Gestion de Projet

M1 Technologies de l'Internet

Analyse d'articles scientifiques

Auteurs :
Bagre Gnebehi
Josuah Perrot
N'tanouan Kouame-Kodia

Tuteurs :
Marie-Noëlle Bessagnet
Annig Lacayrelle
Christian Sallaberry

Remerciements

Nous tenons à remercier nos tuteurs pour leur pédagogie et leur encadrement.
Madame Lacayrelle pour son soutien, sa clarté sa disponibilité et ses conseils concernant la rédaction de ce rapport. Monsieur Sallaberry pour nous avoir remis sur de bonnes pistes quand nous nous égarions. Et enfin, Madame Bessagnet pour avoir assuré la coordination et le suivi de ce projet.

Table des matières

Première partie

Introduction

Nous assistons à un accroissement prodigieux des publications scientifiques disponibles au format numérique, que ce soit à l'échelle nationale ou internationale.

C'est dans ce cadre que nous sommes intervenus pour répondre à l'appel d'offre de la maîtrise d'ouvrage.

L'objectif est de proposer une méthodologie semi-automatique pour l'analyse de l'évolution dans le temps et dans l'espace d'un ensemble de publications scientifiques et des thématiques concernées.
L'intérêt de ces travaux est notamment d'appuyer les scientifiques dans leur travail de veille en mettant en avant l'évolution des thématiques au fil du temps, selon les lieux des conférences et les lieux des laboratoires d'affiliation des auteurs.

En découle les problématiques suivantes :

- Existe-t-il des outils qui se prêtent aisément à l'analyse d'articles scientifiques ?
- De quelle manière pouvons-nous réaliser ces analyse avec des données non-structurées ?

Ce rapport est organisé comme suit :
Premièrement, nous définirons clairement les demandes, leurs contextes et la méthode adoptée à travers la section Gestion de projet.
Deuxièmement, nous présenterons les connaissances qu'il nous a fallu maîtriser pour ce sujet.
Troisièmement, nous détaillerons le circuit de réalisation du projet.
Nous finirons par le bilan et le retour d'expérience de ce projet.

Deuxième partie

Gestion du Projet

I Cahier des charges

Sur la base de plus de 1000 articles de recherche relatifs à la conférence annuelle EGC [1] de 2004 à 2018. La maîtrise d'ouvrage souhaite élargir un travail de recherche [KBS+16] mené en 2014-2015 qui a permis de calculer et d'analyser divers critères liés à de l'information provenant de ces articles : les thèmes traités, les relations entre auteurs, les localisations géographiques des auteurs, l'évolution des thèmes selon les années, etc...
Aujourd'hui, les données concernent les articles des conférences EGC de 2004 à 2018.
Ce corpus est donné dans un fichier csv. La maîtrise d'ouvrage souhaite d'une part mener un autre type d'analyse sur le corpus entier et d'autre part intégrer dans leur base les articles de 2016 à 2018, en automatisant l'intégration des caractéristiques supplémentaires.

Information thématique	Nom ville conférence Noms villes auteurs Noms auteurs Titre article Résumé article Session Domaine
Information spatiale	Coordonnées villes auteurs Coordonnées ville conférence
Information temporelle	Année conférence
Information plein texte	Termes titre article Termes résumé article

FIGURE 1 – Modèle synthétique des données indexées pour chaque article

Le corpus d'articles de 2004 à 2015 a donc été traité et intègre les caractéristiques ci-dessus(I). Toutes ces informations sont stockées dans une base ElasticSearch [2].

Des outils d'analyse sont proposés (TXM, Iramuteq) et un outil de structuration de données (Grobid).

Pour cela, il est nécessaire de connaître le domaine et l'ouvrage afin de savoir quelle information correspond à quelle dimension. Une fois cet apprentissage fait, nous pouvons construire des règles dans une chaîne de traitement permettant d'analyser chaque information en fonction de sa dimension.

1. Extraction et Gestion des Connaissances (EGC) est un événement annuel réunissant des chercheurs et praticiens de disciplines relevant de la science des données et des connaissances.
2. Elasticsearch est un moteur de recherche et d'analyse RESTful distribué

En résumé, les caractéristiques du projet sont :

- l'apprentissage et la compréhension du domaine considéré ;
- le développement d'une chaîne de traitement en utilisant deux logiciels de textométrie importants pour l'analyse que sont : **TXM et Iramuteq** et l'application **Grobid**
- la mise en place de la visualisation des résultats.

II Méthodologie du travail

1 Méthode agile

Interactions maîtres d'oeuvres

Les différents acteurs se réunissaient au moins une fois par semaine pour faire le bilan des tâches accomplies, établir les tâches restantes , partager leur différentes difficultés , établir le planning et l'ordre du jour de la réunion suivante avec la maîtrise d'ouvrage.
A l'aide d'outils tels que :

- Trello : permet la répartition des tâches
- Overleaf : permet la rédaction collaboratif du mémoire
- Discorde : favorise la communication entre membre
- Google Drive : permet le partage et le stockage des fichiers

le travail collaboratif a été favorisé et l'adaptation aux changements fût efficace.

Interaction maîtrise d'ouvrage et maîtres d'oeuvre

Chaque semaine en moyenne nous avions une réunion avec la maîtrise d'ouvrage pour présenter l'évolution de nos travaux. Des directives découlaient donc de ces réunions.
Tout au long du projet et après chaque réunion, nous devions rédiger les différents comptes rendus des réunions et l'ordre du jour de la prochaine réunion.

2 Diagramme de Gantt prévisionnel

	Risque	Nom de la tâche	Attribuée à	Date de début	Date de fin
1		**Définition du projet**		**22/02/19**	**07/03/19**
2		**Prise de contact**		**07/03/19**	**14/03/19**
3		**Documentation + Prise en main du logiciel**		**14/03/19**	**27/03/19**
4		Grobid	Perrot Josuah	14/03/19	27/03/19
5		Iramuteq	Gnebehi Bagre	14/03/19	27/03/19
6		TXM	Kouamé-Kodia Marilyne	14/03/19	27/03/19
7		**Création de l'Application de péparation de données**		**13/03/19**	**05/04/19**
8		Convertir CSV en JSON	Gnebehi Bagre	13/03/19	18/03/19
9		Convertir JSON en Txt + CSV	Kouamé-Kodia Marilyne	19/03/19	25/03/19
10		Fusionner les JSON	Perrot Josuah	26/03/19	01/04/19
11		Extraire pdf des CSV	Perrot Josuah	03/04/19	05/04/19
12		**Exploitation des logiciels pour obtenir des résultats**		07/04/19	06/05/19
13		Produire des fichiers TEI	Perrot Josuah	07/04/19	15/04/19
14		Produire un lexique informatique	Kouamé-Kodia Marilyne	07/04/19	15/04/19
15		Chronologie des thème par année	Gnebehi Bagre	17/04/19	25/04/19
16		Produire fichier JSON pour ElasticSearch	Perrot + Kouamé-Kodia + Gnebehi	27/04/19	06/05/19
17		**Visualisation des Données**			
18		Création d'un site web	Perrot + Gnebehi + Kouamé-Kodia	06/05/19	17/05/19

FIGURE 2 – Diagramme de Gantt prévisionnel

Dans ce planning, nous avions prévu une première phase de documentation sur les différents concepts et technologies que nous allions utiliser. En parallèle, nous devions nous mettre d'accord avec les responsables du projet sur le travail à effectuer, grâce au cahier des charges. Ensuite, nous avions la phase de développement durant laquelle nous allions mettre en place notre chaîne de traitement et notre interface de visualisation.
Évidemment, ce diagramme ne représente en aucun cas la façon dont s'est réellement déroulé notre projet.

Troisième partie

Cadre de l'analyse

Introduction

Pour cerner l'environnement du projet, nous allons définir quelques connaissances liées au domaine de l'analyse de données textuelles. Ensuite, viendra une description des outils qui ont été nécessaires lors de la phase de développement.

I Définition de la textométrie

L'analyse de données textuelles (ou ADT) est une approche des sciences humaines qui envisage les textes comme des données organisées qui, constituées en corpus, peuvent être analysées indépendamment de leur énonciataire, voire de leur énonciation.Ceci est appelé la textométrie[Wik]. Elle cherche à qualifier les éléments des textes à l'aide de catégories et à les quantifier en analysant leur répartition statistique et est appliquée strictement au lexique.

