Néonaute: mining web archives for linguistic analysis

Sara Aubry, Bibliothèque nationale de France
Emmanuel Cartier, LIPN, University of Paris 13
Peter Stirling, Bibliothèque nationale de France

IIPC Web Archiving Conference
Wellington, 15th November 2018

twitter.com/DLWebBnF

Aims and objectives

- Creation of a search engine prototype (Néonaute) with advanced functionalities:
  - linguistic analysis and indexing of web documents : morphological analysis, Named Entities (NE) detection, topic detection
  - full-text queries, advanced queries (Apache Solr), facet handling linked to indexed metadata
  - multidimensional interactive exploration of results (timeline of word occurrences, cross-filtering of metadata distribution, contexts grouping)
- Two access modes:
  - full access to content inside the BnF library, using the Archives de l'internet Labs interface
  - online access to metadata and textual fragments for the lexical units covered by the project (about 30 000 lexical units)

Aims and objectives

- Project led by two linguistic research laboratories (LIPN, LILPA), studying three use cases:
  - life-cycle tracking of about 30 000 neologisms (Logoscope and Néoveille),
  - comparative life-cycle tracking of French Government Commission Recommended Terms (versus mainly anglicisms),
  - life-cycle tracking of feminized terms
- Based on BnF digital legal deposit collections
  - News sites 2010-2017: c. 100 sites crawled daily
  - Homepage plus all articles linked from it (1 click)
- For the BnF: allow new uses of web archive collections and propose full-text searching on the collection of news sites

Legal context and framework

- Access is controlled under legal deposit, intellectual property and data protection legislation
  - accessible onsite in BnF research library reading rooms and in a regional library network
  - users can search/view/cite but not download documents
- Aim to allow analysis of web archive collections while respecting the relevant legislation
- Signature of a research agreement by the BnF and partner institutions in the project
  - List the data and metadata to which researchers have access
  - Conditions of use, both of data onsite and exported metadata and results
  - Define organisational aspects and responsibilities of all parties

Organisational questions

- Research engineer based almost full-time at the BnF
  - Other meetings with research team and project sponsors as needed
- Meetings and exchanges with BnF staff
  - Content curators: collections scope and content
  - Crawl operators: how the collections are built
  - Metadata and format specialists: how the data is described and stored
  - Technical support: how the data can be accessed and parsed
- Use of agile methodology and specifically Scrum project management (also used for IT projects at BnF)
  - Shared monthly sprints with daily or weekly checkpoints
  - Initial planning and review at the end

Shaping the news collection

- BnF web archive: 31 billion URLs, 5.4 million W/ARCs, 965TB
- News sites 2010-2017: 1 billion URLs, 63 000 W/ARCs, 13TB
  - Use collection building procedures
- Define representative subsets to smooth out processes: a week, a month (1%), a month per year (10%)
- Identify relevant documents for different purposes:
  - BnF: document and give comprehensive access to the collection
  - Research team: narrow it to a research corpus

Full-text indexing the news collection

- Tools:
  - webarchive-discovery 3.0.0-SNAPSHOT / WarcIndexer component to process the W/ARCs
  - Apache Solr 7.2.1 for full-text index (and search)
  - Netsearch Archon/Arktika to pilot and monitor indexing processes
- Infrastructure:
  - 1 Lenovo Systems x3650 Intel Xeon 2.5Ghz x 12 cores, 256 GB RAM, 4 TB SSD
- Challenge: define a comprehensive and relevent index modele (schema) with storage concerns
  - Focus on text and metadata, give up on images and links

{BnF


```
"id":"eWiQcz7skBiNVRq90yNP6A==/20171127090246",
"wayback_date":20171127090246,
"title":"Les dna.fr - Actu du jour - Région - Sports - France-Monde - Faits-divers - Strasbourg",
"content_first_bytes":"3c 21 44 4f 43 54 59 50 45 20 68 74 6d 6c 3e 0d 0a 3c 68 74 6d 6c 20 6c 61 6e 67 3d 22 66 72 22",
"ssdeep_hash_bs_192":"68pABSvJ6dAhAMTAHXl1MxY/PTAg9Y7FLoJbNuleeHOkUwr7ChWACeZ5dWkbmHoS",
"ssdeep_hash_bs_384":"68p0SvFmXHMxYXd9CVpTrc55dSzduSH",
"_version_":1594344178372837381,
"content":"Le Journal Pour Sortir Avis de décès Météo sur 5 jours Contactez-nous > Je m'abonne Menu les Dernières Nouvelles
"status_code":200,
"content_type_norm":"html",
"type":"Web Page",
"source_file":"BnF-25002-28-20171127090218-00000-ciblee_2017_fogg128.bnf.fr.warc.gz",
"content_language":"fr",
"index_time":"2018-03-08T04:58:14.412Z",
"content_type":"text/html",
"collections":["actualités"],
"host":"dna.fr",
"public_suffix":"fr",
"content_length":217633,
"source_file_offset":7392377,
"url_path":"/",
"crawl_year":2017,
"url_type":"slashpage",
"content_encoding":"UTF-8",
"crawl_date":"2017-11-27T09:02:46Z",
"content_text_length":10846,
"collection":["collecte ciblée"],
"url":"http://www.dna.fr/",
"record_type":"response",
"domain":"dna.fr",
"hash":"sha1:V3UQ3GQAMTTGOJEVHDJXIFQXT4M5IMXD",
"content_type_tika":"text/html; charset=UTF-8"},
```


