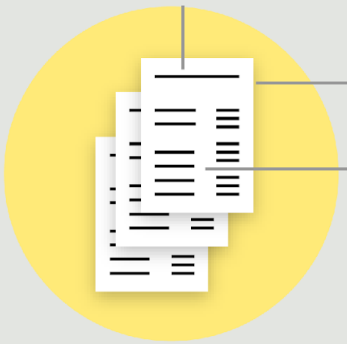
Detecting Text Reuse in Newspapers Data with Passim

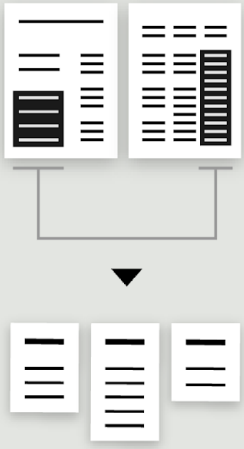
Hacking the News — DHN 2018
Helsinki 5-6 March

impresso Media Monitoring of the Past



facsimiles

extract articles

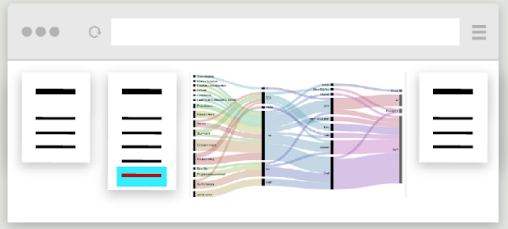


1 source selection

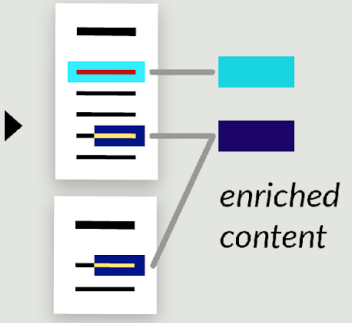
different archives
different languages

4 co-designed web interfaces

digital source criticism
exploratory visualisations
advanced search features



3 web services



2 data processing

extract text,
parse and annotate
articles

2.1 preprocessing

text formatting
OCR correction
metadata mngmt.

2.2 text mining

named entities
lexical analysis
word alignment
text similarity
topic modelling

2.3 annotation

Data curation

impresso
Media
Monitoring
of the Past

Text Reuse — working definition

Text reuse is the **meaningful reiteration of text**, usually beyond the simple repetition of common language.

Such a broad concept can naturally be understood at different levels and studied in a large variety of contexts.

http://dharchive.org/paper/DH2014/Panel-106.xml

Contexts:

- publishing/teaching → plagiarism
- literary studies → intertextuality (quotes, allusions, paraphrases)
- historical newspapers → text reprinting, copy+paste, information spreading

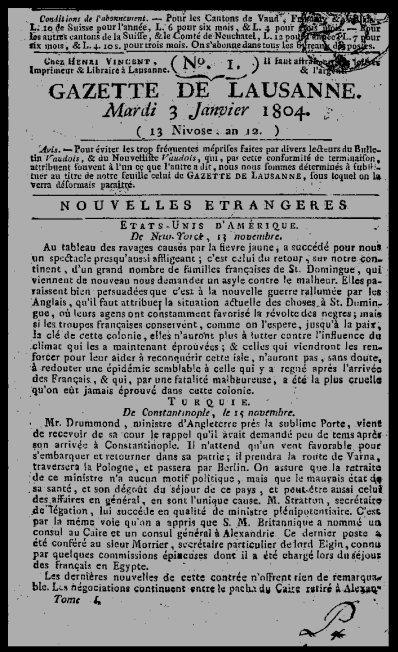
impresso
Media
Monitoring
of the Past

Text Reuse — Tools

Passim	D. Smith et al.	Java / Scala	https://www.etrap.eu/research/tracer/
Tracer	M. Büchler	Java	https://www.etrap.eu/research/tracer/
MatchMaker*	R. Snyder	Python	https://github.com/JSTOR-Labs/matchmaker
Tesserae*	N. Coffee et al.	Perl	https://github.com/tesserae/tesserae
…			

impresso
Media
Monitoring
of the Past

LeTemps newspaper(s)



GAZETTE DE LAUSANNE.
Mardi 3 Janvier 1804.
(13 Nivose, an 12.)

