Les techniques de fouille de texte avec R pour l'analyse de candidatures Campus France

Martial Phélippé-Guinvarc'H

Les techniques de fouille de texte avec R pour l'analyse de candidatures Campus France

Martial Phélippé-Guinvarc'h *

18 novembre 2020

Résumé

Pour un responsable de diplôme, le recrutement des étudiants constitue un enjeu majeur. Avec l'émergence des dossiers électroniques et la standardisation des mentions, il semble que les candidats répliquent leur candidature sur plus de Masters et le nombre de dossiers reçus augmente. Certaines mentions comme Monnaie, Banque, Finance, Assurance offrent des contenus très différents d'une université à l'autre. En répliquant sa candidature sur plusieurs Masters de même mention, le candidat anticipe peut-être les mêmes contenus et les mêmes prérequis.

Ce document de travail utilise R pour analyser les candidatures de Campus France avec des techniques empruntées au *text and data mining* pour explorer les facteurs déterminants d'acceptation et de refus des candidatures. La première étape est de transformer le dossier pdf reçu en une donnée exploitable. Une première analyse descriptive explore les variables du dossier comme l'avis du Service de Coopération et d'Action Culturelle, le niveau moyen de l'étudiant ou le niveau de langue. Une analyse lexicale de la lettre de motivation révèle de fortes différences entre les avis favorables et défavorables. Enfin, l'analyse du sentiment des avis du Service de Coopération et d'Action Culturelle se révèle peu significative.

Introduction

Pour un responsable de diplôme, le recrutement des étudiants constitue un enjeu majeur. L'objectif est non seulement d'avoir de bons étudiants mais aussi d'en avoir suffisamment pour constituer une promotion. Avoir de bons étudiants est essentiel pour réduire le taux d'échec et pour augmenter l'employabilité. Même si n'est pas systématique, le niveau et la bonne adéquation des candidats se voient sur comportement général de la promotion en cours et dans son rapport avec le corps enseignant. C'est aussi souvent plus gratifiant d'enseigner à une promotion motivée et studieuse. La réussite des bons étudiants constitue aussi un facteur de notoriété utile à long terme.

Comme le précise Bian and Malet (09 Jun. 2018), les universités françaises sont influencées par le modèle anglo-américain de l'enseignement supérieur et se trouvent en concurrence. Mais de très nombreux candidats à l'international parlent français et elles offrent des perspectives d'étude non-anglophone. Mais, dans tous les pays, l'accueil d'étudiants étrangers résulte principalement d'une politique publique d'immigration des étudiants. [1] D'autres critères, comme les frais de scolarité d'un étudiant étrangers, les conditions d'allocation des bourses d'études ou les programmes de langue ne relèvent pas d'un responsable de formation. De fait, sa marge de manœuvre pour augmenter l'attractivité de la formation est faible.

Dans le monde, le nombre d'étudiants qui étudient à l'étranger est passé de 2,1 à 4,1 millions entre 2001 et 2016. Le nombre d'étudiants accueillis en France a largement augmenté passant de 132k étudiants internationaux en 1991 à plus 225k en 2016. En économie, il proviennent en majorité du continent africain [Bian and Malet, 09 Jun. 2018]. La France accueille environ 8% d'entre eux Farrugia and Bhandari [2018].

Une des difficultés du recrutement est la gestion du nombre de dossier reçus. Au niveau du responsable du diplôme, il se gère par la communication sur le diplôme et par les conditions d'accès affichées. Un nombre trop faible le mets en difficulté, il faut alors arbitrer entre volume et qualité. Ce document traite de la situation opposée, l'arrivée d'un trop grand nombre de dossier. Souvent la gestion électronique des dossiers facilite la réplication du dossier sur différentes formations, souvent sur une même mention, le

*Actuaire et Maître de Conférence à l'Institut du Risque et de l'Assurance, Laboratoire GAINS (Groupe d'Analyse des Itinéraires et Niveaux Salariaux), Université du Mans. L'auteur remercie Nguyen Thi Thanh Huyen, Maître de conférences, avec qui il co-dirige le Master Monnaie, Banque, Finance, Assurance, pour ses remarques pertinentes qui ont fait progresser l'analyse.

1. Par exemple, dans son article, Sá and Sabzalieva montre le lien entre les politiques publiques et l'évolution du nombre d'étudiants en provenance de l'étranger dans quatre pays anglophones.

candidat modifiant parfois juste à la marge chaque formation ciblée. De plus, pour les diplômes ou les métiers pluri-disciplinaires ou transverses, un étudiant peut se reconnaître dans un domaine de la formation sans pour autant avoir les prérequis dans les autres.

Aujourd'hui, les candidatures sont électroniques. Le responsable de formation les reçois via deux canaux. Il y a d'une part les candidatures « Campus France » issues de l'étranger et d'autre part celles « ecandidat » issue des universités françaises (et des pays hors campus France). Dans plus de 46 pays, Campus France (aussi appelé Centres pour les Études en France (CEF)) constitue un outil majeur auprès d'étudiants étrangers. Il apporte de l'information sur les formations aux candidats, facilite le passage de tests de langue. Le contrôle des documents et l'entretien relatif au projet personnel de l'étudiant constituent une plus-value majeure du dispositif. Près de 293 établissements d'enseignement supérieur adhérents à la convention CEF. Cela offre donc un large choix de formation à ces étudiants étrangers. Les responsables de formation ont ensuite accès à des dossiers structurés et vérifiés. La présente étude se consacre à l'analyse de ces candidatures Campus France. L'objectif est de vérifier si les prérequis peuvent être aménagées. L'idéal serait de pouvoir dissuader les candidats refusables sans décourager la candidats acceptables.

Aujourd'hui, le master MBFA propose les prérequis suivants (rentrée 2020) :

Langue : B2 minimum en compréhension orale, en structure de la langue et en compréhension écrite.

Diplôme : Licence d'économie, Licence mathématiques appliquées en sciences sociales (MASS), Licence d'économie-gestion ou une équivalence avec un de ces trois diplômes. Les licences pro Assurance Banque, les L3 CCA ou les L3 marketing n'offrent pas les prérequis pour suivre le Master.

Niveau : une moyenne supérieure à 12/20 sur chacun des trois derniers relevés de notes semestriels.

Précisions : nous vérifierons attentivement que l'étudiant a bien suivi, en L2 et en L3, un volume suffisant de cours économiques et quantitatifs (d'économie générale, de micro et de macro-économie, d'économétrie (et/ ou statistiques) et de mathématiques financières).

L'étude économétrique est réalisée sur R, en utilisant des techniques empruntées au *text and data mining* pour explorer les facteurs déterminants d'acceptation et de refus des candidatures. Les techniques mises en œuvre s'inspire essentiellement du livre de Silge and Robinson et des travaux de Roquebert. La première étape vise à transformer le dossier pdf reçu en une donnée exploitable. Une première analyse descriptive explore les variables du dossier comme l'avis du Service de Coopération et d'Action Culturelle, le niveau moyen de l'étudiant ou le niveau de langue. Une analyse lexicale permettra de voir si la lettre de motivation contient des différences significatives entre les avis favorables et défavorables. Enfin, l'analyse du sentiment des avis du Service de Coopération et d'Action Culturelle permettra de voir si son commentaire apporte des nuances statistiquement pertinentes en plus de son avis.

1 La construction de la base de données

Cette section présente les méthodes d'importation des données. Le package *pdftools* permet la manipulation des pdf sous R et *tesserac* est un outils de reconnaissance des caractères.

```
> library(pdftools)
> library(tesseract)
```

Les responsables de formation reçoivent le dossier électronique du candidat sous la forme d'un fichier PDF structuré. La première page est générée par l'établissement pour saisir l'avis, les pages suivantes sont générés à partir d'un formulaire Campus France, les pages suivantes sont la concaténation de fichiers scannés, comprenant un justificatif d'identité, le CV et une lettre de motivation, les relevés de notes, les diplômes, justificatifs de stages et d'emplois et lettre de recommandation.

Dans cette études, 198 dossiers sont analysés, dont 55 sont acceptés. Il y a 157 candidatures 2020, 21 candidatures 2019 et 20 candidatures 2018. Les candidatures 2018 et 2019 sont des candidatures acceptées. Dans cette étude, nous utilisons la partie issue du formulaire campus France et pas les pièces jointes adossées. L'information y est structurée, synthétique et contient l'avis du Service de Coopération et d'Action Culturelle (SCAC).