About 30 fields related to:

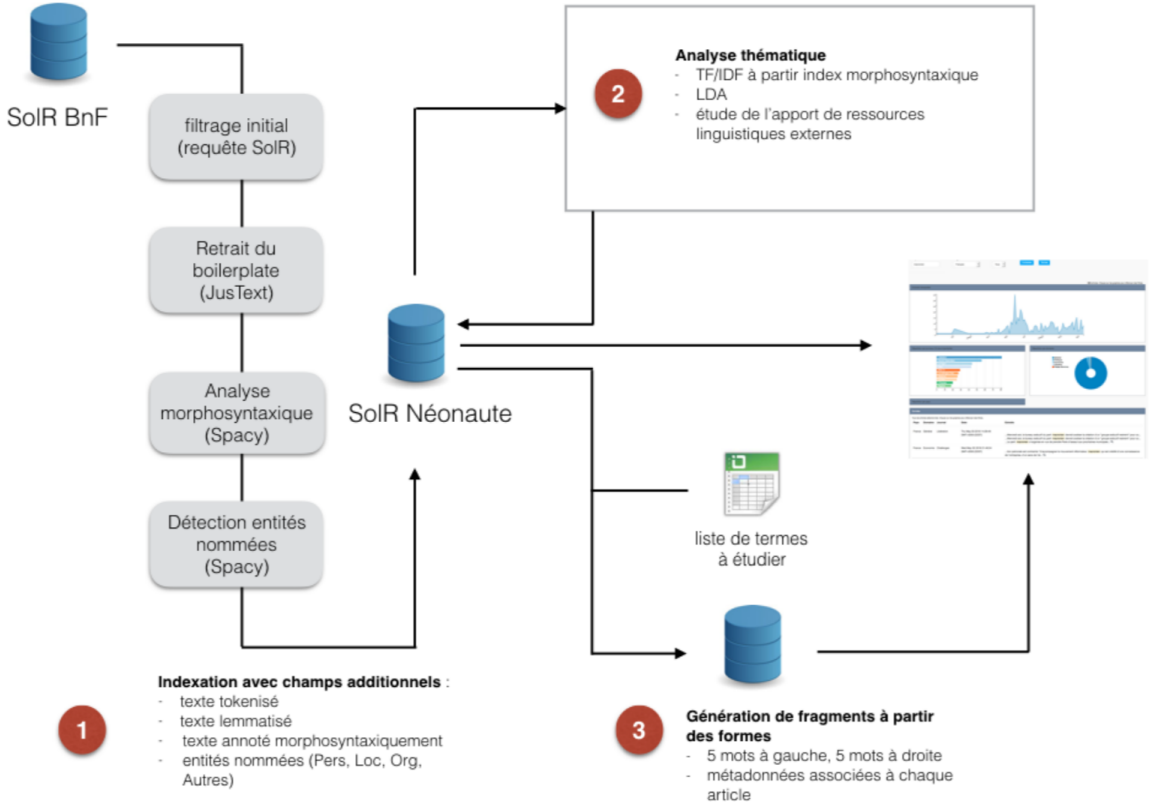
- extracted content (content, title,…)
- content analysis (content_text_length, content_language,…)
- URL analysis (domain, url_type,…)
- format (content_type_tika, content_encoding,…)
- date (crawl_date, crawl_year,…)
- other technical informations

Index: 314.5 mio URLs, 1.03 TB, 2 segments, 5 days

Giving access to the news collection

- Search applications:
  - AILABS 2.1
  - 2 Apache Solr
- Browse and display via OpenWayback
- Infrastructure:
  - 4 Lenovo Systems x3650 Intel Xeon 2,4Ghz x 10 cores, 32 GB RAM, 1.6 TB SSD

Néonaute architecture



- Identify relevant documents
- Define and apply lingustic analysis processes

Filtering documents

- Narrow the collection to a corpus of relevant documents
- Objective: keep « content pages » (homepages and articles), exclude scripts, images, legal information, etc.
- Solution: use a Solr query:
  - content_text_length > 1
  - content_language:fr
  - content_type_norm:(html OR pdf)
  - domain: (list of domains)
- Remove duplicates by grouping URLs and selecting the first occurrence
- Result: reduction to 10% of the whole collection

Boilerplate removal

- Objective: keep the main textual contents from web pages (remove navigation links, headers, footers, side-information)
- Solutions:
  - Get read-only access to the W/ARCs
  - Retrieve HTML code directly from the W/ARCs (and not the index)
  - Use of Justext as a boilerplate removal tool
  - Reject empty documents after processing

{BnF Page archivée le : 26 décembre 2010 à 11:05 GMT
Permalien : http://pfvarchivesinternet.bnf.fr/20101226110509/http://www.di
Accueil | Recherche initiale
Capture 1 sur 1

le 25/12/2010 18:38

Météo

Après la neige, froid glacial ce dimanche

Le temps de dimanche sera ensoleillé mais glacial, selon le bulletin émis par Météo France, qui a levé la vigilance orange pour la neige. Avec un ciel dégagé et un sol enneigé, les températures deviendront glaciales dans la nuit avec -9 à -13 degrés.

