kadaoui_alexandre_extraction_contenu_PDF

Alexandre

23/12/2020

R Markdown

Dans le cadre de notre partiel, nous devons réaliser un total de 12 travaux retracant notre parcours et notre travail durant les 30 heures de cours.
Le travail à faire est le suivant :
- Une entête comportant un titre, un lien Github avec le ou les noms des auteurs.
- Une synthese de ce travail
- Un extrait commenté avec des parties de codes clé avec explication et commentaire.
- Une évalutation du travail avec nos 5 criteres.
- Une conclusion du travail

Definition des 5 critères de notations :

1) Effort de présentation :
2) Le knit est réalisable et bien présenté.
3) Explications simples et efficaces.
4) Le Code reproductible à d'autres DataFrame avec facilité.
5) Description des fonctions utilsés et du raisonnement.

Extraction et manipulation du contenu d'un PDF

Travail réalisé par " chaymae-data " le 20/11/2020.

GitHub: https://github.com/chaymae-data/PSBX (https://github.com/chaymae-data/PSBX)

Synthese :

R possède des fonctionnalités nous permettant d'extraire des données d'un pdf, que ce soit du textes ou des données numériques
Ce tutoriel présente donc comment extraire ces données puis comment les retraiter afin d'obtenir un résultat exploitable pour de l'analyse sémantique (dans le cas de texte ici)

Extrait commenté du code :

Extraire le contenu du fichier PDF via R

Méthode 1: package **pdftools**

```
#install.packages(pdftools)
library(pdftools)
```

```
## Using poppler version 0.73.0
```

```r
#install.packages(pdftools)
library(pdftools)
download.file("https://www.btboces.org/Downloads/I%20Have%20a%20Dream%20by%20Martin%20Luther%20King%20Jr.pdf","I%20Have%20a%20Dream%20by%20Martin%20Luther%20King%20Jr.pdf", mode = "wb")
text <- pdf_text("I%20Have%20a%20Dream%20by%20Martin%20Luther%20King%20Jr.pdf")
text1 <- strsplit(text, "\n")
cat(text[1])
```

```
##                                                      I HAVE A DREAM
##                                                    Martin Luther King, Jr.
## I am happy to join with you today in what will go down in history as the greatest demonstration for freedom in the
## history of our nation.
## Five score years ago, a great American, in whose symbolic shadow we stand today, signed the Emancipation
## Proclamation. This momentous decree came as a great beacon light of hope to millions of Negro slaves who had been
## seared in the flames of withering injustice. It came as a joyous daybreak to end the long night of their captivity.
## But one hundred years later, the Negro still is not free. One hundred years later, the life of the Negro is still sadly
## crippled by the manacles of segregation and the chains of discrimination. One hundred years later, the Negro lives on a
## lonely island of poverty in the midst of a vast ocean of material prosperity. One hundred years later, the Negro is still
## languished in the corners of American society and finds himself an exile in his own land. And so we've come here today
## to dramatize a shameful condition.
## In a sense we've come to our nation's capital to cash a check. When the architects of our republic wrote the magnificent
## words of the Constitution and the Declaration of Independence, they were signing a promissory note to which every
## American was to fall heir. This note was a promise that all men, yes, black men as well as white men, would be
## guaranteed the "unalienable Rights" of "Life, Liberty and the pursuit of Happiness." It is obvious today that America has
## defaulted on this promissory note, insofar as her citizens of color are concerned. Instead of honoring this sacred
## obligation, America has given the Negro people a bad check, a check which has come back marked "insufficient funds."
## But we refuse to believe that the bank of justice is bankrupt. We refuse to believe that there are insufficient funds in the
## great vaults of opportunity of this nation. And so, we've come to cash this check, a check that will give us upon demand
## the riches of freedom and the security of justice.
## We have also come to this hallowed spot to remind America of the fierce urgency of Now. This is no time to engage in
## the luxury of cooling off or to take the tranquilizing drug of gradualism. Now is the time to make real the promises of
## democracy. Now is the time to rise from the dark and desolate valley of segregation to the sunlit path of racial justice.
## Now is the time to lift our nation from the quicksands of racial injustice to the solid rock of brotherhood. Now is the time
## to make justice a reality for all of God's children.
## It would be fatal for the nation to overlook the urgency of the moment. This sweltering summer of the Negro's
## legitimate discontent will not pass until there is an invigorating autumn of freedom and equality. Nineteen sixty-three is
## not an end, but a beginning. And those who hope that the Negro needed to blow off steam and will now be content will
## have a rude awakening if the nation returns to business as usual. And there will be neither rest nor tranquility in America
## until the Negro is granted his citizenship rights. The whirlwinds of revolt will continue to shake the foundations of our
## nation until the bright day of justice emerges.
## But there is something that I must say to my people, who stand on the warm threshold which leads into the palace of
## justice: In the process of gaining our rightful place, we must not be guilty of wrongful deeds. Let us not seek to satisfy
## our thirst for freedom by drinking from the cup of bitterness and hatred. We must forever conduct
```

```
our struggle on the
## high plane of dignity and discipline. We must not allow our creative protest to degenerate into ph
ysical violence. Again
## and again, we must rise to the majestic heights of meeting physical force with soul force.
## The marvelous new militancy which has engulfed the Negro community must not lead us to a distrust
of all white
## people, for many of our white brothers, as evidenced by their presence here today, have come to re
alize that their
## destiny is tied up with our destiny. And they have come to realize that their freedom is inextrica
bly bound to our
## freedom.
## We cannot walk alone.
```