II Outils

1 Grobid

GROBID [3] est une bibliothèque d'apprentissage automatique permettant d'extraire, d'analyser et de restructurer des documents bruts tels que PDF en des documents structurés codés en TEI, avec un accent particulier sur les publications techniques et scientifiques. L'utilité de Grobid dans notre cas est son outils "process header" permettant de structurer un document PDF en document XML.TEI.

Grobid est un logiciel open source disponible sur navigateur depuis n'importe quel système d'exploitation ou par application Python[Lopc] ou Java.

Appréhender Grobid ne fût pas compliqué car il existe une large documentation[Lopa] et des tutoriels[Lopb] pour installer son propre serveur et utiliser au mieux les fonctionnalités qu'il possède.

2 TXM

L'intérêt de TXM pour notre étude, est de faire une analyse lexicale et grammaticale du corpus de sorte à ressortir les concepts les plus parlants par année.
TXM est un logiciel Open Source pour la TextométrieI.
L'apprentissage de l'outil passe par une lecture approfondie du manuel[Heia] et de vidéos relatant les différents usages possibles de TXM[Heib].
«La plate-forme TXM combine des techniques puissantes et originales pour l'analyse de grands corpus de textes au moyen de composants modulaires. Elle a été initiée par le projet ANR Textométrie qui a lancé une nouvelle génération de recherches textométriques, en synergie avec les technologies de corpus et de statistique actuelles (Unicode, XML, CQP [4] et R).» [5]

Elle peut prendre en entrée plusieurs types de données comme présenté ci-dessous :

3. https://grobid.readthedocs.io/en/latest/
4. https://portal.clarin.nl/node/4066
5. http://textometrie.ens-lyon.fr/

FIGURE 3 – Les différents formats que supportent TXM

L'une des techniques puissantes dont dispose TXM est CQL qui facilite les recherches pour l'analyse approfondie grands corpus de texte.

Les requêtes de CQL

CQL est l'acronyme de Corpus Query Language, il s'agit d'un langage d'expression de requêtes.
Une expression CQL est une chaîne de caractères exprimant un motif linguistique - un mot, ou une suite de mots - défini en fonction de formes graphiques, de formes lemmatisées ou de catégories grammaticales.

FIGURE 4 – Exemple de requête CQL

L'exemple ci-dessus recherche les groupes de mots commençant par WEB. On cherche à capter des mots comme *Web Sémantique, Web analytique , etc....*

Quelques notions pour l'apprentissage de CQL

- " nation.* " : on récupère ici toutes les déclinaisons des mots qui commencent par "nation" et on les comptabilise.
  On obtient des mots tels que *nations, nationales, nationalité, etc....*

- " *.patri.* " : on récupère ici tous les mots qui ont patri comme base de mots et on les comptabilise.

  On obtient des mots tels que *compatriote , rapatriés, etc....*

- "[word = "pays | nation"]" : on recherche les mots pays ou nation .

— "[word = "pays | nation"][frpos="ADJ"]" : On recherche des groupes de mots commençant par pays ou nation et suivis d'un adjectif (on arrive à déterminer la nature des mots avec frpos, grâce à Treetagger [6].

— "[frpos = "NOM"][word = "français.*"]" : on recherche les groupes de mots commençant par la base *français* et commençant par un nom.
On peut obtenir comme résultat *société française , peuple français etc...*

— "[ ][ ][ ]" : On recherche toutes les suites de trois mots.

— "temps [ ] travail" : On recherche tous les groupes de mots commençant par temps et se terminant par travail.
On peut obtenir ce genre de résultat *temps de travail*.

6. consiste à associer aux mots d'un texte les informations grammaticales correspondantes

3 IRaMuTeQ

IRaMuTeQ[eBG] (pour « Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires ») est un logiciel libre et ouvert d'analyse de données textuelles ou de statistique textuelle qui fonctionne en interface avec le langage R. [7]

Il s'agit d'un logiciel qui repose sur l'analyse statistiques de corpus de textes et sur des tableaux individus/caractères, en s'appuyant sur la méthode de classification de Max Reinert (classification hiérarchique descendante sur un tableau croisant les formes pleines et des segments de texte).
Le Logiciel dispose d'un dictionnaire composé de mots et d'expressions utiles pour la reconnaissances des mots ou expression et présente d'une fonctionnalité «nuage de mots» capable de fournir une visualisation textuelle de l'ensemblle des mots.

Analyse statistique

Cette analyse présente les statistiques générales du corpus d'entrée à savoir le nombre de mots existant dans le corpus,le nombre de texte, les mots aussi appelé formes existants et leur occurences dans le corpus.

Analyse de similitude (ADS)

Cette analyse va calculer, par exemple la co-occurrence(combien de fois les éléments vont apparaître en même temps).
C'est une théorie qui émane de la théorie des graphes.

Classification méthode Reinhert

Cette méthode se présente sous forme de dendrogramme.
L'objectif est de regrouper des « mondes lexicaux » et de mettre en évidence les thématiques générales du corpus.

7. https://fr.wikipedia.org/wiki/Iramuteq_(logiciel)

Quatrième partie

Développement

Introduction

L'ensemble de notre travail repose sur les données concernant les articles des conférences EGC de 2004 à 2018 et avec des méta-sessions [8] allant de 1 à 9.

I Processus de réalisation du projet

Au vu du cahier des charges qui nous est imposé, un circuit de réalisation du projet a été établi comme le montre l'image ci-dessous :

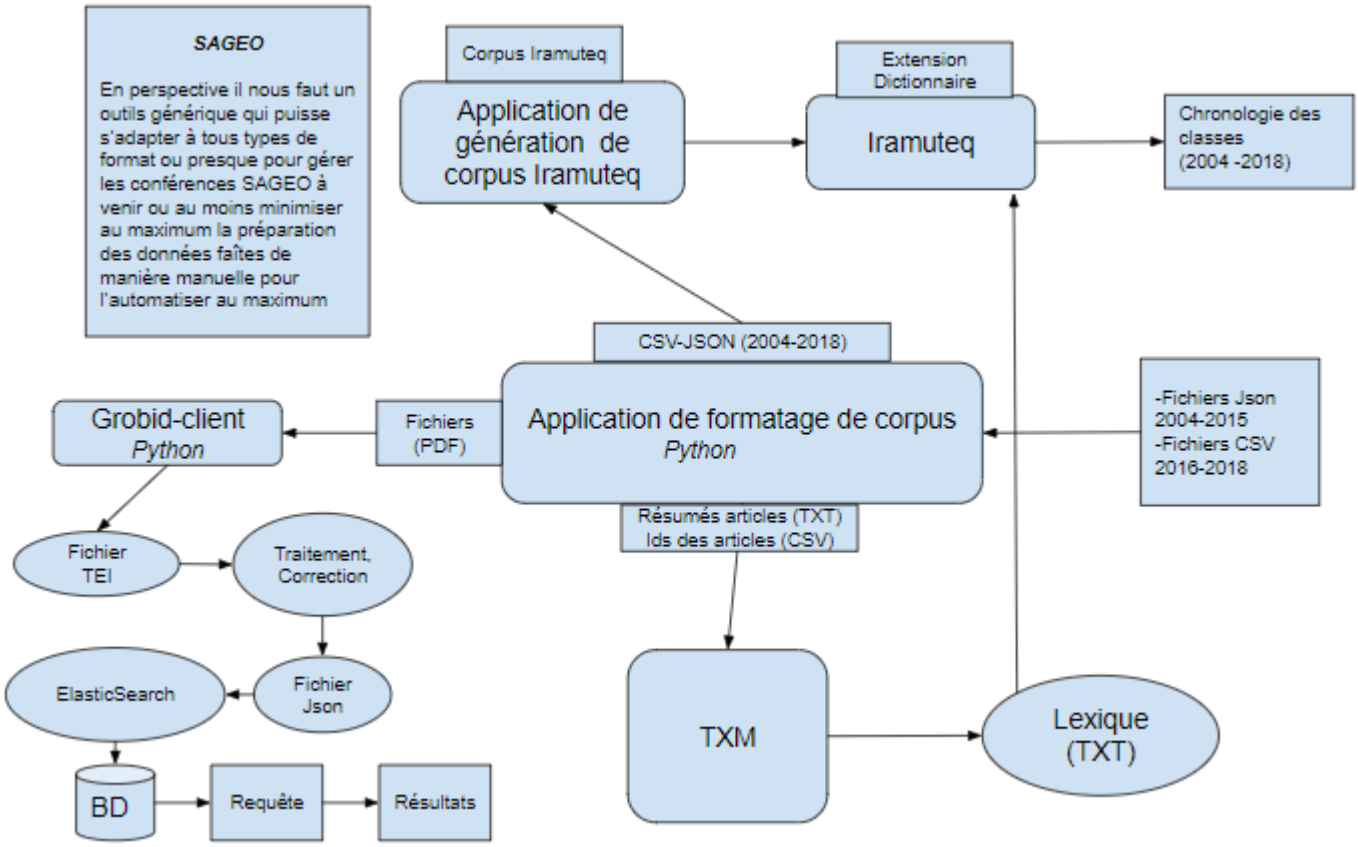


FIGURE 5 – Schéma de réalisation du projet

Comme le montre le schéma, nous partons des fichiers JSON et CSV données par la maîtrise d'ouvrage II, nous lui appliquons une transformationIII qui va être fourni pour analyse avec les outils TXM V et IRAMUTEQ VI.
De ce formatage on extrait les fichiers PDF dans le but de structurer les informations qui y sont(avec XML) avec Grobid IV. Au moment de la rédaction de ce rapport, les maîtres d'oeuvre travaillent sur le traitement/correction des fichiers TEI avec l'outil XSLT et l'intégration des fichiers JSON dans ElasticSearch.