En Alsace, les températures pourraient atteindre demain matin les -15 degrés selon les prévisions. (Photo

Le mercure descendra localement à -15 degrés du nord-est à la Bourgogne et à la Franche-Comté, -4 à -8 degrés en général ailleurs et près de 0 à la côte.
En Alsace
, les températures pourraient atteindre demain matin les -15 degrés selon les prévisions.

Il ne dégèlera pas en journée sur un grand quart nord-est, des Ardennes et du nord-est à l'est du Centre à l'Auvergne et Rhône-Alpes avec -1 à -6 degrés et avoisineront 0 à 4 degrés ailleurs avec 4 à 6 sur le Sud-est, 7 à 12 de la Côte d'Azur à la Corse.

Côté ciel, de la Corse à la Provence et aux Alpes, puis de la Provence à la Corse l'après-midi, le temps sera encore instable et localement orageux avec de la neige vers 1.000 mètres mais également temporairement jusqu'à la côte vers la mi-journée. Plus au nord, sur les Savoie, le temps s'éclaircira. Mistral et tramontane, encore forts près du golfe du Lion et le long du Rhône avec des rafales proches de 90 à 100 km/h le matin, dégageront le ciel.

Des frontières du nord et de la baie de Seine au Centre, les plaques de nuages bas, doublées de nombreux brouillards givrants le matin, lâcheront parfois des flocons. Elles se morcelleront lentement au fil de la journée, laissant filtrer le soleil par endroits.

Partout ailleurs, un soleil d'hiver brillera dans une ambiance glaciale, une fois les bancs de brouillards givrants dissipés, avec un vent de sud qui se renforcera l'après-midi sur le Finistère, avec des rafales à 70 km/h.

ARTICLES LES PLUS...
LUS | COMMENTÉS
30 cm de neige et de nombreux arbres
Fin de la vigilance orange en Alsace
Après la neige, froid glacial ce dimanche
Yenga va signer
Décidément maudit...

DERNIERS COMMENTAIRES
RE: Cote d'Ivoire
Ici il ne s'agit que de faire respecter la volonté du peuple qui a élu Alassane Ouattara comme président et qui a été reconnu...
styles | 26/12/2010 | 11:20
RE: Original
Vous voulez qu'il les approuve,ou quoi? Il est bien dans son role,en disant cela.A Pie XII,on reprochait de n'avoir rien dit...
toni | 26/12/2010 | 10:11
hiver
Comme je l'ai deja dit,il ne manque plus que la taxe carbone à payer pour cause de réchauffement climatique
toni | 26/12/2010 | 10:04
Cote d'Ivoire
Jacques.halon a parfaitement raison,comme je l'ai deja dit,nous avons eu notre revolution,les EU leur guerre de

PUBLICITÉ

A LIRE AUSSI
25/12/2010 16:12
Alsace / L'alerte orange levée samedi en début d'après-midi
30 CM DE NEIGE ET DE NOMBREUX ARBRES ARRACHÉS
25/12/2010 11:39
Episode neigeux sur l'Alsace
VOS PHOTOS EN LIGNE SUR DNA.FR
24/12/2010 04:09
L'alsacien et les Alsaciens sur le réseau social américain
FACEBOOK PREND LANGUE AVEC L'ALSACE
25/12/2010 09:31
Neige et verglas
FIN DE LA VIGILANCE ORANGE EN ALSACE

VOS COMMENTAIRES
Poster un commentaire

Merci, pour vos infos, mais mon petit coin, Vosges du nore, et souvent oublié, les conditions météos, sont bcp pénibles, que les infos officiels

chantal.hetzel@live.fr | 26/12/2010 | 08:12
Répondre | Signaler un abus

enfin un vrai hiver comme nous n'en avons plus eu depuis longtemps - ce jour de Noël, le parc de l'Orangerie était magnifique, féérique, et le lac sera bientôt gelé. ne nous plaignons pas car le froid vif tuera aussi des microbes néfastes ....

bob_67 | 25/12/2010 | 21:54
Répondre | Signaler un abus

PERSÉCUTÉS
RETARD
Haut de page

En temps normal, les transports en commun coutent (hors abonnement) deux fois plus cher que le déplacement en voiture et cela...
riri67 | 26/12/2010 | 09:0

ABONNEMENTS
Toutes les offres web et PDF
Toutes les offres papier

LES VIDÉOS
Les mets du Marché de Noël

Que faut-il manger au marché de Noël?

