

Faculté des sciences et techniques MARRAKECH

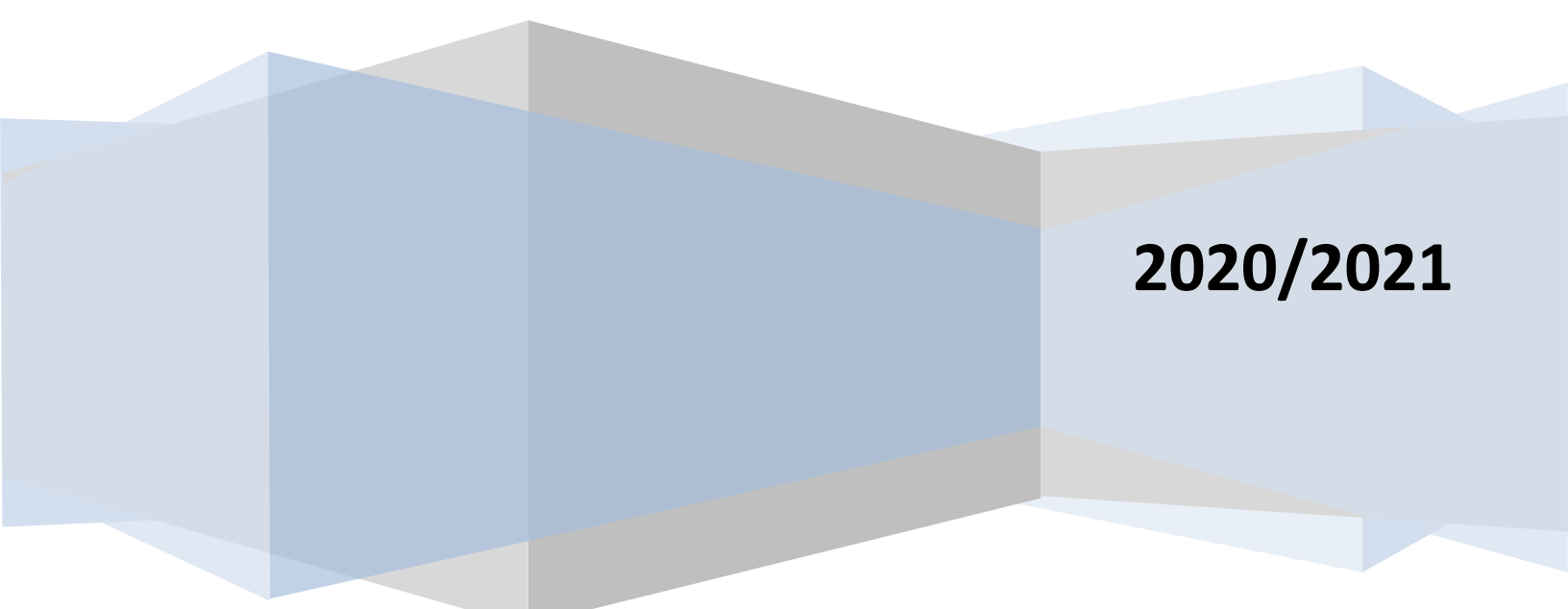
# Compte rendu Mini-Projet : Mining Mail Box

-

Master SDAD

Yassine ELMRHARI

Enseignante : Mme K. Bouzaachane



**2020/2021**

# Plan

## I. Introduction

## II. Simulation d'un projet Data Mining Phase 1

1. Compréhension du Marché
2. Compréhension des données
3. Préparation des données

## III. Simulation d'un projet Data Mining Phase 2

1. Récapitulation
2. Modélisation
3. Evaluation

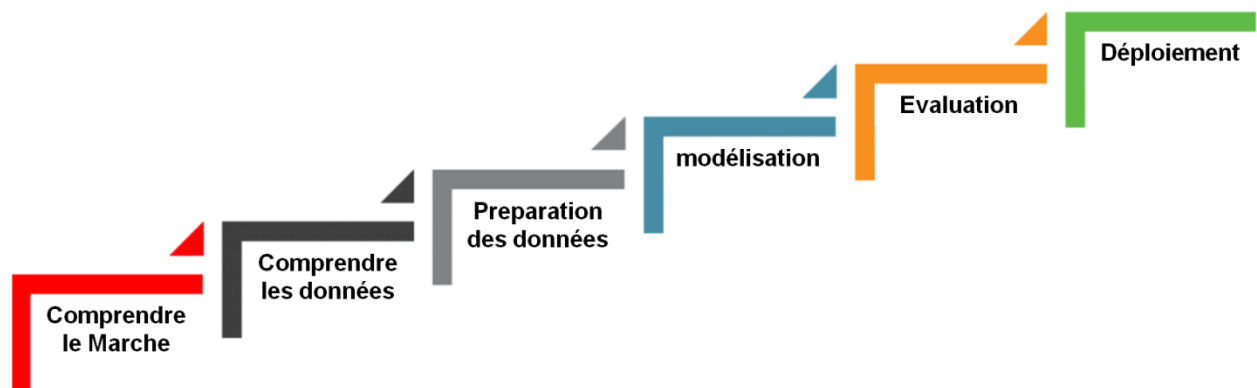
## IV. Conclusion

# Introduction

L'exploration des données ou Data Mining, consiste à explorer et à analyser des données volumineuses afin de **découvrir des règles et des modèles pertinents**.

