

Lab sheet 7c: working with NASA data

Load necessary packages

```
library(jsonlite)
library(dplyr)
library(tidyr)
library(widyr)
library(tidytext)
library(ggplot2)
library(igraph)
library(ggraph)
```

Get the NASA data

```
# metadata <- fromJSON("https://data.nasa.gov/data.json")
```

```
load("metadata.rda")
names(metadata$dataset)
```

```
## [1] "accessLevel"      "landingPage"
## [3] "bureauCode"       "issued"
## [5] "@type"            "modified"
## [7] "references"        "keyword"
## [9] "contactPoint"     "publisher"
## [11] "identifier"        "description"
## [13] "title"             "programCode"
## [15] "distribution"      "accrualPeriodicity"
## [17] "theme"             "license"
## [19] "citation"          "temporal"
## [21] "spatial"           "language"
## [23] "graphic-preview-description" "graphic-preview-file"
## [25] "data-presentation-form"    "release-place"
## [27] "series-name"               "creator"
## [29] "dataQuality"               "editor"
## [31] "issue-identification"      "describedBy"
## [33] "describedByType"           "rights"
## [35] "systemOfRecords"
```

```
class(metadata$dataset$title)
```

```
## [1] "character"
```

```
class(metadata$dataset$description)
```

```
## [1] "character"
```

```
class(metadata$dataset$keyword)
```

```
## [1] "list"
```

Make tibble using the fields title, identifier, description, and keyword

```
nasa_title <- tibble(id = metadata$dataset$identifier,
                    title = metadata$dataset$title)
nasa_title
```

```
## # A tibble: 22,235 x 2
##   id                                     title
##   <chr>                                <chr>
## 1 urn:nasa:pds:context_pds3:data_set:data_set.ro-e-rpcm~ "ROS~
## 2 urn:nasa:pds:context_pds3:data_set:data_set.near-a-rss~ "NEA~
## 3 urn:nasa:pds:context_pds3:data_set:data_set.nh-a-leisa~ "NEW~
## 4 urn:nasa:pds:context_pds3:data_set:data_set.ro-c-rsi-1~ "ROS~
## 5 urn:nasa:pds:context_pds3:data_set:data_set.ear-a-3-rdr~ "AST~
## 6 C2350113375-LARC_ASDC                                "NAR~
## 7 urn:nasa:pds:context_pds3:data_set:data_set.vg2-j-mag-4~ "VOY~
## 8 C2600303267-ORNL_CLOUD                                "Fir~
## 9 C1633993919-GES_DISC                                  "Sou~
## 10 C1577484501-LARC_ASDC                                "CAT~
## # i 22,225 more rows
```

```
nasa_desc <- tibble(id = metadata$dataset$identifier,
                   desc = metadata$dataset$description)
nasa_desc %>%
  select(desc) %>%
  sample_n(5)
```

```
## # A tibble: 5 x 1
##   desc
##   <chr>
## 1 Planetary nomenclature, like terrestrial nomenclature, is used-
## 2 This is a Rosetta Radio Science data set, collected during the-
## 3 not available
## 4 This dataset includes high-resolution (~5 m) gridded estimates-
## 5 Data presented in this data set were collected during an inten-
```

```
nasa_keyword <- tibble(id = metadata$dataset$identifier,
                      keyword = metadata$dataset$keyword) %>%
  unnest(keyword)
nasa_keyword
```

```
## # A tibble: 114,995 x 2
##   id                                     keyword
##   <chr>                                <chr>
## 1 urn:nasa:pds:context_pds3:data_set:data_set.ro-e-rpcm~ earth
## 2 urn:nasa:pds:context_pds3:data_set:data_set.ro-e-rpcm~ unknown
## 3 urn:nasa:pds:context_pds3:data_set:data_set.ro-e-rpcm~ intern~
## 4 urn:nasa:pds:context_pds3:data_set:data_set.near-a-rs~ near e~
## 5 urn:nasa:pds:context_pds3:data_set:data_set.near-a-rs~ eros
## 6 urn:nasa:pds:context_pds3:data_set:data_set.nh-a-leis~ vega
## 7 urn:nasa:pds:context_pds3:data_set:data_set.nh-a-leis~ new ho~
## 8 urn:nasa:pds:context_pds3:data_set:data_set.ro-c-rsi~ intern~
## 9 urn:nasa:pds:context_pds3:data_set:data_set.ro-c-rsi~ 67p/ch~
## 10 urn:nasa:pds:context_pds3:data_set:data_set.ear-a-3-r~ satell~
## # i 114,985 more rows
```

