R: How to Prepare Data for LDA/Text Analysis

Asked today Modified today Viewed 13 times

0

I am working with the R programming language.

I would like to perform BTM (Bitopic Term Model - a variant of LDA (Latent Dirichlet Analysis) for small text datasets) on some text data. I am following this tutorial over here: https://cran.r-project.org/web/packages/BTM/readme/README.html

When I look at the dataset ("brussels_reviews_anno") being used in this tutorial, it look something like this (I can not recognize the format of this data!):

```
library(udpipe)
library(BTM)
data("brussels_reviews_anno", package = "udpipe")

 head(brussels_reviews_anno)

    doc_id language sentence_id token_id      token      lemma  upos xpos
1 32198807       es           1        1       Gwen       gwen  NOUN  NNP
2 32198807       es           1        2        fue        ser  VERB   VB
3 32198807       es           1        3        una         un   DET   DT
4 32198807       es           1        4  magnifica  magnifica  NOUN   NN
5 32198807       es           1        5 anfitriona anfitriono   ADJ   JJ
6 32198807       es           1        6          .          . PUNCT    .
```

My dataset ("my_data") is in the current format - I manually create a text dataset for this example using reviews of fast food restaurants found on the internet:

```
my_data = structure(list(id = 1:8, reviews = c("I guess the employee decided to buy
their lunch with my card my card hoping I wouldn't notice but since it took so long to
run my car I want to head and check my bank account and sure enough they had bought
food on my card that I did not receive leave. Had to demand for and for a refund
because they acted like it was my fault and told me the charges are still pending even
though they are for 2 different amounts.",
                                               "I went to McDonald's and they charge me 50
for Big Mac when I only came with 49. The casher told me that I can't read correctly
and told me to get glasses. I am file a report on your casher and now I'm mad.",
                                               "I really think that if you can buy
breakfast anytime then I should be able to get a cheeseburger anytime especially since
I really don't care for breakfast food. I really like McDonald's food but I preferred
tree lunch rather than breakfast. Thank you thank you thank you.",
                                               "I guess the employee decided to buy their
lunch with my card my card hoping I wouldn't notice but since it took so long to run my
car I want to head and check my bank account and sure enough they had bought food on my
card that I did not receive leave. Had to demand for and for a refund because they
acted like it was my fault and told me the charges are still pending even though they
are for 2 different amounts.",
                                               "Never order McDonald's from Uber or Skip
or any delivery service for that matter, most particularly one on Elgin Street and
Rideau Street, they never get the order right. Workers at either of these locations
don't know how to follow simple instructions. Don't waste your money at these two
locations.",
                                               "Employees left me out in the snow and
wouldn’t answer the drive through. They locked the doors and it was freezing. I asked
the employee a simple question and they were so stupid they answered a completely
```

```
different question. Dumb employees and bad food.",
                                          "McDonalds food was always so good but ever
since they add new/more crispy chicken sandwiches it has come out bad. At first I
thought oh they must haven't had a good day but every time I go there now it's always
soggy, and has no flavor. They need to fix this!!!",
                                          "I just ordered the new crispy chicken
sandwich and I'm very disappointed. Not only did it taste horrible, but it was more bun
than chicken. Not at all like the commercial shows. I hate sweet pickles and there were
two slices on my sandwich. I wish I could add a photo to show the huge bun and tiny
chicken."
)), class = "data.frame", row.names = c(NA, -8L))
```

Can someone please show me how I can take my dataset and transform it in such a way that I can perform BTM analysis on this data and create a visualization similar to the visualizations in this tutorial?

Thanks!

Additional References:

- https://rforanalytics.com/11-7-topic-modelling.html

r text nlp visualization text-mining Edit tags

Share Edit Follow Close Flag

asked 16 hours ago

stats_noob
3,537 2 8 37

1 Answer

1

The class of `brussels_reviews_anno` is just a regular `data.frame`. That structure is generated by the function `udpipe()` from the package **udpipe**.