Charles Aznavour en dédicace à la Librairie Kléb à Strasbourg
Charles Aznavour en dédicace à la librairie Klébe

Le spectacle de Noël d'Osthouse plébiscité lors d sa première représentation, vendredi.
Léon le magnifique

voir toutes les vidéos

Named entities detection

- Objective: detect named entities in the articles (persons, locations, organisations, others) to index them in specific fields
- Solution: evaluation of existing tools:
  - criteria: free of charge, ease of use, availability of language models, quality of extraction, processing performances
  - seven tools evaluated, reduced to four after first two criteria : Spacy (Honnibal and Montani, 2017), Sem (Dupont and Tellier, 2014), Open NER (Garcia-Pablos, 2013), Stanford Core NLP (Finket et al., 2005)
  - Sem is the best tool as quality is concerned but very slow processing capabilities => Spacy

Evaluation		Spacy	Sem	OpeNER	CoreNLP
Vrai	Positif	80	69	62	62
	Négatif	22	45	44	8
Faux	Positif	38	8	6	72
	Négatif	20	29	38	37
Rappel		**0,80**	0,70	0,62	0,63
Précision		0,68	**0,90**	**0,91**	0,46
F-mesure		0,73	**0,79**	0,74	0,53

"ne_per_dpi":"Chilling|1 Président_Costa|1 Broniowski|1 Saisie|1 Ilham_Aliyev|1 Cour|3 Khojaly|2 Premièrement|1 §|4 Lettre_Droits|2
"ne_misc_dpi":"Article_17_CEDH'|1 Web|1 Sect|1 Prix_international_de_la_liberté_de_la_presse|1 §_102|1 Pour_un_exemple_récent|1 Conv
"ne_org_dpi":"CPDH|2 Conférence_de_presse_annuelle|1 Cour_du_district_de_Yasamal|1 Archives|1 Cour_européenne_des_droits_de_l'homme|
"ne_loc_dpi":"Art|1 Fatullayev|1 Iran|1 §|8 Karabagh|3 §_88|2 Saadoon|1 §_126|1 Req|2 Cour|2 Mufdhi|1 Azerbaïdjan|6 A|1 §_86|1 Al|1

Morphological analysis

- Objective: analyse all words of the articles to associate them with a grammatical category and a matching lemma (Part-Of-Speech tagging)
- Solution: use of spacy
  - natural language processing tool suite
  - perform large-scale information extraction tasks
  - extract tokens, lemmas and lemmas_tags


```
"tokens":"Archives de la catégorie : ' Article 17 CEDH ' Le rédacteur en chef de deux journaux publiés en Azerbaïdjan , Eynulla Fatu
"lemmas":"archives de le catégorie : ' article 17 cedh ' le rédacteur en chef de deux journal publier en azerbaïdjan , eynulla fatul
"lemmas_tags":"archives/NOUN__Gender=Fem|Number=Plur de/ADP___ le/DET__Definite=Def|Gender=Fem|Number=Sing|PronType=Art catégorie/NC
```