Avis. — Pour éviter les trop fréquentes méprises faites par divers lecteurs du Bulletin *Vaudois*, & du Nouvelliste *Vaudois*, qui, par cette conformité de termination, attribuent souvent à l'un ce que l'autre a dit, nous nous sommes déterminés à substituer au titre de notre feuille celui de GAZETTE DE LAUSANNE, sous lequel on la verra désormais paroître.

NOUVELLES ETRANGERES

ETATS-UNIS D'AMÉRIQUE.

De New-York, 13 novembre.

Au tableau des ravages causés par la fievre jaune, a succédé pour nous un spectacle presqu'aussi affligeant ; c'est celui du retour, sur notre continent, d'un grand nombre de familles françoises de St. Domingue, qui viennent de nouveau nous demander un asyle contre le malheur. Elles paraissent bien persuadées que c'est à la nouvelle guerre rallumée par les Anglais, qu'il faut attribuer la situation actuelle des choses à St. Domingue, où leurs agens ont constamment favorisé le révolte des negres ; mais si les troupes françoises conservent, comme on l'espere, jusqu'à la paix, la clé de cette colonie, elles n'auront plus à lutter contre l'influence du climat qui les a maintenant éprouvées ; & celles qui viendront les renforcer pour leur aider à reconquérir cette isle, n'auront pas, sans doute, à redouter une épidémie semblable à celle qui y a regné après l'arrivée des Français, & qui, par une fatalité malheureuse, a été la plus cruelle qu'on eût jamais éprouvé dans cette colonie.

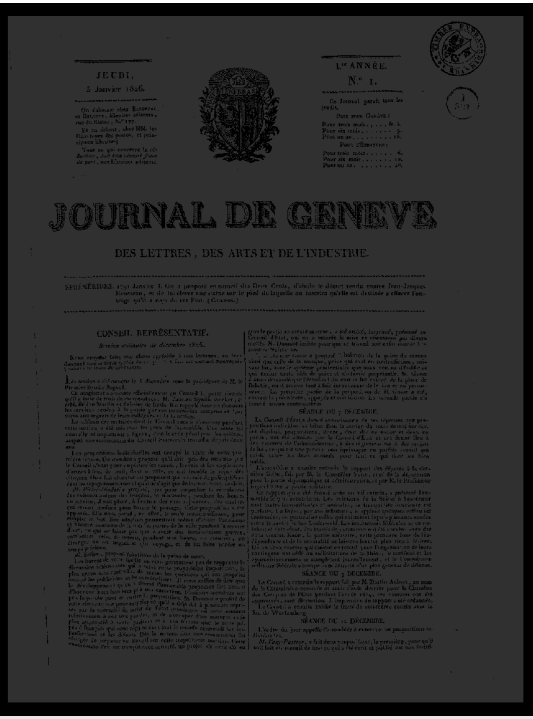
TURQUIE.

De Constantinople, le 14 novembre.

Mr. Drummond, ministre d'Angleterre près la sublime Porte, vient de recevoir de sa cour le rappel qu'il avait demandé peu de tems après son arrivée à Constantinople. Il n'attend qu'un vent favorable pour s'embarquer et retourner dans sa patrie ; il prendra la route de Varna, traversera la Pologne, et passera par Berlin. On assure que la retraite de ce ministre n'a aucun motif politique, mais que le mauvais état de sa santé, et son dégoût du séjour de ce pays, et peut-être aussi celui des affaires en général, en sont l'unique cause. M. Stratton, secrétaire de légation, lui succède en qualité de ministre plénipotentiaire. C'est par la même voie qu'on a appris que S. M. Britannique a nommé un consul au Caire et un consul général à Alexandrie. Ce dernier poste a été conféré au sieur Morrier, secrétaire particulier de lord Elgin, connu par quelques commissions épineuses dont il a été chargé lors du séjour des français en Egypte.

Les dernières nouvelles de cette contrée n'offrent rien de remarquable. Les négociations continuent entre le pacha du Caire retiré à Alexan

Tome 4.



JOURNAL DE GENEVE

DES LETTRES, DES ARTS ET DE L'INDUSTRIE.

CONSEIL REPRÉSENTATIF.

SÉANCE DU 2 DÉCEMBRE.

SÉANCE DU 3 DÉCEMBRE.