Il est considéré comme une discipline dans le domaine d'études de la science des données et se distingue de l'analyse prédictive, qui décrit des données historiques, tandis que l'exploration des données vise à **prédire des résultats futurs**.

Les techniques d'exploration des données sont utilisées pour **créer des modèles d'apprentissage automatique** qui **alimentent les applications d'intelligence artificielle modernes**.



# Simulation d'un projet Data Mining

## Phase 1

### 1. Compréhension du Marché

La première étape consiste à définir les objectifs du projet et à déterminer comment l'exploration des données peut aider à atteindre l'objectif souhaité.

#### Sujet de la simulation:

Une grande entreprise d'agroalimentaire souhaite étudier le type de contenu échangé entre ses employés afin de détecter les comportements ne faisant pas partie de la politique de l'entreprise (non-professionnels).

#### Les objectifs du projet:

- Filtrer les emails en se basant sur leurs fichiers de signature.
- Faire le nettoyage du contenu des emails de chaque fonction.
- Générer les fréquences des mots.
- Réaliser un programme de prédiction de la nature du contenu des emails avec un apprentissage basée sur les résultats.

## 2. Compréhension des données

Les données sont collectées à partir de toutes les sources de données applicables à cette étape. Une première visualisation des données afin d'explorer les propriétés des ceux la pour s'assurer qu'elles contribueront à la réalisation des objectifs de l'entreprise.

Message-ID	Date	From	To	Subject	X-From	X-To	X-cc	X-bcc	X-Folder	X-Origin	X-FileName	content	user	at_1_level_Cat_1_level_0
<18782981.1>	2001-05-14 23:39	frozenset{}	pf frozenset{}	tim.belden@enr	Phillip K Allen	Tim Belden	<Tim Belden/Enron@EnronXC>	Phillip_Allen	Allen-P		pallen (Non-P	Here is our	allen-p	
<15464986.1>	2001-05-04 20:51	frozenset{}	pf frozenset{}	Joel Re:	Phillip K Allen	John J Lavorato	<John J Lavorato/ENRONX>	Phillip_Allen	Allen-P		pallen (Non-P	Traveling to	allen-p	
<24216240.1>	2000-10-18 10:00	frozenset{}	pf frozenset{}	le: Re: test	Phillip K Allen	Leah Van Arsdall		Phillip_Allen	Allen-P		pallen.nsf	test successfu	allen-p	
<13505866.1>	2000-10-23 13:13	frozenset{}	pf frozenset{}	randall.gay@eni	Phillip K Allen	Randall L Gay		Phillip_Allen	Allen-P		pallen.nsf	Randy, Can y	allen-p	
<30922949.1>	2000-08-31 12:07	frozenset{}	pf frozenset{}	gr Re: Hello	Phillip K Allen	Greg Piper		Phillip_Allen	Allen-P		pallen.nsf	Let's shoot fo	allen-p	
<30965995.1>	2000-08-31 11:17	frozenset{}	pf frozenset{}	gr Re: Hello	Phillip K Allen	Greg Piper		Phillip_Allen	Allen-P		pallen.nsf	Greg, How ab	allen-p	
<16254169.1>	2000-08-22 14:44	frozenset{}	pf frozenset{}	david.l.johnson@	Phillip K Allen	david.l.johnson@enron.com	John Shafer	Phillip_Allen	Allen-P		pallen.nsf	Please cc the	allen-p	
<17819699.1>	2000-07-14 13:59	frozenset{}	pf frozenset{}	jo Re: PRC revie	Phillip K Allen	Joyce Teixeira		Phillip_Allen	Allen-P		pallen.nsf	any morning l	allen-p	
<20641191.1>	2000-10-17 09:26	frozenset{}	pf frozenset{}	m Re: High Spee	Phillip K Allen	Mark Scott		Phillip_Allen	Allen-P		pallen.nsf	1. login: pal	allen-p	
<30795301.1>	2000-10-16 13:44	frozenset{}	pf frozenset{}	zir FW: fixed for	Phillip K Allen	zimam@enron.com		Phillip_Allen	Allen-P		pallen.nsf	-----	allen-p	
<33076797.1>	2000-10-16 13:42	frozenset{}	pf frozenset{}	bu Re: FW: fixed	Phillip K Allen	"Buckner, Buck" <buck.buckner@honeyw>		Phillip_Allen	Allen-P		pallen.nsf	Mr. Buckner,	allen-p	
<25459584.1>	2000-10-13 13:45	frozenset{}	pf frozenset{}	stagecoachmam	Phillip K Allen	stagecoachmama@hotmail.com		Phillip_Allen	Allen-P		pallen.nsf	Lucy, Here ar	allen-p	
<13116875.1>	2000-10-09 14:16	frozenset{}	pf frozenset{}	ke Consolidated	Phillip K Allen	Keith Holst		Phillip_Allen	Allen-P		pallen.nsf	-----	allen-p	
<2707340.10>	2000-10-09 14:00	frozenset{}	pf frozenset{}	ke Consolidated	Phillip K Allen	Keith Holst		Phillip_Allen	Allen-P		pallen.nsf	-----	allen-p	
<2465689.10>	2000-10-05 13:26	frozenset{}	pf frozenset{}	david.delaingey@	Phillip K Allen	David W Delaingey		Phillip_Allen	Allen-P		pallen.nsf	Dave, Here ar	allen-p	
<1115198.10>	2000-10-05 12:55	frozenset{}	pf frozenset{}	pa Re: 2001 Mar	Phillip K Allen	Paula Harris		Phillip_Allen	Allen-P		pallen.nsf	Paula, 35 mil	allen-p	
<19773657.1>	2000-10-04 16:23	frozenset{}	pf frozenset{}	in Var, Reportin	Phillip K Allen	Ina Rangel		Phillip_Allen	Allen-P		pallen.nsf	-----	allen-p	
<7391389.10>	2001-05-04 18:26	frozenset{}	pf frozenset{}	tim.heizenrader	Phillip K Allen	Tim Heizenrader	<Tim Heizenrader/Enron>	Phillip_Allen	Allen-P		pallen (Non-P	Tim, mike gri	allen-p	
<12759088.1>	2000-10-03 16:30	frozenset{}	pf frozenset{}	pa Westgate	Phillip K Allen	pallen70@hotmail.com		Phillip_Allen	Allen-P		pallen.nsf	-----	allen-p	
<29177675.1>	2000-10-03 16:15	frozenset{}	pf frozenset{}	in Meeting re: S	Phillip K Allen	Ina Rangel		Phillip_Allen	Allen-P		pallen.nsf	-----	allen-p	
<24729148.1>		frozenset{}	pf frozenset{}	bs_stone@yaho	Phillip K Allen	bs_stone@yahoo.com		Phillip_Allen	Allen-P		pallen.nsf	Brenda, Pleas	allen-p	
<17610321.1>		frozenset{}	pf frozenset{}	st Re: Not busin	Phillip K Allen	Steve Touchstone	<STouchstone@natsou>	Phillip_Allen	Allen-P		pallen.nsf	I think Fletc	allen-p	
<26575732.1>		frozenset{}	pf frozenset{}	bs Re: Original S	Phillip K Allen	"BS Stone" <bs_stone@yahoo.com>	@ EN/	Phillip_Allen	Allen-P		pallen.nsf	Brenda, Pleas	allen-p	
<15294346.1>		frozenset{}	pf frozenset{}	lki San Juan Inde	Phillip K Allen	Lkuch@mh.com		Phillip_Allen	Allen-P		pallen.nsf	-----	allen-p	
<25140503.1>	2000-05-20 12:50	frozenset{}	pf frozenset{}	jel San Juan Inde	Phillip K Allen	Jeffrey T Hodge		Phillip_Allen	Allen-P		pallen.nsf	Liane, As we	allen-p	
<19034252.1>	2000-09-26 16:28	frozenset{}	pf frozenset{}	kh Investment St	Phillip K Allen	kholst@enron.com		Phillip_Allen	Allen-P		pallen.nsf	-----	allen-p	
<719350107>	2000-09-26 16:26	frozenset{}	pf frozenset{}	nk Investment St	Phillip K Allen	pallen70@hotmail.com		Phillip_Allen	Allen-P		pallen.nsf	-----	allen-p	

