Northeastern University

INFO 6105
Data Sci Eng Mth & Tools
Lecture 1 NLP Lab

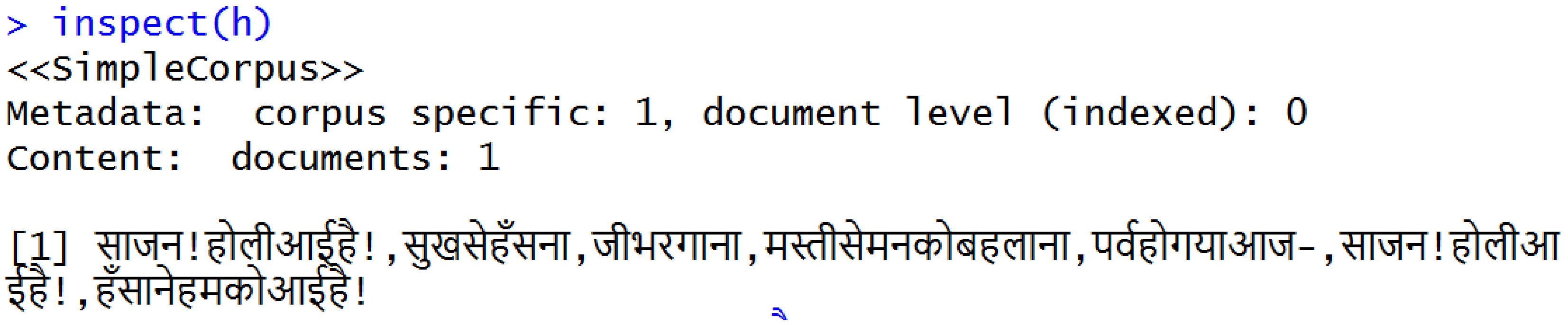
8 January 2020

tm

- **R text mining library**
- **`install.packages(‘tm”)`**
- **`library(tm)`**

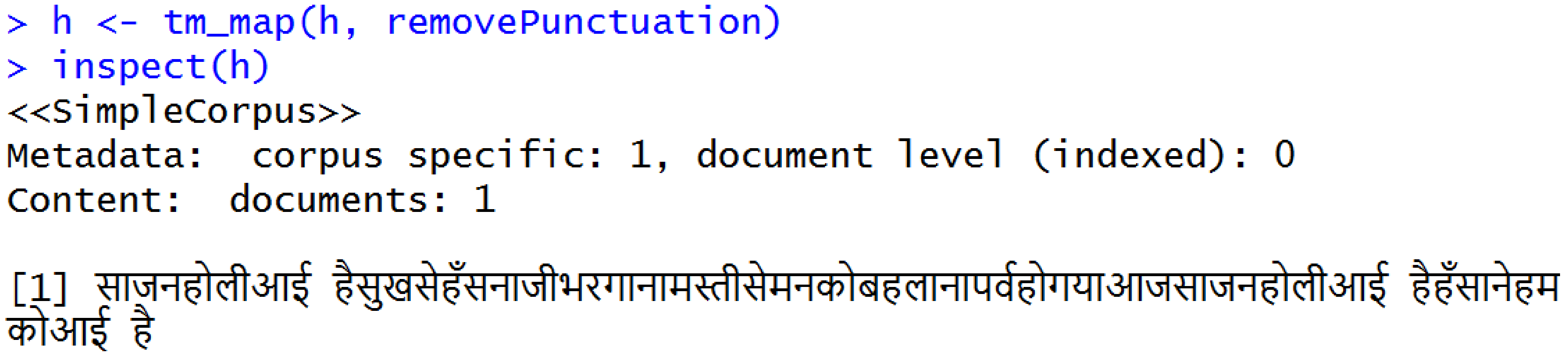
Load hindi text & remove punctuation

- Assume `hindi.txt` contains Unicode for poem in hindi
  - Contained in RStudio Session folder
- `h <- Corpus(VectorSource(readLines("hindi.txt", n=1, encoding="UTF-8")))`



```
> inspect(h)
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1] साजन!होलीआईहै!,सुखसेहँसना,जीभरगाना,मस्तीसेमनकोबहलाना,पर्वहोगयाआज-,साजन!होलीआ
ईहै!,हँसानेहमकोआईहै!
```