À partir de la deuxième page, le fichier PDF est stocké sous forme d'image et la conversion native en texte via acrobat reader ou via ghostscript produisent un fichier texte illisible. D'abord le fichier pdf est comme scanné par *pdftools* (converti sous forme d'image) puis transformé en texte par *tesseract* qui effectue une reconnaissance des caractères. Les lettres et les groupes de lettres utilisés diffèrent d'une langues à l'autre, il s'avère utile de télécharger le dictionnaire français.

```
> if(is.na(match("fra", tesseract_info()$available)))
+ tesseract_download("fra")
> french <- tesseract("fra")
```

Cette première partie de code convertit les pages 2 à 8 de chaque pdf en image png, applique la reconnaissance des caractères et transforme le tout en une table de texte. Dans le code suivant *pdf* désigne le nom du fichier pdf à importer, l'opération est répétée sur les 198 dossiers.

```
>   doc<-pdf_ocr_text(pdf,  pages=2:8, language = "fra", dpi = 300)
>   doc<- paste (c(doc[[1]],doc[[2]],doc[[3]],doc[[4]],
+                     doc[[5]],doc[[6]],doc[[7]]),collapse  = "\n")
>   datadoc<- data.frame(x=strsplit(doc, '\n',"", fixed = TRUE))
>   colnames(datadoc)<-c("List")
```

L'exercice est ensuite de mettre en œuvre les fonctions R qui exploitent les expressions régulières et qui permettent de localiser, substituer ou extraire du texte. Dans notre cas, c'est généralement la recherche de mots clés qui détermine la lecture d'un champ à extraire, par exemple pour l'âge et le sexe du candidat, qui sur le pdf se trouvent sur la même ligne. La code suivant extrait la ligne qui contiennent le mot clé "Date de naissance". Comme le sexe ne contient que deux modalités, la présence du mot "Masculin" ou non sur cette ligne permet de renseigner le champs *'Sexe'* et d'enrichir la base de données. Pour l'âge, la stratégie retenue est d'extraire les caractères numériques dès lors qu'il y a en 4 consécutifs.

```
>   Dossier_Age_Rows <- grep("Date de naissance", datadoc$List)
>   if (regexpr("Masculin", datadoc[Dossier_Age_Rows[1],"List"])>0) {
+                         CEF[l,"Sexe"]="M"} else {CEF[l,"Sexe"]="F"}
>     o<-regexpr("[:0-9:]{4}", datadoc[Dossier_Age_Rows[1],"List"])
>   CEF[l,"Age"]<-substring(datadoc[Dossier_Age_Rows[1],"List"],o,o+3)
>
```

Le package *reshape2* permet de manipuler la base de données, en particulier pour la restructurer.

2 Analyse des données

Quelques statistiques descriptives

Dans cette section, nous faisons quelques statistiques élémentaires. Les packages *reshape2* et *dplyr* permettent de manipuler la base de données, respectivement pour la restructurer (transposer) et pour réaliser et répéter des calculs.

Les packages *ggplot2* et *scales* visent la construction de graphiques. *ggplot2* est une library très complète pour réaliser un graphique et *scale* le complète pour déterminer les graduations des axes et le placement des légendes.

```
> library(reshape2)
> library(dplyr)
> library(ggplot2)
> library(scales)
> #transposition de la table pour les variables 'texte'
> Categorial <- melt (CEF[, sapply(CEF, class) == 'character'],
+                  id.vars=c("Avis"), na.rm = TRUE,
+                 factorsAsStrings = TRUE)
> #calculs des fréquence des occurences
> var_select = c( "variable","Avis","value")
> Categorial2 = plyr::count(Categorial, vars= var_select)
> Categorial2<-Categorial2[!(Categorial2$variable %in%
+                              c("Motivation"
+                              ,"Com SCAC Cursus"
+                              ,"Com SCAC Entretien")),]
> # format de l'axe
> french_percent <- number_format(
+   accuracy = NULL,
+   scale = 100,
+   decimal.mark = ",",
+   suffix = " %"
+ )
> # création d'une fonction pour réaliser un graphique sur une
> # variable qualitative
> GraphiqueDescriptif<- function (var) {
```

```
+   Categorial4<-Categorial2[Categorial2$variable==var,]
+   select<-Categorial4$Avis=="Accepté"
+ Categorial4$freq[select]<-Categorial4$freq[select]/
+                           sum(Categorial4$freq[select])
+ Categorial4$freq[!select]<-Categorial4$freq[!select]/
+                           sum(Categorial4$freq[!select])
+      g<-ggplot(Categorial4, aes(fill =Avis,y=value, x=freq))+
+                  geom_col(position = "dodge")+
+                  xlab("")+   ylab("")+
+                  scale_x_continuous(labels = french_percent)
+  #   labs(title="Analyse variable de type Catégorie", subtitle=var)
+      print(g)
+ }
```

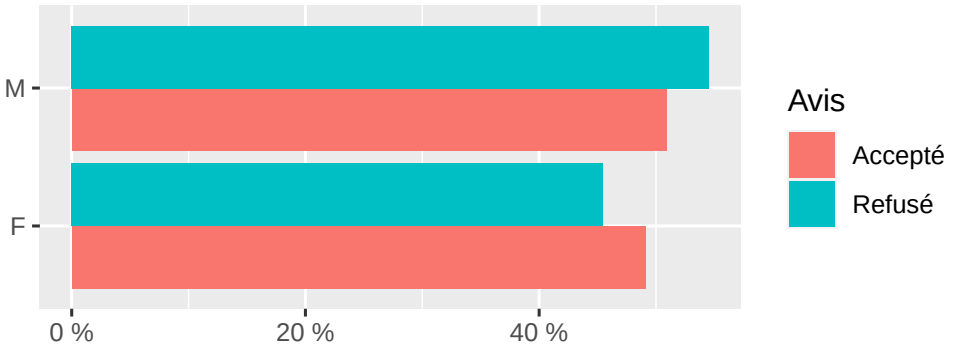

FIGURE 1 – Répartition Homme-Femme

La figure 1 révèle une inégalité homme-femme pour les candidatures Campus France. La sélection des candidats ne semble pas rééquilibrer ce résultat. Ce résultat est bien sur à comparer à la situtation dans l'emploi, notamment vers les métiers visés par le master. La référence des métiers sur l'assurance est fournie par . et le master vise les fonctions de

— 05I - Contrôle et surveillance du portefeuille
— 07A - Management des risques
— 02A - Marketing stratégique et études
— 08E - Études économiques, financières et statistiques

L'assurance est un secteur d'activité majoritairement féminin avec un taux d'employé femme respectivemnt de 74%, 62%, 66% et 61% dans les classes de métiers 05 (Gestion des contratsou prestations), 07 (Gestion et maîtrisedes risques internes), 02 (Marketing) et 08 (Pilotage économique,comptable et financier). En master 1, nous avions 13 femmes pour 11 hommes à l'inscription 2019 et 12 sur 24 à l'inscription 2018. Ce décalage, même modéré, invite à réfléchir à la manière de rendre plus attractive la formation aux étudiantes.

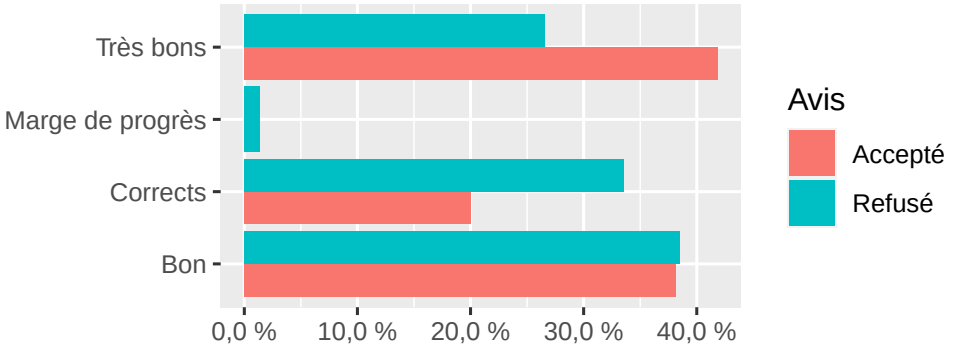

FIGURE 2 – Fréquence Avis du SCAC sur les résultats académiques du candidat

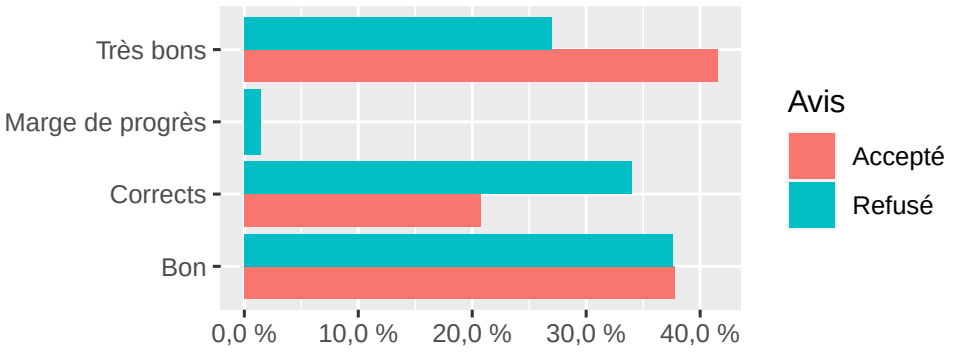

FIGURE 3 – Fréquence Avis du SCAC sur le cursus du candidat

Les figures 2 et 2 sont étonnamment similaires. Il semble que le SCAC évalue globalement le niveau de l'étudiant en fonction de son parcours et de ses notes et reporte la même information dans les deux champs.