## Division des emails selon leurs fichiers de signature et ajout d’une colonne de type de fonction :

```
#Groupage des emails par fichier de signature (departement) (14)
#pallen emails / La fonction technique
emailspallen = filter(emails, grepl("^pallen",emails$X.FileName, ignore.case = TRUE))
emailspallen['Function']='Technical'
#jarnold emails / La fonction commerciale
emailsjarnold = filter(emails, grepl("^jarnold",emails$X.FileName, ignore.case = TRUE))
emailsjarnold['Function']='Commercial'
#rbadeer emails / La fonction financière
emailsrbadeer = filter(emails, grepl("^rbadeer",emails$X.FileName, ignore.case = TRUE))
emailsrbadeer['Function']='Finantial'
#sbaile2 emails / La fonction de sécurité
emailssbaile2 = filter(emails, grepl("^sbaile2",emails$X.FileName, ignore.case = TRUE))
emailssbaile2['Function']='Security'
#ebass emails / La fonction de comptabilité
emailsebass = filter(emails, grepl("^ebass",emails$X.FileName, ignore.case = TRUE))
emailsebass['Function']='Comptability'
#sbeck emails / La fonction administrative
emailssbeck = filter(emails, grepl("^sbeck",emails$X.FileName, ignore.case = TRUE))
emailssbeck['Function']='Administration'
#rbenson emails / La fonction direction et administration générale
emailsrbenson = filter(emails, grepl("^rbenson",emails$X.FileName, ignore.case = TRUE))
emailsrbenson['Function']='General direction'
#mcarson2 emails / La fonction achat
emailsmcarson2 = filter(emails, grepl("^mcarson2",emails$X.FileName, ignore.case = TRUE))
emailsmcarson2['Function']='Achat'
```