Méthode 2: package **tm** (text mining)

```
#install.packages("tm")
library(tm)
```

```
## Loading required package: NLP
```

Importation de documents et corpus

```
docs <- getwd()
my_corpus <- VCorpus(DirSource(docs, pattern = ".pdf"), readerControl = list(reader = readPDF))
```

```
inspect(my_corpus) #lecture du document
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 1
##
## [[1]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 9388
```

```
writeLines(as.character(my_corpus[[1]])) #affichage du contenu du document
```

```
##                                                      I HAVE A DREAM
##                                                    Martin Luther King, Jr.
## I am happy to join with you today in what will go down in history as the greatest demonstration fo
r freedom in the
## history of our nation.
## Five score years ago, a great American, in whose symbolic shadow we stand today, signed the Emanci
pation
## Proclamation. This momentous decree came as a great beacon light of hope to millions of Negro slav
es who had been
## seared in the flames of withering injustice. It came as a joyous daybreak to end the long night of
their captivity.
## But one hundred years later, the Negro still is not free. One hundred years later, the life of the
Negro is still sadly
## crippled by the manacles of segregation and the chains of discrimination. One hundred years later,
the Negro lives on a
## lonely island of poverty in the midst of a vast ocean of material prosperity. One hundred years la
ter, the Negro is still
## languished in the corners of American society and finds himself an exile in his own land. And so w
e've come here today
## to dramatize a shameful condition.
## In a sense we've come to our nation's capital to cash a check. When the architects of our republic
wrote the magnificent
## words of the Constitution and the Declaration of Independence, they were signing a promissory note
to which every
## American was to fall heir. This note was a promise that all men, yes, black men as well as white m
en, would be
## guaranteed the "unalienable Rights" of "Life, Liberty and the pursuit of Happiness." It is obvious
today that America has
## defaulted on this promissory note, insofar as her citizens of color are concerned. Instead of hono
ring this sacred
## obligation, America has given the Negro people a bad check, a check which has come back marked "in
sufficient funds."
## But we refuse to believe that the bank of justice is bankrupt. We refuse to believe that there are
insufficient funds in the
## great vaults of opportunity of this nation. And so, we've come to cash this check, a check that wi
ll give us upon demand
## the riches of freedom and the security of justice.
## We have also come to this hallowed spot to remind America of the fierce urgency of Now. This is no
time to engage in
## the luxury of cooling off or to take the tranquilizing drug of gradualism. Now is the time to make
real the promises of
## democracy. Now is the time to rise from the dark and desolate valley of segregation to the sunlit
path of racial justice.
## Now is the time to lift our nation from the quicksands of racial injustice to the solid rock of br
otherhood. Now is the time
## to make justice a reality for all of God's children.
## It would be fatal for the nation to overlook the urgency of the moment. This sweltering summer of
the Negro's
## legitimate discontent will not pass until there is an invigorating autumn of freedom and equality.
Nineteen sixty-three is
## not an end, but a beginning. And those who hope that the Negro needed to blow off steam and will n
ow be content will
## have a rude awakening if the nation returns to business as usual. And there will be neither rest n
or tranquility in America
## until the Negro is granted his citizenship rights. The whirlwinds of revolt will continue to shake
the foundations of our
## nation until the bright day of justice emerges.
## But there is something that I must say to my people, who stand on the warm threshold which leads i
nto the palace of
## justice: In the process of gaining our rightful place, we must not be guilty of wrongful deeds. Le
t us not seek to satisfy
## our thirst for freedom by drinking from the cup of bitterness and hatred. We must forever conduct
```

```
our struggle on the
