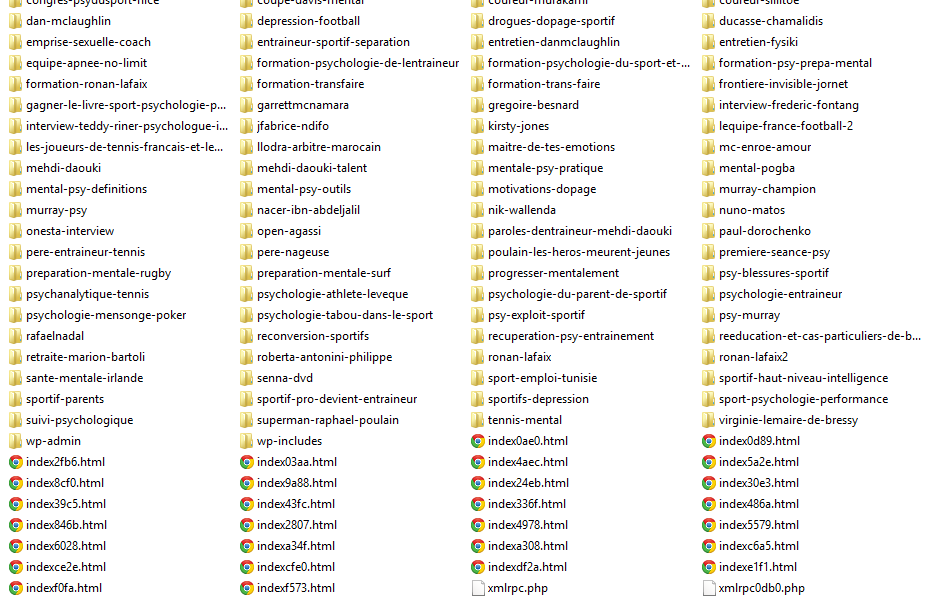
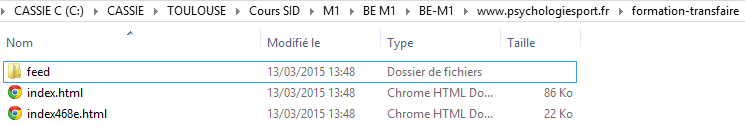
Après avoir récupéré tous les articles du web grâce au logiciel HTTRACK, nous devions trier ces résultats pour en retirer les articles utiles.



En étudiant les résultats retournés, sous forme de dossiers (un dossier par article), nous avons remarqué que seuls les fichiers HTML nommés index.html étaient utiles.

Les fichiers en questions n’étaient pas stockés dans un dossier spécifique.

Par exemple, le fichier index.html pouvaient être présent directement à l’ouverture du dossier correspondant à l’article.