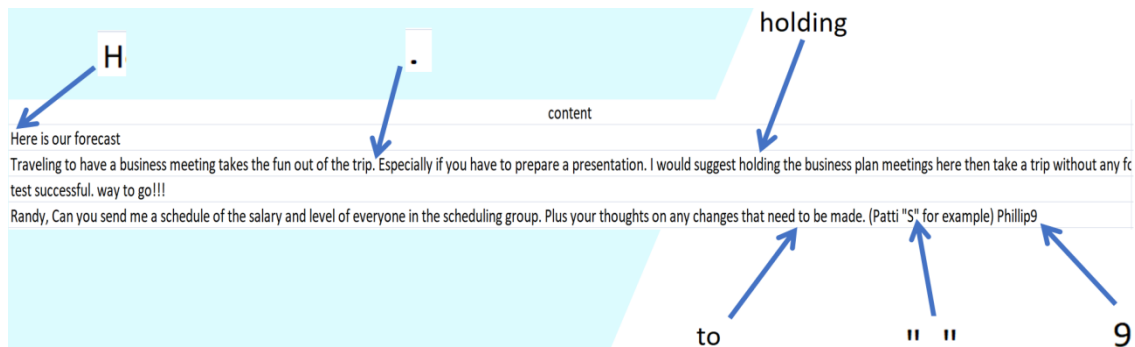
X.Origin	X.FileName	content	user	Function
Allen-P	pallen (Non-Privileged).pst	Here is our forecast	allen-p	Technical
Allen-P	pallen (Non-Privileged).pst	Traveling to have a business meeting takes the fun out of th...	allen-p	Technical
Allen-P	pallen.nsf	test successful. way to go!!!	allen-p	Technical
Allen-P	pallen.nsf	Randy, Can you send me a schedule of the salary and level ...	allen-p	Technical
Allen-P	pallen.nsf	Let's shoot for Tuesday at 11:45.	allen-p	Technical
Allen-P	pallen.nsf	Greg, How about either next Tuesday or Thursday? Phillip	allen-p	Technical
Allen-P	pallen.nsf	Please cc the following distribution list with updates: Phillip ...	allen-p	Technical
Allen-P	pallen.nsf	any morning between 10 and 11:30	allen-p	Technical
Allen-P	pallen.nsf	1. login: pallen pw: ke9davis I don't think these are required ...	allen-p	Technical
Allen-P	pallen.nsf	----- Forwarded by Phillip K Allen/HOU/ECT ...	allen-p	Technical

### Signature ➔ Fonction

```
pallen emails ==> La fonction technique
jarnold emails ==> La fonction commerciale
rbadeer emails ==> La fonction financière
sbaile2 emails ==> La fonction de sécurité
ebass emails ==> La fonction de comptabilité
sbeck emails ==> La fonction administrative
rbenson emails ==> La fonction direction et administration générale
mcarson2 emails ==> La fonction achat
scorman emails ==> La fonction finance et comptabilité
cdean emails ==> La fonction logistique
jderric emails ==> La fonction marketing et commerciale
fermis emails ==> La fonction production
dfarmer emails ==> La fonction recherche et développement
dfossum emails ==> La fonction ressources humaines
```

## Nettoyage du contenu des emails pour la préparation au traitement :

```
#Create corpus function
Transform_to_corpus = function(email) {
  corpus = Corpus(VectorSource(email$content))
  # convert the text to lowercase
  corpus = tm_map(corpus, content_transformer(tolower))
  corpus = tm_map(corpus, PlainTextDocument)
  # remove all punctuation from the corpus
  corpus = tm_map(corpus, removePunctuation)
  # remove numbers
  corpus = tm_map(corpus, removeNumbers)
  # remove all English stopwords from the corpus
  corpus = tm_map(corpus, removeWords, stopwords("en"))
  # remove special_chars
  specialchars = content_transformer(function(x) gsub("[^[:alnum:]]//'", " ", x))
  corpus = tm_map(corpus, specialchars)
  # stem the words in the corpus
  corpus = tm_map(corpus, stemDocument)
}
```



## Génération des tableaux Mot/Fréquence et prendre le top 20 :

```
#Generation des mots/frequences
for (i in list) {
  a = paste("Words",i,sep = "")
  b = paste("corpus",i,sep = "")
  assign(a,Generate_words(get(b)))
}

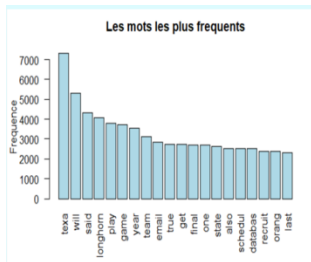
#Top 20 words per email
for (i in list) {
  a = paste("Top",i,sep = "")
  b = paste("Words",i,sep = "")
  assign(a,head(get(b),20))
}
```



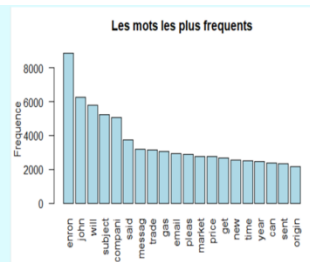


## Génération des barplots:

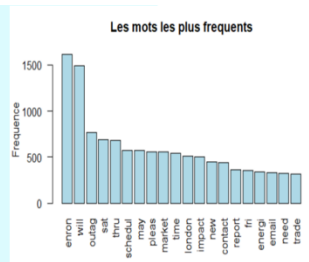
```
#Generation des barplot
for (i in list) {
  a = paste("Top",i,sep = "")
  barplot(get(a)$Frequence,las = 2, names.arg = get(a)$Mot,
          col = "lightblue", main = "Les mots les plus frequents",
          ylab = "Frequence")
}
```



