

TPE

MISE EN PLACE D'UN MODÈLE D'EXTRACTION D'INFORMATION DE DONNÉES TEXTUELLES EN GENOMIQUE

Réalisé par: APEDO Kodzo Sitsofe Degnon , Promo 21, SIM

Encadrants:

Mr Pierre Larmande

Mme Bich Hai Ho

Mr Ho Tuong Vinh

INSTITUT FRANCOPHONE INTERNATIONAL (IFI)

Hanoi, 03-04/01/2018

DÉVELOPPEMENT D'OUTILS DE RECHERCHE D'INFORMATION EN GÉNOMIQUE LE RIZ «ORYZA SATIVA»

MISE EN PLACE D'UN MODÈLE D'EXTRACTION D'INFORMATION DE DONNÉES TEXTUELLES EN GENOMIQUE

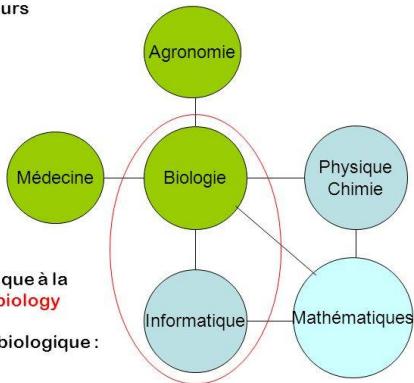
PLAN DU TRAVAIL

- 1 CONCEPT ET DOMAINE D'ÉTUDE
- 2 INTRODUCTION
- 3 ANALYSE DU SUJET
- 4 ÉTAT DE L'ART
- 5 PROBLÉMATIQUE
- 6 PROPOSITION DE SOLUTION
- 7 IMPLEMENTATION
- 8 TESTS ET VALIDATION
- 9 CONCLUSION ET PERSPECTIVES
- 10 RÉFÉRENCES

BIO-INFORMATIQUE

Qu'est ce que la bioinformatique ?

Domaine de recherche passionnant
qui interagit avec plusieurs
disciplines



Application de l'informatique à la
biologie : **computational biology**

Analyse de l'information biologique :
bioinformatics

CONCEPT ET DOMAINE D'ÉTUDE

WEBCRAWLER

Web-crawler : Web-crawler est un méta-moteur qui rassemble les meilleurs résultats d'autres moteurs de recherche.

INDEXATION

L'indexation automatique de documents est un domaine de l'informatique et des Sciences de l'information et des bibliothèques qui utilise des méthodes logicielles pour organiser un ensemble de documents et faciliter ultérieurement la recherche de contenu dans cette collection.

ANALYSE DE DONNÉES TEXTUELLES

Analyse de données textuelles : La recherche qualifiée des éléments du texte à l'aide de catégories et de la recherche quantifiée en analysant la répartition statistique des éléments du texte.

INTRODUCTION

CONTEXTE DU MOTEUR DE RECHERCHE

Documents

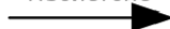


Indexation



Moteur de recherche

Recherche



Résultats



Besoin
d'information

Requête



CONTEXTE DU TEXT MINING

Les résultats de recherches biologiques ou biomédicales sont des documents, articles et bibliographies consignés au sein des vastes bases de données spécialisées sous différentes formats selon les projets des différentes communautés.

Ces bases de données se trouvent sur des plateformes logées sur le Net. Cependant, il n'existe pas d'outils automatiques permettant des les explorer et d'en extraire des informations pertinentes.

Alors la connaissance et l'extraction d'information dans ses bases représente un grand défi.

DOMAINES D'ÉTUDE

- Bio-informatique
- Indexation
- Fouille de données :Text Mining

LES EXIGENCES DU PROJET

- Conception d'un modèle d'extraction d'informations dans un texte de données biologique
- Les exigences(Language R et ses librairies)

ANALYSE DU SUJET

TRAVAUX PRATIQUES A RÉALISER

- Conception du modèle d'analyse de texte
- Validation du modèle
- Analyse du résultat


DIFFICULTÉS PRÉVUES

Outils existants et comparaison des résultats

LES DIFFÉRENTS POINTS ET DOMAINES DONT PARLE LE SUJET

- Bioinformatique
- La génétique végétale
- WebCrawling
- Traitement Automatique du Langage Naturel

GRAMENE



Portails Gramene



Navigateur Génome
Parcourir les génomes avec des annotations, des variations et des outils comparatifs



Plant Reactome
Parcourir et analyser les voies métaboliques et réglementaires




Outils
Des outils pour traiter nos données et le vôtre



Expression végétale ATLAS 
Parcourez les résultats d'expression des plantes chez EBI ATLAS




EXPLOSION
Consulter nos génomes avec une séquence d'ADN ou de protéines



Gramene Mart
Une interface de requête avancée alimentée par BioMart



Track Hub Registry 
Une collection globale centralisée de centres de pistes accessibles au public



Sensibilisation et formation
Ressources pédagogiques et webinaires



Téléchargements en vrac
Téléchargement FTP de nos données



Archiver
Outils et données héritées (marqueurs, chemins Cyc, etc.)