JOURNAL DE GENÈVE
et Gazette de Lausanne

LE QUOTIDIEN SUISSE D'AUDIENCE INTERNATIONALE

LE DERNIER NUMÉRO

Au revoir

Une journée pour en finir

SAMEDI LITTÉRAIRE

Les artistes célèbrent la fin

LA DER DES DER

Un cahier pour vous

Fin de partie

3 Jan 1804
La Gazette de Lausanne

5 Jan 1826
Journal de Genève

1991 Le Nouveau Quotidien; Fusion GDL and JDG

18 Mar 1998 Le Temps

2008 Digitization

1800 1900 2000

impresso
Media
Monitoring
of the Past

Passim

Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers

David A. Smith, Ryan Cordell, and Abby Mullen*

American Literary History, vol. 27, no. 3, pp. E1–E15
http://dx.doi.org/10.1093/alh/ajv029

Step 1: search of candidate document pairs

Step 2: local document alignment

Step 3: passage clustering

http://github.com/dasmiq/passim

impresso
Media
Monitoring
of the Past

Passim — Algorithm

Step 1: search of candidate document pairs

Goal: reduce total number of comparisons to perform

1.1 **shingling**:

- efficient document indexing via n-grams
- document → set of 5-word sequences (5-grams)
- singleton n-grams are filtered out (> 50%)

1.2 **extraction and filtering of candidate pairs**

- suppress repeated n-grams within same series
- suppress n-grams leading to > 5k document pairs
- filter out document pairs with < 5 shared n-grams

Step 2: local document alignment

Step 3: passage clustering

5-grams

devant le verdict de l'expert sur la responsabilité et réclame, en raison de la violence

1. devant le verdict de l'expert
2. le verdict de l'expert sur
3. verdict de l'expert sur la
4. de l'expert sur la responsabilité
5. …
6.

