

(Re)catégorisation automatique des Résumés et des Séquences des notices documentaires INA

Bénédicte Pincemin



Ce document est publié sous licence Creative Commons Attribution 4.0 International
<https://creativecommons.org/licenses/by/4.0>

Avertissement : Dans ce document, les noms des corpus et la forme précise des requêtes CQL, en particulier celles utilisant la structuration des corpus, **correspondent à l'état des corpus en février 2018**. Ces éléments ne sont plus directement utilisables tels quels pour les dernières versions des corpus dont les structures ont été améliorées, **un travail d'adaptation est nécessaire pour la plupart des requêtes** complexes.

Expérimentations de BP sur le corpus TXM AFNOTICES en février 2018

Attention : les requêtes de 2018 ont été effectuées sur une première version du corpus, les décomptes pourraient être mis à jour :

- en mettant à jour les requêtes (ex. < sujet > -> < notice >) ;
- En se focalisant sur le sous-corpus des notices de type sujet (qui correspond à notre principal corpus de travail, en lien avec les choix faits pour le corpus AFVOIXOFFV02).

Critère 1 : recherche d'un vocabulaire caractéristique, lié aux plans

Requête pour repérer les séquences enchaînées au résumé :

```
<resume>[]* [word="- *(VG|GP|PP|PM|PANO|Pano|PA|PG|GPP|PR|DP|VA|TRAV(EL)?|VP|VSG|VE|ZAV)"] []* </resume> |
<resume>[]* [word="- "] []* [word="vue|plans?"%c] []* </resume> |
<resume>[]* [word="vue|plans?"%c] []* [word="- "] []* </resume> |
<resume>[]* [word="Vue|Plans?" ] []* [word="Vue|Plans?" ] []* </resume> |
<resume>[]* [word="plan"%c] [word="par"%c] [word="plan"%c] []* </resume>
|
<resume>[]* [word="relevé|détail"%c] [word="des"%c] [word="plans"%c]
[]* </resume>
```

Donc on teste :

- Les abréviations caractéristiques de désignation de plan (la liste n'est sans doute pas complète, ici c'est une première recherche indicative, on n'avait pas encore de vue de traitement systématique et aussi complet que possible)
- La coprésence d'au moins un tiret **et** d'au moins un mot dans l'ensemble *vue*, *plan*,

plans (si je n'ai pas mis *vues* au pluriel c'est sans doute que cela paraissait moins fiable)

- Le mot *Vue* ou *Plan(s)* en début de "phrase" (i.e. avec une majuscule), au moins **deux** fois
- La présence d'une expression comme *plan par plan*, ou *relevé des plans*, ou *détail des plans*.

Critère 2 : Le résumé contient du vocabulaire caractéristique des séquences ET le champ séquences est vide

Même chose sur une seule ligne, avec condition sur l'absence de séquences :

```
((<resume>[ ]*[word="- *(VG|GP|PP|PM|PANO|Pano|PA|PG|GPP|PR|DP|VA|TRAV(EL)?|VP|VSG|VE|ZAV)"] ]*</resume>)|(<resume>[ ]*[word="- "[ ]*[word="vue|plans?"%c] ]*</resume>)|(<resume>[ ]*[word="vue|plans?"%c] ]*[word="- "[ ]*</resume>)|(<resume>[ ]*[word="Vue|Plans?" ]*[word="Vue|Plans?" ]*</resume>)|(<resume>[ ]*[word="plan"%c][word="par"%c][word="plan"%c] ]*</resume>)|(<resume>[ ]*[word="relevé|détail"%c][word="des"%c][word="plans"%c] ]*</resume>))[!sequences]*</sujet>
```

[Rq. dans AFNOTICES actuel, remplacer le </sujet> terminant la requête par </notice>]

6516 occurrences -> **plus de 6000 sujets sans séquences mais avec un résumé ressemblant à des séquences, c'est énorme, non ?**

(69 occ. avec sequences non vide)

Et si l'on se restreint à

```
<resume>[ ]*[word="- *(VG|GP|PP|PM|PANO|Pano|PA|PG|GPP|PR|DP|VA|TRAV(EL)?|VP|VSG|VE|ZAV)"] ]*</resume>[!sequences]*</sujet>
```

[Rq. dans AFNOTICES actuel, remplacer le </sujet> terminant la requête par </notice>]

on a encore 6139 occ.