Remove unnecessary words

```
nasa_title <- nasa_title %>%  
  unnest_tokens(word, title) %>%  
  anti_join(stop_words)
```

```
## Joining with `by = join_by(word)`
```

```
nasa_desc <- nasa_desc %>%  
  unnest_tokens(word, desc) %>%  
  anti_join(stop_words)
```

```
## Joining with `by = join_by(word)`
```

```
nasa_title %>%  
  count(word, sort = TRUE)
```

```
## # A tibble: 12,549 x 2
```

```
##   word      n  
##   <chr>  <int>  
## 1 v1.0    6184  
## 2 data    4627  
## 3 2       4153  
## 4 rosetta 4031  
## 5 1       3941  
## 6 orbiter 3887  
## 7 3       3794  
## 8 67p     2676  
## 9 ges     1719  
## 10 disc   1718  
## # i 12,539 more rows
```

```
my_stopwords <- tibble(word = c(as.character(1:10),  
                                "v001", "0.5", "v1", "v1.0", "v2.0", "0.4", "r2022.0", "67p", "v03", "12", "13",  
                                "v003", "v004", "v005", "v006", "v7"))
```

```
nasa_title <- nasa_title %>%  
  anti_join(my_stopwords)
```

```
## Joining with `by = join_by(word)`
```

```
nasa_desc <- nasa_desc %>%  
  anti_join(my_stopwords)
```

```
## Joining with `by = join_by(word)`
```

Grouping and counting keywords

```
nasa_keyword %>%  
  group_by(keyword) %>%  
  count(sort = TRUE)
```

```
## # A tibble: 9,104 x 2
```

```
## # Groups:   keyword [9,104]
```

```
##   keyword      n  
##   <chr>      <int>  
## 1 earth science 9762  
## 2 atmosphere   4248
```

```
## 3 international rosetta mission      3806
## 4 67p/churyumov-gerasimenko 1 (1969 r1) 2977
## 5 land surface                      2201
## 6 oceans                          1926
## 7 spectral/engineering              1580
## 8 biosphere                        1399
## 9 atmospheric water vapor           1354
## 10 mars                            1321
## # i 9,094 more rows
```

Capitalize the keywords

```
nasa_keyword <- nasa_keyword %>%
  mutate(keyword = toupper(keyword))
```

Some usefull graphs

```
title_word_pairs <- nasa_title %>%
  pairwise_count(word, id, sort = TRUE, upper = FALSE)
```

```
title_word_pairs
```

```
## # A tibble: 189,748 x 3
##   item1 item2      n
##   <chr> <chr> <dbl>
## 1 rosetta orbiter 3672
## 2 ges     disc   1717
## 3 rosetta rsi    1295
## 4 orbiter rsi    1295
## 5 rosetta comet 1138
## 6 orbiter comet 1138
## 7 rosetta escort 1084
## 8 orbiter escort 1084
## 9 comet   escort 1024
## 10 rsi     comet   847
## # i 189,738 more rows
```