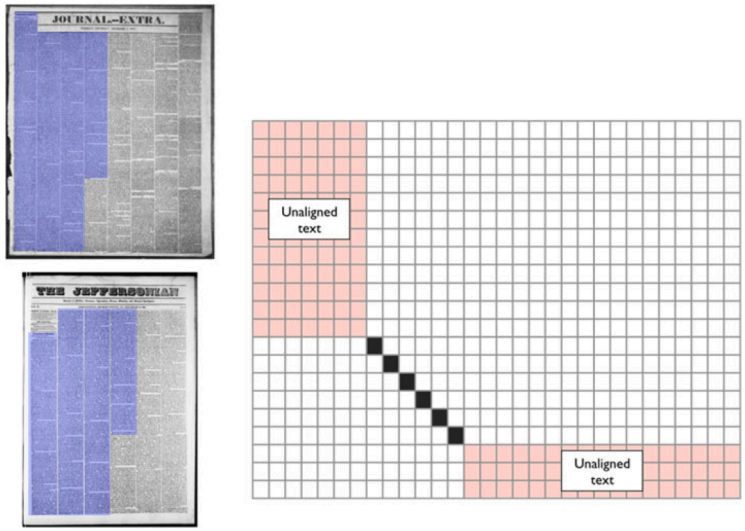
impresso
Media
Monitoring
of the Past

Passim — Algorithm

Step 1: search of candidate document pairs

Step 2: local document alignment

Goal: given a set of document pairs, obtain a set of aligned passage pairs



*source of illustration: Smith et al. 2015, Fig. 4, p. E9

Step 3: passage clustering

impresso
Media
Monitoring
of the Past

Passim — Algorithm

Step 1: search of candidate document pairs

Step 2: local document alignment

Step 3: passage clustering

Goal: group similar passages into clusters

- “single-link” clustering
- overlap threshold 50%

impresso
Media
Monitoring
of the Past

Passim — Experiment on GDL/JDG

Data Preparation

- XML → JSONlines format
- size of data:
    - 200 years
    - 15 Gb

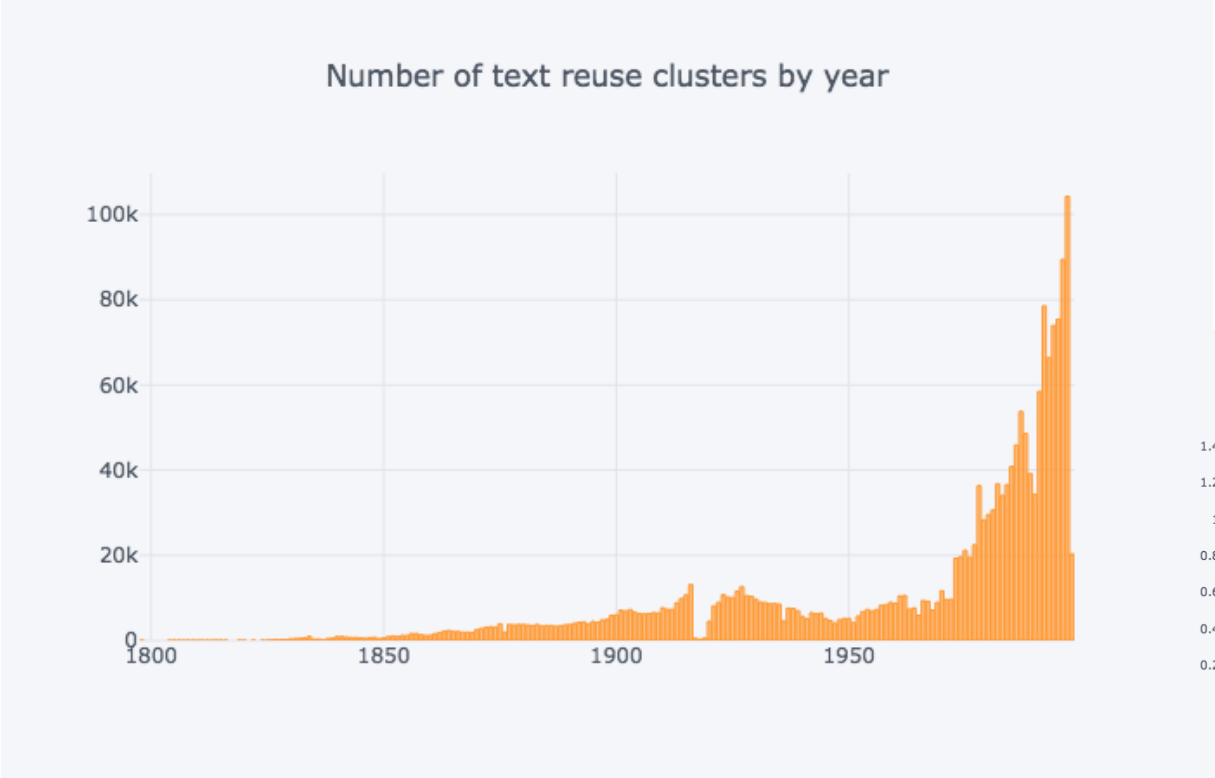
```
/GDL-1841-5.jsonl | jq "."
{
  "text": "<full_text>''100 LOTERIE D'ARGENT DE FRANCFORT SUR-LE-MEIN. Elle est divisée en 6 classes et compos
ée d'uncapital de Un Million 822,500 florins a\"Empire, au Louis d'or à 11 fl. 2600 billets, dont 13,500 lots
elS 4 primes, outreun grand nombre de billets franc » Les principaux lots sont : 2 lots de fl. 100,000 chacun.
 1 lot de 50,000. 2 lots de 25,000 chacun. 2 lots de 20,000 chacun. Ilot de 15,000. 1 lot de 12,000. 4 lots de
 10,000 chacun. 1 lot de 6,000. 5 lots de 5,000 chacun. 98 lots de 4000 à 1000.94 lots de 600 à 300.5880 lots
de 250 à 100 florins, non compris un grand nombre de moindres gains. Cette loterie est établie et garantie par
 le gouvernement de la ville libre de Francfortbasée sur les principes les plus loyaux et avantageux pour les
joueurs, elle jouit d'une confiance et d'un crédit général et bien mérité. Le plan, offert gratis aux amateurs
, en contient le bilan exact et les conditions ultérieures ; la' première classe sera tirée les 9 et 10 Juin p
rochain ; un billet entier coûte 6 fl ., un demi 3 fl ., un tiers 2 fl ., un quart 1 fl. \" 50, payables compt
ant. Le soussigné offre son entremise aux personnes qui voudront s'y intéresser ; il désignera dans le but de
faciliter les paiemens des mises, aux persoonnes à qui cela pourra convenir une maison de commerce, se trouvan
t le plus à leur portée. J .-B. ZUNDEL i à Schaffouse.</full_text>",
  "page_no": [
    "5"
  ],
  "name": "Untitled Article",
  "date": 1841,
  "series": "GDL",
  "id": "GDL-1841-05-21-a_Ar00504"
}
```