NCBI

PubMed Entrez BLAST OMIM Taxonomy Structure

NCBI

SITE MAP

BLAST info
BLAST overview

Frequently Asked Questions

BLAST Program Selection Guide
Revised: 4/25/02

Description of BLAST Services

Subscribe to BLAST-Announce

News/Noteworthy

BLAST course

What's NEW in BLAST®

NEW March 5th 2002: New database linkouts from BLAST results. Results of a BLAST search will now link sequences from the BLAST results page to the NCBI LocusLink and UniGene databases. Links to additional databases coming soon

Nucleotide BLAST ?

- [Standard nucleotide-nucleotide BLAST \[blastn\]](#)
- [MEGABLAST](#)
- [Search for short nearly exact matches](#)

click

Protein BLAST ?

- [Standard protein-protein BLAST \[blastp\]](#)

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡|≡ ↺ 🔍 ↻

APPRENTISSAGE AUTOMATIQUE NLP(TALN)

- **Erick Alphonse et Al. en 2009**, on eu a faire un projet similaire nommé **CADERIGE** mais avec un problème lié a la taille des données a mettre en entrée pour le modèle.
- Ils se sont appuyer sur les travaux de **Freitag, 1998, Califf et Al. 1998, Craven et Al 1999**
- Traitement par Racinisation
- création de patron
- Utilisation du logiciel Asium pour la classification hiérarchique
- Annotation des données en créant une connaissance préalable dans les données.
- Projet limité a leur traitement.

MOTIVATION

- Les biologistes exploitent les bases de données biologiques en soumettant des requêtes constituées de mots-clefs a des moteurs de recherche qui retournent des données textuelles.
- Il leur est également possible d'utiliser les références présentes dans les bases de données génomiques sous la forme d'hyperliens vers les bases de données bibliographiques.
- Afin d'extraire les connaissances recherchées, il leur faut identifier les résumés ou les paragraphes pertinents de ses références. Une telle démarche est manuelle, répétitive et coûteuse dans le temps.
- Une requête avec un chromosome comme donnée d'entre peut générer 1500 a 2500 résumés.
- La mise a jour très régulière des informations contenues dans les bases de données empêchent le stockage des informations recueillies.

PROPOSITION DE SOLUTION

TRAVAIL A FAIRE

Fouilles de données recueillies par le moteur de recherche en vue d'extraire des informations plus adéquats dans le lot du grand sac de mot constitué et trouver la bonne formule pour améliorer les paramètres de recherches du moteur conçu ainsi que l'interprétation des résultats en vue qu'ils servent d'appui aux packages extraites.

LES BESOINS POUR ATTEINDRE L'OBJECTIF

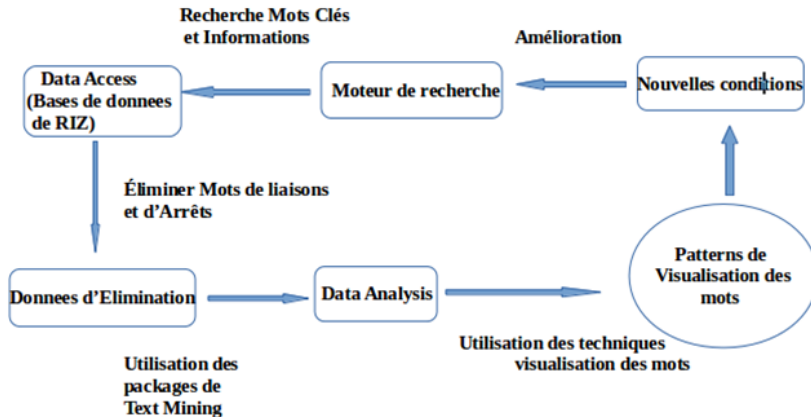
- Langage R pour l'analyse de données textuelles.
- Logiciel RStudio

BUT VISE

- Efficacité du moteur par rapport aux informations recueillies.
- La pertinence et satisfaction des résultats
- Appui et ajout d'informations aux packages obtenus en CSV