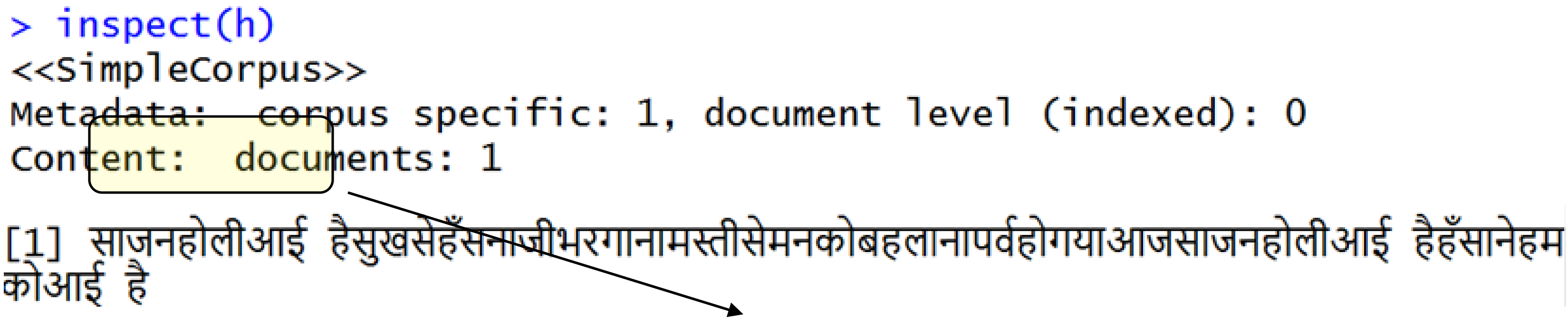
```
> h <- tm_map(h, removePunctuation)
> inspect(h)
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1] साजनहोलीआई हैसुखसेहँसनाजीभरगानामस्तीसेमनकोबहलानापर्वहोगयाआजसाजनहोलीआई हैहँसानेहम
कोआई है
```

Works?

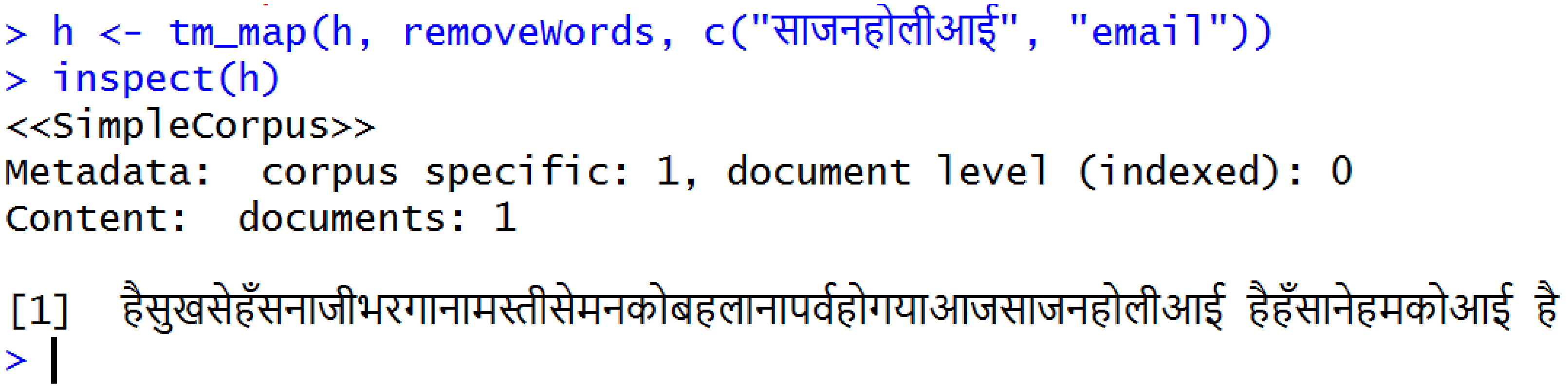
- Hmm..
- Maybe try different encoding?
  - How about `"UCS-2LE"` ?