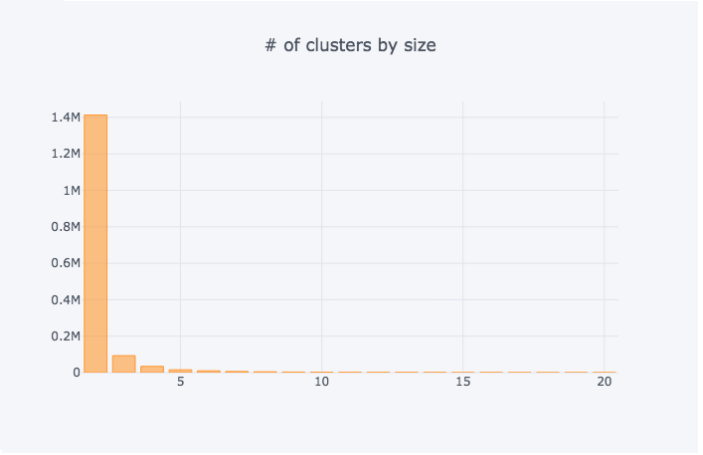
Technical Setup

- one cluster machine (node)
    - 48 cores, ~280Gb RAM
- run with default parameters
    - used 36 cores
    - **3.5 hours** to complete
    - given 100Gb to spark executor/driver

impresso
Media
Monitoring
of the Past

Passim – Output





impresso
Media
Monitoring
of the Past

Passim — Output examples

GDL 03/12/1863

— Mardi il est arrivé un accident, sans suites fâcheuses, au bateau à vapeur parti d'Ouchy pour Genève à 2 heures 20 minutes. Le bateau était arrêté pour le débarquement el l'embarquement devant Nyon, lorsque tout à coup on entendit une détonation, un nuage de fumée (ou de vapeur ?) sor tit de la machine et le bâtiment subit une violente secousse. Nous ne savons pas d'une manière précise en quoi consiste la rupture qu'il y a eu ; on parle d'une clavette cassée. Quoi qu'il en soit, le bateau était mis dans l'impossibilité de continuer sa route ; heureusement les voyageurs qui devaient aller plus loin ont pu prendre le train qui passe à Nyon à 5 heures 3 minutes el sont arrivés à bon port, sans autre mal qu'un moment de frayeur. On

JDG 05/12/1863

— Mardi il est arrivé un accident, sans suites fâcheuses, au Guillaume-Tilt, parti d'Ouchy pour Genève a 2 heures 20 minutes. Le bateau était arrêt é Pour le débarquement devant Nyou, lorsque tout à coup on entendit une détonation, un nuage de fu- n) p e (ou de vapeur ?) sortit de la hiadiine, et le hû liment subit une violente secousse. Nous ne savons pas d'une manière précise en quoi consiste la rup'" re qu'il y a eu ; on parle d'une clavette cassée. Quoi qu'il en soii, | e bateau était misdans l'impostJuilite de continuer sa route ; h'ureuiipnient les v < y 'geursqui devaient aller plus l <> iri ont pu piendre le train qui passe à Nyon à 5 heures 3 minutes, et sont arrivés ù bon poil, sans aulre mal qu'un moment de frayeur. Le

impresso
Media
Monitoring
of the Past

Passim — Output examples

GDL 26/05/1900

un journal a prétendu que des documents relatifs à cette affaire avaient été détournés en vue d'un renouvellement de l'agitation 'dont nous avons tant souffert. Le fait est-il vrai ? Le gouvernement, s'il est vrai, a-til pris des mesures ou entend-il en prendre pour prévenir toute émotion nouvelle ? (Applaudissements.) Le général DE GALLIPPET monte à la tribune au milieu d'une grande attention et dit : L'autre jour, à la Chambre, répondant à M. Alphonse Humbert, j'ai dit que je ne connaissais pas le » documents dont on a parlé, que ces documents n'existaient pas. A ce moment j'avais le droit de tenir un paroil langage. Il était strictement conforme à la vérité. **J'ai le regret de dire aujourd'hui que je me suis trompé. Le lendemain du jour où je m'exprimais ainsi, j'avais un entretien avec le chef d'étatmajor général, et, là, j'avais la douleur d'apprendre non seulement que les documents existaient, mais qu'ils avaient été divulgués par un officier du ministère de la guerre.** (Vive