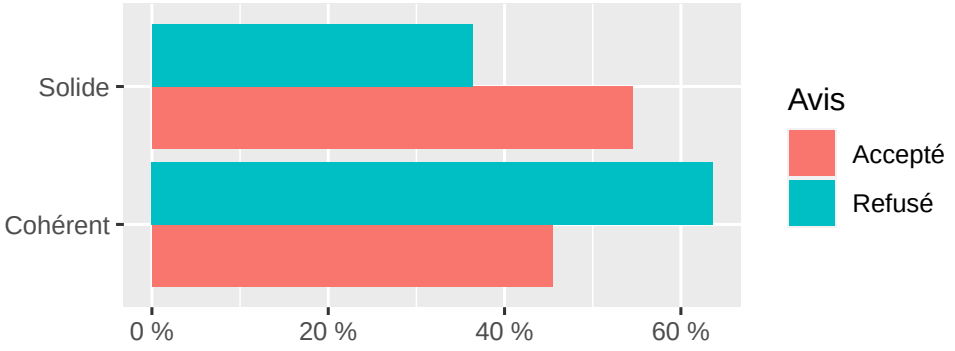

FIGURE 4 – Fréquence Avis du SCAC sur le projet du candidat

La figure 4 confirme deux résultats attendus. Les avis « correct » ou « insuffisants » ne parviennent pas jusqu'au responsable de la formation. De plus, un avis du SCAC « Solide » augmente les chances d'être accepté.

L'outil Campus France permet aux responsables de formation de remplir un champ à destination du SCAC. Nous avons communiqué depuis 2018 la même information aux candidats et aux SCAC (celles indiquées en introduction). Nous estimons que ce résultat pourrait être plus marqué parce qu'il n'est pas rare qu'un avis « Solide » concerne un étudiant de très bon niveau qui n'a pas les prérequis ou n'a pas un objectif professionnel cohérent avec le MBFA du Mans. Nous pouvons donc envisager de revoir notre communication sur les prérequis, les objectifs et le contenu de la formation auprès du SCAC.

La figure 5 montre que les candidatures respectent le minimum de niveau B2. Au delà de B2, la langue n'est pas le critère et cela se voit par le peu de différences entre les résultats sur les deux avis.

Le niveau de Langue est essentiel pour deux raisons. La première raison est pédagogique. Pour un étudiant non francophone, la compréhension des cours, la rédaction des copies et des rapports, l'interaction avec les membres de la promotion et la participation en cours dépendent fortement du niveau en langue. Tous ces éléments impactent la réussite académique. Bian and Malet [09 Jun. 2018].

Plus important, elle impacte aussi l'accès à l'emploi en France. Quand l'étudiant préfère réaliser un stage dans son pays et ambitionne d'y retourner travailler à l'issue de sa formation, l'impact est mineur. Par expérience, le niveau de langue est le premier critère pour l'accès à l'emploi en France. Retenir des étudiants avec un niveau insufisant en français posent des difficultés. Ils peinent souvent pour trouver un stage ou une alternance (le Master 2 est ouvert à l'alternance).

La figure 6 présente les résultats par mention en L3. Mais ce sont les valeurs manquantes et les non renseignés qui prédominent. C'est donc une information peu exploitable.

```
>  CEF$`YBAC.3`[CEF$Cohorte==2019] <- CEF$`YBAC.3`[CEF$Cohorte==2019]+1
>  CEF$`YBAC.3`[CEF$Cohorte==2018] <- CEF$`YBAC.3`[CEF$Cohorte==2018]+2
> ggplot(CEF[CEF$`YBAC.3`>0,], aes(x=`YBAC.3`, color=Avis,  fill=Avis)) +
```

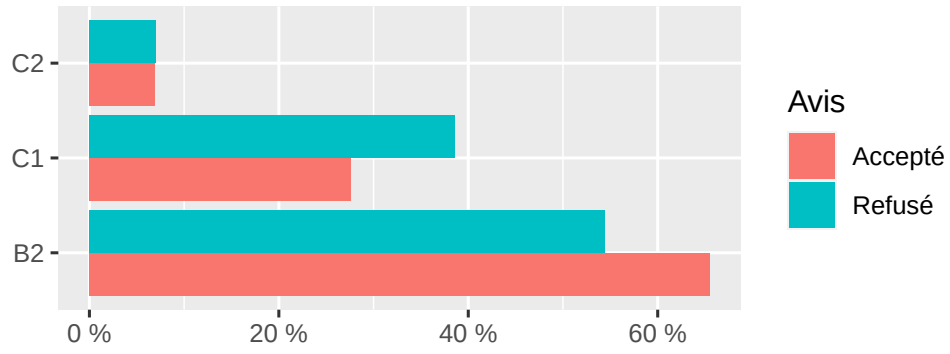

FIGURE 5 – Répartition du TCF, niveau de français

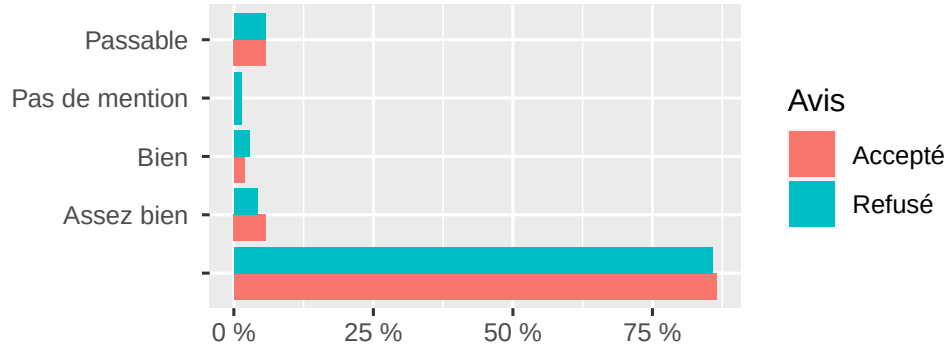

FIGURE 6 – Répartition des mentions obtenues en L3

```
+    geom_bar(aes(y = ..prop..),alpha=0.6,position="dodge")  +
+   xlab("Année d'obtention de la L3")+ ylab("")+
+   scale_y_continuous(labels = french_percent)
```

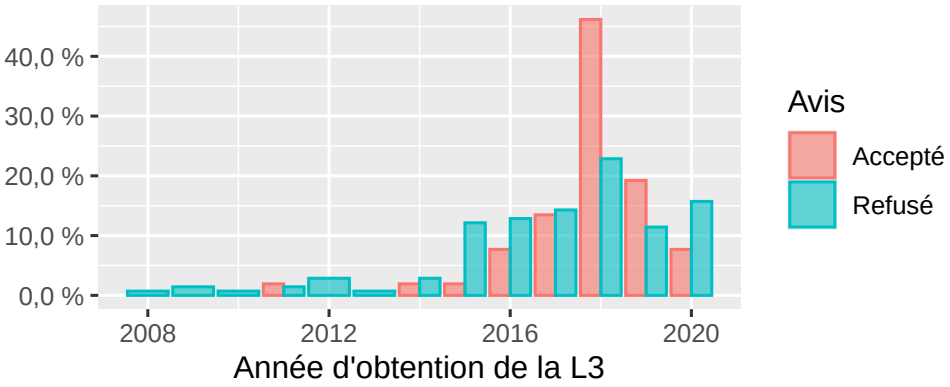

FIGURE 7 – Répartition de l'année d'obtention de la L3

La figure 7 montre l'année d'obtention de la L3. Pour la cohorte 2019, l'année a été incrémentée de 1 pour avoir l'équivalence. Comme le processus Campus France est long, les notes du S6 ne sont jamais disponibles pour les étudiants encore en L3, souvent mêmes les notes du S5 non plus. Ces dossiers sont donc souvent jugés trop incertains pour être réellement acceptés. Dans la grande majorité des cas, les étudiants étrangers acceptés ont leur licence depuis un ou deux ans. Ce que font les candidats dans l'intervalle est déterminant. S'ils poursuivent leur études, les spécialités de poursuite d'étude viennent confirmer ou infirmer l'adéquation du MBFA avec leurs objectifs. Si les notes sont disponibles, elles permettent de vérifier leur niveau. S'ils travaillent, on regarde si le métier exercé offre une expérience directement utile pour le master.