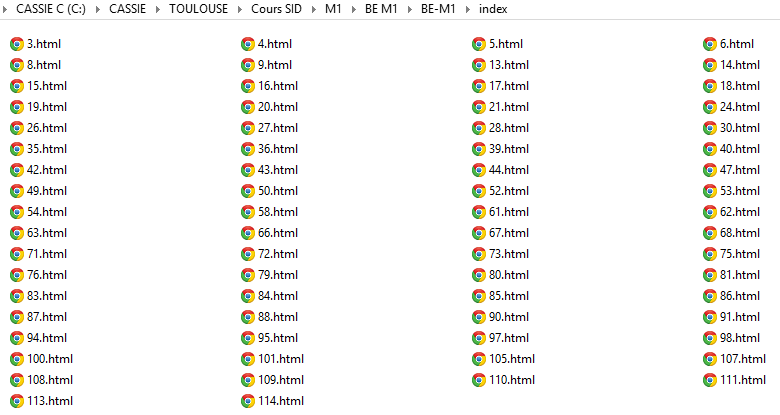
Dossier correspondant à l’article

Fichier HTML à ne pas récupérer

Fichier HTML à récupérer

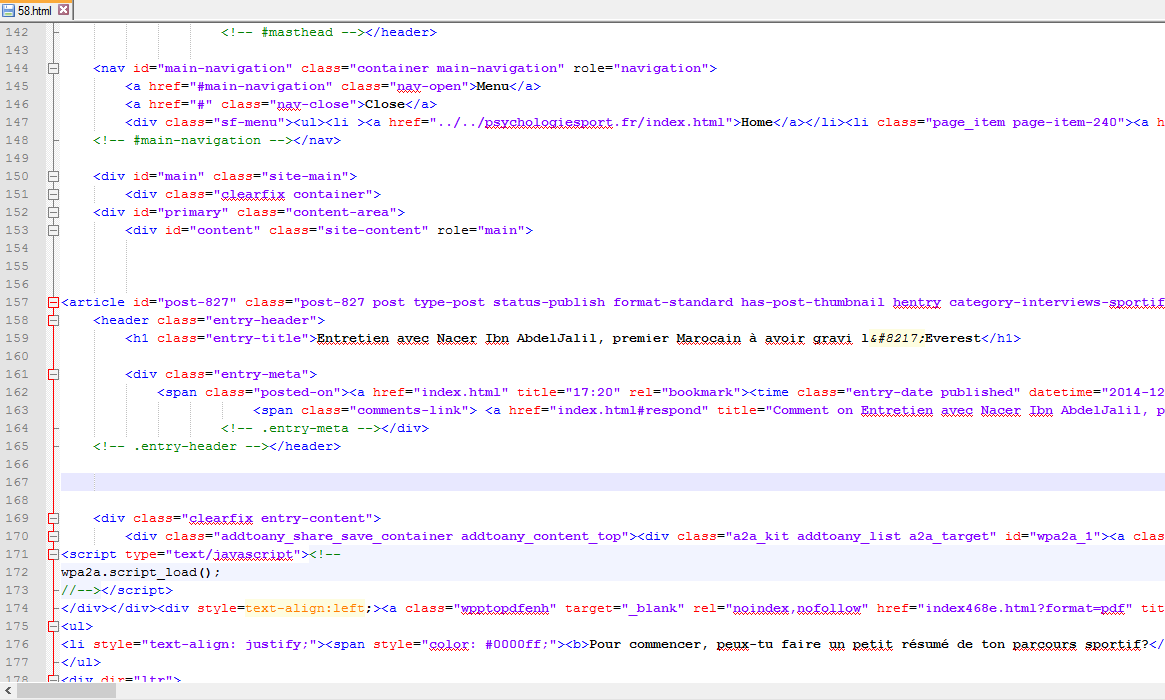
D’une autre manière cet index.html pouvait se retrouver dans un sous-dossier ou encore un sous-sous-dossier. L’emplacement était aléatoire. De plus, dans chaque dossier, un sous-dossier nommé « feed » est présent, avec, à l’intérieur un fichier « index.html » vide à ne pas garder.

Un programme perl a donc été codé : « lebon.pl ». Ce programme parcourt tous les articles (donc tous les dossiers) et récupère le bon fichier « index.html » pour le stocker dans un autre dossier qui contiendra tous les « index.html » de chaque article. Par ailleurs, chaque fichier récupéré est renommé avec un nombre incrémentant pour chaque « index.html ».



Les fichiers index.html rapatriés et renommés, une récupération du texte utile de ces pages était primordiale.

La forme de nos fichiers HTML ressemble à l’image suivante :



Nous avons réussi à identifier que le texte des articles (souligné en orange) était stocké entre les balises <article>…</article> (encadré rouge sur la figure).

Un programme « Extract.pl » a été mis à notre disposition pour extraire le texte entre certaines balises dans un fichier .html. Ce programme nous a servi de base et, par suite, l’approprier pour notre cas (fichier « ExtracModif.pl »). Pour simplifier les extractions des informations désirées et ne pas augmenter le risque de perte d’informations, nous avons réduits tous les fichiers HTML sur une seule et même ligne grâce au programme perl « ligne.pl ».

Les titres (balises <title>…</title>), contenus des articles (<article>…</article>) ainsi que les dates de publication (<time>…</time>) ont été stockés dans un fichier texte résultat. Ces résultats serviront à une future analyse sémantique des articles.

FUSIONS DES BASES DE DONNEES

Le sujet de notre analyse empruntant plusieurs axes thématiques tels que le sport ou encore la psychologie, nous a permis de récupérer des articles répartis sur plusieurs bases de données, plus précisément depuis trois bases :

* Pubmed : spécialisée dans la santé
* PubPsych : spécialisée dans la psychologie
* ProQuest : spécialisé dans la psychologie

Or, les formats des données récupérées différaient suivant la base de données.

|  |  |  |
| --- | --- | --- |
| Pubmed | PubPsych | ProQuest |
| IDENTIFICATEURS |  |  |

Notre objectif était donc de :

* convenir d’un paramétrage commun aux trois bases de données
* écrire un programme perl (« bd.pl ») pour récupérer les identificateurs (certains devaient être renommés) et leurs informations associées
* Fusionner les trois bases de données en une commune pour faciliter l’intégration sur le logiciel « Tétralogie »

Lors de la fusion des trois bases nous voulions tout de même pourvoir différencier les articles venant des différentes bases. Pour ce faire, nous avons convenu d’adopter un code spécifique à chaque base identifiant chaque article.

Les numéros articles provenant de la base PubMed ne changent pas tandis que ceux de PubPsych et ProQuest ont respectivement les caractères ‘PP’ et ‘PQ’ ajoutés.

|  |  |  |
| --- | --- | --- |
| PubMed | PubPsych | ProQuest |
| PMID- 25648198  PMID- 25631930  PMID- 25617538 | PMID- PP91  PMID- PP92  PMID- PP93 | PMID- PQ201  PMID- PQ202  PMID- PQ203 |

Ainsi, grâce à cette convention nous pouvons savoir de quelle base de données provient l’article en question.