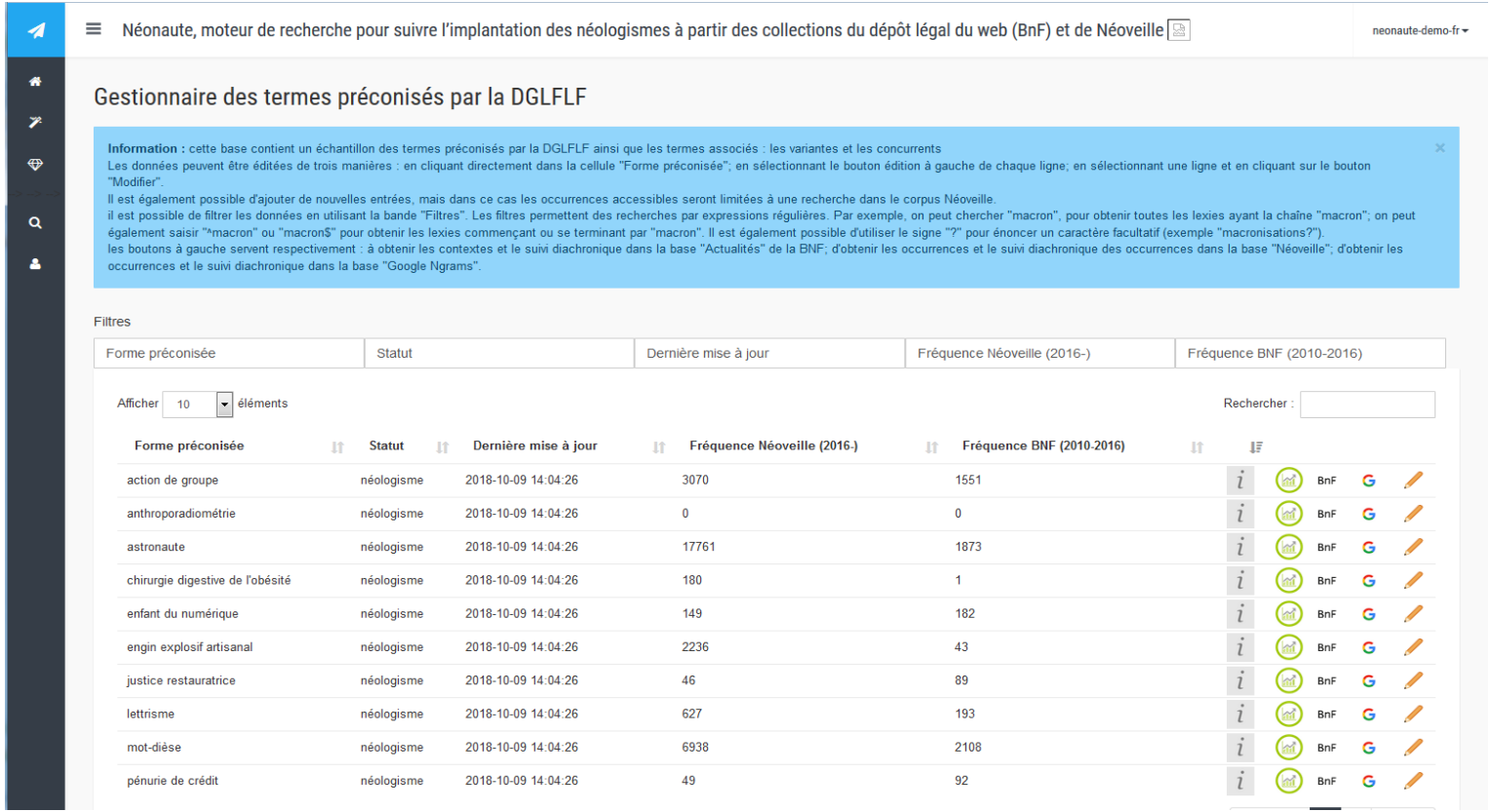
Extraction of fragments

- Objective: generate article fragments which contains specifical lexical items
- Solution: develop scripts to anonymise the data and extract 5 words before and after each lexical item

```
3       fact-checker
4       nuitsdeboutistes
5       post-minijupe
6       anti-surendettement
7       micro-festival
8       sur-dramatisation
9       pseudo-bouddhique
10      hollandienne
11      couverture-socle
12      domiciliateur
13      broumroumroum
14      antijihadistes
15      quasi-administration
16      demi-tweet
17      anti-drone
18      kwassas-kwassas
```

List of terms	# of Lexical Items	# of Fragments	Size of Fragments
DGLFLF recommended terms	191	213 495	114 MB
Neologisms	31 461	19 235 875	9 GB
Feminized Terms	5 368	01 107 319	76 GB
Total	37 020	220 556 689	85 GB

Néonaute interface



(BnF

enfant du numérique	néologisme	2018-10-09 14:04:26	149	182	BnF

Termes préconisés

Terme	Statut	Fréquence BNF	Fréquence Néoveille	Date dernière mise à jour
enfant du numérique	néologisme	0	5	2018-10-09 14:04:26
enfants du numérique	variante-N	83	5	2018-10-09 14:04:26
Total		**83**	**10**	

Termes concurrents

Terme	Statut	Fréquence BNF	Fréquence Néoveille	Date dernière mise à jour
digital native	concurrent	1	128	2018-10-09 14:04:26
enfant de l'ère digitale	concurrent	0	0	2018-10-09 14:04:26
enfant de l'ère numérique	concurrent	0	0	2018-10-09 14:04:26
enfant du digital	concurrent	0	2	2018-10-09 14:04:26
natif digital	concurrent	0	0	2018-10-09 14:04:26
natif du numérique	concurrent	98	8	2018-10-09 14:04:26
natif numérique	concurrent	0	1	2018-10-09 14:04:26
natifs digitaux	variante-C	0	0	2018-10-09 14:04:26
Total		**99**	**139**	