8. Les conférences sont organisé en session.Chaque session étant des sous-parties du thème. Les méta-sessions sont des ensembles de sessions.

II Corpus

Pour débuter notre travail il nous a été fourni plusieurs données sur les conférences de 2004 à 2015 :

— des fichiers au format JSON concernant les conférences de 2004 à 2015 issu de trois répertoire différents.

— la liste des articles , leur identifiant, leur résumé etc...

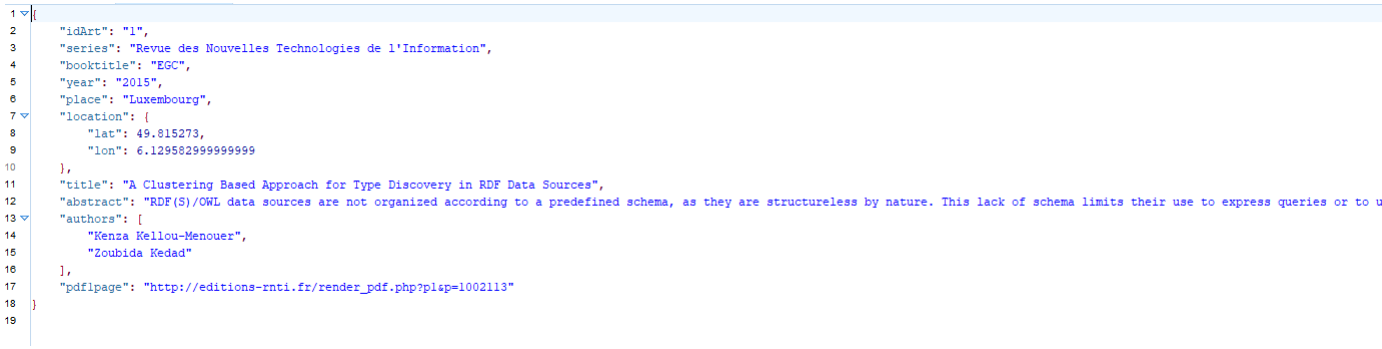


FIGURE 6 – Format des fichiers du répertoire 1

— la liste des articles , leur identifiant et la méta-session correspondante.

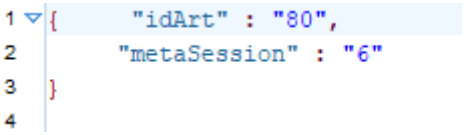


FIGURE 7 – Format des fichiers du répertoire 2

— la liste des articles , leur identifiant et les coordonnées géographiques des auteurs correspondant.

```
{
    "idArt": "2",
    "placeAut": [
        {
            "place": "Clermont-Ferrand",
            "country": "France",
            "location": {
                "lat": 45.77722199999999,
                "lon": 3.0870250000000397
            }
        },
        {
            "place": "Paris",
            "country": "France",
            "location": {
                "lat": 48.856614,
                "lon": 2.3522219000000177
            }
        },
        {
            "place": "Pavia",
            "country": "Italie",
            "location": {
                "lat": 45.1847248,
                "lon": 9.158206899999982
            }
        }
    ]
}
```


FIGURE 8 – Format des fichiers du répertoire 3

— des fichiers CSV contenant les conférences de 2016 à 2018 et leur méta-session.

Id	series	booktitle	year	title	abstract	authors	pdf1page	pdfarticle	MS
1104	Revue des N	EGC	2016	A Relevant P	Les système	Nouha Othm	http://editic	http://editic	2
1105	Revue des N	EGC	2016	Adaptation c	Cette thèse	Julio Cesar D	http://editic	http://editic	13

FIGURE 9 – Format des conférences et leur méta-session

— un fichier CSV contenant les coordonnées géographiques des conférences de 2016 à 2018.

place	Latitude	Longitude
Reims	49.258329	4.031696

FIGURE 10 – Format des coordonnées géographique

Au regard de ces données fournies dans des formats différents et étant inadaptées aux logiciels qui nous ont été imposée par la maîtrise d'ouvrage nous avons décidé de proposer une application capable de préparer correctement les données pour leur exploitations par les différents outils vus précédemment.

III Application de préparation des données

En corrélation avec la chaîne de préparation des données de chaque outils et ayant à disposition une application Python nous avons alors décidés de mettre en place des options de préparation des données pour chaque outils.
Cela nous offre donc la possibilité de traiter rapidement un gros
volume d'informations et ainsi passer ces informations dans chaque outils cités précédemment.
L'enjeu repose ainsi sur le format qu'attend chaque outils en paramètre pour produire des données résultats.
Mais également de l'importance de garder une librairie structurée de l'ensemble des informations.

1 Fonctionnalités implémentées

L'application Python dispose d'un menu permettant de choisir chaque options spécifiques à chaque outils.

```python
#Menu de navigation
if reponse == 'a':
    #CSV vers XML avec recuperation des PDF en ligne
    print("Transformation CSV -> XML")
    #nomFichier = reponse + ".csv
    data = open("Data.csv" , "r")

    #indice des colonnes contenant les informations qui nous interesse
    indiceColPDF = 8
    indiceColMS = 9

    #Param definissant les delimiteurs dans le document CSV
    #print(nomFichier)
    reader = csv.reader(data,delimiter=';')

    #Boucle principale
    #Etape 2 : Trier ce CSV pour ne recuperer que les lignes avec les bons MS
    #Etape 3 : Parmis ces lignes on ne recup que les valeurs lignes par lignes
    #de la colonne contenant le lien vers le fichier PDF
    #Etape 4 : (a)On va recuperer un par un les PDF
    #les telecharger et (b) les passer individuellement sur GroBid TEI
    #pour obtenir un .XML par ligne
    #Etape 5 : Ces .XML vont ensuite etre transformes au format .JSON
    for row in reader:
        if(row[indiceColMS] != 'MS'):
            if float(row[indiceColMS]) >= 1 and float(row[indiceColMS]) <= 9:
                valeur = row[indiceColPDF]
                #apres cela on a plus que les lignes triees par MS
                #print(row)test recup des lignes
                #Etape 3 :
                #valeur contient desormais l'adr du lien PDF du CSV
                print(valeur)
                #Etape 4 (a):
                chaineNom = "pdfArticle"+row[0]+".pdf"
                print(chaineNom)
                chaineNom = "./in/"+chaineNom
                urllib.urlretrieve(valeur,chaineNom)
```

FIGURE 11 – Partie a) de l'application python

Fusion des fichiers Json :
Compte-tenu des données brutes fournies par la maîtrise d'ouvrage qui contenaient trois types de fichiers , nous avons proposé à la maîtrise d'ouvrage , une fusion de ces fichiers pour un traitement optimal.
On obtient donc un fichier concaténant ces trois types de fichiers.


```
{
"year": "2014",
"metaSession": "6",
"pdflpage": "http://editions-rnti.fr/render_pdf.php?pl&p=1001945",
"pdfarticle": "",
"abstract": "Dans cet article, nous proposons un nouveau descripteurspatio-temporel appelé ST-SURF pour l'analyse et la
"title": "Classification des actions humaines basée sur les descripteurs spatio-temporels",
"placeAut":[{
"place":"Paris",
"country":"France",
"location":{
    "lat":"48.856614",
    "lon":"2.3522219"
}
}],
"series": "Revue des Nouvelles Technologies de l'Information",
"location": "{u'lat': 48.117266, u'lon': -1.6777926}",
"place": "Rennes",
"booktitle": "EGC",
"idArt": "80",
"authors": "['Sameh Megrhi','Azeddine Beghdadi','Wided Souidène']"
}
```


FIGURE 12 – Format des fichiers fusionnées

Ainsi pour chaque outils on obtient un format unique et traitable pour chaque outils.