## high plane of dignity and discipline. We must not allow our creative protest to degenerate into physical violence. Again
## and again, we must rise to the majestic heights of meeting physical force with soul force.
## The marvelous new militancy which has engulfed the Negro community must not lead us to a distrust of all white
## people, for many of our white brothers, as evidenced by their presence here today, have come to realize that their
## destiny is tied up with our destiny. And they have come to realize that their freedom is inextricably bound to our
## freedom.
## We cannot walk alone.
##
## And as we walk, we must make the pledge that we shall always march ahead.
## We cannot turn back.
## There are those who are asking the devotees of civil rights, "When will you be satisfied?" We can never be satisfied as
## long as the Negro is the victim of the unspeakable horrors of police brutality. We can never be satisfied as long as our
## bodies, heavy with the fatigue of travel, cannot gain lodging in the motels of the highways and the hotels of the cities.
## We cannot be satisfied as long as the negro's basic mobility is from a smaller ghetto to a larger one. We can never be
## satisfied as long as our children are stripped of their self-hood and robbed of their dignity by signs stating: "For Whites
## Only." We cannot be satisfied as long as a Negro in Mississippi cannot vote and a Negro in New York believes he has
## nothing for which to vote. No, no, we are not satisfied, and we will not be satisfied until "justice rolls down like waters,
## and righteousness like a mighty stream."¹
## I am not unmindful that some of you have come here out of great trials and tribulations. Some of you have come fresh
## from narrow jail cells. And some of you have come from areas where your quest -- quest for freedom left you battered
## by the storms of persecution and staggered by the winds of police brutality. You have been the veterans of creative
## suffering. Continue to work with the faith that unearned suffering is redemptive. Go back to Mississippi, go back to
## Alabama, go back to South Carolina, go back to Georgia, go back to Louisiana, go back to the slums and ghettos of our
## northern cities, knowing that somehow this situation can and will be changed.
## Let us not wallow in the valley of despair, I say to you today, my friends.
## And so even though we face the difficulties of today and tomorrow, I still have a dream. It is a dream deeply rooted in
## the American dream.
## I have a dream that one day this nation will rise up and live out the true meaning of its creed: "We hold these truths to
## be self-evident, that all men are created equal."
## I have a dream that one day on the red hills of Georgia, the sons of former slaves and the sons of former slave owners
## will be able to sit down together at the table of brotherhood.
## I have a dream that one day even the state of Mississippi, a state sweltering with the heat of injustice, sweltering with
## the heat of oppression, will be transformed into an oasis of freedom and justice.
## I have a dream that my four little children will one day live in a nation where they will not be judged by the color of their
## skin but by the content of their character.
## I have a dream today!
## I have a dream that one day, down in Alabama, with its vicious racists, with its governor having his lips dripping with the
## words of "interposition" and "nullification" -- one day right there in Alabama little black boys and black girls will be able
## to join hands with little white boys and white girls as sisters and brothers.
## I have a dream today!
## I have a dream that one day every valley shall be exalted, and every hill and mountain shall be ma
de low, the rough
## places will be made plain, and the crooked places will be made straight; "and the glory of the Lor
d shall be revealed and
## all flesh shall see it together."2
## This is our hope, and this is the faith that I go back to the South with.
## With this faith, we will be able to hew out of the mountain of despair a stone of hope. With this
faith, we will be able to
## transform the jangling discords of our nation into a beautiful symphony of brotherhood. With this
faith, we will be able
##
## to work together, to pray together, to struggle together, to go to jail together, to stand up for
freedom together,
## knowing that we will be free one day.
## And this will be the day -- this will be the day when all of God's children will be able to sing w
ith new meaning:
## My country 'tis of thee, sweet land of liberty, of thee I sing.
## Land where my fathers died, land of the Pilgrim's pride,
## From every mountainside, let freedom ring!
## And if America is to be a great nation, this must become true.
## And so let freedom ring from the prodigious hilltops of New Hampshire.
## Let freedom ring from the mighty mountains of New York.
## Let freedom ring from the heightening Alleghenies of Pennsylvania.
## Let freedom ring from the snow-capped Rockies of Colorado.
## Let freedom ring from the curvaceous slopes of California.
## But not only that:
## Let freedom ring from Stone Mountain of Georgia.
## Let freedom ring from Lookout Mountain of Tennessee.
## Let freedom ring from every hill and molehill of Mississippi.
## From every mountainside, let freedom ring.
## And when this happens, when we allow freedom ring, when we let it ring from every village and ever
y hamlet, from
## every state and every city, we will be able to speed up that day when all of God's children, black
men and white men,
## Jews and Gentiles, Protestants and Catholics, will be able to join hands and sing in the words of
the old Negro spiritual:
##          Free at last! Free at last!
##          Thank God Almighty, we are free at last!3
```