Below I provide a working example, with the exclusion of the path where I save the language model, that shows how to replicate a similar data structure.

Please keep in mind that `udpipe()` does a lot of stuff. The reason why you see many more columns in the final data.frame `out` is because I did not tweak any parameters of the function nor simply deleted any of the columns.

Overall, to get started with `BTM()` you need to tokenize your textual data. That's one of the things you can do with the package **udpipe**.

Hope this helped!

```
library(udpipe)
library(BTM)

data("brussels_reviews_anno", package = "udpipe")
head(brussels_reviews_anno)
#>     doc_id language sentence_id token_id      token      lemma  upos xpos
```

```
#> 1 32198807       es          1        1       Gwen       gwen  NOUN  NNP
#> 2 32198807       es          1        2        fue        ser  VERB   VB
#> 3 32198807       es          1        3        una         un   DET   DT
#> 4 32198807       es          1        4  magnifica  magnifica  NOUN   NN
#> 5 32198807       es          1        5 anfitriona anfitriono   ADJ   JJ
#> 6 32198807       es          1        6          .          . PUNCT    .
```

```
my_data = structure(list(id = 1:8, reviews = c("I guess the employee decided to buy
their lunch with my card my card hoping I wouldn't notice but since it took so long to
run my car I want to head and check my bank account and sure enough they had bought
food on my card that I did not receive leave. Had to demand for and for a refund
because they acted like it was my fault and told me the charges are still pending even
though they are for 2 different amounts.",
                                                "I went to McDonald's and they charge me
50 for Big Mac when I only came with 49. The casher told me that I can't read correctly
and told me to get glasses. I am file a report on your casher and now I'm mad.",
                                                "I really think that if you can buy
breakfast anytime then I should be able to get a cheeseburger anytime especially since
I really don't care for breakfast food. I really like McDonald's food but I preferred
tree lunch rather than breakfast. Thank you thank you thank you.",
                                                "I guess the employee decided to buy
their lunch with my card my card hoping I wouldn't notice but since it took so long to
run my car I want to head and check my bank account and sure enough they had bought
food on my card that I did not receive leave. Had to demand for and for a refund
because they acted like it was my fault and told me the charges are still pending even
though they are for 2 different amounts.",
                                                "Never order McDonald's from Uber or
Skip or any delivery service for that matter, most particularly one on Elgin Street and
Rideau Street, they never get the order right. Workers at either of these locations
don't know how to follow simple instructions. Don't waste your money at these two
locations.",
                                                "Employees left me out in the snow and
wouldn’t answer the drive through. They locked the doors and it was freezing. I asked
the employee a simple question and they were so stupid they answered a completely
different question. Dumb employees and bad food.",
                                                "McDonalds food was always so good but
ever since they add new/more crispy chicken sandwiches it has come out bad. At first I
thought oh they must haven't had a good day but every time I go there now it's always
soggy, and has no flavor. They need to fix this!!!",
                                                "I just ordered the new crispy chicken
sandwich and I'm very disappointed. Not only did it taste horrible, but it was more bun
than chicken. Not at all like the commercial shows. I hate sweet pickles and there were
two slices on my sandwich. I wish I could add a photo to show the huge bun and tiny
chicken."
)), class = "data.frame", row.names = c(NA, -8L))
```

```
# download a language model
udpipe_download_model("english-ewt", model_dir = "~/Desktop/")
#> Downloading udpipe model from
https://raw.githubusercontent.com/jwijffels/udpipe.models.ud.2.5/master/inst/udpipe-ud-