Grobid :

Informations recherchées : Contenu des articles d'un ou plusieurs articles format PDF
Fichiers générés : fichiers d'informations structurées au format TEI.XML pour chaque article.

TXM :
Informations brutes : Ensemble de tous les fichiers json fusionné au nombre de 1330.
Informations recherchées : Contenu des articles au format TXT et un fichier CSV
Fichiers générés : Corpus contenant les résumés en format TXT.

Iramuteq :
Informations brutes : Ensemble de tous les fichiers json fusionné au nombre de 1330.
Informations recherchées : L'ensemble des résumés des articles.
Fichiers générés : corpus texte au format iramuteq.

2 Préparation des données

Pour Grobid

La préparation des données pour Grobid se fît à travers plusieurs étapes.
Celle-ci reposait en effet sur la manipulation des annexes au format .CSV qui nous avaient été fournies.
Le but était donc d'en extraire un ou plusieurs articles afin de télécharger ceux-ci automatiquement au format PDF.
Ainsi il nous été possible de créer une librairie de fichiers structurés comprenant les informations liées à chaques articles de conférences (thème, noms auteurs, résumé, ...).
Afin d'automatiser ce processus il a donc été nécessaire de développer une application et nous avons choisis de développer celle-ci grâce au langage Python.
Cette automatisation était en effet obligatoire car il n'était pas possible depuis l'API Web (http ://cloud.science-miner.com/grobid/) de traiter un grand nombre d'articles en simultanés et cela représentait un temps de travail important pour passer chaque articles un par un de chaque années sur cette API Web.

Pour TXM

La préparation des données en TXM fût la plus fastidieuse . En effet TXM nécessite une réelle préparation des données qui se doit d'être robuste. Concernant l'importation des fichiers, notre étant de travailler sur les conférences EGC de 2004 à 2018, nous avons choisi de nous orienter de façon arbitraire vers une importation au format TXT+CSV 2, ceci étant propice à l'exploitation des résumés que nous voulons mettre en relation avec les années des conférences.
La préparation des données consiste à élaborer un corpus qui est en fait un dossier contenant deux types de fichiers :
— l'un .txt qui contient les résumés des différents articles des conférences
— l'autre .csv qui décrit pour chaque article l'année associée.

NB : Selon l'intérêt on aurait pu rajouter d'autres champs au csv qui aurait pu faire sens de sorte à orienter notre analyse sur d'autres aspects (par exemple montrer les thèmes les plus abordés par auteurs , on aurait alors ajouter un champs *author*.)

Pour IraMuteQ

Les corpus ayant par exemple plus de 1000 texte sont considérés comme étant gros. Créer un corpus à partir de ces textes manuellement serait long et pénible.
C'est dans ce contexte que nous décidé de créer une petite application écrite en javascript permettant de générer automatiquement un corpus texte selon le modèle IraMuteQ.
Cette application prend en entrée l'ensemble de tous les 1330 fichiers JSON et le corpus est généré selon l'encodage demandé par IraMuteQ.

```
if ((data.year == "2018") && (data.abstract != null) && (data.metaSession >=1 && data.metaSession <= 9)  && (info[2]=="French")) {
    $('#zone').append('**** *annee_'+data.year+" "+" "+"*VilleConf_"+data.place+ '<br>' + data.abstract + '<br>');
    console.log(info[2]);
}
if ((data.year == "2017") && (data.abstract != null) && (data.metaSession >=1 && data.metaSession <= 9) && (info[2]=="French")) {
    $('#zone').append('**** *annee_'+data.year+" "+" "+"*VilleConf_"+data.place+ '<br>' + data.abstract + '<br>');
    console.log(info[2]);
}
```

FIGURE 13 – Application de Génération de corpus IraMuteQ

Tous ce travail de préparation de donnée permettra de produire une analyse propre à chaque outil.

IV Grobid

Grobid est une API Web ou Python ou Java
Celle-ci permet d'exploiter plusieurs fonctionnalités tel que l' analyse des documents au format PDF pour en extraire des annotations lexicales.
Cependant, dans notre cas, nous nous intéresserons à son outils de structuration de fichiers PDF au format TEI.XML (le process header).

1 Structuration TEI.XML

Une fois le programme créer il été ainsi facile d'appeler grâce à un serveur local utilisant Gradle l'API de Grobid
. Il ne restait donc plus qu'à faire une automatisation de cet appel pour chaque article. On récupère ensuite chaque fichiers PDF pour ensuite un par un les traiter avec Grobid et ainsi obtenir une librairie de fichiers structurés au format TEI.XML.
Exemple de fichier TEI.XML obtenus :

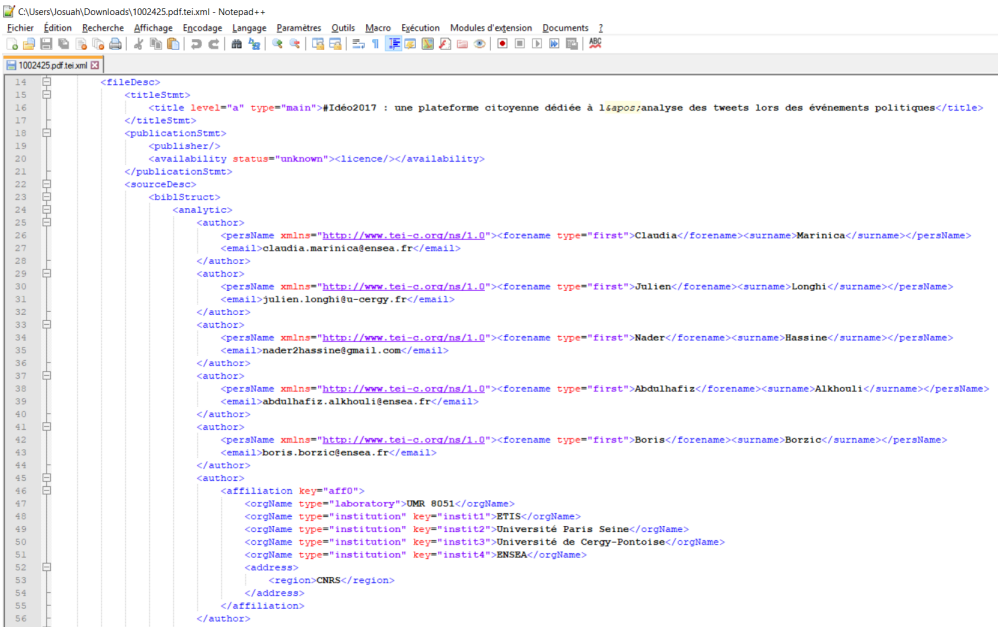


FIGURE 14 – Exemple de fichier TEI.XML généré pour un article

En exploitant ces fichiers TEI-XML issus de Grobid, nous nous sommes donc intéressés au fait d'extraire depuis certaines balises des informations précises.
Dans notre cas une balise nommée "settlement" nous permet pour chaque articles de récupérer le nom de la ville de son ou ses auteur(s) pour ensuite utiliser un service web afin de récupérer des informations comme la latitude et longitude du lieu nommé.

2 Structuration JSON

Le but dans cette tâche visait à automatiser au maximum la création des fichiers JSON grâce aux TEI.XML obtenus précédemment.
Celà nous a ainsi permet de compléter la librairie de fichier JSON de 2004 à 2015 déjà existante qui nous avait été fournies au départ de notre projet.
La difficulté relevait donc de générer des JSON grâce aux informations spécifiques obtenues précédemment dans les articles et de les structurer de la même manière que les JSON qui avaient été mis à notre disposition.

V TXM

Imposé par la maîtrise d'ouvrage car permettant une analyse sémantique, lexicale, grammaticale de corpus. TXM est un logiciel qui permet aussi de voir la proportion des mots, leurs degrés de spécificité, leurs co-occurrences [9], leur concordance [10] et leur progression dans le temps.
La mise en oeuvre des traitements possibles avec TXM permet de déterminer quels seraient les données résultats. Lors de cette recherche, plusieurs pistes ont été avec le temps explorées puis utilisées ou pas pour rendre des résultats pertinents.

