Read US patent data

```
library(tidyverse)
library(xml2)
```

I’ve saved one file from the US Patent data at this link (https://bulkdata.uspto.gov/data/patent/grant/redbook/fulltext/2019/) and the folowing code will get you started with working on it.

[Note, if you han’t come across xml before, you might want to look at an introductory tutorial, e.g. https://www.w3schools.com/xml/default.asp (https://www.w3schools.com/xml/default.asp) , or something more comprehensive: https://www.ibm.com/developerworks/xml/tutorials/xmlintro/xmlintro.html (https://www.ibm.com/developerworks/xml/tutorials/xmlintro/xmlintro.html)]

After reading the documentation for `xml2` , the obvious thing to try is:

```
p <- read_xml("Data/ipg190115.xml")
```

Unfortunately this gives an error:

Error in doc_parse_file(con, encoding = encoding, as_html = as_html, options = options) : XML declaration allowed only at the start of the document [64]

You’ll need to investigate the file by reading it in with `read_lines` (this will take some time):

```
lines <- read_lines("Data/ipg190115.xml")
```

You should find the file is actually the concatentation of a large number (4137) of individual xml files, each starting with an xml declaration. So if you’re going to use the xml protocol you’ll have to split it up.

The following code chunk does that, then uses `xml2` functions on each separate file, to extract the text in the first node.

```
xml_declaration <- "<?xml version=\"1.0\" encoding=\"UTF-8\"?>"
# Find which lines have an xml declaration
start_patent <-
  lines %>%
  str_which(fixed(xml_declaration))
# Find the last line in each xml section.
# Note the use of the default parameter to deal with the edge case.
stop_patent <- lead(start_patent, 1, default =length(lines) + 1) - 1
# Set up a vector to hold one claim for each separate patent in our file.
claims <- character(length(start_patent))
# For each patent, read it in as xml and finde the first <claim-text> node.
for (p in seq.int(length(start_patent))) {
  pat <- paste(lines[start_patent[p]:(stop_patent[p])], collapse = "")
  patx <- read_xml(pat)
  claims[p] <- xml_text(xml_find_first(patx, ".//claim-text"))
}
```

You now have a large vector containing the text of the first claim in each patent. There are rather a lot of them. We can look at the first six, and then a random sample:

```
set.seed(123)
head(claims)
```

```
## [1] "The ornamental design for a chocolate, as shown and described."
## [2] "The ornamental design for a cracker, as shown and described."
## [3] "The ornamental design for a necktie, as shown and described."
## [4] "The ornamental design for a jockstrap having light-emitting stripes, as sh
own and described."
## [5] "The ornamental design for a lower lumbar support legging, as shown and des
cribed."
## [6] "The ornamental design for a double seam yoke, as shown and described."
```

```
sample(claims, 10)
```

```
##  [1] "1. A vehicle traveling control apparatus comprising:a setter configured t
o set, when controlling a lane change, a lane-change route on a basis of a route p
arameter of a zone in which the lane change is performed and a route parameter of
a lateral movement of the lane change, the lane-change route serving as a route th
rough which an own vehicle is to proceed to an adjacent lane from a current target
route of the own vehicle, the lane change allowing for movement from a traveling l
ane to the adjacent lane, the traveling lane being a lane along which the own vehi
cle travels, the adjacent lane being positioned next to the traveling lane;a detec
tor configured to detect a road surface pattern formed by an irregular part on a s
urface of a road to which the lane change is to be made during the lane change; an
da determiner configured to determine whether the own vehicle interferes, on the l
ane-change route, with the road surface pattern, and instruct, on a basis of a res
ult of the determination, correction of the lane-change route set by the setter,wh
erein, in a case where determination is made that the own vehicle does not interfe
re with the road surface pattern, the determiner permits the lane change through t
he lane-change route, andwherein, in a case where determination is made that the o
wn vehicle interferes with the road surface pattern, the determiner instructs the
setter to correct the lane-change route,wherein, in a case where determination is
made that the own vehicle interferes with the road surface pattern, the determiner
instructs correction of the route parameter of the zone,wherein, in a case where d
etermination is made that the own vehicle interferes with the road surface pattern
even if the correction of the route parameter of the zone is made to a limit, the
determiner instructs correction of the route parameter of the lateral movement by
returning the route parameter of the zone to initial value, andwherein, in a case
where determination is made that the own vehicle interferes with the road surface
pattern even if the correction of the route parameter of the lateral movement is m
ade to a limit, the determiner instructs prohibition or cancelation of the lane ch
ange."
##  [2] "1. An apparatus, for angularly aligning an antenna disposed at a geograph
ical location, comprising:a plurality of reticle members, each reticle member havi
ng a reticle; anda plurality of reference members, each adjustably engaged with an
associated one of the plurality of reticle members;wherein each of the plurality o
f reference members comprises an associated template having a reference mark posit
```