Requête de contrôle et recherche de cas limites

Requête de contrôle partiel :

```
<resume>[word!="- "[ ]*[word="vue|plans?" ]*[word!="- "[ ]*</resume>
```

Glose : **on cherche un résumé contenant le mot *vue*, ou *plan*, ou *plans*, mais sans aucun tiret.**

Remarque : à l'époque de ces tests, on n'avait pas dans TXM l'information sur les **retours à la ligne**, qui semblaient un élément de signature caractéristique du champ séquences : on pouvait rechercher cette information en revenant au fichier source (tableur) fourni par l'INA.

On en trouve :

Ex. de cas limite (sans retours chariot dans xlsx, donc a priori OK comme résumé) :

Identifiant de la notice : AFE04016111, Date de diffusion : 26/03/1958

Au Palais de l'Europe, se tenait du 19 au 21 mars 1958, la première session de travail du nouveau Parlement européen (Assemblée unique des Communautés européennes ou Assemblée européenne). Différents plans sur les députés composants l'assemblée et du public dans les tribunes. Victor LAROCK, ministre belge des Affaires étrangères ouvre la première séance de travail. Accolade et poignée de main entre Robert SCHUMAN et Luciano GRANZOTTO BASSO, lors de la passation de pouvoir entre les deux hommes. Robert SCHUMAN, devient président de la nouvelle assemblée.

Ex. de cas limite (avec retours-chariot dans xlsx, donc a priori mauvais) :

Identifiant de la notice : AFE85008896

Date de diffusion : 16/11/1960

Source (le tableur xlsx INA) :

Le point sur le vote des électeurs américains qui ont choisi le candidat démocrate John KENNEDY.

Différents plans de citoyens américains remplissant leur devoir électoral.

Le président sortant, EISENHOWER, vote à Gettysburg. Le couple NIXON vote près de Los Angeles et KENNEDY à Boston.

Dans la nuit, les panneaux lumineux affichent les différents résultats. Et Pierre SALINGER lit les résultats sur les téléscripteurs.

Extrait déclaration de Richard NIXON, après sa défaite, son épouse à ses côtés.

Les titres des journaux dont le "NEW YORK MIRROR" annonce la victoire de KENNEDY.

Sur une estrade John KENNEDY, entouré de Jackie, de Rose et de Joseph adresse quelques mots au public.

(Edition dans TXM identique, à observer en faisant une CONCORDANCE sur

<notice_identifiantdelanotice="AFE85008896">[]

Puis double-clic sur la ligne de résultat.)

Ex. de cas limite (2) (avec retours-chariot dans xlsx, donc a priori mauvais) :

Identifiant de la notice : AFE85009043, Date de diffusion : 05/04/1961

Source :

Rencontre à la Maison blanche de John KENNEDY et Andréi GROMYKO, alors que l'affaire laotienne assombrit le paysage diplomatique.

Différents plans des deux hommes discutant. Les photographes les mitraillant .