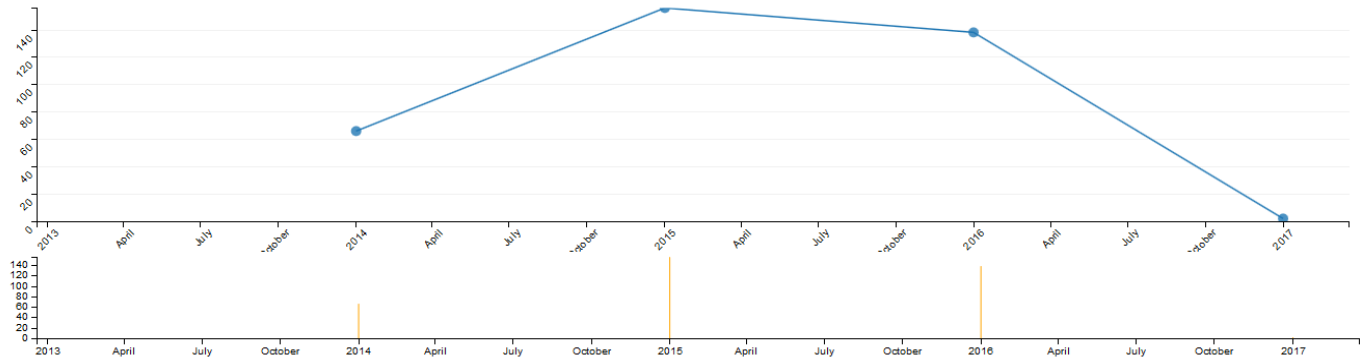
{BnF

enfant du numérique	néologisme	2018-10-09 14:04:26	149	182

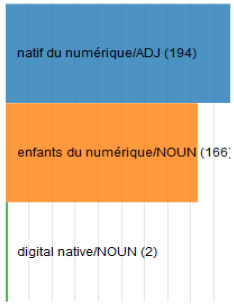
364

Evolution temporelle globale

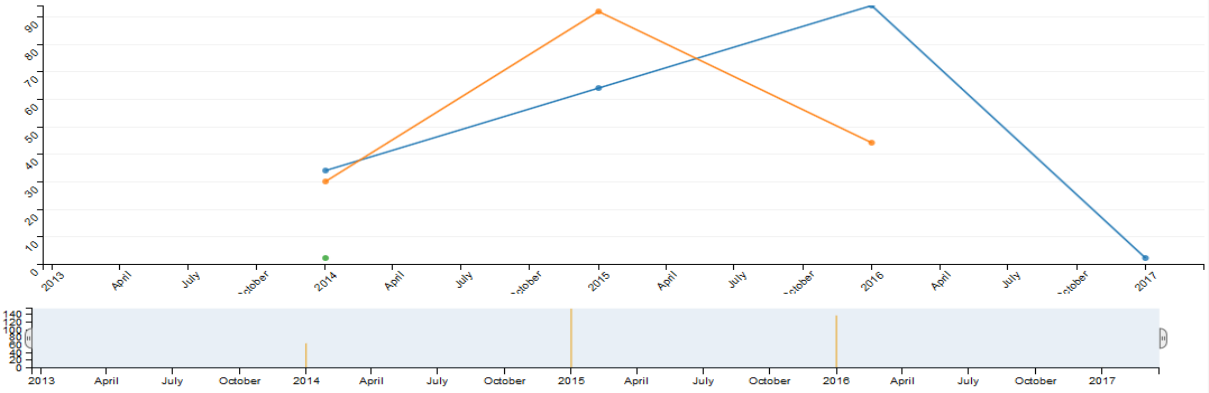
362 sélectionnés sur **364** articles | Réinitialiser