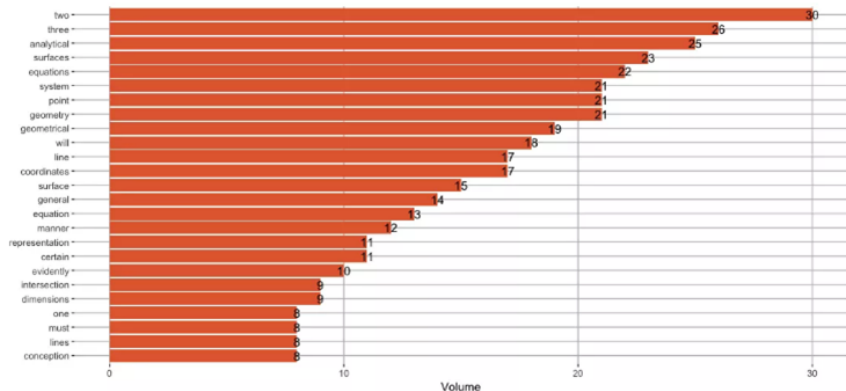
PROPOSITION DE SOLUTION

CYCLE DE TRAITEMENT DE NOTRE PROJET



PROPOSITION DE SOLUTION

GRAPHE DE FRÉQUENCE DE MOTS



www.thinkr.fr

PROPOSITION DE SOLUTION

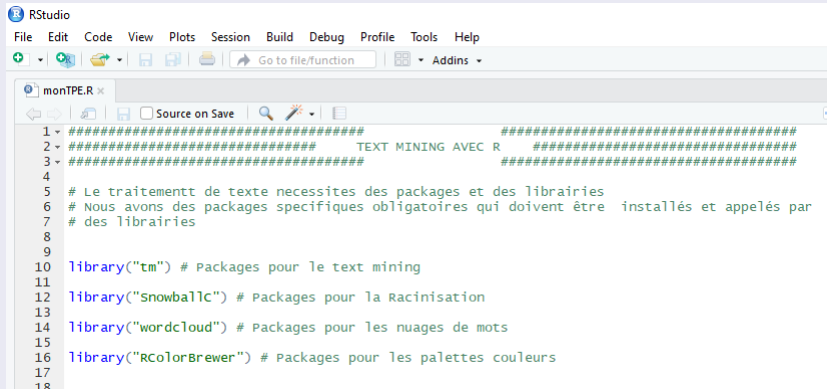
PLANNING DU TRAVAIL

Travaux	Descriptions	Durées estimées
Conception des modèles	Descriptions des paramètres d'évaluation	2semaines
Implémentation du modèle	-Codage pour la réalisation du programme -simulation avec des données créer par nous même	6semaines
Tests et validation	Simulation avec des vraies données provenant du moteur	3semaines
Rédaction du rapport final	Assemblages des sous projets et soumissions après accord avec les encadrants	3semaines

IMPLEMENTATION

LES DIFFÉRENTES ÉTAPES DE NOTRE IMPLEMENTATION

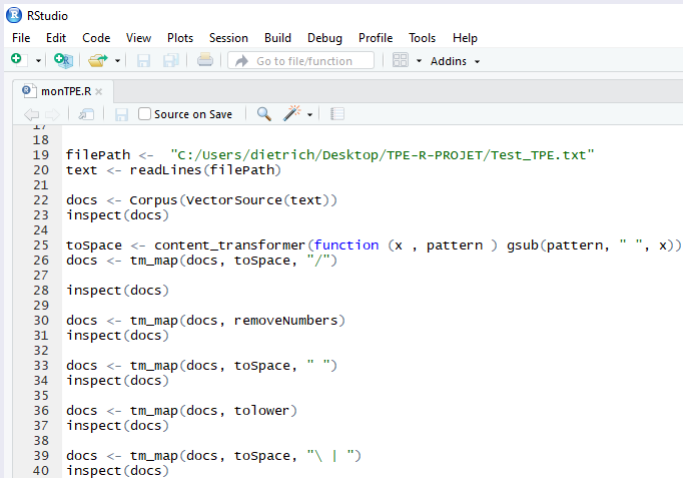
Comme dit plus haut, nous aurons besoin de l'utilisation du logiciel R et de ses Librairies.



```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
+ Go to file/function Addins
monTPE.R x
Source on Save
1 #####
2 ##### TEXT MINING AVEC R #####
3 #####
4
5 # Le traitement de texte necessite des packages et des librairies
6 # Nous avons des packages specifiques obligatoires qui doivent être installés et appelés par
7 # des librairies
8
9
10 library("tm") # Packages pour le text mining
11
12 library("snowballc") # Packages pour la Racinisation
13
14 library("wordcloud") # Packages pour les nuages de mots
15
16 library("RColorBrewer") # Packages pour les palettes couleurs
17
18
```

