
Autoři: ČTK, iDNES.cz
Související
Nic než zlato. Novozélanďané vědí, že jejich země jinou medaili neuznává
Japonci vzývají další zázrak. A co Irové? Fascinující podívaná začíná
Sport v roce 2019
20. 9. - 2. 11. MS v ragby (Japonsko)
26. 12. - 5. 1. MS hokejistů do 20 let (Ostrava a Třinec)
Témata: MS v ragby 2019 v Japonsku, Ragby
Vstoupit do diskuse
Sdílet na Facebooku

demo: https://nlp.fi.muni.cz/projects/justext/

Ukázka identifikace tématu webové stránky

LIDOVKY.cz

DOMOV SVĚT BYZNYS SPORT KULTURA CESTOVÁNÍ RELAX DESIGN DOBRÁ CHUŤ NÁZORY LIDÉ VIDEO SPECIÁLY DNEŠNÍ LN

Kurdové a Turecko se vzájemně obvinili z nedodržování příměří v Sýrii

Následuje článek: Johnson doufá, že hlasování přinese rozuzlení brexitu. Situaci ale komplikuje pozměňovací návrh

DAMAŠEK/ANKARA Turecko a syrští Kurdové se navzájem obvinili z nedodržování příměří na severu Sýrie. Bylo ujednáno ve čtvrtek a má platit 120 hodin. Kurdové v té době mají stáhnout své oddíly z území u turecké hranice. Turecký prezident Recep Tayyip Erdogan v pátek pohrozil, že turecká armáda ofenzívu v úterý večer obnoví, jestliže Kurdové zónu nevyklidí. Turecký ministr obrany v sobotu tuto hrozbu zopakoval.

V sobotu turecké ministerstvo obrany oznámilo, že turecká armáda dohodu vyjednanou ve čtvrtek v celém rozsahu plní. „Teroristé přesto v minulých 36 hodinách zaútočili čtrnáctkrát," uvádí se v prohlášení úřadu, který slovem teroristé označuje syrské kurdské oddíly YPG.

USA sjednaly s Tureckem příměří, protože nevěřily, že napadení Kurdové oblast udrží, tvrdí diplomat

Ministr obrany Hulusi Akar řekl, že Turecko přerušilo boje, aby se mohli kurdští bojovníci stáhnout. „Jejich zbraně budou zabaveny a jejich posty zničeny. Pokud se to nestane, budeme dále bojovat," řekl Akar.

Stanice Al-Džazira má informace, že do vyjednávání s Kurdy vstoupili syrští povstalci, kteří Turecko podporuje od začátku syrské války a kteří se turecké ofenzívy zúčastnili. Podle Al-Džaziry jsou nyní tito bojovníci u Rás al-Ajnu a snaží se přimět arabsko-

jusText

- Plain text -

Damašek/Ankara Turecko a syrští Kurdové se navzájem obvinili z nedodržování příměří na severu Sýrie. Bylo ujednáno ve čtvrtek a má platit 120 hodin. Kurdové v té době mají stáhnout své oddíly z území u turecké hranice. Turecký prezident Recep Tayyip Erdogan v pátek pohrozil, že turecká armáda ofenzívu v úterý večer obnoví, jestliže Kurdové zónu nevyklidí. Turecký ministr obrany v sobotu tuto hrozbu zopakoval.

V sobotu turecké ministerstvo obrany oznámilo, že turecká armáda dohodu vyjednanou ve čtvrtek v celém rozsahu plní. „Teroristé přesto v minulých 36 hodinách zaútočili čtrnáctkrát," uvádí se v prohlášení úřadu, který slovem teroristé označuje syrské kurdské oddíly YPG.

Ministr obrany Hulusi Akar řekl, že

Get labels

scikit learn

- Topics -

#	label	score
1	turecko	1.000
2	sýrie	0.970
3	nepokoje, konflikty a válka	0.881
4	diplomacie	0.426
5	zločin, zákon a spravedlnost	0.107
6	národní vláda	0.053
7	irák	0.028
8	ozbrojené síly	0.022
9	eu	0.020
10	místní úřad	0.012

Děkuji za pozornost