Ex. de résumé avec retours-chariot et de token -DP [problème de tokenisation dans TXM pour ce corpus, noté pour le corriger, qui fait que le tiret n'est pas détaché et donc "passe inaperçu" pour la requête]

Identifiant de la notice : AFE04016127, Date de diffusion : 02/04/1958

Source :

Monsieur Georges Duhamel,Président de l'Alliance Française reçoit à l'aéroport d'Orly une délégation de maires des grandes villes d'Amérique Latine,arrivés de Bogota en avion "Caravelle" d'Air France

Personnalités présentes: MM Garcia, Ministre de l'air du Pérou, Roza,Ambassadeur, Montes,attaché militaire, Ribero,Maire de Lima et Madame Devane de l'ambassade

-DP dans l'avion et vue aérienne

Edition dans TXM :

Monsieur Georges Duhamel, Président de l'Alliance Française reçoit à l'aéroport d'Orly une délégation de maires des grandes villes d'Amérique Latine, arrivés de Bogota en avion " Caravelle " d'Air France Personnalités présentes : MM Garcia, Ministre de l'air du Pérou, Roza, Ambassadeur, Montes, attaché militaire, Ribero, Maire de Lima et Madame Devane de l'ambassade-DP dans l'avion et vue aérienne

Rq. tokenisation : l'/ambassade/-DP

Ex. de cas particulier (résumé mais impropre) :

Identifiant de la notice : AFE85008262, Date de diffusion : 15/04/1959

source :

Pas de détection dans le registre des plans par plan (page manquante);le texte dit par CLAUDE DAUPHIN n'a pas été relevé (absent du livre des commentaires)

Ex. de cas particulier (résumé = méta séquence) :

Identifiant de la notice : AFE08001181, Date de diffusion : 16/01/1968

mêmes plans qu'au 68012 + 2 autres plans de bateaux exposés

Ex. de cas particulier (résumé dans style séquences)

Identifiant de la notice : AFE86003095, Date de diffusion : 18/05/1945

- 2 VG plongeantes sur la ville en ruines- Divers plans de quartiers de la ville en ruines
Images de Saint-Nazaire après les destructions de la Seconde Guerre mondiale-

Noter que ce résumé est suivi d'un champ séquences non vide, et même beaucoup plus long :

2 VA de l'Arsenal détruit

- 3 Plans d'ateliers détruits dans l'Arsenal

- 3 PM d'un remorqueur coulé près d'un quai de l'Arsenal.

- 2 VG d'une partie de la ville en ruines

- 2 VG de la gare en ruines

- 2 VG d'une chicane construite par les allemands

- 2 VG de blockhaus sur lesquels sont peintes des portes et des fenêtres

- 2 VG des installations portuaires détruites

- 2 VG d'un hangar pour sous-marin protégé par des "dents de dragon "

Quelques hypothèses écartées

Faudrait-il s'appuyer sur les retours à la ligne ? (repérer tout simplement les résumés avec plusieurs paragraphes ?) : NON

On peut le vérifier avec la requête :

```
<resume>( <p>[ ]+</p>){2, }</resume><sequences>[ ]
```

-> 436 notices avec un champ résumé en plusieurs paragraphes et un champ séquences

non vide => les documentalistes peuvent faire des paragraphes dans le résumé.

Est-ce qu'un contenu séquences prend toujours la forme d'une liste à tirets ?

NON

```
<sequences>[word!=" - "]+<sequences>
```

On décompte 231 séquences non vides sans aucun tiret.

Est-ce que la collision résumé/séquence serait liée à des lignes blanches ? : PAS

SÛR, ne semble pas toujours le cas.

[Là j'avoue que je ne sais plus de quelles lignes blanches je parlais en février 2018...]

Tout champ séquences vide pourrait être rempli avec le contenu du résumé du même sujet ?

Hypothèse : quand il n'y a qu'un seul champ renseigné, c'est (généralement ?) le résumé mais avec un contenu de type séquences)

Observation : NON

```
<resume>[word!=" - *(vg|p[pmg]|gpp?|plans?|vue|pano(ramique)?|trav(el(l?ing)?))"%c & word!="V[APEMFR]|P[RALE]|DP|VSG|ZA[VR]|GVG|TGP"]+</resume>[!sequences]*</notice>
```

On obtient 1713 notices avec des séquences vides et un résumé qui ressemble plus à un résumé qu'à des séquences.