ioned thereon according to the geographical location of the antenna and the antenna is angularly aligned when each reference mark of each template is aligned with the reticle associated with the reference mark."

## [3] "1. A system of formation sampling while drilling comprising:a suction pup joint provided to pump out and clean formation fluid;a sampling pup joint provided to sample and store the formation fluid;a setting pup joint provided between the suction pup joint and the sampling pup joint and provided to establish a fluid channel between the formation fluid, the suction pup joint and the sampling pup joint; andan electrohydraulic high voltage connection device provided between the setting pup joint and the sampling pup joint,wherein the suction pup joint is provided with a first electric motor, a first piston cylinder and a first electromagnetic valve in the interior thereof which are connected, the first electric motor being provided to supply operation power of pumping out and cleaning to the suction pup joint, andwherein the sampling pup joint is independent of the suction pup joint and provided with a second electric motor, a second piston cylinder, a second electromagnetic valve and a sample storage room in the interior thereof which are connected, the second electric motor being provided to supply operation power of sampling and storing to the sampling pup joint."

## [4] "1. A method of transmitting a message from a sender to a recipient through a server displaced from the recipient, including the steps at the server of: receiving the message at the server from the sender, transmitting the message from the server to the recipient, and providing for a transmission of a reply to the sender through the server of the message by the recipient;wherein the message is provided with a unique identification by the server and wherein the reply from the recipient through the server to the sender is provided on the basis of this unique identification of the message by the server; andwherein the reply by the recipient through the server provides for an identification of each of a plurality of recipients on the basis of individual identifications related to the unique identification of the message and wherein the message from the sender to the recipient is provided in a particular format at the server and wherein the reply includes a request from the recipient to receive proof of transmission or delivery of the reply and wherein the server responds to the request in the reply to provide the proof of the transmission or delivery of the reply to the sender and wherein the recipient provides a fictional destination address and wherein the destination address is at the server and wherein a database associated with the server stores the identity of the message and the identity and address of the sender and wherein the reply includes an identification of the message and the name and address of the sender and wherein the server parses the message and the name and address of the sender from the fictional destination address and directs the reply to the sender at the sender's address."

## [5] "1. A method for automatically selecting between broadcast and recorded content for inclusion in playlists based on current whereabouts of users, the method comprising:receiving a user request to generate a playlist of media assets for presentation on a display device, wherein the display device is positioned within a viewing area;detecting a location of a user relative to the viewing area within which the display device is positioned and presents the playlist;determining that the location of the user is away from the viewing area within which the display device is positioned;calculating a length of time required for the user to reach the viewing area by:creating a graphical representation of an area between the location and the viewing area;associating a first point in the graphical representation with a first node and a second point in the graphical representation with a second node; anddetermining a path from the location to the viewing area through either the first node or the second node based on an amount of time required to travel the path through either the first node or the second node;determining the length of time

based on the path;comparing the length of time to a threshold length of time;based on determining that the length of time is less than or equal to the threshold length of time, selecting a broadcast media asset for inclusion in the playlist presented by the display device positioned within the viewing area; andbased on determining that the length of time is greater than the threshold length of time, selecting a recorded media asset for inclusion in the playlist presented by the display device positioned within the viewing area."