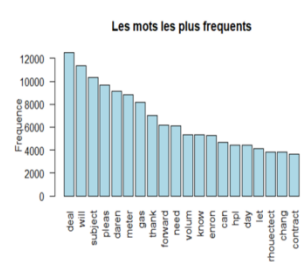
La fonction logistique



La fonction commerciale



La fonction direction et  
administration générale

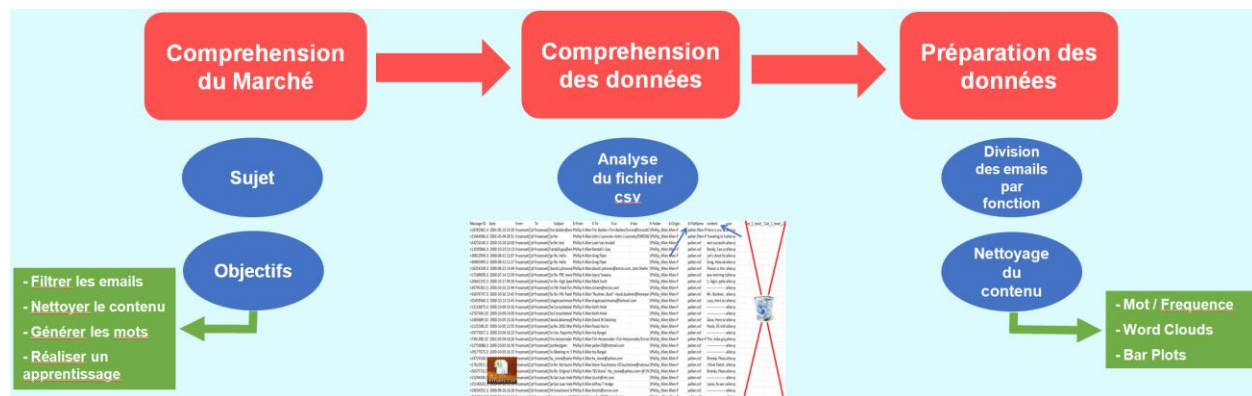


La fonction recherche et  
développement

# Simulation d'un projet Data Mining

## Phase 2

### 1. Récapitulation :

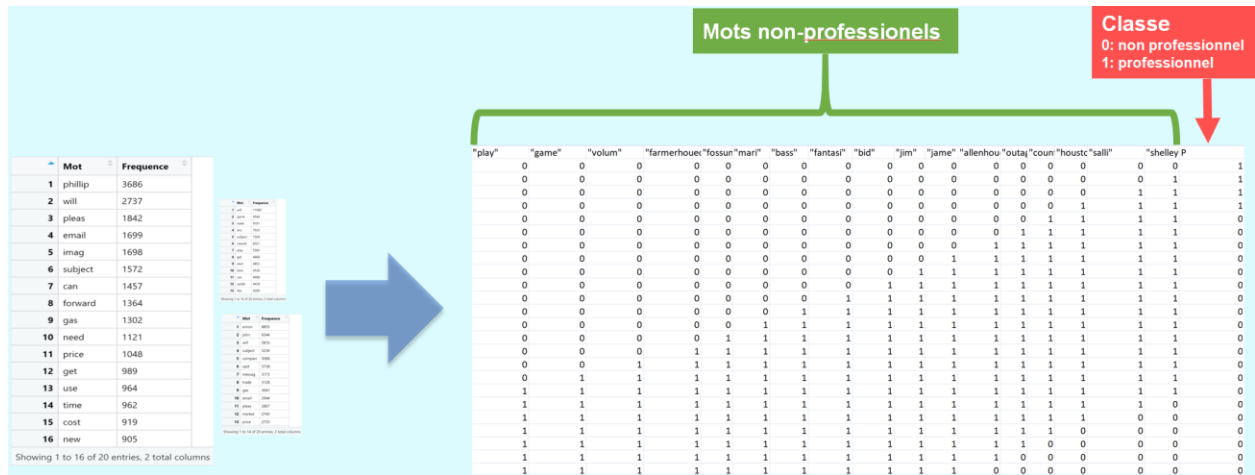


### 2. Modélisation :

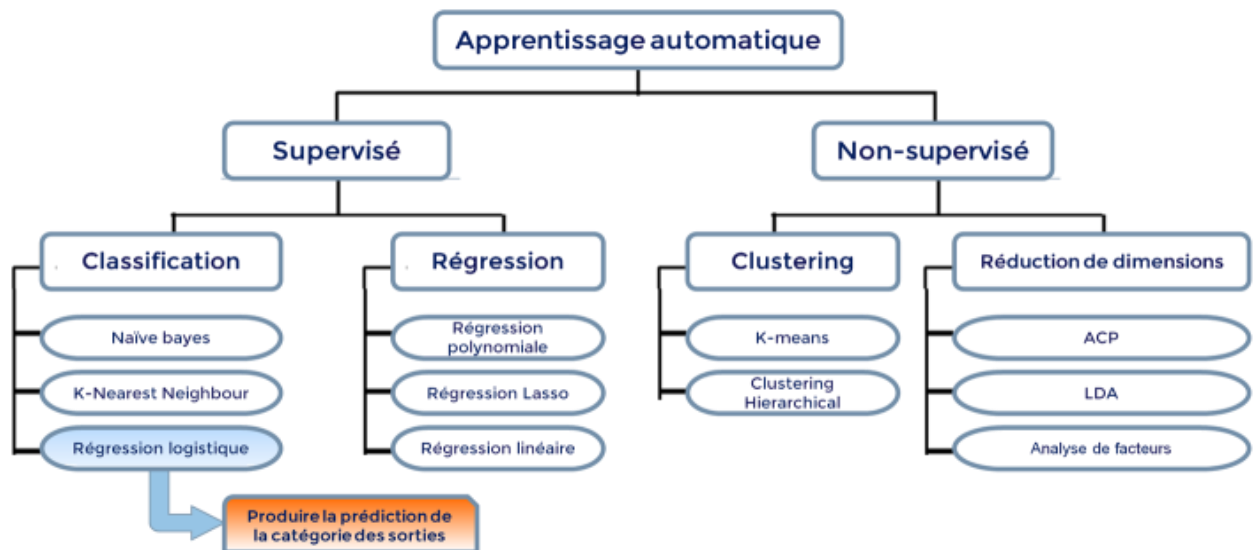
Pour générer un modèle adapté au projet, Il est impératif de suivre les étapes clés suivantes:

- Collecter les données d'apprentissage.
- Déterminer l'algorithme au besoin.
- Faire toutes les optimisations pour assurer le bon fonctionnement.
- Commencer l'apprentissage.