IMPLEMENTATION

CHARGEMENT DES DONNÉES



```
17
18
19 filePath <- "C:/Users/dietrich/Desktop/TPE-R-PROJET/Test_TPE.txt"
20 text <- readLines(filePath)
21
22 docs <- Corpus(VectorSource(text))
23 inspect(docs)
24
25 tospace <- content_transformer(function(x, pattern) gsub(pattern, " ", x))
26 docs <- tm_map(docs, tospace, "/")
27
28 inspect(docs)
29
30 docs <- tm_map(docs, removeNumbers)
31 inspect(docs)
32
33 docs <- tm_map(docs, tospace, " ")
34 inspect(docs)
35
36 docs <- tm_map(docs, tolower)
37 inspect(docs)
38
39 docs <- tm_map(docs, tospace, "\\ | ")
40 inspect(docs)
41
```

IMPLEMENTATION

LES TRAITEMENTS

```
monTPE.R x
Source on Save
65      , "for", "is", "us", "be"
66      "are", "the", "The", "be"
67      , "been", "from", "have"
68      , "being", "to", "of", "that"
69      , "was", "were", "but"))
70 inspect(docs)
71
72 docs <- tm_map(docs, removePunctuation)
73 inspect(docs)
74
75 docs <- tm_map(docs, stripwhitespace)
76 inspect(docs)
77
78 #docs <- tm_map(docs, stemDocument)
79 #inspect(docs)
80
81 dtm <- TermDocumentMatrix(docs)
82 inspect(dtm)
83
84 m <- as.matrix(dtm)
85
86 v <- sort(rowSums(m), decreasing=TRUE)
87
88 d <- data.frame(word = names(v), freq=v)
89
90 m <- as.matrix(dtm)
91
92 inspect(dtm)
93
```

LES TRAITEMENTS

```
93
94 head(d, 300)
95
96 set.seed(1234)
97   wordcloud(words = d$word, freq = d$freq, min.freq = 1,
98             max.words=300, random.order=FALSE, rot.per=0.35,
99             colors=brewer.pal(8, "Dark2"))
100
101
102 findFreqTerms(dtm, lowfreq = 1)
103
104 head(d, 100)
105
106 findAssocs(dtm, terms = "protein", corlimit = 0.3)
107
108 barplot(d[1:30,]$freq, las = 2, names.arg = d[1:30,]$word,
109         col = "lightblue", main = "Most frequent words",
110         ylab = "word frequencies")
111
112
113
114 ~ #####
115 ~ ##### FIN DU PROGRAMME #####
116 ~ #####
117
```

16:27 [Untitled] ↕

TRAITEMENT SUR LE CORPUS

- Dans les différents traitements de notre travail, nous avons eu à traiter notre texte selon les blocs de textes et obtenus par le moteur de recherche pour finalement avec un seul résultat à la sortie vu que les blocs de textes sont les sortie ou représentent la sortie d'une seule entrée.
- Nous avons ensuite enlever les mots et ponctuations inutiles.
- Nous avons créer un nuage de mots nous permettant de façon visuelle, de voir les mots les plus fréquents dans une apparence de taille.

MATRICE DES FRÉQUENCES

```
Non-/sparse entries: 5828/115822
Sparsity           : 95%
Maximal term length: 96
Weighting          : term frequency (tf)
Sample            :

                Docs
Terms   10 12 17 2 31 33 34 39 45 46
cell      0 0 0 1 0 3 1 12 5 0
expression 0 1 1 1 0 7 3 0 1 1
gene      5 2 0 0 5 1 2 1 2 3
genes     0 6 3 1 5 0 2 0 0 1
mutant    0 0 0 2 0 0 0 3 0 7
plant     0 1 0 1 1 0 3 2 3 1
rice      4 5 1 7 1 1 8 3 3 3
wall      0 0 0 1 0 0 0 10 6 0
which     1 0 2 1 1 0 1 1 3 2
with      2 3 1 1 3 0 2 1 2 0
```