Removing stopwords from hindi frame



```
> inspect(h)
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1] साजनहोलीआई हैसुखसेहँसनाजीभरगानामस्तीसेमनकोबहलानापर्वहोगयाआजसाजनहोलीआई हैहँसानेहम
कोआई है
```

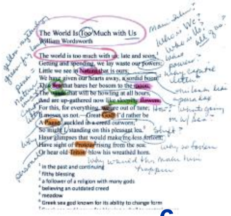


```
> h <- tm_map(h, removeWords, c("साजनहोलीआई", "email"))
> inspect(h)
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1]  हैसुखसेहँसनाजीभरगानामस्तीसेमनकोबहलानापर्वहोगयाआजसाजनहोलीआई हैहँसानेहमकोआई है
> |
```

udpipe

- **`Udpipe`** provides language-agnostic ‘tokenization’ and ‘parts of speech tagging’, of raw text in many languages, including Chinese and Hindi.
- `library(udpipe)`
- `model <- udpipe_download_model(language = "english")`
- `# When you download the language, you will see the associated filename download from GitHub, pass that filename in the next command below..`
- `udmodel_english <- udpipe_load_model(file = 'english-ud-2.0-170801.udpipe’)`
- `#Now annotate your corpus or sentence (or haiku)`
- `s <- udpipe_annotate(udmodel_english,` “An old silent pond... A frog jumps into the pond, splash! Silence again.”)
- `x <- data.frame(s)`
- `colnames(x)`



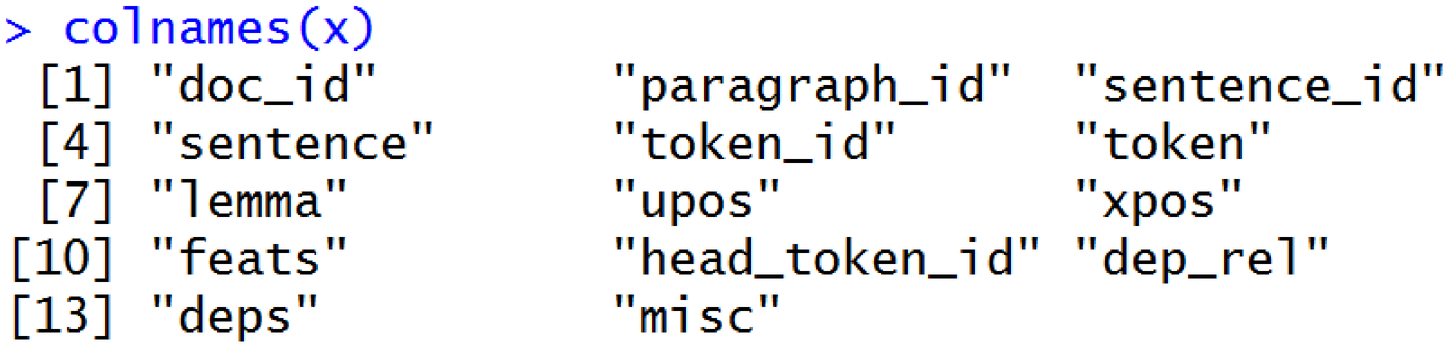
The World Is Too Much with Us
William Wordsworth

The world is too much with us; late and soon,[1]
Getting and spending, we lay waste our powers;
Little we see in Nature that is ours;
We have given our hearts away, a sordid boon![2]
This Sea that bares her bosom to the moon;
The winds that will be howling at all hours,
And are up-gathered now like sleeping flowers;
For this, for everything, we are out of tune;
It moves us not.—Great God! I'd rather be
A Pagan[3] suckled in a creed outworn;[4]
So might I, standing on this pleasant lea,[5]
Have glimpses that would make me less forlorn;
Have sight of Proteus[6] rising from the sea;
Or hear old Triton[7] blow his wreathed horn.

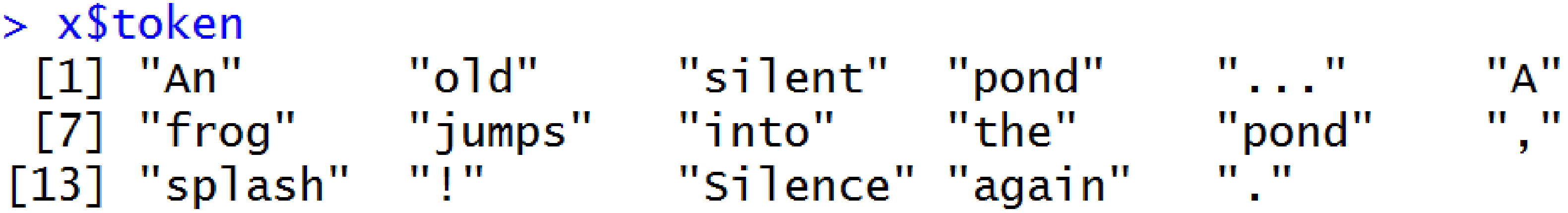
[1] in the past and continuing
[2] filthy blessing
[3] a follower of a religion with many gods
[4] believing an outdated creed
[5] meadow
[6] Greek sea god known for its ability to change form