9. présence simultanée de deux ou de plusieurs mots dans le même énoncé (la phrase, le paragraphe, l'extrait).
10. Lien existant entre plusieurs mots

1 Étude de la progression

Dans l'optique de pouvoir mettre en évidence l'apparition de certains concepts dans le temps, nous nous orientons vers l'étude de la progression qui permet d'afficher l'évolution d'un ou de plusieurs motifs contenus dans le corpus au fil du temps.

Elle produit au choix un graphique cumulatif ou de densité qui permettent de visualiser une évolution des motifs qu'on veut mettre en valeur.
Comme le montre la figure suivante , on veut montrer la présence du mot Web dans le corpus entre 2004 et 2015. Le graphique cumulatif parcourt le corpus et fait un saut dès qu'il voit apparaître le mot WEB. Dans notre exemple le mot WEB apparaît trois fois.

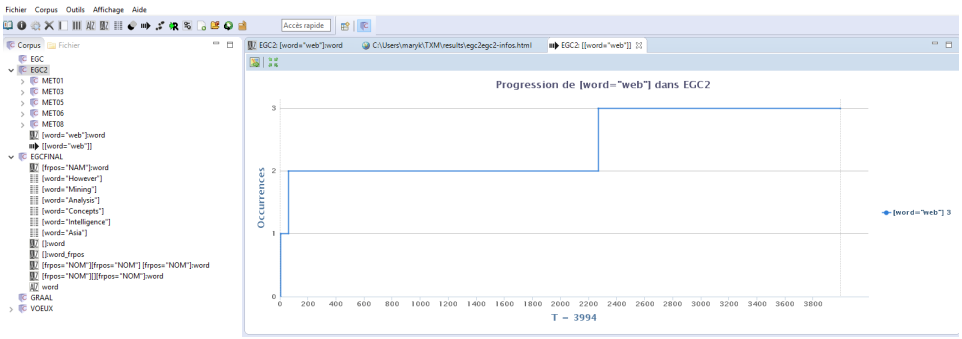


FIGURE 15 – Exemple de progression cumulatif du mot WEB dans le corpus test

Cette piste fût abandonnée car l'analyse du graphique n'était pas propice à des conclusions car superposant les mots de façon aléatoire sans tenir compte de la chronologie, qui est un des critères importants que nous souhaitons mettre en évidence.

2 Les tables lexicales

L'exploitation de cette fonctionnalité devait nous permettre de référencer les fréquences des mots (apparition des mots) par année et ainsi de voir par méta-sessions**??** les mots qui prédominent.
Une table lexicale réunit dans un tableau les fréquences des différentes unités lexicales d'une partition. Pour évaluer cette piste, nous avons composé un corpus test composé de 30 résumés, sur lequel on fait des sous corpus par méta-sessions**??**.
Comme le montre notre figure2, ce corpus test comprend cinq méta-sessions représentés dans ce corpus test.

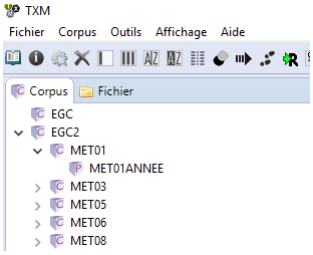


FIGURE 16 – Exemple de divisions de corpus

Dans le figure ci-dessus le corpus test est EGC2. Il est ainsi divisé en sous corpus qui sont MET01 ... MET08 qui eux même sont subdivisés en partitions que sont MET01ANNEE ... MET08ANNEE.
Pour comprendre la différence entre un sous corpus et une partition, on peut imaginer qu'un sous corpus est une coupe horizontale d'un ensemble de données(dans notre cas on prend que les articles d'une méta-sessions données) et une partition une coupe verticale (de chaque méta-sessions on les divise par année).
Cette piste fût aussi mise de côté étant donné que l'analyse proposée par l'outil Ira-MuteQ était plus complète faisant en plus des analyses plus approfondie avec des similitudes , des associations, etc...

3 Données résultats

Après exploration de différentes pistes, le choix des requête CQL s'est avéré étant une plus value pour le travail qu'on souhaite mener.
C'est alors naturellement que nos données résultats ont été produit à l'aide de CQL pour capter certains termes informatiques qui pourraient améliorer l'analyse du corpus.

CQL2 est un langage formel, avec un lexique et une syntaxe d'opérateurs, qui forment un métalangage permettant de combiner des éléments pour la recherche de motifs structurés.

On dispose d'une infinité de possibilités avec les requêtes CQL qui vont nous servir à relever certains termes importants qui peuvent être captés que par une recherche par analyse lexicale.
On a alors choisi ce mécanisme pour établir un lexique qui sera introduit dans l'outil IRAMUTEQ afin de pouvoir enrichir son dictionnaire et permettre une meilleure analyse du corpus.

Les résultats CQL

Avec les requêtes CQL , nous souhaitons extraire les noms propres et des groupes de mots composés pertinents de notre corpus .

Pour les noms propres

— "/p{Lu}.* " : cette requête demande tous les mots commençant par une majuscule.
Elle ne fut pas concluante quant à la très grande quantité de faux positifs [11].

— " [frpos = "NAM"] " : cette requête demande tous les noms propres.
Celle-ci donne de meilleurs résultats mais a quand même nécessité un tri (retrait des faux positifs).
Sur l'ensemble du corpus : cette requête a donné 734 noms propres, qui, après un tri manuel a donné 596 noms propres.

Pour les groupes de mots composés

11. résultat déclaré positif, là où il est en réalité négatif

— ” [ ][ ][ ]” : cette requête récupère tous les groupes de mots composés de trois mots. Nous avons conclu que cette requête était très infructueuse car elle avait plus de 45000 résultats.

— ” [frpos=”NOM”][frpos=”NOM”][frpos=”NOM”] ” : cette requête est pratiquement la même que la première , la subtilité est qu’elle précise la catégorie grammaticale des mots.
Ces résultats ne sont pas très concluant car contenant beaucoup de faux négatifs [12] avec un résultat de 383.

— ”[frpos=”NOM”][ ][frpos=”NOM”]” : cette requête récupère tous les groupes de mots composés de trois mots en précisant qu’on veut cette fois deux noms liés par n’importe quel mot.
Les résultats sont alors de meilleures qualités. On obtient 4149 résultats.

4 Conclusion

Finalement, l’outil TXM a permis d’obtenir deux fichiers .txt (nomsPropres.txt et nomsComposes.txt) qui sont insérés dans IraMuTeQ dans le but d’enrechir son dictionnaire et permettre une meilleure analyse du corpus. En effet grâce à TXM, IraMuteQ pourra reconnaître des mots propres à certains concepts informatique.
Au vue de la volonté de départ qui était de pouvoir faire une analyse directement sur l’outil TXM, il n’a pas tout à fait rempli sa mission par rapport aux attentes du départ2.

12. résultat déclaré négatif, là où il est en réalité positif

VI Iramuteq

1 Extension du dictionnaire IraMuteq

Le dictionnaire de base du logiciel IraMuteQ est certes utilisable mais inadapté à l'ensemble des textes que nous lui soumettons.
En effet, certains mots techniques sont absents de ce dictionnaire ou encore il existe des mots composés que le dictionnaire ne prend pas en compte.
Pour une meilleure analyse , il est donc obligatoire et nécessaire de procéder à cette extension du dictionnaire d'IraMuteQ afin d'avoir des classes[1] plus significatives.

2 Génération d'un corpus IraMuteQ

Les documents donnés en entrées sont des corpus de textes encodé de la manière suivante :

```
**** *annee_2015  *VilleConf_Luxembourg
texte texte texte
texte texte texte
**** *annee_2015  *VilleConf_Luxembourg
texte texte texte
texte texte texte
```

FIGURE 17 – **Exemple de corpus IraMuteQ**

A noter que :

******** : introduit chaque texte
***annee_2015** : crée une variable **annee** ayant la valeur **2015**.

Nous avons nettoyé le corpus en déconcaténant certains éléments de phrases.
En effet il y a certains résumés de conférences qui avait des mots concaténés.
Ces mots concaténés ne seront pas pris en compte dans IraMuteQ et pourtant pourraient être bien utiles.
Le corpus étant complètement propre, l'analyse portera sur les résumés des conférences par année et par ville de conférences afin de suivre l'évolution des thématiques au fil du temps.