On regarde aussi quels niveaux de responsabilité ont été confié au candidat car le sens des responsabilités permet d'aborder l'apprentissage avec plus de sérieux, de résilience et d'efficacité.

```
>  CEF$YBAC[CEF$Cohorte==2019] <- CEF$YBAC[CEF$Cohorte==2019]+1
>  CEF$YBAC[CEF$Cohorte==2018] <- CEF$YBAC[CEF$Cohorte==2018]+2
> CEF$BL <- CEF$`YBAC.3`-CEF$YBAC
> ggplot(CEF, aes(x=BL, color=Avis,  fill=Avis)) +
+   geom_bar(aes(y = ..prop..),alpha=0.6,position="dodge")  +
+   xlab("Durée entre le Bac et l'obtention de la L3")+ylab("")+
+   scale_y_continuous(labels = french_percent)
```

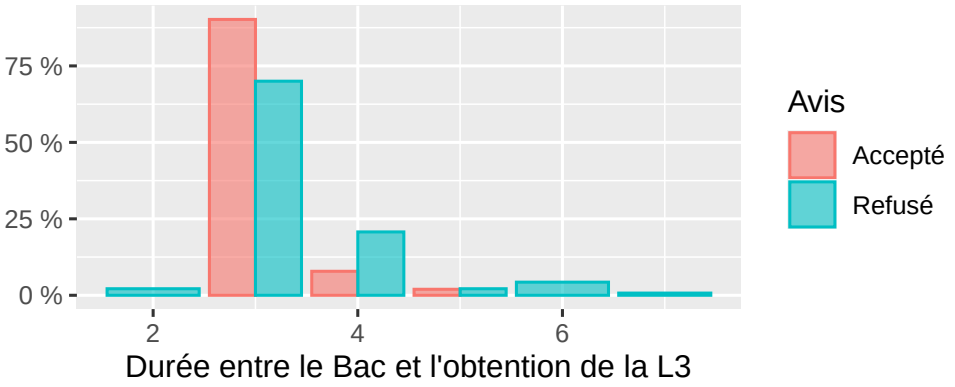

FIGURE 8 – Durée entre le BAC et l'obtention de la L3

La figure 8 exprime la différence entre l'année de l'obtention de la Licence et l'obtention du Bac. S'il n'y a pas de redoublement (ou de réorientation), elle est généralement de trois ans et parfois quatre ans. Pour la cohorte 2019, l'année du BAC a également été incrémentée de 1 pour avoir l'équivalence. Ce graphique prouve que les redoublements pénalisent le candidat. Aucune candidature affichant plus de 4 ans pour obtenir sa licence n'a été acceptée. À l'issue de ce constat, nous pourrions envisager de refuser toute candidature avec redoublement à l'université.

Cela soulève une question : les réorientations impactent-ils positivement ou négativement sur la candidature ? Un candidat qui obtient des notes correctes sur l'orientation abandonnée prouve sa résilience bien que les cours suivis ne correspondent pas à ses ambitions. S'il progresse significativement dans sa nouvelle voie, nous imaginons que celui-ci se passionne maintenant pour ce qu'il apprend. Dans le cas contraire, nous imaginons un candidat insatisfait chronique peu capable de s'intéresser ou de faire preuve de recul. Quelque soit la réorientation, nous espérons que l'étudiant l'explique dans sa candidature si elle est récente.

L'ensemble des quatre densités en figure 9 représentent les moyennes respectivement obtenues en Licence 2 & 3 et en Master 1 & 2. Seules respectivement 102, 99, 70 et 28 valeurs sont utilisées. En premiers lieux, les moyennes sont pas systématiquement renseignées, ensuite certains pays utilisent des notes sur 100 ou sur 5, qui sont donc exclues de l'analyse. De plus, un étudiant en L3 aura une valeur non renseignés pour la L3 non achevée et pour les niveaux supérieurs, ce qui explique que le nombre de valeurs prisent en compte diminue avec le niveau après le BAC.

Il est intéressant de noter le changement de configuration entre la licence et le master. Si en Licence, les meilleurs moyennes ont plus de chance d'être pris, ce phénomène s'efface voir s'inverse en Master. En fait, le changement vient plus densité des Refusés, dont la moyenne augmente entre la licence 2 et le master 2. Ce résultat s'explique de deux manières. Premièrement, les masters étant plus spécialisés, l'adéquation des prérequis prend plus d'importance que les notes. Deuxièmement, si l'adéquation est bonne, un excellent niveau diminue l'intérêt ou la possible progression pédagogique que le master MBFA peut apporter à l'étudiant.

Analyse des mots utilisés dans la lettre de motivation

La librarie *tm* est dédiée au text mining. Nous analyserons dans cette section la lettre de motivation des candidats. La première étape est de construire un corpus lexical qui associe à chaque mot utilisé, le nombre d'occurence grâce à la fonction *tm : :corpus*. Cette analyse nécessite quelques précautions d'usage :

— la suppression des caractères spéciaux, des nombres,
— indiquer la langue qui optimise la construction du corpus en fonction de la langue,

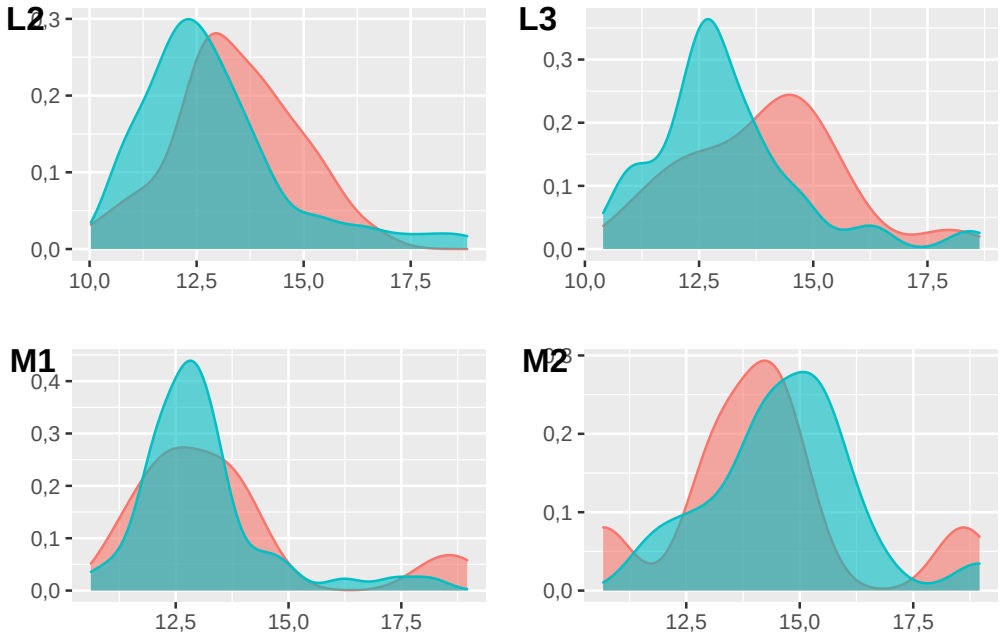

FIGURE 9 – Moyennes obtenues en respectivement en L3 et en L2

— suppression des mots vides (*stopswords*), ils sont tellement communs que les indexer n'apporte aucune information. Ils sont bien sûr différent d'une langue à l'autre.
— la gestion des sigles qui pourraient être confondus avec un mot.
Le code suivant crée la fonction CreationCorpus.

```
> library(tm)
> Avis="Accepté"
> CEF$Motivation<-gsub("CCA","Comptabilité contrôle audit", CEF$Motivation)
> CreationCorpus <-function (Avis){
+   print(Avis)
+ Motivations <- as.character (CEF$Motivation[CEF$Avis==Avis])
+ print(Motivations)
+ Motivations<-gsub("['']`^~\"?’]"," ",Motivations)
+ vs <- VectorSource(Motivations)
+ Motivations<-tm::Corpus(vs, readerControl=
+                   list(reader=readPlain, language="fr", load=TRUE))
+ #retrait des ponctuations
+ Motivations <-tm_map(Motivations,removePunctuation)
+ #retrait des nombres
+ Motivations <-tm_map(Motivations,removeNumbers)
+ #retrait des stopwords (mots outils)
+ Motivations <-tm_map(Motivations,removeWords,stopwords("french"))
+ #retirer les espaces en trop (s'il en reste encore)
+ Motivations <-tm_map(Motivations,stripWhitespace)
+ dtm<-DocumentTermMatrix(Motivations)
+ dtm.matrix <- as.matrix(dtm)
+
+ wordcount <- sort(colSums(dtm.matrix), decreasing=TRUE)
+ return(as.data.frame(wordcount))
+ }
```