Ci-dessous, les conventions adoptées pour les trois bases de données :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | PubMed | PubPsych | ProQuest | Après renommage |
| Titre | TI | TI | Titre | TI |
| Auteur | AU | AU | Auteur | AU |
| Date de publication | DP | PY | Date de publication | DA |
| Langue | LA | LA | Langue | LA |
| Abstract/Résumé | AB | ABHR | Résumé | AB |
| Mots clés | OT+MH | CTEP | Sujet + Identificateur | MC |
| Adresse | AD | CS | Auteur | AD |
| Pays |  | COU |  | PY |
| Email |  | EMAILO | Adresse courriel | EM |

Utilisation de TROPES

Qu’est-ce que le contenu d’un texte ? Ou plus précisément qu’est-ce qui est essentiel pour comprendre le sens d’un texte ? Quels mots ou couples de mots ressortent le plus dans un texte ou dans un groupe te textes ?

Chaque texte, qu’il s’agisse d’un article, d’un ouvrage, d’un discours, contient quelques phrases clés correspondant aux idées qui constituent son ossature, son squelette. Le problème dès lors est d’atteindre ces informations du texte qui contiennent l’essentiel du sens.

D’une autre manière, un texte (article dans notre cas) est écrit d’une certaine manière et est composé d’une façon telle que certains mots ou groupes de mots peuvent être associés (couplés) dans le but de représenter une idée particulière. De plus, la structuration de l’article ces mots et groupes de mots a un effet sur leur signification. Certains d’entre eux auront une haute importance dans la compréhension de l’article tandis que ces mêmes termes retrouvés dans un autre texte auront un poids beaucoup plus faible. C’est pour cette raison que le retrait de termes avec un poids particulièrement fort aura un impact sur l’analyse du corps de texte.

L’analyse de contenu permet de connaître :

* Quels sont les principaux acteurs à l’œuvre dans le texte ?
* Quelle est la structure des relations qui les lient ?
* Quelle est la hiérarchie de ces relations et leur évolution ?

Globalement, l’analyse de contenu consiste à faire apparaître le sens du texte. L’outil Tropes a été conçu pour répondre à toutes ces problématiques. Tropes est un logiciel libre d'analyse sémantique de textes.

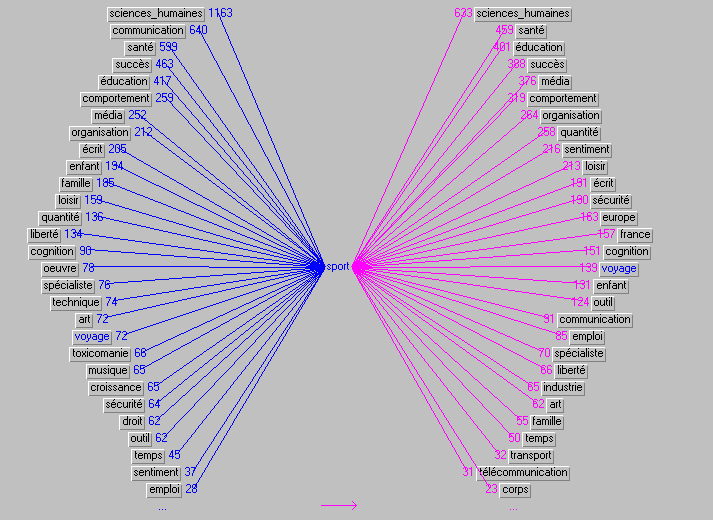
Dans notre cas, nous utiliserons Tropes pour :

* Repérer les liens entre les différents mots significatifs
* Discerner si certains mots ou groupes de mots ont un intérêt particulier
* Répondre à notre problématique de sujet : quelles informations, idées sont identifiées comme significatives par Tropes ? Ces informations pourront-elles nous aider dans notre étude ?

Une analyse textuelle est un atout supplémentaire pour répondre à nos questions. Cette méthode, ajoutée aux analyses effectuées sur l’outil Tétralogie ainsi qu’aux récupérations de données complètera nos résultats obtenus au préalable et permettra de confirmer les idées avancées.

Tropes propose de nombreuses représentations graphiques pour illustrer les résultats obtenus.

Par exemple le schéma ci-dessous présente les termes associés au mot « sport » :



Les liens en bleu et rose correspondent aux termes principaux rencontrés respectivement avant, après le mot « sport ». Le nombre d’apparition se trouve à côté du terme.

Par exemple, le terme « succès » est retrouvé 463 fois avant le mot « sport ».

Tropes a donc jugé utile de faire apparaître ce terme avec « sport » car le nombre d’occurrences est significatif. Il faut noter que cet outil n’analyse pas en profondeur le texte. Des termes, couples de mots sont présentés mais la suite de l’analyse doit être réalisée à la main.

Notre travail sera donc de :

* Garder les couples de termes significatifs. « télécommunication » n’est retrouvé que 31 fois après « sport », ces apparitions sont donc surement purement hasardeuses.
* Se poser les questions utiles pour pousser l’analyse en profondeur pour certains cas. Par exemple il serait intéressant de nous demander pourquoi à la suite du mot « sport », « cognition » apparaît 151 fois. La cognition dans le sport est-elle importante ? Serait-t-elle une réponse possible à la manière dont les sportifs sont entraînés mentalement ?

Cet outil nous a donc été d’une grande utilité dans l’avancement de notre analyse.