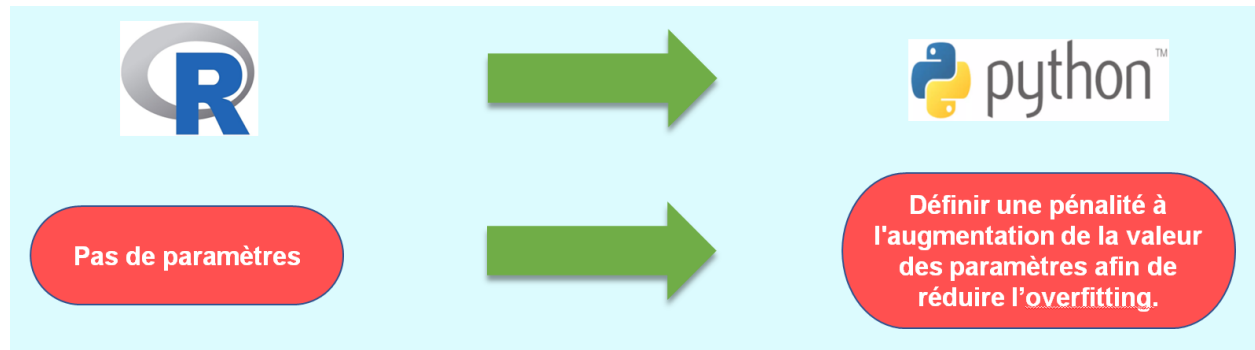
## Collecte des données d'apprentissage :



### Détermination de l'algorithme au besoin :



## Les optimisations :



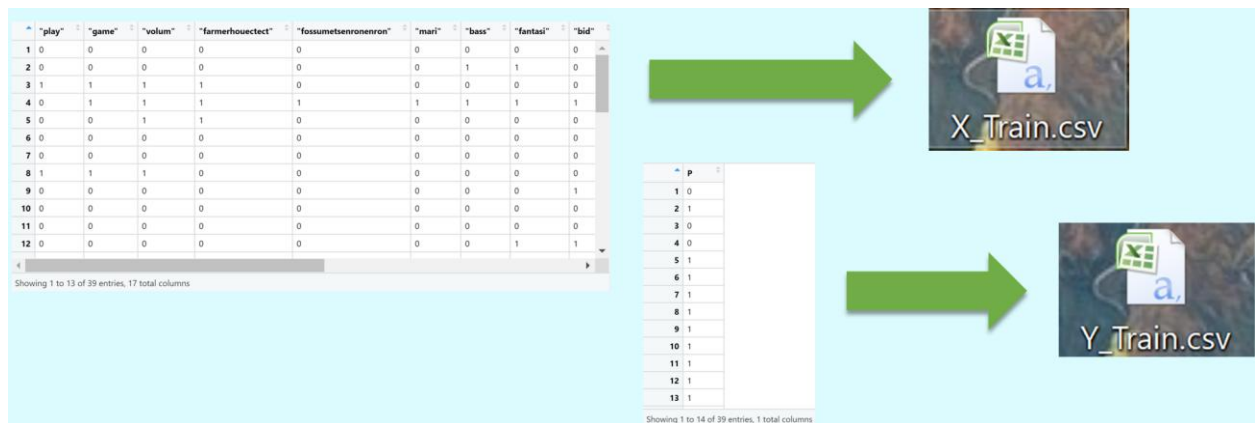
## Préparation des données d'apprentissage :

```
#Importation des donnees d'apprentissage:
library(readxl)
data = read_excel('C:/Users/PC Gamer/Desktop/Training.xlsx')
data$P <- factor(data$P)

#Generation du train/test:
# Set Seed so that same sample can be reproduced in future also
set.seed(101)
# Now Selecting 80% of data as sample from total 'n' rows of the data
sample <- sample.int(n = nrow(data), size = floor(.80*nrow(data)), replace = F)
Train <- data[sample, ]
Test <- data[-sample, ]

#Preparation des donnees:
X_Train = Train[, -ncol(Train)]
Y_Train = Train[, ncol(Train)]
X_Test = Test[, -ncol(Test)]
Y_Test = Test[, ncol(Test)]

write.csv(X_Train, "C:/Users/PC Gamer/Desktop/X_Train.csv", row.names = FALSE)
write.csv(Y_Train, "C:/Users/PC Gamer/Desktop/Y_Train.csv", row.names = FALSE)
write.csv(X_Test, "C:/Users/PC Gamer/Desktop/X_Test.csv", row.names = FALSE)
write.csv(Y_Test, "C:/Users/PC Gamer/Desktop/Y_Test.csv", row.names = FALSE)
```