Pour chaque mot, le code suivant établie le nombre d'occurences dans les lettres de motivation pour les acceptés et les réfusés.

```
> wordcountR<-CreationCorpus("Refusé")
> colnames(wordcountR)<-'Refusé'
> wordcountA<-CreationCorpus("Accepté")
```

```
> colnames(wordcountA)<-'Accepté'
> Diffwordcount<-merge(wordcountR, wordcountA,by=0,
+                      sort = TRUE,all=TRUE)
```

L'algorithme indexe 3211 mots, mais l'analyse n'est pas immédiate. Un mot peut avoir plusieurs sens et plusieurs déclinaisons. Par exemple, les mots actuariat, actuariel, actuarielles, actuaire relèvent de la même science. L'étape suivante vise de regrouper les mots ayant la même racine, grâce à la librairie *SnowballC*. Deux techniques sont possibles : la lemnisation et la racination. La racination tronque un ensemble de suffixes et préfixes adaptés à la langue. La lemnisation considère un dictionnaire beaucoup plus élaboré, capable d'associer les mots "suis" et "sommes" au verbe être. Mais sa mise en œuvre serait plus complexes. Par la fonction SnowballC : :wordsterm, le code suivant donne la racine de chaque mot du corpus, et agrège les occurrences par groupe de mots de même racine.

```
> library("SnowballC")
> Diffwordcount$stem <- wordStem(Diffwordcount$Row.names,
+                                language = "french")
> #length(unique(Diffwordcount$stem))
> DW<-aggregate(Diffwordcount[,c("Accepté","Refusé")],
+               by=list(stem = Diffwordcount$stem), FUN=sum
+               , na.rm=TRUE, na.action=NULL)
> colnames(DW)<- c("stem" ,"AcceptéStem", "RefuséStem")
> DWC<-merge(Diffwordcount,DW, all = TRUE)
> DWC[is.na(DWC)] <- 0
> DWC <- DWC[(DWC[["RefuséStem"]]+DWC[["AcceptéStem"]]>20),]
```

En regroupant les mots ayant la même racine et en limitant l'étude aux mots (la racine commune) comptabilisés plus de 20 fois, le nombre de mots indexés diminue à 238.

Maintenant comment distinguer les mots qui discriminent l'entrée en Master MBFA ? L'*ODDS ratio* ou rapport de cote ou risque relatif rapproché est une statistique simple qui permet de mesurer le lien entre la probabilité d'occurrence d'un mot et l'avis émis. Si la probabilité qu'un mot arrive dans les lettres de motivation d'un candidat « Accepté » est p, et q d'un « Refusé », l'*ODDS ratio* se défini par :

$$\frac{p/(1-p)}{q/(1-q)} = \frac{p(1-q)}{q(1-p)}$$

Nous sélectionnons donc les valeurs très supérieures à 1 ou très inférieures à 1, ici le seuil est fixé à $3/2$ et $2/3$.

```
> # calculs des sommes d'occurences
> A<-sum(DWC[["Accepté"]])
> R<-sum(DWC[["Refusé"]])
> DWC[["ODDS"]]<-DWC[["AcceptéStem"]]/(A-DWC[["AcceptéStem"]])/
+                 (DWC[["RefuséStem"]]/(R-DWC[["RefuséStem"]]))
> DWC <- DWC[(DWC[["ODDS"]]>3/2 | DWC[["ODDS"]]<2/3),]
> DWC$PA <-DWC[["AcceptéStem"]]/A
> DWC$PR <-DWC[["RefuséStem"]]/R
>
> #write.xlsx(DWC,"CEF_MotivationWordList.xlsx",col.names = TRUE)
```

Parmi ces 68 mots trouvés significatifs statistiquement, nous sélectionnons les mots qui ont sens unique. Par exemple, le mot « intérêt » est trouvé significatif pour les refusés et le mot « intéressant » pour les acceptés. Il n'ont pas la même racine car, suivant le contexte l'intérêt peut avoir deux sens différents. De plus, les deux sens ont un rapport avec leur candidature. Montrent-t-il leur intérêt pour la formation et leurs capacités en mathématique financière ? Il y a aussi les mots qui n'ont pas de réelle pertinence seuls. Par exemple, le mot « métier » a un sens pour un responsable de formation dans son contexte, en particulier en notant de quel métier il s'agit. Ce genre de mots ajouterait de la confusion dans l'analyse, mieux vaut bâtir des conclusions sur les mots avec un sens unique.

```
> ShortListWort <- c("risque","quantitatif","economie",
+                    "passion","crédit","espère",
+                    "détermination","économétrie","statistique",
+                    "programmation","aspirations","compétences",
+                    "gestion","entreprise", "réputation",
+                    "comptabilité", "dynamisme","renommée")
> DWC<-DWC[(DWC$Row.names %in% ShortListWort),]
```

```
> DWC<-DWC[order(DWC$ODDS),]
> dfplot <- as.data.frame(melt(DWC[,c("Row.names","PA","PR")]))
> dfplot$word <- factor(dfplot$Row.names,
+                       levels=dfplot$Row.names[1:(nrow(dfplot)/2)])
```

La figure 10 montre les probabilités d'occurrence des mots préalablement sélectionnés de la lettre de motivation qui discriminent le plus entre Accepté et Refusé.

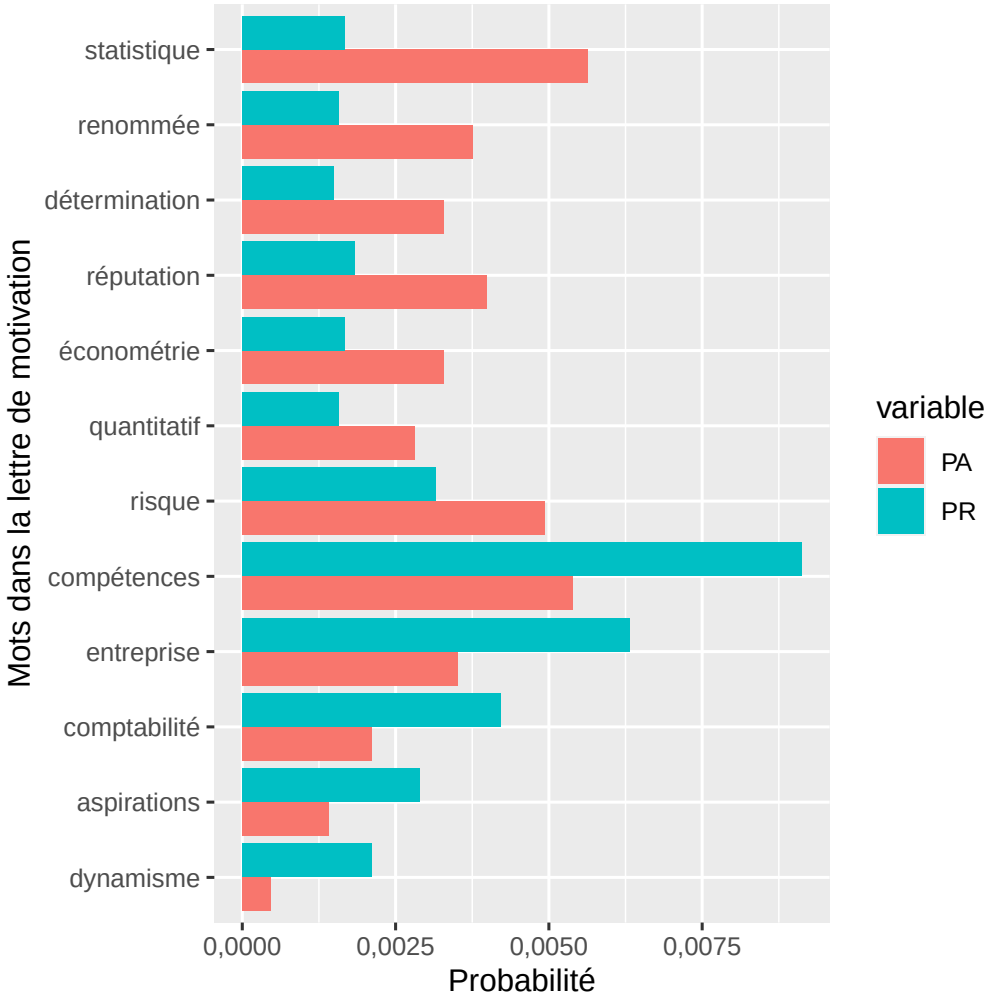

FIGURE 10 – Liste des mots qui discriminent le plus entre Accepté et Refusé (mots triés par ODDS)