Pistes à étudier

Vocabulaire caractéristique + séquences vide +

Paragraphes multiples et courts ou/et phrases nominales

Eléments de construction d'une requête sur les notations de valeur de plan

Plusieurs pistes +/- parallèles, qu'on peut éventuellement recroiser (si si, recroiser des parallèles ;-)

Requête construite dans le cadre de la recherche de Franck sur les plans

Sans doute la plus complète et la plus synthétique, s'appuyait vraisemblablement sur les trois techniques et résultats ci-dessus

```
[word=" - *(vg|p[pmg]|gpp?|plan|vue|pano(ramique)?|trav(el(l?ing)?))"%c | word="V[APEMFR]|P[RALE]|DP|VSG|ZA[VR]|GVG|TGP"]
```

Par un calcul de spécificités

Repérage des abréviations de description de séquences à partir d'un calcul de spécificités faisant ressortir les mots statistiquement globalement sur-représentés dans les séquences par rapport au reste. Pour construire le sous-corpus, on entre la requête suivante dans l'onglet avancé :

<sequences>[]+</sequences>

Par un INDEX sur les abréviations en majuscules

Voir aussi l'INDEX de :

" - " "[A-Z]{2,}"

Remarques :

DG = De Gaulle

GD : toutes les occ. sont avec PANO (ex. PANO horizontal GD)

PC = abréviation (poste de commandement, parti communiste...)

Liste construite par Jean en avril 2018

C'était un exercice que s'était donné Jean en application du document tutoriel d'exploitation du corpus des Notices avec TXM.

Sujet : RE: Tutoriel exploitation corpus Notices AF avec TXM, Was: [ANTRACT]

Proposition de communication au GDR ISIS (29 mars)

Date : Thu, 5 Apr 2018 16:39:18 +0000

De : Jean Carrive

Pour : Serge Heiden , Benoit Huet , 'Sylvain Meignier'

Copie à : 'Pascale Goetschel' , 'Franck Mazuet' , 'Bernard Merialdo' , 'Pincemin Bénédicte' , 'Vincent Jousse'

Merci Serge pour ces précisions. Avec la requête suivantes :

```
"-" "\p{Lu}+" "d.*|et"
```

Qui cherche à représenter les expressions comme « - VG de (l'arrivée du général...) » ou « VG diverses (d'enfants jouant) », j'obtiens 448 réponses. Il y a bien sûr pas mal de réponses qui ne sont pas pertinentes mais le nombre est raisonnable et peut être traité à la main. Je trouve ainsi des codes que je ne connaissais pas comme « VSG » (Exemple VSG des partisans avec un drapeau).

Voici en pièce jointe et sur sharedocs (lien), un document excel avec dans différents onglets :

- L'export TXM des résultats
- Un tableau croisé dynamique pour avoir des valeurs uniques des codes (pas très utile)
- Une sélection des codes qui me paraissent pertinents (37 codes)

[BP : le fichier est rangé sur le sharedocs à cet endroit :

SP2... > TXM > Corpus > Corpus notices documentaires

Et il se nomme 20180405 Export txm nomenclature plans.xls]

C'est peut-être intéressant effectivement d'ajouter les précisions au tutoriel.

Bonne soirée,

Jean

Résultat pour la sélection de codes :

CODE	Commentaire
CG	
CPL	Vue en CPL, PM en CPL
DP	
DV	
GFPP	
GO	
GP	
GPP	
GVG	
ITW	
PA	
PANO	
PANORAMA	
PANORAMIQUE	
PE	
PG	
PL	
PLR	
PM	
POP	
PP	
PPL	
PR	
PRL	
PSG	
TGP	
TRAV	
TRAVEL	
VA	
VE	
VF	
VG	
VM	
VP	
VR	
VSG	
ZAR	souvent: (avec ZAR)