Main Idea?
Who is We? Who is Us? all of us?
What powers?
How?
What's going on w/ Sea?
Shift?
Why so forlorn
Why would this make him happier
yellow - mythology
pink = nature
Why
personal

Annotating (continued)

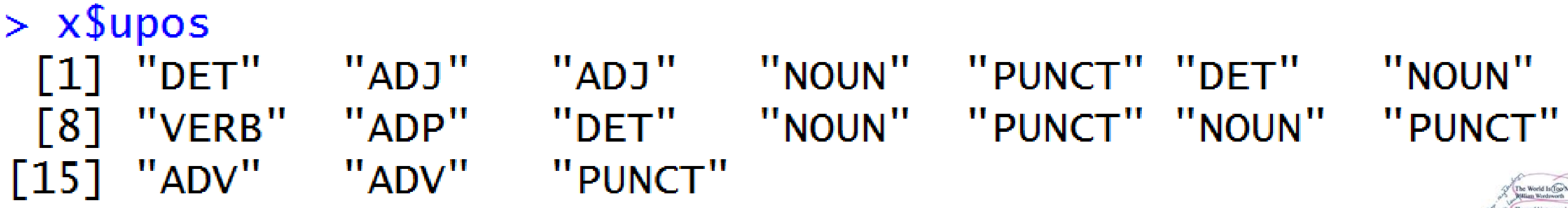


```
> colnames(x)
 [1] "doc_id"        "paragraph_id"  "sentence_id"
 [4] "sentence"      "token_id"      "token"
 [7] "lemma"         "upos"          "xpos"
[10] "feats"         "head_token_id" "dep_rel"
[13] "deps"          "misc"
```

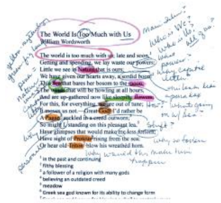


```
> x$token
 [1] "An"      "old"     "silent"  "pond"    "..."     "A"
 [7] "frog"    "jumps"   "into"    "the"     "pond"    ","
[13] "splash"  "!"       "Silence" "again"   "."
```

- **And your Universal Parts of Speech (UPOS):**



```
> x$upos
 [1] "DET"   "ADJ"   "ADJ"   "NOUN"  "PUNCT" "DET"   "NOUN"
 [8] "VERB"  "ADP"   "DET"   "NOUN"  "PUNCT" "NOUN"  "PUNCT"
[15] "ADV"   "ADV"   "PUNCT"
```



Getting parts of speech (PoS): Verbs

- `verbs <- subset(x, upos %in% c("VERB"))`
- `stats$token`

And now..

- **You can do a much better text analysis since you know about tokens *and their roles (grammar)* in the text..**

Unicode package

- `install.packages("utf8")`
- `library(utf8)`

Example: Greek

```
library(udpipe)
udmodel <- udpipe_download_model(language =
"greek")
udmodel_greek <- udpipe_load_model(file =
'greek-ud-2.0-170801.udpipe')
s <- udpipe_annotate(udmodel_greek, "Πενθώ τόν
ήλιο καί πενθώ τα χρόνια που έρχονται. Χωρίς εμάς καί
τραγουδώ τ’ άλλα πού πέρασαν. Εάν είναι αλήθεια. Μιλημένα
τα σώματα καί οί βάρκες πού έκρουζαν γλυκά.. Οί κιθάρες πού
αναβόσβησαν κάτω από τα νερά")
x <- data.frame(s)
colnames(x)
utf8_print(unlist(x$token))
x$upos
verbs <- subset(x, upos %in% c("VERB"))
utf8_print(unlist(verbs$token))
```

Example: Hindi

```
library(udpipe)
udmodel <- udpipe_download_model(language = "hindi")
udmodel_hindi <- udpipe_load_model(file = 'hindi-ud-2.0-170801.udpipe')
s <- udpipe_annotate(udmodel_hindi, "जंगल में मोर नाचा किस ने देखा ?")
x <- data.frame(s)
colnames(x)
utf8_print(unlist(x$token))
x$upos
verbs <- subset(x, upos %in% c("VERB"))
utf8_print(unlist(verbs$token))
```