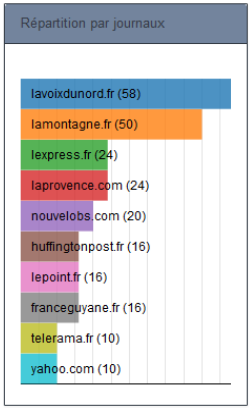
Répartition par partie du discours

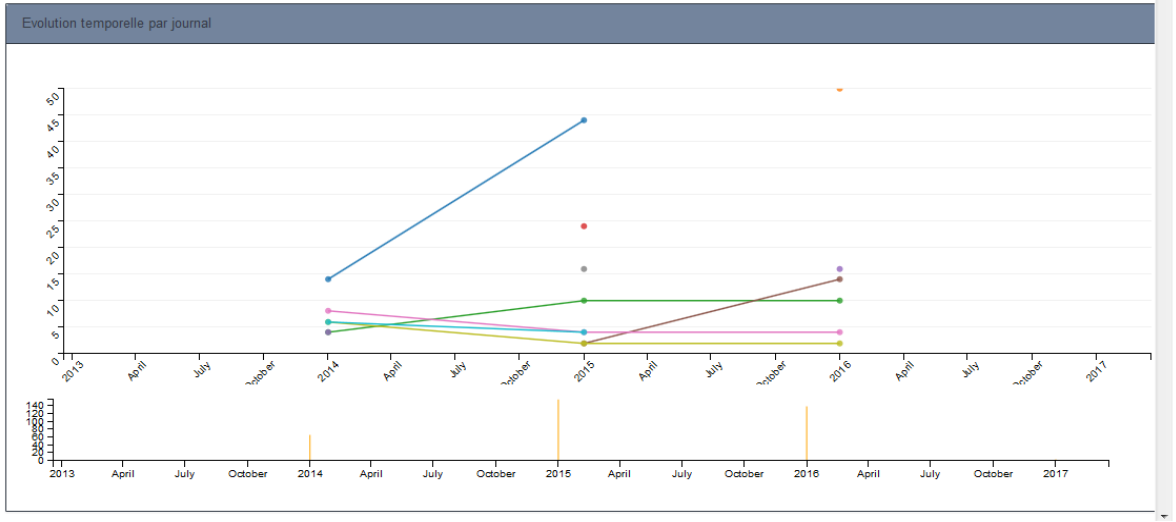


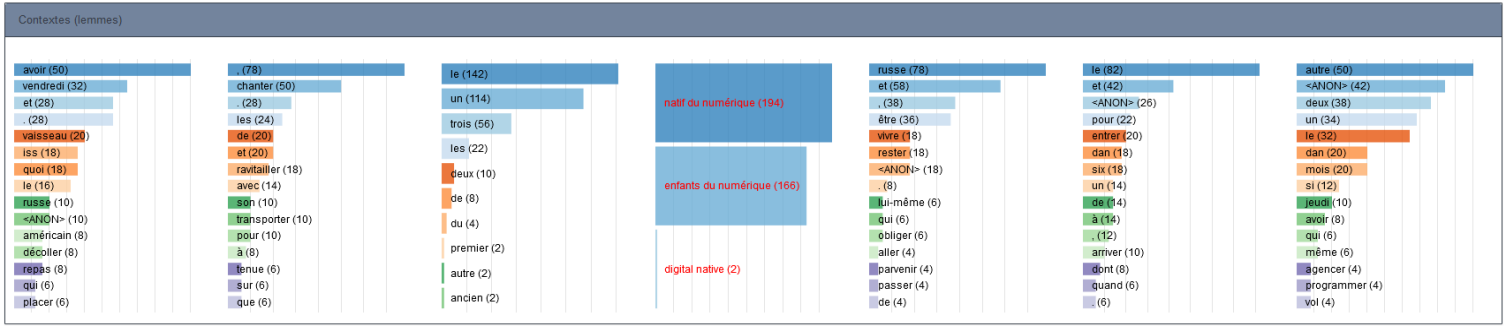
Evolution temporelle par lemme/pos



{BnF


Répartition par journaux
lavoixdunord.fr (58)
lamontagne.fr (50)
lexpress.fr (24)
laprovence.com (24)
nouvelobs.com (20)
huffingtonpost.fr (16)
lepoint.fr (16)
franceguyane.fr (16)
telerama.fr (10)
yahoo.com (10)


Evolution temporelle par journal
2013
April
July
October
2014
April
July
October
2015
April
July
October
2016
April
July
October
2017


Contextes (lemmes)
avoir (50)
vendredi (32)
et (28)
, (28)
vaisseau (20)
iss (18)
quoi (18)
le (16)
russe (10)
<ANON> (10)
américain (8)
décoller (8)
repas (8)
qui (6)
placer (6)
, (78)
chanter (50)
. (28)
les (24)
de (20)
et (20)
ravitailler (18)
avec (14)
son (10)
transporter (10)
pour (10)
à (8)
tenue (6)
sur (6)
que (6)
le (142)
un (114)
trois (56)
les (22)
deux (10)
de (8)
du (4)
premier (2)
autre (2)
ancien (2)
natif du numérique (194)
enfants du numérique (166)
digital native (2)
russe (78)
et (58)
, (38)
être (36)
vivre (18)
rester (18)
<ANON> (18)
. (8)
lui-même (6)
qui (6)
obliger (6)
aller (4)
parvenir (4)
passer (4)
de (4)
le (82)
et (42)
<ANON> (26)
pour (22)
entrer (20)
dan (18)
six (18)
un (14)
de (14)
à (14)
, (12)
arriver (10)
dont (8)
quand (6)
. (6)
autre (50)
<ANON> (42)
deux (38)
un (34)
le (32)
dan (20)
mois (20)
si (12)
jeudi (10)
avoir (8)
qui (6)
même (6)
agencer (4)
programmer (4)
vol (4)

{BnF

{BnF Archives de l'internet Labs

Accueil | Collections | Services | Aide

Rechercher

» Recherche avancée
» Recherche experte

» Se connecter

Archives de l'internet Labs permet d'explorer une partie des collections issues du dépôt légal de l'internet, constituées depuis 1996, en utilisant la recherche plein texte. *Archives de l'internet Labs* est complémentaire de l'application *Archives de l'internet* qui permet de faire des recherches par adresse URL et par date, de visualiser les pages archivées et de naviguer dans les archives du web français.

Des services complémentaires sont également proposés aux chercheurs dont les travaux sont réalisés sous couvert d'une convention. Trois collections sont actuellement disponibles.

Actualités 2010-2017

Attentats 2015

Incunables 1996-2000

© BnF 2018 | Informations légales | ailabs 2.1.0 (pfo)

BnF Archives de l'internet Labs

Accueil | Collections | Services | Aide

Rechercher

» Recherche avancée
» Recherche experte

» Mes recherches
» Mes corpus

P. Stirling
» Se déconnecter

Recherche avancée (?)

Mot(s) : macronisme — Rechercher un ou plusieurs mots ou une "expression exacte"

Exclure le(s) mot(s) : (facultatif) — Exclure les documents où ce(s) mot(s) apparaissent

Filtres (facultatifs)

Collection(s) : ☐ toutes ☑ actualités ☐ attentats 2015 ☐ incunables 1996-2000

Date de collecte : — Limiter la recherche à une date (aaaa, mm/aaaa ou jj/mm/aaaa)

Ou période : de à — Limiter la recherche à une période (aaaa, mm/aaaa ou jj/mm/aaaa)

Hôte, nom de domaine, extension : — Limiter la recherche à un hôte (gallica.bnf.fr), un nom de domaine (bnf.fr) ou une extension (fr)