MATRICES DES FRÉQUENCES

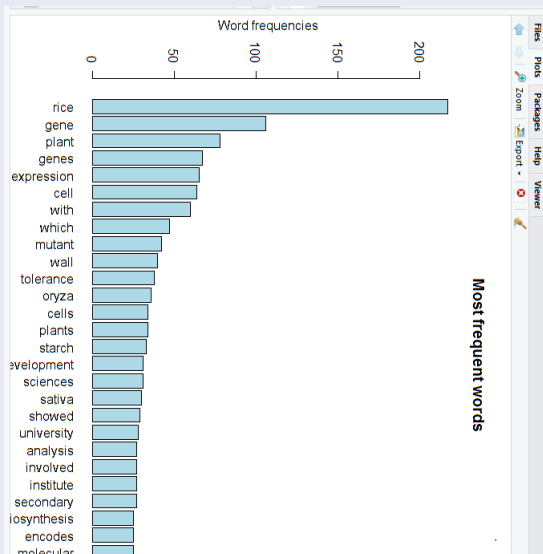
La matrice des fréquences des mots nous renseigne sur l'allure et la combinaison des mots dans le corpus, nous pouvons voir les mots les plus fréquents, leur taux dans les corpus et les liens avec d'autres mots selon les fréquences pour savoir s'ils sont ensemble pour exprimer ou expliquer un thème ou décrivent un élément ou s'ils sont séparés.

NUAGE DE MOTS



ANALYSE DES RÉSULTATS

FRÉQUENCE DES MOTS DU CORPUS



RÉSULTATS OBTENUS

Les mots les plus fréquents sont : RICE, GENE, EXPRESSION, CELL, WITH, MUTANT, UNIVERSITY, etc. ces mots nous renseignent sur le contenu essentiel des résultats provenant des abstracts des articles qui ont servi à constituer le texte que nous avons. Le résultat obtenu nous permet de comprendre que les gènes découlant de ce riz ont des gènes mutants qui n'ont plus la même constitution chromosomique.

NOTRE POINTS DE VUE

Le résultat recueilli par le moteur est renvoyer comme un tableau contenant du texte dont les colonnes sont : Title, Year, Journal, Affiliation et Abstract.

La majorité du texte vient de l'abstract des articles et des documents faisant liens dans la recherche des informations.

Nous pouvons toutefois dire que les informations selon leur pertinences ne résumant pas ou n'englobe pas les informations sur les deux formats des gènes dans les bases RapDB ou MSU, mais plutôt que les informations sont pris séparément ne permettant pas de mailler la connaissance découlant des deux normes sur les gènes du riz *Oryza Sativa*.

CONCLUSION

- Notre projets rentre dans un cadre où pour comprendre et appréhender les informations contenues dans un texte, surtout dans notre cas avec les informations concernant la génomique, il est primordiale de connaître le domaine de la biologie.
- Précisons que nous avons fait juste un résumé avec le résultat.
- L'information sortie serait plus approfondi dans un autre traitement s'il faille s'appuyer sur tous les mots du texte mais en présence d'un texte de plusieurs pages, nous ne pouvons aucunement les traiter ensemble.

PERSPECTIVES

Les informations descendues à notre niveau nous ont permis de comprendre comme tout profane en biologie de comprendre des concepts cachés dans les abstracts. Il serait plus avantageux à un biologiste d'avoir plus d'informations sur les gènes et les formules de leurs mutations. Ainsi, nous dirons que les perspectives futures de notre projet seraient d'améliorer le résultat en extrayant également les symboles et les formules des chromosomes pour plus d'utilités aux biologistes et aux scientifiques.



Iris Eshkol, Jean-Yves Antoine (Eds.)

24e Conférence sur le Traitement Automatique des Langues Naturelles

<https://taln2017.cnrs.fr> (TALN) Orléans, France – 26-30 juin 2017



Mayer U.

Protein Information Crawler (PIC) :

extensive spidering of multiple protein information resources for large protein sets. Proteomics 2008; 8 :42–4.



Bare JC, Shannon PT, Schmid AK, et al

The Firegoose

two-way integration of diverse data from different bioinformatics web resources with desktop applications. BMC Bioinformatics 2007; 8 :456

RÉFÉRENCES



N. Alexandrov, S. Tai, W. Wang, L. Mansueto, K. Palis, R.R. Fuentes, V.J. Ulat, D.Chebatarov, G. Zhang, Z. Li, R. Mauleon, R.S. Hamilton, K.L. McNally

SNP-Seekdatabase of SNPs derived from 3000 rice genomes,
Nucleic Acids Res. 43(2015) D1023– D1027



Erick Alphonse, Sophie Aubin, Philippe Bessieres, Gilles Bisson.

Extraction d'information appliquée au domaine biomédical
Apprentissage et traitement automatique de la langue. Juin 2004, 2004.