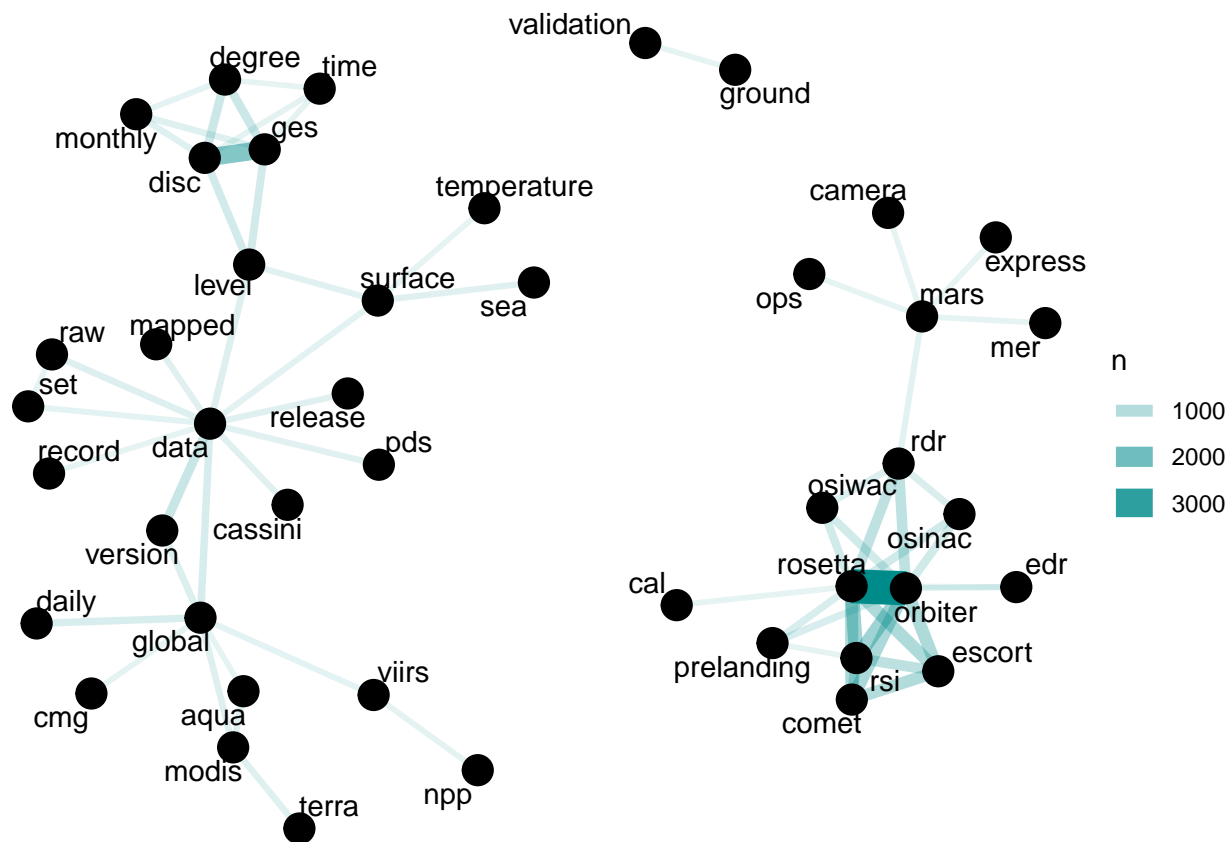
```
desc_word_pairs <- nasa_desc %>%
  pairwise_count(word, id, sort = TRUE, upper = FALSE)
```

```
desc_word_pairs
```

```
## # A tibble: 7,957,269 x 3
##   item1 item2      n
##   <chr> <chr> <dbl>
## 1 data  set      9359
## 2 data  time     4951
## 3 data  phase    4889
## 4 data  mission  4713
## 5 data  level    4555
## 6 data  version  4519
## 7 data  instrument 4506
## 8 set   phase    4259
## 9 data  global   4092
```

```
## 10 data product      4022
## # i 7,957,259 more rows

set.seed(1234)
title_word_pairs %>%
  filter(n >= 250) %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr") +
  geom_edge_link(aes(edge_alpha = n, edge_width = n), edge_colour = "cyan4") +
  geom_node_point(size = 5) +
  geom_node_text(aes(label = name), repel = TRUE,
    point.padding = unit(0.2, "lines")) +
  theme_void()
```



```
keyword_pairs <- nasa_keyword %>%
  pairwise_count(keyword, id, sort = TRUE, upper = FALSE)
```