3 Données résultats

IRaMuTeQ affiche les résultats de chaque analyse et génère un répertoire ou dossier dans lequel il place des fichiers résultats.

Analyse statistique

Le but de cette analyse est de visualiser les formes ou mots que le logiciel reconnaît, leur occurrence dans l'ensemble des texte, et leur classement par ordre décroissant.
Ce type d'analyse permet d'avoir plus de lisibilité en matière de compréhension et d'analyse du texte.
Il faut d'abord choisir de lemmatiser ou non les formes/mots et paramétrer les catégories de paramétrage de la lemmatisation

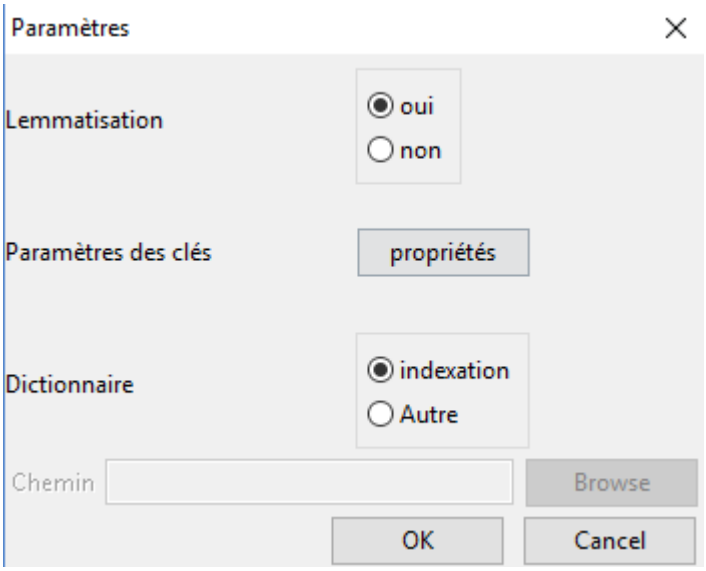


FIGURE 18 – Paramétrage de la lémmatisation

La lemmatisation[3] permet de réduire les verbes à leurs formes infinitives.
Le logiciel fait une lemmatisation à l'aide de ses dictionnaires.
Le paramétrage des clés 3 est une propriété qui permet de modifier les clés d'analyse par catégories et de différencier le traitement de certaines formes.
Dans notre cas, nous choisissons le paramétrage des clés par défaut car il est adapté à l'analyse qu'on souhaite réaliser.

3. les occurrences sont réduites à leur racine : les verbes sont ramenés à l'infinitif, les noms au singulier et les adjectifs au masculin singulier.

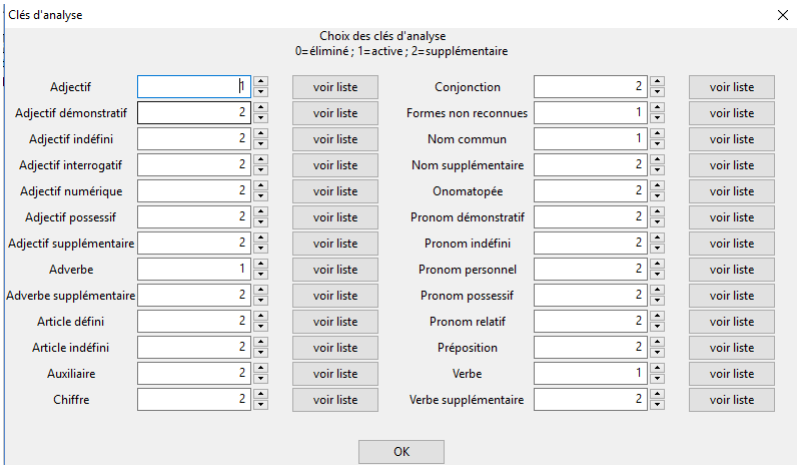


FIGURE 19 – Paramétrage des clés d'analyses

— Ce qui est mis en actif par défaut (codé 1) : adjectifs, adverbes, formes non reconnues, noms communs et verbes.
— Ce qui est mis en supplémentaire par défaut (codé 2) : mots outils.
Attention l'option « voir liste » affiche des exemples qui ne correspondent pas aux mots du corpus analysé.

Un mot qui n'est pas dans le dictionnaire est mis dans la catégorie Formes non reconnues. Une fois le paramétrage validé (OK), IRaMuTeQ affiche les résultats et génère un répertoire (ou dossier) dans lequel il place des fichiers résultats.

•1er onglet :
Le résumé est la description générale du corpus (nombre de textes du corpus, d'occurrences, de formes. . .)

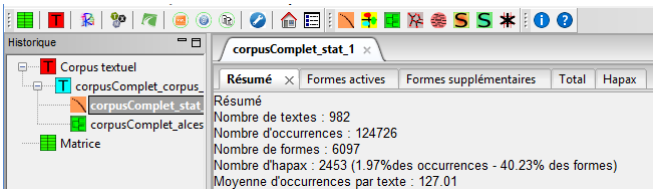


FIGURE 20 – Résumé

•2em onglet :
Les formes actives sont des liste de formes/mots actifs (avec leur catégorie grammaticale) par fréquences décroissantes.

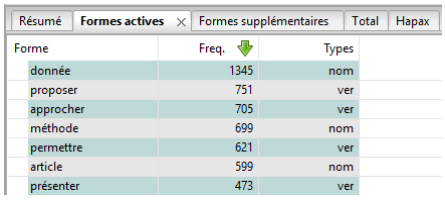
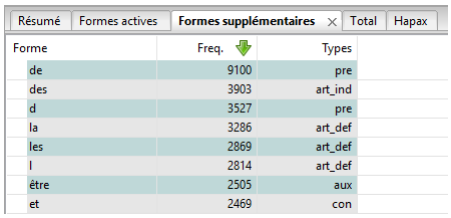


FIGURE 21 – Forme active

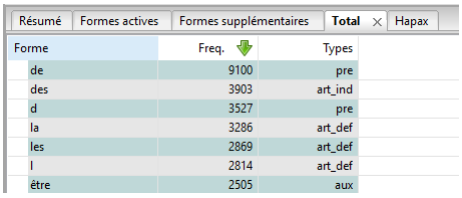
•3em onglet :
Les formes supplémentaires concerne la liste des formes/mots supplémentaires par fréquences décroissantes



Forme	Freq.	Types
de	9100	pre
des	3903	art_ind
d	3527	pre
la	3286	art_def
les	2869	art_def
l	2814	art_def
être	2505	aux
et	2469	con

FIGURE 22 – Forme supplémentaire

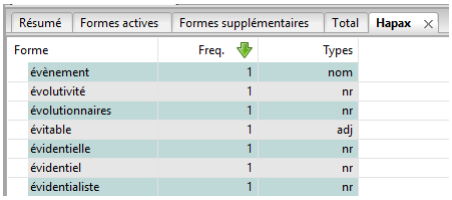
•4em onglet :
Le total est l'ensemble des mots par fréquences décroissantes



Forme	Freq.	Types
de	9100	pre
des	3903	art_ind
d	3527	pre
la	3286	art_def
les	2869	art_def
l	2814	art_def
être	2505	aux

FIGURE 23 – Forme supplémentaire

•5em onglet :
L'Hapax montre les mots du corpus présents une seule fois.



Forme	Freq.	Types
évènement	1	nom
évolutivité	1	nr
évolutionnaires	1	nr
évitable	1	adj
évidentielle	1	nr
évidentiel	1	nr
évidentialiste	1	nr

FIGURE 24 – Forme supplémentaire

4 Nuage de mots 3

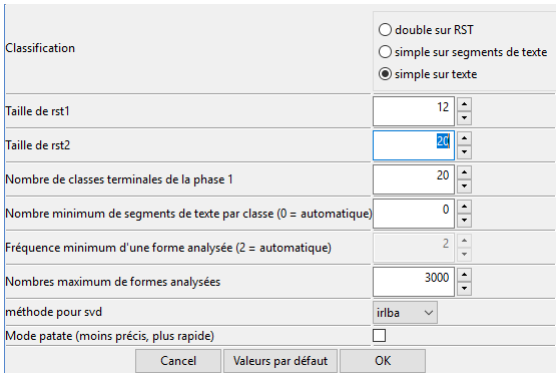
Cette analyse est assez simple et représente juste une illustration à partir des données que l'on utilise. Le but étant de découvrir l'ensemble des mots avec le plus grand nombre d'occurrences déterminées par la taille de la police.

L'avantage est qu'on peut voir dans l'ensemble les mots dominants.