Example: Chinese

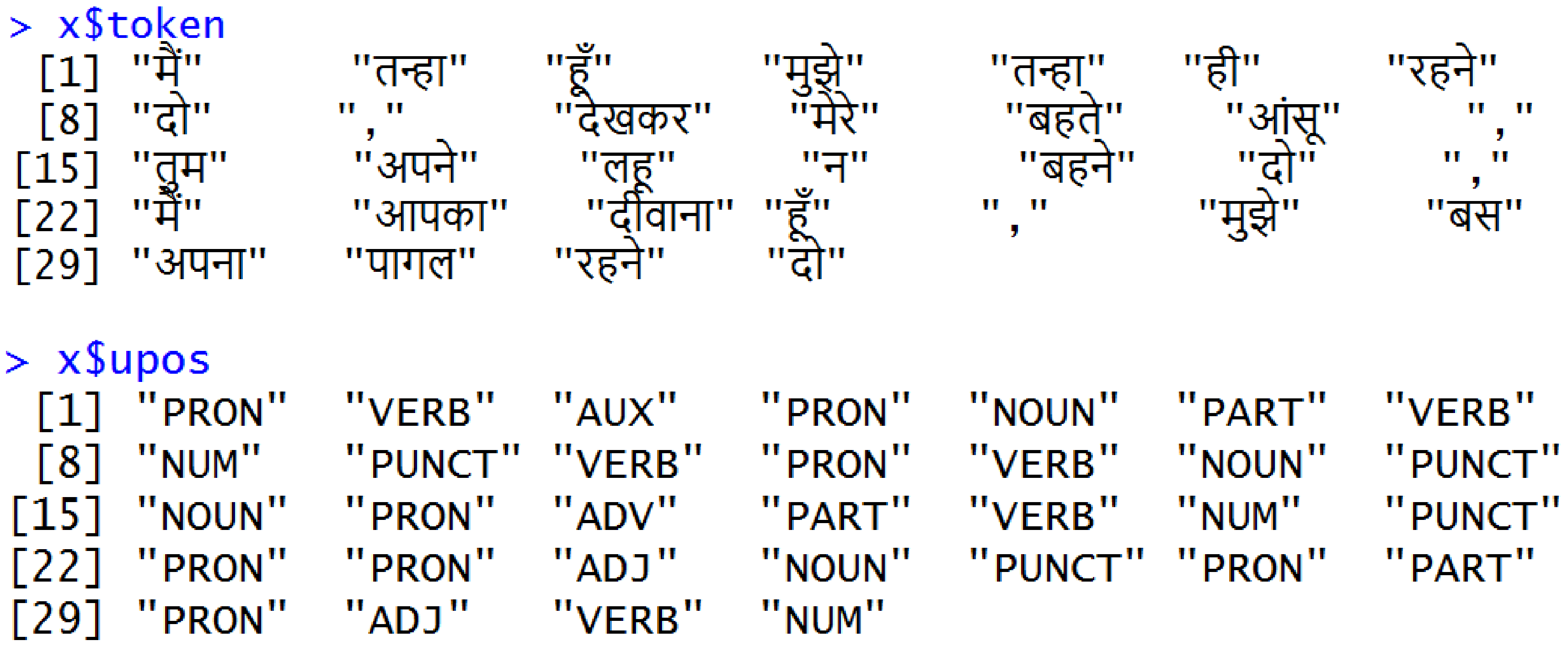
```r
udmodel <- udpipe_download_model(language = "chinese")
udmodel_zhongwen <- udpipe_load_model(file = 'chinese-ud-2.0-170801.udpipe')
s <- udpipe_annotate(udmodel_zhongwen, "授人以鱼不如授人以渔")
x <- data.frame(s)
colnames(x)
utf8_print(unlist(x$token))
x$upos
verbs <- subset(x, upos %in% c("VERB"))
utf8_print(unlist(verbs$token))
```