## Alimentation du model :

```
#Importing csv
X_Train = pd.read_csv (r'C:/Users/PC Gamer/Desktop/X_Train.csv')
Y_Train = pd.read_csv (r'C:/Users/PC Gamer/Desktop/Y_Train.csv')
X_Test = pd.read_csv (r'C:/Users/PC Gamer/Desktop/X_Test.csv')
Y_Test = pd.read_csv (r'C:/Users/PC Gamer/Desktop/Y_Test.csv')

#Convert to numpy array
X_Train = np.array(X_Train)
Y_Train = np.array(Y_Train)
X_Test = np.array(X_Test)
Y_Test = np.array(Y_Test)
```

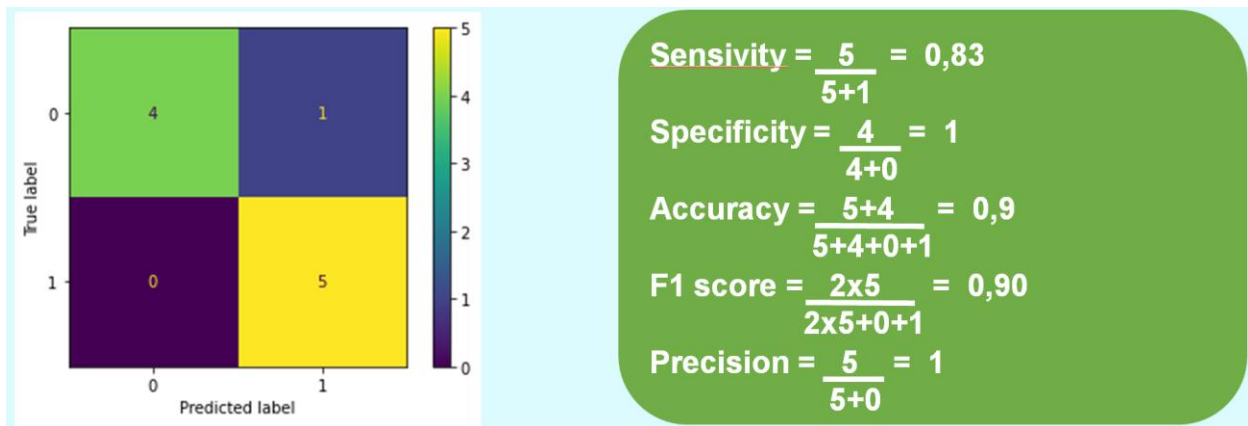


## 3. Evaluation :

### Evaluation du model :

```
#Initialise the model
model = LogisticRegression(C=100) #plus C est grande plus on a la regularisation

#Fit the model
model.fit(X_Train,Y_Train)
pred = model.predict(X_Test)
score = metrics.accuracy_score(Y_Test,pred)
conf = metrics.plot_confusion_matrix(model, X_Test, Y_Test)
```



## Prédiction de la nature du contenu des emails à partir du model :

Exemple d'emails :

### Non-professionnel

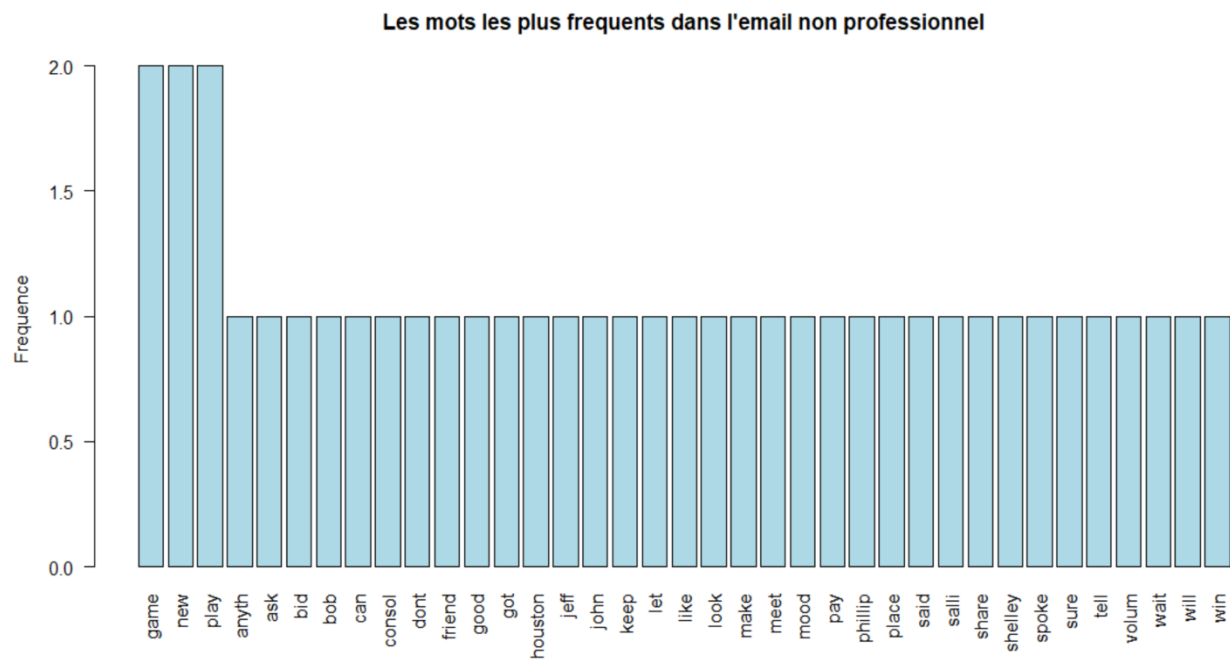
Bob, I spoke to Jeff. He said he would not pay anything. I am waiting for John to be in a good mood to ask. Can we meet to **play** some **games**? I got the new ps5 console and I am looking for friends to **play** with me and share the new **games**. Let's place a **bid** on who will win. Keep your volume down, I don't like **houston** and **salli**. Make sure to tell **shelley**. Phillip