2.5-191206/english-ewt-ud-2.5-191206.udpipe to ~/Desktop//english-ewt-ud-2.5-
191206.udpipe
#>  - This model has been trained on version 2.5 of data from
https://universaldependencies.org
#>  - The model is distributed under the CC-BY-SA-NC license:
https://creativecommons.org/licenses/by-nc-sa/4.0
#>  - Visit https://github.com/jwijffels/udpipe.models.ud.2.5 for model license
details.
#>  - For a list of all models and their licenses (most models you can download with
this package have either a CC-BY-SA or a CC-BY-SA-NC license) read the documentation at
?udpipe_download_model. For building your own models: visit the documentation by typing
vignette('udpipe-train', package = 'udpipe')
#> Downloading finished, model stored at '~/Desktop//english-ewt-ud-2.5-191206.udpipe'
#>      language                                  file_model
#> 1 english-ewt ~/Desktop//english-ewt-ud-2.5-191206.udpipe
#>
```

```
url
#> 1
https://raw.githubusercontent.com/jwijffels/udpipe.models.ud.2.5/master/inst/udpipe-ud-
2.5-191206/english-ewt-ud-2.5-191206.udpipe
#>   download_failed download_message
#> 1           FALSE               OK

# load in the environment
eng_model = udpipe_load_model("~/Desktop/english-ewt-ud-2.5-191206.udpipe")

# apply the tokenization
out = udpipe(my_data$reviews, object = eng_model)
head(out)
#>   doc_id paragraph_id sentence_id
#> 1   doc1            1           1
#> 2   doc1            1           1
#> 3   doc1            1           1
#> 4   doc1            1           1
#> 5   doc1            1           1
#> 6   doc1            1           1
#>
sentence
#> 1 I guess the employee decided to buy their lunch with my card my card hoping I
wouldn't notice but since it took so long to run my car I want to head and check my
bank account and sure enough they had bought food on my card that I did not receive
leave.
#> 2 I guess the employee decided to buy their lunch with my card my card hoping I
wouldn't notice but since it took so long to run my car I want to head and check my
bank account and sure enough they had bought food on my card that I did not receive
leave.
#> 3 I guess the employee decided to buy their lunch with my card my card hoping I
wouldn't notice but since it took so long to run my car I want to head and check my
bank account and sure enough they had bought food on my card that I did not receive
leave.
#> 4 I guess the employee decided to buy their lunch with my card my card hoping I
wouldn't notice but since it took so long to run my car I want to head and check my
bank account and sure enough they had bought food on my card that I did not receive
leave.
#> 5 I guess the employee decided to buy their lunch with my card my card hoping I
wouldn't notice but since it took so long to run my car I want to head and check my
bank account and sure enough they had bought food on my card that I did not receive
leave.
#> 6 I guess the employee decided to buy their lunch with my card my card hoping I
wouldn't notice but since it took so long to run my car I want to head and check my
bank account and sure enough they had bought food on my card that I did not receive
leave.
#>   start end term_id token_id    token    lemma upos xpos
#> 1     1   1       1        1        I        I PRON  PRP
#> 2     3   7       2        2    guess    guess VERB  VBP
#> 3     9  11       3        3      the      the  DET   DT
#> 4    13  20       4        4 employee employee NOUN   NN
#> 5    22  28       5        5  decided   decide VERB  VBD
#> 6    30  31       6        6       to       to PART   TO
#>                                       feats head_token_id dep_rel deps misc
#> 1 Case=Nom|Number=Sing|Person=1|PronType=Prs             2   nsubj <NA> <NA>
#> 2           Mood=Ind|Tense=Pres|VerbForm=Fin             0    root <NA> <NA>
#> 3                  Definite=Def|PronType=Art             4     det <NA> <NA>
#> 4                              Number=Sing               5   nsubj <NA> <NA>
#> 5           Mood=Ind|Tense=Past|VerbForm=Fin             2   ccomp <NA> <NA>
#> 6                                       <NA>             7    mark <NA> <NA>
```

Created on 2022-09-20 by the reprex package (v2.0.1)

Share Edit Follow Flag

answered 3 hours ago

Francesco Grossetti
1,475 8 17