Nous pourrions penser que la lettre de motivation est un exercice factice, le candidat se fixant comme unique objectif de rédiger ce que le responsable de la formation espère lire. Mais au global, ces statistiques montrent sa pertinence. Nous trouvons deux types de mots significatifs, ceux qui désignent l'orientation et ceux qui désignent la motivation. Si le « Refusé » est « dynamique », qu'il a des « aspirations » et souhaite acquérir des « compétences », l'« Accepté » lui exprime son « espoir », sa « détermination » et choisi une formation en fonction de sa « réputation/renommée ».

Le lien entre l'avis et les mots exprimant l'orientation est également intéressant. Si le « Refusé » vise la « comptabilité », le« crédit », la « gestion » et la gestion d'« entreprise », l'« Accepté » lui vise l'analyse « quantitative », le « risque », l' « économétrie », l' « économie » et les « statistiques ». Il est logique que nous ayons accordé un avis favorable aux candidats qui font références dans leur lettre de motivation aux mots clés de la formation. Dans ce cas, nous déduisons que l'étudiant c'est réellement informé sur le master et que ces matières correspondent à ses aptitudes et aspirations.

Malheureusement, cette méthode n'indexe pas tous les mots clés, comme « SAS »,« R », « SQL » ou « Python ». Ces éléments poseraient des difficultés lors de l'élaboration du corpus, parce que sont des sigles (parfois inférieurs à trois lettres) ou parce que ce ne sont pas des mots français.

2.1 Analyse des sentiments du SCAC

Les service de Coopération et d'Action Culturelle rédigent deux commentaires plus détaillés, l'une sur l'entretien et l'autre sur le cursus du candidat. L'objet de cette section est d'analyser ces commentaires. Un commentaire très positif impacte-t-il la décision du responsable de la formation ? La méthode de base est de comparer les commentaires à des lexiques de mots retenus comme positifs et de mots retenus comme négatifs en français.

Plusieurs lexiques libres sont disponibles. Le premier est lié à la librairie Orange Demšar et al. [2013] développé par Bioinformatics Lab de l'Université de Ljubljana, Slovénie, en collaboration avec la communauté open source. Le deuxième, noté Feel Abdaoui et al. [2017] est issue du projet ADVANSE : ADVanced Analytics for data SciencE. D'autres lexiques détaillent les sentiments selon une typologie : l'angoisse et l'excitation, le dégoût et la crainte, la joie et la tristesse, la surprise et la confiance.

```
> CEF[["Com_SCAC_Entretien"]]<-gsub("CCA",
+              "Comptabilité contrôle audit",
+              CEF[["Com_SCAC_Entretien"]])
> CEF[["Com_SCAC_Cursus"]]<-gsub("CCA",
+              "Comptabilité contrôle audit",
+              CEF[["Com_SCAC_Cursus"]])
>   Feel<- read.csv("FEEL.csv",header = TRUE ,
+              encoding = "UTF-8",sep = ";")
>   Biolabp<- read.csv("positive_words_fr.txt",
+              header = FALSE , encoding = "UTF-8",sep = ";")
>   Biolabn<- read.csv("negative_words_fr.txt",
+              header = FALSE, encoding = "UTF-8",sep = ";")
>     Avis="Accepté"
>   CommentSCACAccepte <- as.character (
+            rbind(CEF[["Com_SCAC_Entretien"]][CEF$Avis==Avis],
+            CEF[["Com_SCAC_Cursus"]][CEF$Avis==Avis]))
>    FeelAp <-unlist(lapply(Feel$word[Feel$polarity=="positive"],
+            function(x) grep(x, CommentSCACAccepte, fixed = TRUE)))
>   FeelAn <-unlist(lapply(Feel$word[Feel$polarity=="negative"],
+             function(x) grep(x, CommentSCACAccepte, fixed = TRUE)))
>    BiolabAp <-unlist(lapply(Biolabp$V1,
+             function(x) grep(x, CommentSCACAccepte, fixed = TRUE)))
>   BiolabAn <-unlist(lapply(Biolabn$V1,
+             function(x) grep(x, CommentSCACAccepte, fixed = TRUE)))
>   Avis="Refusé"
>   CommentSCACRefuse <- as.character (
+           rbind(CEF[["Com_SCAC_Entretien"]][CEF$Avis==Avis],
+           CEF[["Com_SCAC_Cursus"]][CEF$Avis==Avis]))
>    FeelRp <-unlist(lapply(Feel$word[Feel$polarity=="positive"],
+          function(x) grep(x, CommentSCACRefuse, fixed = TRUE)))
>   FeelRn <-unlist(lapply(Feel$word[Feel$polarity=="negative"],
+         function(x) grep(x, CommentSCACRefuse, fixed = TRUE)))
>    BiolabRp <-unlist(lapply(Biolabp$V1,
+         function(x) grep(x, CommentSCACRefuse, fixed = TRUE)))
>   BiolabRn <-unlist(lapply(Biolabn$V1,
+         function(x) grep(x, CommentSCACRefuse, fixed = TRUE)))
```

Il ressort que les avis positifs obtiennent les scores de 0,583 (Biolab) et 0,781 (Feel) contre 0,564 (Biolab) et 0,777 (Feel) pour les avis négatifs. Avec les deux lexiques, nous constatons peu d'écart du sentiment du SCAC entre les avis positifs et négatifs. Les conclusions du SCAC sur le projet donnent des résultats plus marqués (voir fig 4).

3 Ajustement du process de recrutement

Le premier changement pour 2021, déjà mis en œuvre en 2020 pour le canal ecandidat, est de demander un entretien vidéo différé de 5-6 minutes aux candidats présélectionnés. Cette méthode a été choisie parce qu'elle est légère en terme d'organisation et peu chronophage. La convocation à un entretien (présentiel ou distanciel) permet un vrai échange avec le candidat et une évaluation de sa motivation, mais nécessite

beaucoup d'organisation et de temps. Notre choix est un compromis pragmatique entre l'entretien direct et une acceptation sur le seul dossier. Nous avons obtenu un taux de retour de 73,4% en 2020. L'expérience montre que l'exercice révèle tout de même la personnalité du candidat, la qualité de son français et son degré de connaissance du Master. De plus cela lui donne l'occasion de confirmer sa motivation.

Le deuxième changement est plus directement lié à cette étude. Avec Mme Nguyen Thi Thanh Huyen, co-directrice du Master, nous demanderons aux candidats de renseigner un questionnaire complémentaire.

Ces évolutions visent un premier objectif : motiver l'étudiant à réexaminer l'offre de formation du Master MBFA du Mans pour remplir le questionnaire ou pour préparer l'entretien vidéo différé. Pour sa candidature en ligne, le MBFA du Mans est peut-être une candidature répliquée. L'étudiant a une préférence pour un autre master, il a préparé sa candidature avec soin pour celui-ci. Mais pour sécuriser son avenir académique, il décide de postuler aussi au Mans, mais peut-être sans avoir réellement examiné son contenu.

Pourquoi ? Quel est l'intérêt pour les responsables de la formation ?

— en se renseignant à nouveau, le candidat pourra réfléchir à son adéquation avec la formation, de s'auto-sélectionner. L'objectif est de réduire le nombre de candidatures inadaptées qui occasionnent des pertes de temps. Le questionnaire [2] calcule un score un score représentatif de l'adéquation entre le profil du candidat et les objectifs professionnels et pédagogiques du Master pour l'aider dans sa réflexion.