JDG 26/05/1900

d'un journal est exact, prétendant que des documents relatifs à /' « Affaire » ont été détournés en vue du renouvellement de cette affaire ? — t e général do Galliffet répond qu'il a lo regret de devoir dire qu'il s'est trompé mardi en disant à la Chambre que ces documents n'existaient pas. **Il ignorait alors leur existence, mais le lendemain il eut la douleur d'apprendre, dans un entretien avec le chef de l'état-major, non seulement que ces documents existaient, mais qu'ils avaient été divulgués par un officier du ministère de la guerre.** L'officier

impresso
Media
Monitoring
of the Past

Passim — Output examples

GDL 28/04/1980

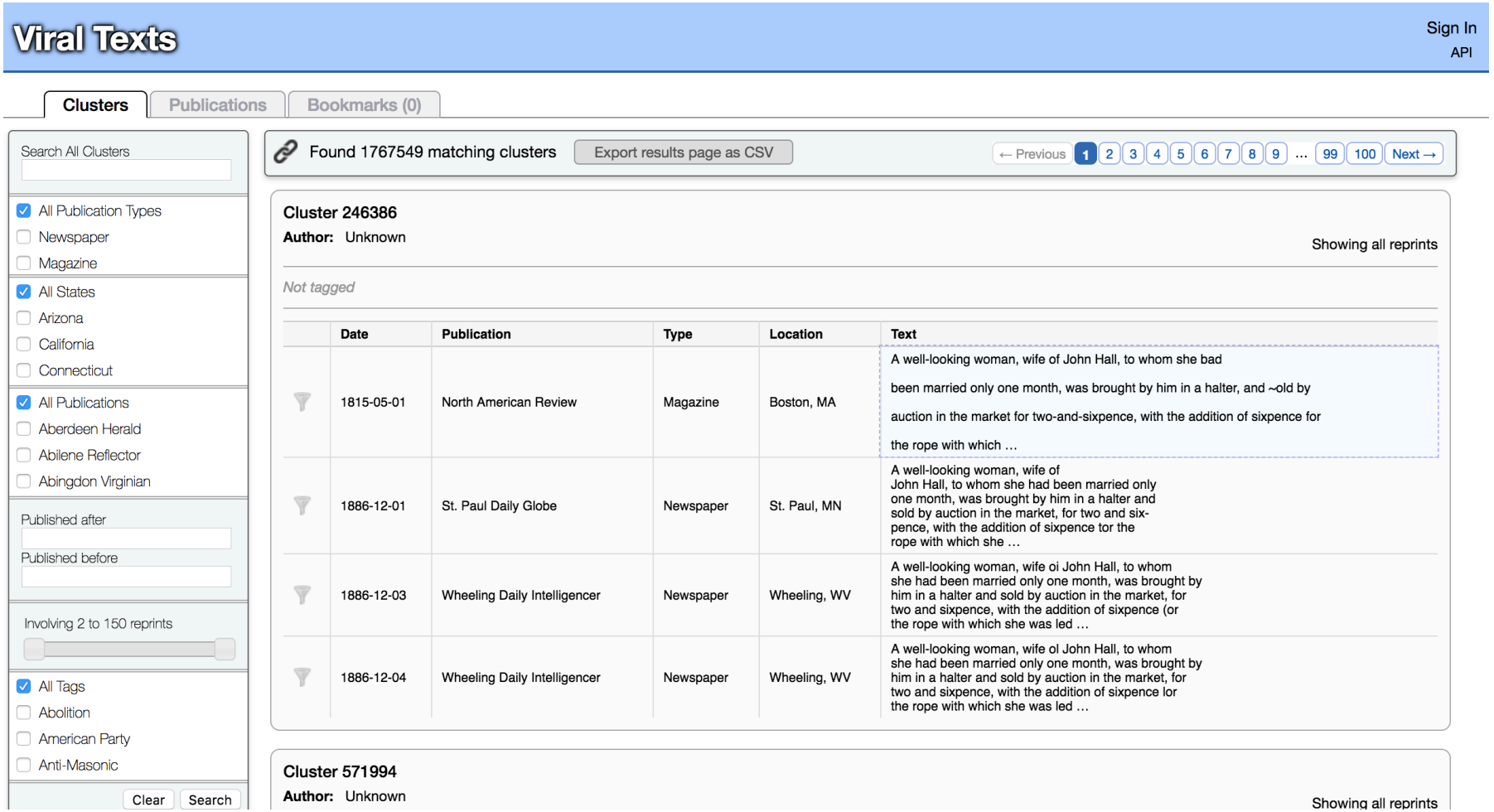
Faux chèque et faux cheval Lausanne, 27 (ATS).-On a confirmé dimanche soir à l'ATS, de source autorisée, une affaire d'escroquerie d'environ trois millions de dollars qui s'est passée en automne dernier à Lausanne et que le quotidien genevois « La Suisse » a révélée dimanche. Signé au siège lausannois d'une banque suisse, un contrat de vente portait sur un cheval de bronze considéré par le vendeur comme une antiquité grecque absolument unique. Après avoir acquis cette pièce rare d'un antiquaire suisse, ce vendeur, un Italien associé à une société ayant son siège à Panama, l'avait revendue à un riche arabe. Mais, peu après la signature du contrat, la banque fut informée que le chèque-trois millions de dollars-était faux. Par la suite, on devait apprendre que le cheval de bronze était faux, lui aussi. Les vendeurs auraient eu le temps de toucher près d'un million de dollars à New York, mais les deux autres millions, destinés à une suite d'intermédiaires, auraient pu être bloqués. La justice vaudoise tente depuis cinq mois de débrouiller cette affaire particulièrement compliquée, qui aurait des ramifications en Italie, au Koweït, en Suisse, à Londres et à New York, écrit encore « La Suisse ». Plusieurs personnes ont été appréhendées et l'une d'elles est toujours détenue à Lausanne, a-t-on appris dimanche soir