## [6] "The ornamental design for a computer, as shown and described."

## [7] "1. A fiber optic cable bundle, comprising:an inner layer of at least one subunit fiber optic cable; andan outer layer of a plurality of subunit fiber optic cables helically stranded about the inner layer, wherein the inner layer comprises a single subunit and the outer layer comprises five subunits, wherein the outer layer of subunits comprises the exterior perimeter of the bundle and the bundle is free of external binders and a jacket, wherein a helical lay length of the outer layer is between 40-60 mm, and wherein each subunit fiber optic cable comprises:at least one optical fiber;a layer of loose tensile strength members surrounding the at least one optical fiber; anda polymeric subunit jacket surrounding the layer of loose tensile strength members; andwherein the bundle is free of a glass-reinforced plastic (GRP) strength member."

## [8] "1. At least one machine accessible storage medium having code stored thereon, the code, when executed on a machine, to cause the machine to:establish a secure logical connection between a remotely located computer and the machine over a network;based on a registration corresponding to the remotely located computer, deliver a web page from the machine to the remotely located computer across said logical connection, said web page comprising scripts for performing a search on the remotely located computer; andcause, using the machine and without intervention of a user, the search to be executed at the remotely located computer via a network browser of the remotely located computer."

## [9] "1. An image forming apparatus comprising:a plurality of developing units to develop a visible image from an electrostatic latent image through a developer;a transfer member provided to transfer the visible image developed by the plurality of developing units onto a printing medium and to rotate in a first direction;a sensing unit including at least one sensor disposed opposite the transfer member, and at least one window disposed between the at least one sensor and the transfer member and corresponding to the at least one sensor, to sense the developer transferred onto the transfer member; anda shutter device having a shutter movably provided with a displacement in the first direction of rotation of the transfer member, wherein the shutter is to open and close the at least one window."

## [10] "1. An air conditioning system for conditioning a plurality of rooms within an interior of a building, the air conditioning system comprising:a single outdoor unit comprising:a compressor;a condenser; anda condenser fan associated with the condenser that moves air to cool the condenser;a refrigerant flow pathway comprised of a plurality of refrigerant conduits having a common refrigerant flow path portion and at least two divergent flow path portions, a first divergent flow path portion that delivers refrigerant to a first evaporator configured to operate at a first evaporator pressure and a second divergent flow path portion that delivers refrigerant to a second evaporator such that the first evaporator and second evaporator are in parallel connection with one another;at least one throttling device wherein a single throttling device is positioned along a common refrigerant flow path portion when a single throttling device is used and wherein a first throttling device is positioned along the first divergent flow path portion and a second throttling device is positioned along the second divergent flow path portion when two or more throttling devices are employed; andat least a first indoor air handling unit providing cooling to a first room within the interior of the building and a sec

```
ond indoor air handling unit providing cooling to a second room within the interio
r of the building and wherein the first indoor air handling unit comprises the fir
st evaporator and a first indoor air handling unit fan configured to deliver cooli
ng to the first room and the second indoor air handling unit comprises the second 
evaporator and a second indoor air unit handling fan configured to deliver cooling
to the second room; andwherein the compressor provides all of a compression of the
refrigerant used in the refrigerant flow pathway and the compressor is incapable o
f simultaneously supplying refrigerant to both the first evaporator and the second
evaporator at their full cooling capacity while both the first and second evaporat
ors are operating at the same time, and wherein the first room and second room are
separate rooms."
```

Challenge 1

If you just want the text from all the nodes you could use a single regular expression. What is it? (Hint: the answer isn't "chocolate".)

```
search_pattern <- "chocolate"
x <- lines %>%
  str_extract_all(search_pattern) %>%
  unlist()
sample(x,10)
```

```
##  [1] "chocolate" "chocolate" "chocolate" "chocolate" "chocolate"
##  [6] "chocolate" "chocolate" "chocolate" "chocolate" "chocolate"
```

Challenge 2

Extract a unique identifier for each claim and then create a data frame with two columns: the claim ID, and the claim text.

Then you can use the code from the "Romeo & Juliet" exercise to investigate word frequencies.

Challenge 3

This data is rather big to store and manipulate in memory - especially if you want to download load many of the files.

Investigate storing the data in an SQL database and sending queries that will operate on the database server rather than having to load everything into memory. Hint: MonetDBLite and RSQLite are packages that avoid you having to install and run a separate database server on your machine.

Challenge 4

Some of the text contains HTML or XML tags. Can you remove them?