— donner au candidat l'occasion d'adhérer à la formation, d'y voir ses avantages pour son projet professionnel et de renforcer sa motivation. L'objectif est de mieux préparer les étudiants aux exigences du Master, de les convaincre que le choix du Mans ouvre sur l'emploi. Un certain nombre de questions visent donc la réflexion de l'étudiant sur son projet professionnel, encouragent la recherche d'information sur le diplôme et vise la promotion de la formation.

Le score d'adéquation avec le MBFA du Mans

Un score, même bien construit, ne peut détecter toute la richesse d'une candidature, mais il offre une indication à l'étudiant pour qu'il s'auto-sélectionne. L'idée est de faire une moyenne entre trois notes :

1. une sur le niveau académique,
2. une sur l'adéquation des prérequis,
3. une dernière sur l'adéquation avec le projet professionnel.

Le principe de la note sur le niveau académique est de se baser sur les notes semestrielles de l'étudiant et de majorer/minorer cette moyenne en fonction des informations complémentaires. Nous demandons s'il a obtenu cette moyenne en première ou seconde session et en combien d'année il a eu sa licence. Nous demandons s'il a effectué son année au Mans, dans l'UE ou hors de l'UE. De fait, le lieu ne traduit pas dans ce contexte un écart de niveau, mais une sécurité prise pour faire face à notre ignorance du réel niveau d'exigence académique de la formation d'origine du candidat. L'impact est forfaitaire et ne traduit pas les disparités existantes au sein de ces formations. Demander plus de détail sur cette question serait sans doute plus équitable, mais il serait impossible d'être exhaustif et l'écart de note sur chaque formation serait peut être inutilement perçu comme de la discrimination. Cette moyenne des notes minorées/majorée est diminuée d'un tiers de son écart type afin d'avantager les résultats réguliers.

C'est surtout sur l'adéquation entre les prérequis et le master MBFA que le score doit être judicieusement sélectif. La deuxième note évalue donc si les modules suivis offrent les prérequis pour suivre le master MBFA du Mans. La difficulté est que l'intitulé de la licence n'est pas suffisant pour en connaître le contenu, en particulier les licences Gestion & Éco-gestion. Nous procédons donc en deux étapes.

Quelle est la discipline dominante de votre dernier diplôme (L3 le plus souvent, le M1 parfois) ? La réponse donne un point de départ à la note :

Économie	9.5
Économétrie	10.0
MAAS	10.0
Mixte Économie/Gestion	8.5
Gestion	7.5
Mathématique	8.5
Informatique	8.0
LP Banque/Assurance	6.0

Même s'il commence à 9,5, une licence d'économie est assurée d'avoir au moins 10 grâce aux questions suivantes. À l'opposé, les Licences Pro n'ont pas les prérequis pour intégrer le Master. Mais en commençant à 6, nous sommes assurés d'une note inférieure à 10. Comme l'objectif est de mesurer l'adéquation et aucunement d'humilier qui que ce soit, il n'est pas utile de mettre une note de départ inférieure.

2. Il a été optimisé pour être rempli sur Acrobat Reader, le calcul se fait sans stockage de données sur un serveur.

Le candidat précise quels sont les cours qu'il a suivi en L3 en économie et en statistiques. Chaque Oui rapporte des points. Plus la matirère est utile pour suivre le master, plus la bonification est importante.

Économie industrielle	0,50
Économétrie	0,75
Économique du risque	0,75
Optimisation	0,50
Micro économie	0,50
Théorie des jeux	0,50
Probabilité & Mesure	0,75
Statistiques descriptives	0,125
Recherches opérationnelles	0,50
Macro économie	0,25
Économie agricole/ publique	0,125
Mathématiques financières	0,25
Inférence statistique	0,50
Séries temporelles	0,50
Équilibre général	0,25

Un des piliers de la formation est l'informatique, plus particulièrement SAS. Il est donc intéressant de qualifier la culture informatique de l'étudiant selon les critères suivants :

Le logiciel est	Point
un langage de programmation est	1,00
orienté Statistiques	1,00
orienté visualisation de la données	0,30
orienté Survey/Plan d'Expérience	0,25
orienté Économétrie	0,50
orienté cartographie (modèle/visualisation)	0,75
orienté Big Data	1,00
un logiciel de Requête/Base de Données	0,25
un logiciel de Gestion/Comptabilité	0,10

La valeur du point reflète l'adéquation avec les compétences informatiques du Master. Nous soumettons une liste de 182 logiciels, établie essentiellement à partir de listes de popularité comme celle des logiciels les plus utilisés dans l'enseignement des statistiques [3] ou comme la liste Tiobe par exemple. SAS, R, Python, Excel/Calc/Numbers obtiennent respectivement 3,75, 3,25, 3,75 et 0,8 points. Même si c'est très bien d'avoir utilisé SAS ou un autre langage de programmation en économétrie, le niveau d'expérience ou d'acquisition d'un savoir faire sur le logiciel est à prendre en compte. Cinq niveaux sont proposés à l'étudiant : Débutant, Basique, Opérationnel, Avancé, Expert ou Formateur. La note du logiciel est ensuite multiplié par la réponse sur le niveau (respectivement 0,7, 1,2, 2, 3,5, et 5). Le score de l'adéquation logiciel est alors la somme de cinq produits.

La note d'adéquation du parcours est la note obtenue en renseignant la discipline dominante augmentée des points sur les modules suivis en licences plus le score de l'adéquation logiciel affecté d'une pondération de 7,5%.

La troisième note représente l'adéquation avec le projet professionnel. Nous soumettons une liste de métiers que nous avons trouvés dans les lettres de motivation des candidats et les intitulés des métiers ciblés à l'issue de la formation. Nous leur demandons de choisir, entre 1 et 5 métiers et entre 1 et 3 lieux d'exercice qui vous conviendrait le mieux. La note d'adéquation est la note moyenne des métiers multipliée par le facteur moyen présenté dans les tables suivantes.

3. hal-00913110

Intitulé des Métiers (extrait)	Note
Chargé d'étude économique	20
Chargé d'étude marketing quantitatif	20
Chargé d'étude Pilotage économique	20
Analyste Know Your Custumer	18
Chargé d'étude	18
Ingénieur économique	16
SAS Data Manager	16
Gestion des risques financiers	15
Chargé d'étude statistique	15
Responsable lutte contre la fraude	15
Actuaire	14
Chercheur	14
Gestionnaire sinistre	10
Directeur d'agence	9
Trader	9
Conseiller en patrimoine	8
Conseiller financier	8
Comptabilité générale	5
Contrôle de gestion	5
Juriste	1

Où ? (extrait)	facteur
Agence bancaire	0,25
Agence d'assurance	0,30
Assurances (services centraux)	1,00
Caisses de retraites ou de santé	1,00
Instituts de prévoyance	1,00
Réassurances	1,00
Salle de marché	0,70

Ces trois notes d'adéquation sur 20 (niveau académique, adéquation du parcours, adéquation du projet professionnel) sont pondérés respectivement par 3,5, par 5 et par 1 pour former le score final. Les notes et le parcours ont globalement la même importance, mais les matières les plus en adéquation font aussi partie des matières jugées les plus difficiles et où les étudiants obtienne plus difficilement de très bonne note. L'adéquation au projet professionnel intègre faiblement le score car la connaissance du monde de l'emploi et la maturité sur le sujet est généralement faible.

Des questions qui visent la réflexion

Nous posons en premier lieux des questions classiques, comment l'étudiant a connu la formation et où a-t-il trouvés les informations. En demandant s'il a visionné les vidéos qui décrivent le Master[4], l'objectif est de susciter la curiosité ou l'envie de la visionner car il préfère naturellement répondre « Oui » à ce type de question. De même, quand on lui demande les critères qui déterminent si oui ou non, il candidate à un master, nous proposons les spécificités qui mettent en valeur le master MBFA du Mans comme l'alternance et les certifications TOSA et SAS. Nous lui demandons à la fin quels sont pour lui les facteurs de réussite. Finir sur cette question positive place l'étudiant vers l'avenir et l'invite à se responsabiliser ou à devenir acteur de sa réussite académique. De même, en demandant s'il a consulté les syllabus[5] (ou description) des modules, il est implicitement invité à regarder attentivement les matières qui l'attende s'il est admis en Master. Des questions sur leur ambition salariale les invite à réfléchir sur le fait que le salaire d'un jeune diplômé diffère d'une formation à l'autre. Nous avons indiqué le salaire médian observé formellement pour les alternants en 2020, et le niveau de salaire net mensuel moyen que nous avons obtenu informellement auprès d'étudiants en deuxième année d'exercice.