5 Classification 3

L'objectif ici est de pourvoir donner une chronologie des classes par année. La classification de Reinert permet de classer les formes dans des classes de formes regroupées. Ces mêmes classes peuvent alors être représentées à l'aide de différents arbres, comme ici avec des phylogrammes [13] 5 des classes lexicales. Ce diagramme fournit la liste des formes les plus associées pour chaque classe.
A noter qu'une forme peut se retrouver dans plusieurs classes différentes.
Il s'agit pour nous ici de retenir les classes les plus importante et les plus significatives de notre corpus.

Implémentation de la méthode de classification «Alceste» de Max Reinert



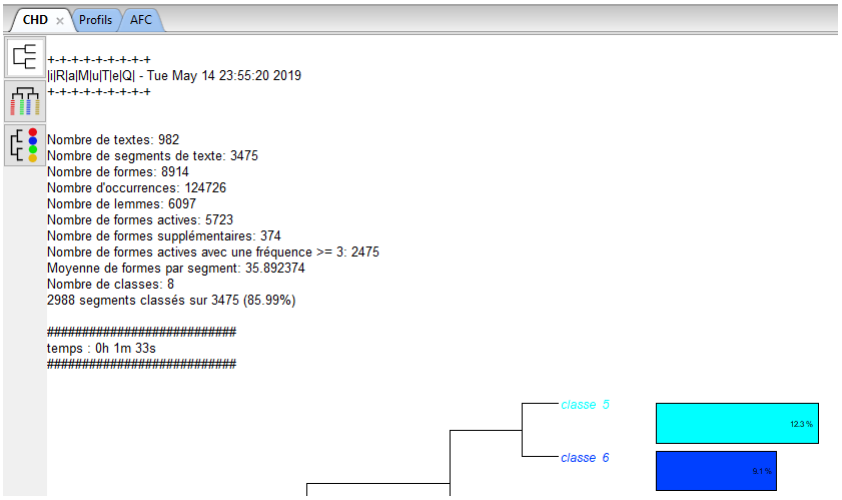
Dans notre exemple, on a choisi d'opérer une classification simple sur textes car les textes sont très courts.
Par défaut, la méthode propose de découper les textes en segments de textes en

13. https://fr.wikipedia.org/wiki/Phylogramme

fonction du nombre de formes actives.
Modifier le nombre de classe de la phase terminale comme vu sur l'image ci dessus modifie le nombre de classes.
Ceci serait avantageux pour l'étude car cela permettra de voir un plus grand nombre de formes associés aux classes.

Sortie résultats de la classification :

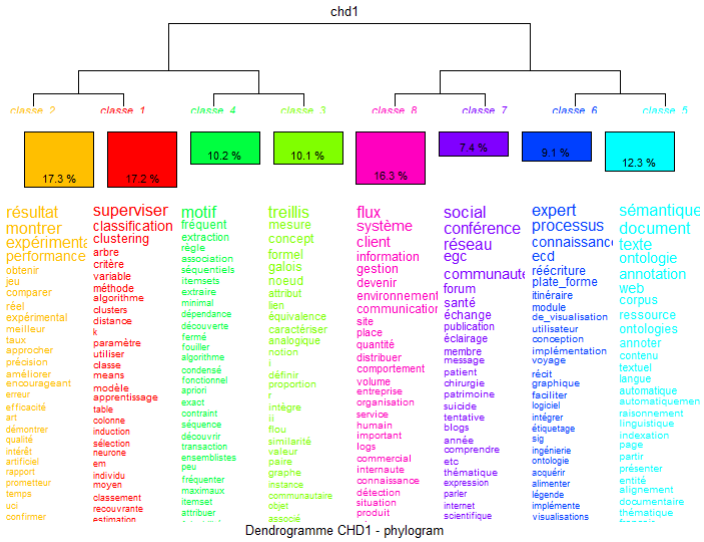
Onglet CHD



On trouve un résumé des résultats (nombre de textes, de formes, de classes, le pourcentage de textes classés et le dendrogramme).

Pour chaque classe, on trouve les formes/mots les plus associés (effectifs, pourcentage) .

Pour chaque classe on affiche les mots spécifiques des classes pour aider à leur interprétation comme montre le graphique suivant :



Les classes 3,4,7 et 6 étant peu significatives pour nous procédons à un tri en les enlevant de l'analyse.
En effet derrière chaque classe se trouve un thème qui lui est associé et qu'on peut cerner grâce aux concordancier [14].
Par exemple la figure ci-dessous montre les textes associés au formes de la classes.

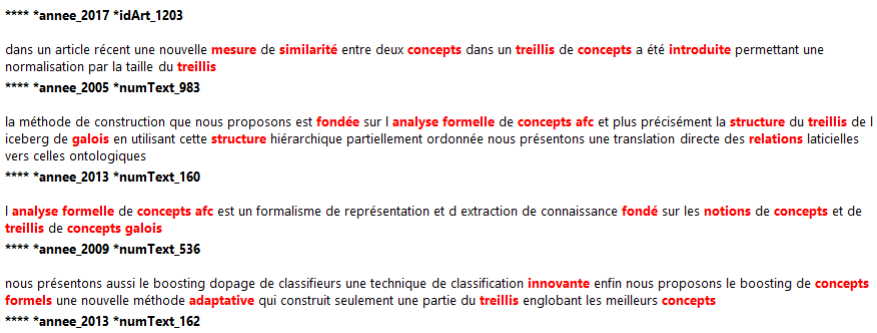


FIGURE 25 – Concondorcier de la classe 3

Il est difficile de trouver le ou les thèmes associés à cette classe étant donné les grosses différences thématique entre les textes.
A partir donc des classes 1,2,5 et 8 le logiciel nous permet de générer un nouveau corpus avec les même textes mais cette fois en ce basant sur ces classes.
On obtient le graphique suivant : Grâce à l'option concordancier il est possible de mieux comprendre le sens derrière une classe .
On ressort don les thèmes liés à chacune des classes.
— Classe 1 : Reprend les thèmes liés aux données textuels , leur sémantique.
— Classe 2 : Recherche d'images
— Classe 3 : Méthode de classification avec arbre
— Classe 4 : Intérêt des expérimentations sur des jeux de données.
— Classe 5 : Mécanisme de traitement de l'information
— Classe 6 : Stockage de données

14. fonctionnalité du logiciel permettant de visualiser dans les textes du corpus les mots associés à la classe

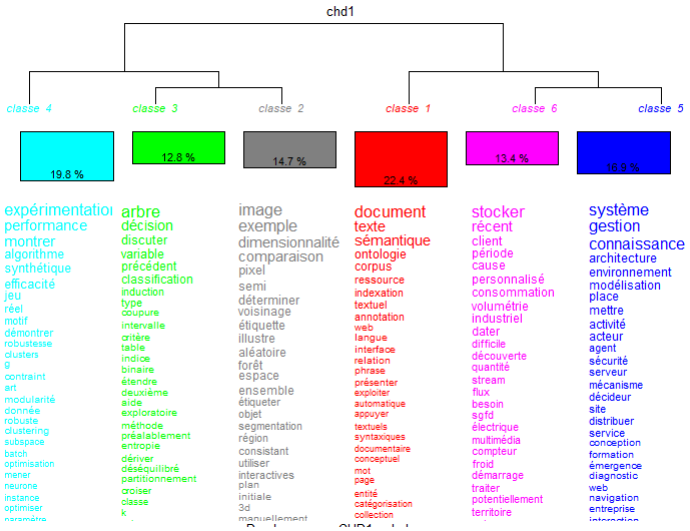
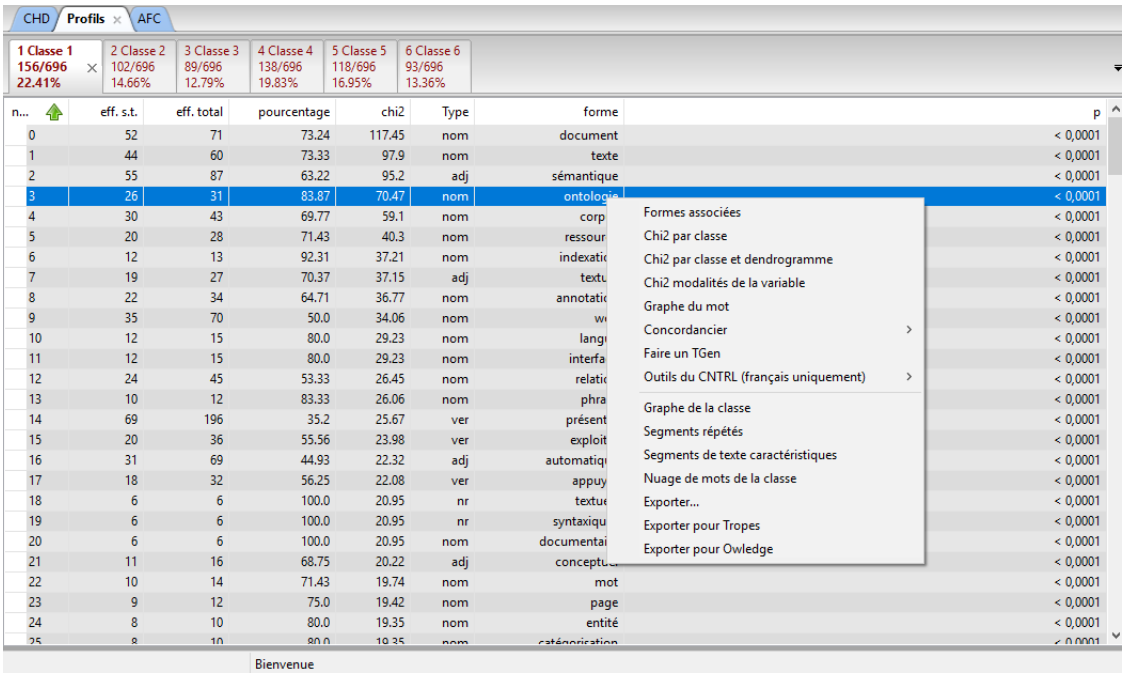