JDG 28/04/1980

Faux chèque et faux cheval GROSSE ESCROQUERIE À LAUSANNE Lausanne, 27 (ATS).-On a confirmé dimanche soir à l'ATS, de source autorisée, une affaire d'escroquerie d'environ trois millions de dollars qui s'est passée en automne dernier à Lausanne et que le quotidien genevois « La Suisse » a révélée dimanche. Signé au siège lausannois d'une banque suisse, un contrat de vente portait sur un cheval de bronze considéré par le vendeur comme une antiquité grecque absolument unique. Après avoir acquis cette pièce rare d'un antiquaire suisse, ce vendeur, un Italien associé à une société ayant son siège à Panama, l'avait revendue à un riche arabe. Mais, peu après la signature du contrat, la banque fut informée que le chèque-trois millions de dollars-était faux. Par la suite, on devait apprendre que le cheval de bronze était faux, lui aussi. Les vendeurs auraient eu le temps de toucher près d'un million de dollars à New York, mais les deux autres millions, destinés à une suite d'intermédiaires, auraient pu être bloqués. La justice vaudoise tente depuis cinq mois de débrouiller cette affaire particulièrement compliquée, qui aurait des ramifications en Italie, au Koweit, en Suisse, à Londres et à New York, écrit encore « La Suisse ». Plusieurs personnes ont été appréhendées et rune d'elles est toujours détenue à Lausanne, a-t-on appris dimanche soir

impresso
Media
Monitoring
of the Past

Viral Texts — Cluster explorer



http://viraltexts.northeastern.edu/clusters

impresso
Media
Monitoring
of the Past

Hacking on Text Reuse

Data:

- **~1.6M** text reuse clusters
- **~1.7M** articles from GDL and JDG (1798 → 1998)
- data formats: parquet, CSV, JSON
- IIIF endpoint with page images

Hack ideas:

- cluster explorer:
  - filters: size, date
  - user tagging/classification
  - full text search
  - visualization? (e.g. visual `diff`)
- re-run on more French newspapers?

impresso
Media
Monitoring
of the Past

Thanks for your attention

… and happy hacking!

https://twitter.com/ImpressoProject

https://twitter.com/56k