Nous leur proposons de livrer les critères qui déterminent leur choix de master. La liste de critères provient essentiellement de cette étude de E. Moyou intitulée : "Critères de choix d'un établissement d'études supérieures par les jeunes France 2018" (10 sept. 2019). Nous avons ajouté quelques critères qui différencient le MBFA du Mans de quelques-uns de ses concurrents pour en faire la promotion.

4. Vidéo de présentation du master MBFA
5. MBFA > Programme > Organisation de la formation & Tout déplier

Critère de sélection
Le lieux d'étude : proximité de la famille / des amis
Le lieux d'étude : l'ambiance et la qualité de vie
Le lieux d'étude : la notoriété de l'Université
Le contenu de la formation : La liste des modules
Le contenu de la formation : l'alternance
La qualité de l'enseignement et de l'équipe pédagogique
Les perspectives de carrières (métier, taux d'emploi, salaire)
Le coût de la formation
L'accessibilité (sur concours/dossier/entretien...)
Les taux de réussite/d'obtention du diplôme
Le classement dans les palmarès (Classement Eduniversal, Meilleurs-Masters.com ...)
L'ouverture à l'entreprise (enseignant vacataire, alternances...)
L'environnement de travail (locaux, équipements...)
Les certifications proposées (TOSA, SAS...)

Nous profitons de ce questionnaire pour voir comment est perçue la lutte contre le plagiat et la fraude. Nous l'avons intensifiée jusqu'à systématiser l'usage de logiciels anti-plagiat pour les rapports rendus au cours de la formation, en particulier les mémoires de stages. Nous savons qu'elle contribue à la confiance établie avec les entreprises qui accueillent nos stagiaires et nos jeunes diplômés. Mais l'étude de Guibert and Michaut, qui rend publique une enquête sur le plagiat réalisée à l'université du Québec en Outaouais, prouve qu'elle peut être aussi considérée positivement par les étudiants. En particulier, les étudiants ont rapporté que

— La détection systématique et les sanctions sont des facteurs de justice et d'équité.
— Le plagiat et la fraude sont contraires aux missions de l'Université.
— Les personnes qui font du plagiat ne méritent pas le même diplôme que moi.
— La lutte contre le plagiat favorise l'esprit d'équipe de la promotion.

Le questionnaire nous permettra de discerner l'adhésion des candidats à ces quatre assertions et de comprendre leur perception de notre lutte contre la fraude.

La dernière question porte sur les facteurs de succès à l'université. C'est bien notre perception des chances de réussite du candidat qui sont au cœur de notre décision pour admettre ou refuser un candidat. Autant le faire savoir en conclusion de notre questionnaire. Nous demandons : quels sont pour vous les facteurs les plus déterminants de réussite ou d'échec en Master ?

Facteurs de réussite
La motivation à l'entrée
L'assiduité et le travail scolaire
La qualité des pré-requis antérieures (en L3 principalement)
Le financement des études (parents/bourses/Job étudiant)
La dynamique de la promotion et de l'association étudiante
Le contexte familial, social et la santé de l'étudiant
Le niveau d'exigence des enseignants
Les modalités de contrôle des connaissances
Les outils et qualités pédagogiques des enseignants
Le niveau de français

La liste de propositions reprend le travail de Duguet et al. [2016]. Cette étude fait une revue de littérature sur les facteurs de réussite des étudiants, ceux qui étaient majeurs dans les années 90, et leurs évolutions avec l'internationalisation des recrutements et des évolutions sociales et numériques. L'idéal serait qu'il réfléchissent à leur propre situation au regard de ces critères. Ils ne peuvent pas tout changer à leur situation, mais réaliser quels sont leurs chances de réussite et bien utiliser les leviers à leur portée peut changer le résultat. C'est ce que nous leur souhaitons.

Conclusion

Sur le plan scientifique, cette étude livre un exemple de mise en œuvre des techniques empruntées au *text and data mining* dans la résolution d'un problème pratique. À partir d'image pdf, structurée par une mise en forme standardisée issue du fichier électronique des candidats campus France, nous construisons une base de données sous R. Outre les statistiques descriptives nécessaires, nous décryptons la différence de vocabulaire dans les lettres de motivation et recherchons si les sentiments que le SCAC exprime dans ces commentaires pourrait affiner notre sélection.

L'analyse de ces candidatures confirme qu'un bon niveau général et un bon niveau en langue française sont requis. Cela est confirmé par l'étude de variable comme le nombre d'années pour obtenir la L3 ou les résultats au TCF et les conclusions du SCAC.

La difficulté est d'évaluer simplement et objectivement l'adéquation avec le master MBFA du Mans. Trop de dossiers refusés sont pourtant jugés "Très bon cursus" par le SCAC (26,9%). L'avis du SCAC est globalement conforme au niveau de l'étudiant mais discrimine mal les candidatures n'ayant pas les prérequis pour viser le Master d'économie/économétrie MBFA du Mans.

La bonne surprise vient de la lettre de motivation. Alors qu'elle s'avère parfois factice (rédaction souvent assistée par un proche ou motivation sur-jouée), les statistiques montrent pourtant une différence significative entre les candidatures. Le candidatures acceptée utilisent les mots clés en accord avec le contenu de la formation et un vocabulaire exprimant une motivation plus intense.

Par contre, les SCAC ont deux champs pour exprimer de manière plus détaillée leur avis sur le candidat. Dans ces deux champs, ils rappellent très souvent les éléments clés du dossier (parcours et notes). Cela s'avère très pratique dans les faits pour le responsable de la formation. Dans de nombreux cas, il n'est alors plus nécessaire de voir les détails des relevés de notes ou les autres justificatif pour refuser le candidat. Néanmoins, l'analyse du sentiment exprimé par le SCAC ne révèle pas de différence significative entre les candidats acceptés et refusés.

Rappelons l'objectif de cette étude : décourager les étudiants qui ne peuvent être retenus de postuler sans décourager les étudiants qui nous intéresse. L'idée a donc été de construire un questionnaire complémentaire qui calcule un score qui reflète l'adéquation du candidat avec le Master. Il prend en compte nos prérequis et les résultats de cette étude. Il y a des questions qui n'influencent pas le score. Comprenons que le questionnaire vise à s'assurer que l'étudiant s'informe réellement sur la formation et en particulier sur ses spécificités, ses qualités et ses valeurs. Le but est de le faire adhérer au projet pédagogique proposé, pour qu'un étudiant motivé le soit encore plus et qu'il se prépare en vu de le réussir.

Références

Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. Feel : a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3) :833–855, 2017.

Cui Bian and Régis Malet. *Language(s) of Power and Power of Language(s) in International Student Recruitment in French Heis in the Context of Internationalization*, pages 83 – 99. Brill | Sense, Boston, USA, 09 Jun. 2018. ISBN 9789463512251. doi : https://doi.org/10.1163/9789463512275_005. URL https://brill.com/view/book/edcoll/9789463512275/BP000014.xml.

Janez Demšar, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, Miha Štajdohar, Lan Umek, Lan Žagar, Jure Žbontar, Marinka Žitnik, and Blaž Zupan. Orange : Data mining toolbox in python. *Journal of Machine Learning Research*, 14 :2349–2353, 2013. URL http://jmlr.org/papers/v14/demsar13a.html.

Amélie Duguet, Marielle Le Mener, and Sophie Morlaix. Les déterminants de la réussite à l'université. quels apports de la recherche en éducation ? quelles perspectives de recherche ? 2016.

Christine Farrugia and Rajika Bhandari. Global trends in student mobility. 2018.

Pascal Guibert and Christophe Michaut. Le plagiat étudiant. *Education et sociétés*, 28(2) :149, 2011. doi : 10.3917/es.028.0149.

Corentin Roquebert. Workshop : Text analysis with r, 2019.

Creso M. Sá and Emma Sabzalieva. The politics of the great brain race : public policy and international student recruitment in australia, canada, england and the USA. *Higher Education*, 75(2) :231–253, mar 2017. doi : 10.1007/s10734-017-0133-1.

J. Silge and D. Robinson. *Text Mining with R : A Tidy Approach*. O'Reilly Media, 2017. ISBN 9781491981603. URL https://books.google.fr/books?id=qtcnDwAAQBAJ.