Professionnel

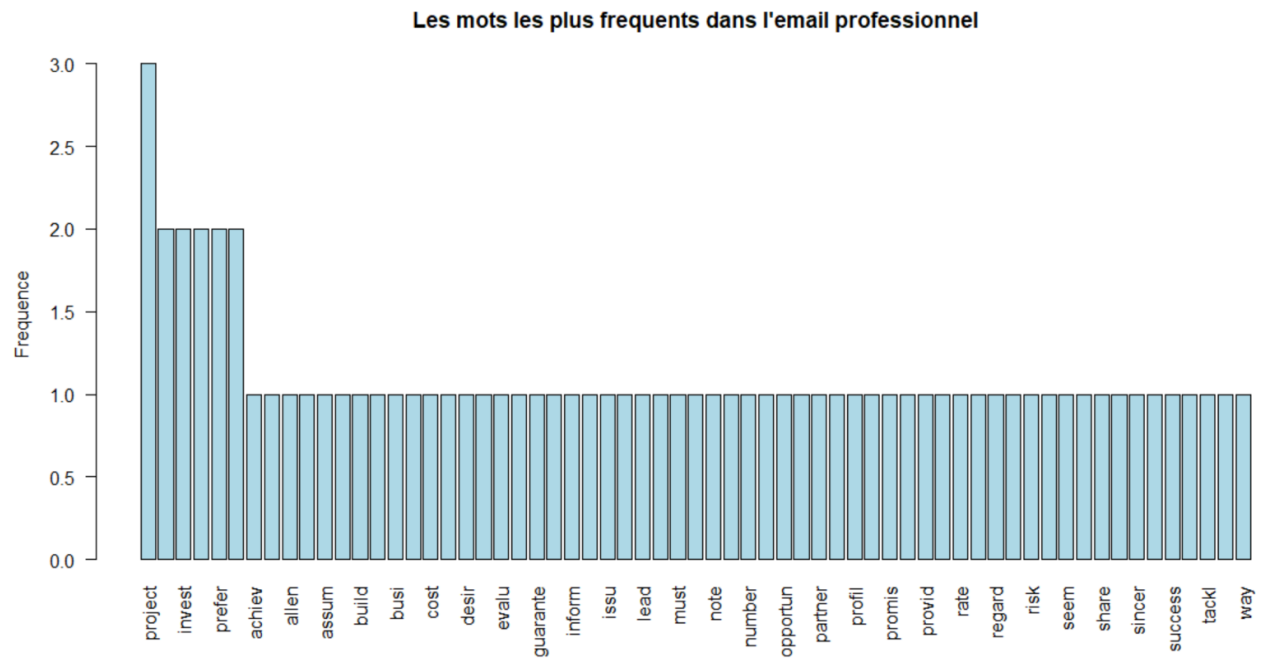
Numbers of several subcontractors used on recent projects. With regard to the proposed investment structure, I would suggest a couple of changes to better align the risk/reward profile between Creekside and the investors. Preferable Investment Structure: Developers guarantee note, not investors. Preferred rate of return (10%) must be achieved before any profit sharing. Builder assumes some risk for cost overruns. Since this project appears so promising, it seems like we should tackle these issues now. These questions are not intended to be offensive in any way. It is my desire to build a successful project with Creekside that leads to future opportunities. I am happy to provide you with any information that you need to evaluate myself or Keith as a business partner. Sincerely, Phillip Allen

Les mots fréquents dans les emails d'exemple :

Non-professionnel



## Professionnel



Génération des vecteurs de prédiction :

Avec la fonction :

```
#Definition de la boucle generatrice:
Generate_vectors = fonction(words){

  #Definition de la liste generatrice:
  Liste=c('play','game','volum','farmerhouectect','fossumetsenronenron','mari','bass','fantasi

  i=1
  vec_gen=c()
  while (i <= length(Liste)) {
    for (j in words$Mot) {
      vec_gen[i]=0
      if(tolower(j)==tolower(Liste[i])){
        vec_gen[i]=1
        break
      }
    }
    i = i+1
  }
  return(vec_gen)
}

#Generation des vecteurs de prediction:
vect_p = Generate_vectors(wordsp)
vect_np = Generate_vectors(wordsnonp)
```



### Non-professionnel

```
> vect_np  
[1] 1 1 1 0 0 0 0 0 1 0 0 0 0 0 1 1 1
```

### Professionnel

```
> vect_p  
[1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Prédiction :

### Non-professionnel

```
prednonp Array of int64 (1,) [0]
```

```
In [14]: affichage(prednonp)  
####Cet email n'est pas professionnel####
```

### Professionnel

```
predp Array of int64 (1,) [1]
```

```
In [12]: affichage(predp)  
####Cette email est professionnel####
```

# Conclusion

Il apparaît que le Data Mining s'appuie sur le constat qu'il existe des connaissances latentes dans les gisements d'informations au sein des entreprises. Il donne reflète ce que les américains appellent la « million dollars décision ». A l'aide des outils intelligents comme les algorithmes du Machine Learning, Il touche à la prévision, à l'optimisation ou encore à la classification.