Hindi

```
> model <- udpipe_download_model(language = "hindi")
Downloading udpipe model from https://raw.githubusercontent.com/jwijf
fels/udpipe.models.ud.2.0/master/inst/udpipe-ud-2.0-170801/hindi-ud-2
.0-170801.udpipe to D:/user/docs/NU/_Info6101/Lecture 2/labs/udpipe/m
odels/hindi-ud-2.0-170801.udpipe
trying URL 'https://raw.githubusercontent.com/jwijffels/udpipe.models
.ud.2.0/master/inst/udpipe-ud-2.0-170801/hindi-ud-2.0-170801.udpipe'
Content type 'application/octet-stream' length 26137581 bytes (24.9 M
B)
downloaded 24.9 MB

> model <- udpipe_load_model(file = "hindi-ud-2.0-170801.udpipe")
> x <- udpipe_annotate(model, " मैं तन्हा हूँ मुझे तन्हा ही रहने दो, देखकर मेरे बहते
 आंसू, तुम अपने लहू न बहने दो, मैं आपका दीवाना हूँ, मुझे बस अपना पागल रहने दो     "
)#hindi poem
> x <- data.frame(x)
>
```

Hindi uPOS

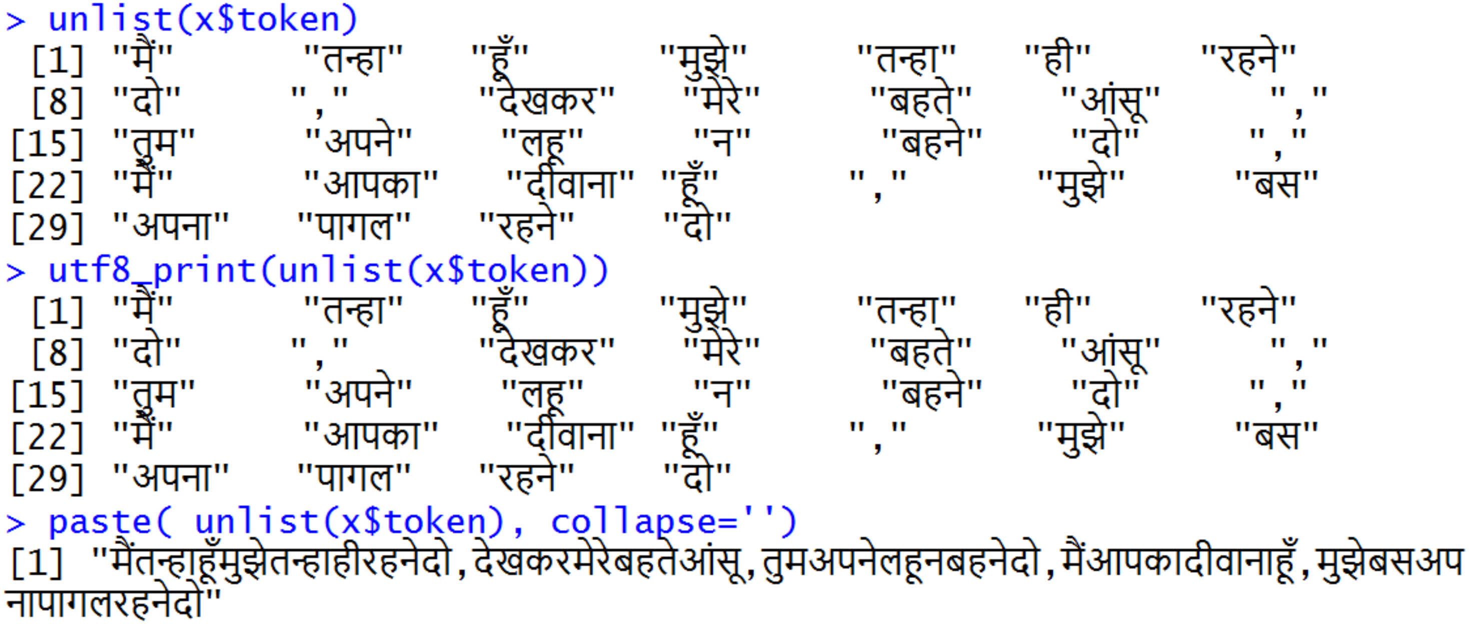


```
> x$token
 [1] "मैं"      "तन्हा"    "हूँ"       "मुझे"     "तन्हा"    "ही"      "रहने"
 [8] "दो"      ","       "देखकर"   "मेरे"     "बहते"    "आंसू"    ","
[15] "तुम"     "अपने"    "लहू"     "न"       "बहने"    "दो"      ","
[22] "मैं"      "आपका"    "दीवाना"  "हूँ"      ","       "मुझे"     "बस"
[29] "अपना"    "पागल"    "रहने"    "दो"

> x$upos
 [1] "PRON"  "VERB"  "AUX"   "PRON"  "NOUN"  "PART"  "VERB"
 [8] "NUM"   "PUNCT" "VERB"  "PRON"  "VERB"  "NOUN"  "PUNCT"
[15] "NOUN"  "PRON"  "ADV"   "PART"  "VERB"  "NUM"   "PUNCT"
[22] "PRON"  "PRON"  "ADJ"   "NOUN"  "PUNCT" "PRON"  "PART"
[29] "PRON"  "ADJ"   "VERB"  "NUM"
```