Format : — Préciser le format de fichier des documents (html, text, pdf, image, audio, video, word, excel, powerpoint, other)

Rechercher

© BnF 2018 | Informations légales | ailabs 2.1.0 (pfo)

BnF Archives de l'internet Labs
Accueil | Collections | Services | Aide
Rechercher
» Recherche avancée
» Recherche experte
» Mes recherches
» Mes corpus
P. Stirling
» Se déconnecter
3 résultats
Rappel de la recherche :
Mot(s) : macronisme
Collection(s) : actualités
Année (inclus) : 2014
modifier | enregistrer | exporter
Trier par : Pertinence
Page 1
10 résultats par page
Enregistrer dans un corpus (0)
Trier les valeurs des facettes : 0-9 | A-Z
Collection (1) inclure | exclure
+ actualités (3)
Année (4) inclure | exclure
+ 2017 (28 981)
+ 2016 (1 964)
+ 2015 (102)
+ 2014 (3)
Nom de domaine (2) inclure | exclure
+ nouvelobs.com (2)
+ lepoint.fr (1)
Extension (2) inclure | exclure
+ com (2)
+ fr (1)
Langue (1) inclure | exclure
+ fr (2)
"Emmanuel Macron, "l'hémisphère droit" de François Hollande - Le Nouvel Observateur"
Archive du 30 août 2014
Format : html
http://tempsreel.nouvelobs.com/politique/20140826.OBS7233/emmanuel-macron-l-hemisphere-droit-de-francois-hollande.html
Voir les 2 captures
cela mais.. la réactivité enfin salvatrice surgira peut-être dans les prochains jours et semaines en pleine lumière.. qui sait.. ce si gentil macronisme de La Finance semble tellement attirer l'attention maintenant
"Commentaires - Tollé après l'annonce de la modulation des allocations familiales | Le Point"
Archive du 17 octobre 2014
Format : html
http://mobile.lepoint.fr/reactions/economie/commentaires-sur-les-allocations-familiales-modulees-selon-le-revenu-a-partir-de-2015-16-10-2014-1873154_28
Page 1

Mes recherches

1 recherches enregistrées

Nom ▼ ▲	Recherche	Critères	Date d'enregistrement ▼ ▲	Relancer	Supprimer
macronisme	avancée	Mot(s) : macronisme Collection(s) : actualités	26/04/2018	relancer	supprimer

{BnF

{BnF Archives de l'internet Labs

Accueil | Collections | Services | Aide

Rechercher

» Recherche avancée
» Recherche experte

» Mes recherches
» Mes corpus

P. Stirling
» Se déconnecter

Corpus macronisme

6 documents
Crée le 26/04/2018
Mis à jour le 12/10/2018

exporter | supprimer | liste des corpus

URL ▼ ▲	Date de l'archive ▼ ▲	Format ▼ ▲	Titre ▼ ▲	Consulter	Supprimer
http://news.google.fr/news?pz=1&js=0&cf=all&ned=fr&hl=fr&q=Emmanuel+Macron&output=rss	09/12/2014	other	Emmanuel Macron - Google Actualités	consulter	supprimer
https://fr.search.yahoo.com/search?p=Emmanuel+Macron&vm=r&woeid=615702&fp=1&ei=UTF-8&nojs=1	28/11/2014	html	Emmanuel Macron - Résultats de recherche Yahoo France	consulter	supprimer
https://mobile.twitter.com/hashtag/macronisme?src=hash	15/05/2017	html	Recherche Twitter - #macronisme	consulter	supprimer
http://tempsreel.nouvelobs.com/economie/20160901.OBS7271/jean-marc-daniel-le-macronisme-est-une-forme-de-blairisme.atom	04/09/2016	other	L'Obs en temps réel : Economie	consulter	supprimer
http://tempsreel.nouvelobs.com/politique/20140826.OBS7233/emmanuel-macron-l-hemisphere-droit-de-francois-hollande.html	30/08/2014	html	Emmanuel Macron, "l'hémisphère droit" de François Hollande - Le Nouvel Observateur	consulter	supprimer
http://www.marianne.net/melenchon-macronisme-est-vision-du-monde-primaire-barbare-100242400.html	01/05/2016	html	Pour Mélenchon, "le macronisme est une vision du monde primaire, barbare"	consulter	supprimer

© BnF 2018 | Informations légales | ailabs 2.1.0 (pfo)

Conclusions and perspectives

- Access to the external Néonaute interface to researchers working on the project
  - Work still ongoing on the three case studies
  - Named Entity Recognition and Topic Detection not fully developed in the project
  - Improvements on the linguistic analysis modules and adaptation of the language models
- Full-text searching and access to content onsite in Archives de l'internet Labs
  - Aim to offer corpus creation and saved searches to all users
  - Integrate aspects of data visualisation in search interface
- Need to simplify organisation of future projects to answer researchers' needs
  - Service rather than co-development
  - "Corpus": four-year BnF project to provide digital corpora to researchers
- Legal questions on use of text and data mining for research purposes

{BnF

Questions?

depot.legal.web@bnf.fr

https://webcorpora.hypotheses.org/