Nettoyage du contenu afin de ne conserver que les mots clefs

Insertion d’un espace séparant les poncutations et le texte afin de protéger les données de texte lors de la suppresion des ponctuations

```
toSpace<-content_transformer(function(x,pattern) {return(gsub(pattern," ",x))})
my_corpus<-tm_map(my_corpus,toSpace,"-")
my_corpus<-tm_map(my_corpus,toSpace,",")
my_corpus<-tm_map(my_corpus,toSpace,"!")
my_corpus<-tm_map(my_corpus,toSpace,"--")
my_corpus<-tm_map(my_corpus,toSpace,"'")
```

Suppression des ponctuations

```
my_corpus<-tm_map(my_corpus, removePunctuation)
```

On uniformise tout le texte en minuscules via la fonction **content_transformer** en utilisant le terme “tolower” (on pourrait tout passer en majuscule en utilisant toupper)

```
my_corpus<- tm_map(my_corpus, content_transformer(tolower))
```

On supprime les mots de liaisons tels que and, or, if , yet (ou “stopwords” en anglais) via la fonction **removewords**

```
my_corpus<- tm_map(my_corpus, removeWords, stopwords("english"))
```

On supprime ensuite les espaces de plus d’un caractère dits “extrêmes” via la fonction **stripwhitespace**

```
my_corpus<- tm_map(my_corpus, stripWhitespace)
```

Stemming: on conserve uniquement la racine des mots en supprimant préfixes et suffixes grâce au package **Snowballc** et sa focntion **stemdocument**

```
#install.packages(Snowballc)
library(SnowballC)

my_corpus<- tm_map(my_corpus, stemDocument)
```

Création d’une matrice document-termes listant la fréquence d’apparition de mots clef par document bia la fonction **DocumentTermMatrix**

```
dtm <- DocumentTermMatrix(my_corpus)
inspect(dtm)
```

```
## <<DocumentTermMatrix (documents: 1, terms: 439)>>
## Non-/sparse entries: 439/0
## Sparsity           : 0%
## Maximal term length: 12
## Weighting          : term frequency (tf)
## Sample             :
##                                                               Terms
## Docs                                                           come day dream
##   I%20Have%20a%20Dream%20by%20Martin%20Luther%20King%20Jr.pdf   10  12    12
##                                                               Terms
## Docs                                                           freedom let
##   I%20Have%20a%20Dream%20by%20Martin%20Luther%20King%20Jr.pdf      20  13
##                                                               Terms
## Docs                                                           nation negro one
##   I%20Have%20a%20Dream%20by%20Martin%20Luther%20King%20Jr.pdf     11    15  13
##                                                               Terms
## Docs                                                           ring will
##   I%20Have%20a%20Dream%20by%20Martin%20Luther%20King%20Jr.pdf   12   27
```

###Analyse des données de texte, présentation sous forme de tableau exploitable

Déterminer fréquence d’apparition de chaque mot-clef dans le corpus via la fonction **colSums**.

```
freq<-colSums(as.matrix(dtm))
```

Affichage des fréquences des mots-clefs ordonnés du plus fréquent au moins fréquent via la fonction **order**

```
ord<-order(freq,decreasing = TRUE)
```

On peut afficher uniquement les mots-clefs les plus utilisés avec la fonction **head**

```
freq[tail(ord)]
```

```
##  whose   wind wither  wrong  wrote    yes
##      1      1      1      1      1      1
```

On peut afficher les mots-clefs les moins utilisés avec la fonction **tail**

```
freq[head(ord)]
```

```
##    will freedom   negro     let     one     day
##      27      20      15      13      13      12
```

Evaluation du travail :

1) Effort de présentation : Le pdf est bien présenté et bien ordonné via des sections bien visibles et reconnaissables

2) Le knit est réalisable et bien présenté : Le knit n’a posé aucun soucis et la présentation est bien mise en place.

3) Explications simples et efficaces : Les explication fournies sont simples et facilement compréhensibles. De plus, le découpage efficace permet de bien appréhender la méthode par étapes.

4) Le Code reproductible à d’autres DataFrame avec facilité :b Grâce au découpage efficace du code en sous-parties et à l’explication de toutes les fonctions et termes requis, le code est facilement reproductible et adaptable à tout autre PDF.

5) Description des fonctions utilsés et du raisonnement : Les fonctions principales utilisées sont les fonctions **pdf_text** pour la récupération du texte et son stockage dans un objet puis les fonctions **content_transformers** et **stemdocument** pour le nettoyage des données afin de ne conserver que les mots clefs et enfin la fonction **DocumentTermMatrix** afin d’obtenir une matrice exploitable des résultats de l’analyse.

Conclusion :

Ce tutoriel illustre de manière simple et très claire les différentes étapes et fonctions nécessaires à l’analyse d’un document pdf via R dans un but d’analyse sémantique.