Printing Unicode to console

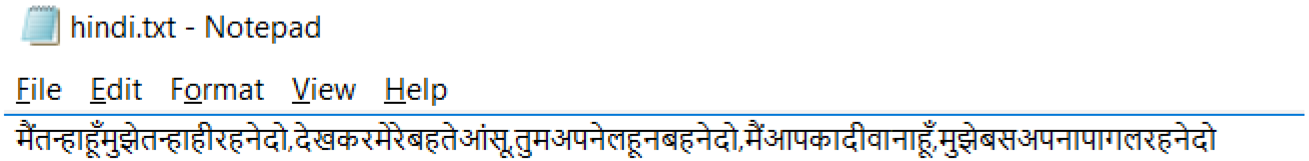
- `install.packages("utf8")`
- `library(utf8)`
- `utf8_print(unlist(x$token))`
- `#concatenating:`
  `paste( unlist(x$token), collapse='')`



```
> unlist(x$token)
 [1] "मैं"       "तन्हा"    "हूँ"         "मुझे"       "तन्हा"   "ही"       "रहने"
 [8] "दो"        ","        "देखकर"     "मेरे"       "बहते"     "आंसू"      ","
[15] "तुम"       "अपने"     "लहू"        "न"          "बहने"     "दो"       ","
[22] "मैं"       "आपका"     "दीवाना" "हूँ"          ","        "मुझे"      "बस"
[29] "अपना"     "पागल"     "रहने"       "दो"
> utf8_print(unlist(x$token))
 [1] "मैं"       "तन्हा"    "हूँ"         "मुझे"       "तन्हा"   "ही"       "रहने"
 [8] "दो"        ","        "देखकर"     "मेरे"       "बहते"     "आंसू"      ","
[15] "तुम"       "अपने"     "लहू"        "न"          "बहने"     "दो"       ","
[22] "मैं"       "आपका"     "दीवाना" "हूँ"          ","        "मुझे"      "बस"
[29] "अपना"     "पागल"     "रहने"       "दो"
> paste( unlist(x$token), collapse='')
[1] "मैंतन्हाहूँमुझेतन्हाहीरहनेदो,देखकरमेरेबहतेआंसू,तुमअपनेलहूनबहनेदो,मैंआपकादीवानाहूँ,मुझेबसअप
नापागलरहनेदो"
```

Printing Hindi Unicode to file

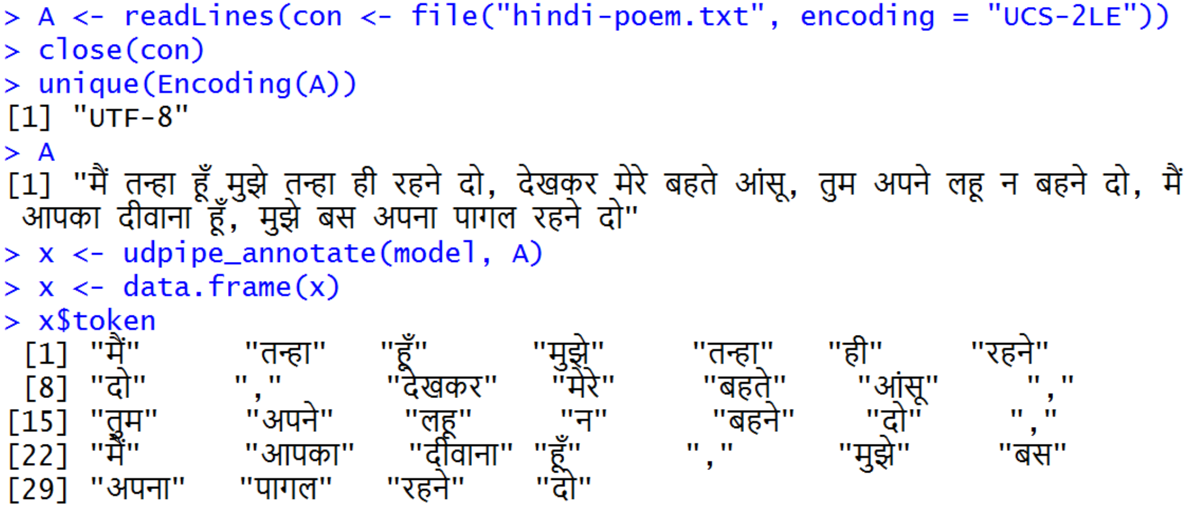
- `writeLines(text = paste( unlist(x$token), collapse=''), con = "hindi.txt", useBytes = T)`