FIGURE 26 – Classification Méthode Reinhert

Onglet Profils



On accède à d'autres menus offrant des aides à l'interprétation des classes :

— Les formes associées : Donne les effectifs dans la classe des formes regroupées dans un lemme

— Chi2 Modalités de la variable : crée un graphique qui représente le chi2 [15] d'association des modalités de la variable sélectionnée à chacune des classes. Nécessite un formatage du type variable_modalité.

— Outils du CNRTL : renvoie sur le site du Centre National de Ressources Textuelles et Lexicales et pour cette forme affiche (définition, étymologie, synonyme) si la langue du corpus est le français)

— Concordancier : propose le concordancier de la (ou des) forme(s) / lemmes sélectionnée(s).

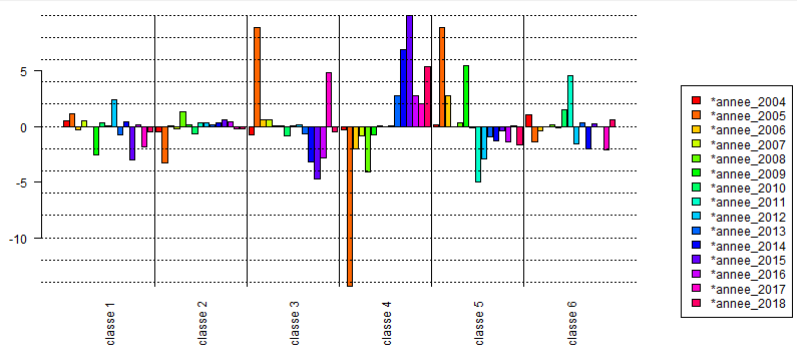
15. fonction Test d'association du Khi deux pour déterminer si deux variables de catégorie sont associées

Ce concordancier est disponible pour les segments de texte de la classe, ceux classées ou tous ceux du corpus.

Nous nous intéressons ici à l'option Chi2 Modalités de la variable très important pour observer une évolution dans le temps des thèmes associés aux classes.

Évolution des thèmes dans le temps

Après une manipulation de l'option Chi2 Modalités de la variable un graphe montrant la présence des classes par année est généré. Il se présente comme suit :



Toute les années en dessous de 0 n'ont pas la classe correspondante dans les résumés des conférences qui leur sont associé. Par exemple, la classe 3 est très représentée en 2005 et en 2017.

6 Conclusion

Nous avons trouvé ce travail très enrichissant, malgré le démarrage lent dû au retard de l'obtention de certains textes. Nous avons pu travailler enfin sur un corpus entier de texte, et non pas un seul . Grâce à la variable année, nous avons pu réaliser des analyses temporelles et créer des sous corpus par classes. L'analyse du sous-corpus par classe reflète bien les thèmes qui ont été les centres d'intérêts des conférences EGC allant de 2004 à 2018.

VII Bilan du développement

Finalement , la réussite de notre analyse a reposé sur le travail conjoint de trois outils différents qui se sont peu à peu complétés et ont permis de dégager des conclusions quant aux thèmes récurrents entre 2004 et 2018.

Cinquième partie

Conclusion

I Bilan du projet

Somme toute, nous avons réussi à faire une analyse chronologique par thème des différents articles en utilisant leur résumé.
Vis à vis du travail demandé I, nous pouvons établir que l'élargissement des travaux[KBS+16] et son automatisation ont été réalisé.
Cette analyse a pris en compte des termes nouveaux et propres à l'environnement informatique. Une structuration parfois une restructuration des fichiers passant par leur transformation a été bénéfique pour accomplir les différentes tâches.
Nous avons pu ainsi dégager une chaîne de traitement générique qui peut être abordé pour tout autre corpus semblable. L'automatisation pratiquement complète du travail a été proposée , validée et abordée de sorte à permettre une amélioration, une réutilisation des concepts. A notre sens l'apprentissage, la compréhension du domaine considérée et le développement d'une chaîne de traitement en utilisant les deux logiciels de textométrie, ont été effectué.
Quant à la visualisation des résultats grâce à un outils, elle n'a pu être réalisé, du fait du délai imparti. Aussi l'application de notre chaîne de traitement à d'autres conférences telles que SAGEO [16] est aussi a envisagé.

II Expériences acquises

La réalisation de ce projet a permis de découvrir un aspect de notre formation très important qui est le travail d'équipe et la collaboration directe avec une maîtrise d'ouvrage.
La découverte intégrale d'un outil informatique qui traite d'un sujet peu connu, qui nécessite d'être appréhendée de façon complète, fût pour nous une expérience très bénéfique. Aussi la découverte et l'apprivoisement de nouvelles perspectives informatiques (python, la gestion de projet informatique) ont été un plus pour parfaire notre formation.
Ainsi nous avons acquis une certaine persévérance dans la recherche de documentation parfois diverses et pêle-mêle , une rigueur dans le travail nécessaire afin de respecter les délais imposés, les consignes et le travail des membres de l'équipe.
Finalement nous avons non seulement appréhender ce qu'était une analyse sémantique, mais en plus nous avons découvert de nouveaux outils , de nouvelles façons de travailler, de nouveaux termes, tout simplement de nouvelles perspectives.

16. la conférence internationale francophone SAGEO est un événement annuel majeur dans le paysage de la Géomatique, de l'analyse Spatiale et des Sciences de l'information Géographique.

Bibliographie

[eBG] Elodie Baril et Bénédicte Garnier. Iramuteq 0.7 alpha 2. http://www.iramuteq.org/documentation/fichiers/Pas%20a%20Pas%20IRAMUTEQ_0.7alpha2.pdf.

[Heia] Serge Heiden. Initiation txm. https://www.youtube.com/watch?v=qNUxOI8vfU4&t=3251s.

[Heib] Serge Heiden. Manuel de txm. http://textometrie.ens-lyon.fr/files/documentation/Manuel%20de%20TXM%200.7%20FR.pdf.

[KBS+16] Eric Kergosien, Marie-Noëlle Bessagnet, Christian Sallaberry, Annig Le Parc-Lacayrelle, and Albert Royer. Analyse géographique de séries de publications : application aux conférences EGC. In *16ème Journées Francophones Extraction et Gestion des Connaissances, EGC 2016, 18-22 Janvier 2016, Reims, France*, pages 371–382, 2016.

[Lopa] Patrice Lopez. Documentation de grobid. https://grobid.readthedocs.io/en/latest/Introduction/.

[Lopb] Patrice Lopez. Manuel d'installation de grobid. https://grobid.readthedocs.io/en/latest/Install-Grobid/.

[Lopc] Patrice Lopez. Répertoire git du service grobid en python. https ://github.com/kermitt2/grobid-client-python.

[Wik] Wikipedia. Adt. https://fr.wikipedia.org/wiki/Analyse_de_donn%C3%A9es_textuelles.