मैंतन्हाहूँमुझेतन्हाहीरहनेदो,देखकरमेरेबहतेआंसू,तुमअपनेलहूनबहनेदो,मैंआपकादीवानाहूँ,मुझेबसअपनापागलरहनेदो

Reading Hindi Unicode from file

- `hindi <- readLines(con <- file("hindi-poem.txt", encoding = "UCS-2LE"))`
  - Other option: `hindi <- readLines(con <- file("hindi-poem.txt", encoding = "UTF-16")) )`
- `close(con)`
- `unique(Encoding(hindi))`
- `x <- udpipe_annotate(model, hindi)`
- `x <- data.frame(x)`



```
> A <- readLines(con <- file("hindi-poem.txt", encoding = "UCS-2LE"))
> close(con)
> unique(Encoding(A))
[1] "UTF-8"
> A
[1] "मैं तन्हा हूँ मुझे तन्हा ही रहने दो, देखकर मेरे बहते आंसू, तुम अपने लहू न बहने दो, मैं
 आपका दीवाना हूँ, मुझे बस अपना पागल रहने दो"
> x <- udpipe_annotate(model, A)
> x <- data.frame(x)
> x$token
 [1] "मैं"      "तन्हा"   "हूँ"      "मुझे"     "तन्हा"   "ही"      "रहने"
 [8] "दो"      ","       "देखकर"   "मेरे"     "बहते"    "आंसू"    ","
[15] "तुम"     "अपने"    "लहू"     "न"       "बहने"    "दो"      ","
[22] "मैं"      "आपका"    "दीवाना"  "हूँ"      ","       "मुझे"     "बस"
[29] "अपना"    "पागल"    "रहने"    "दो"
```

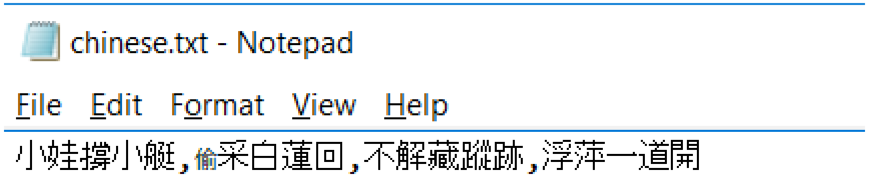
References

- https://www.rdocumentation.org/packages/base/versions/3.5.0/topics/readLines
- https://www.twilio.com/docs/glossary/what-is-ucs-2-character-encoding

Chinese

```
> model <- udpipe_load_model(file = "chinese-ud-2.0-170801.udpipe")
> x <- udpipe_annotate(model, " 小娃撐小艇 ， 偷采白蓮回 ， 不解藏蹤跡 ， 浮萍
一道開    ")#mandarin poem
> x <- data.frame(x)
> x$token
 [1] "小"    "娃撐"  "小艇"  "，"    "偷采"  "白"    "蓮"    "回"    "，"
[10] "不"    "解"    "藏蹤"  "跡"    "，"    "浮萍"  "一"    "道"    "開"
> x$upos
 [1] "PART"  "NOUN"  "NOUN"  "PUNCT" "VERB"  "PROPN" "PROPN"
 [8] "VERB"  "PUNCT" "ADV"   "VERB"  "VERB"  "NOUN"  "PUNCT"
[15] "PROPN" "NUM"   "NOUN"  "VERB"
>
```

- **`writeLines(text = paste( unlist(x$token), collapse=''), con = "Chinese.txt", useBytes = T)`**



chinese.txt - Notepad

小娃撐小艇,偷采白蓮回,不解藏蹤跡,浮萍一道開