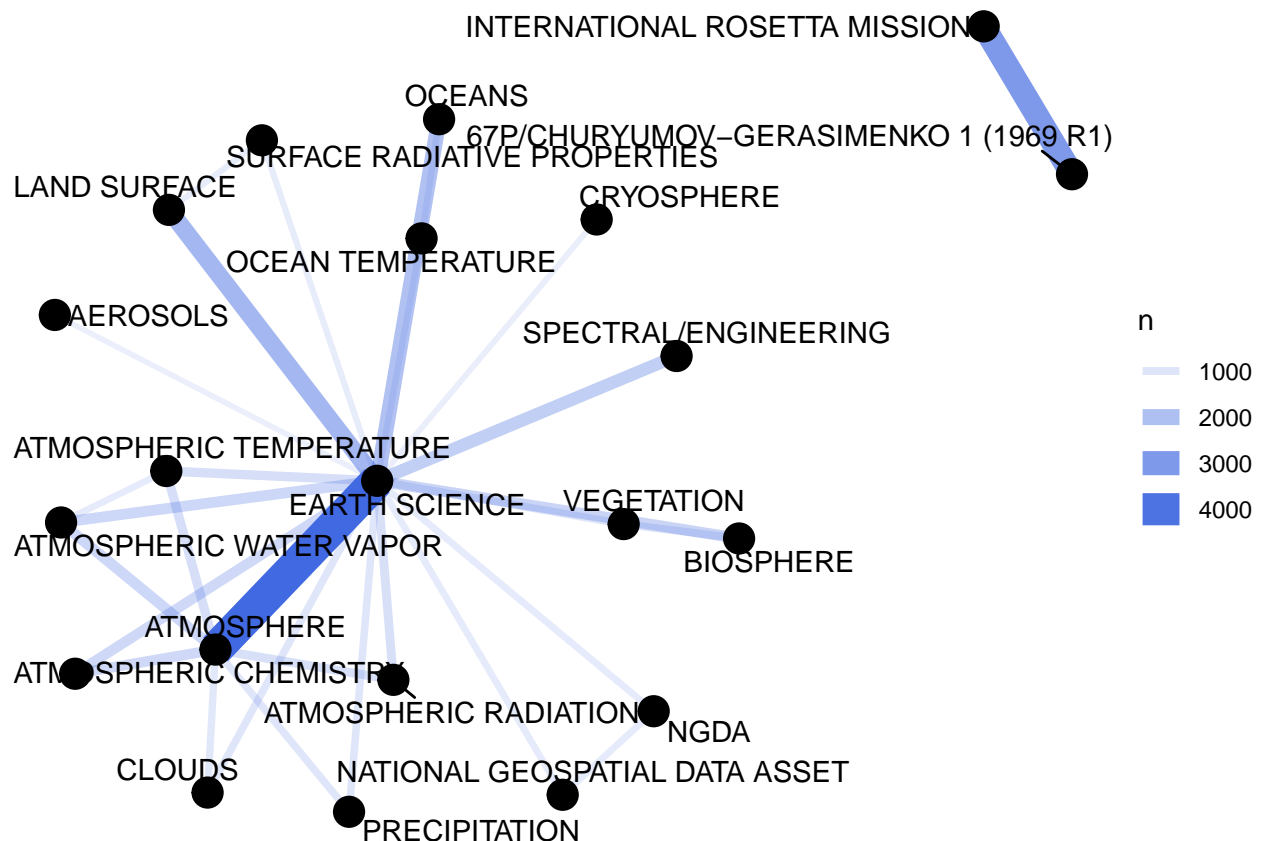
```
keyword_pairs
```

```
## # A tibble: 3,008,145 x 3
```

##	item1	item2	n
##	<chr>	<chr>	<dbl>
##	1 EARTH SCIENCE	ATMOSPHERE	4244
##	2 INTERNATIONAL ROSETTA MISSION	67P/CHURYUMOV-GERASIMENKO~	2971
##	3 EARTH SCIENCE	LAND SURFACE	2201
##	4 EARTH SCIENCE	OCEANS	1923
##	5 EARTH SCIENCE	SPECTRAL/ENGINEERING	1580
##	6 EARTH SCIENCE	BIOSPHERE	1398

```
## 7 EARTH SCIENCE          ATMOSPHERIC WATER VAPOR    1354
## 8 ATMOSPHERE             ATMOSPHERIC WATER VAPOR    1354
## 9 EARTH SCIENCE          ATMOSPHERIC CHEMISTRY      1309
## 10 ATMOSPHERE            ATMOSPHERIC CHEMISTRY      1309
## # i 3,008,135 more rows
```

```
set.seed(1234)
keyword_pairs %>%
  filter(n >= 700) %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr") +
  geom_edge_link(aes(edge_alpha = n, edge_width = n), edge_colour = "royalblue") +
  geom_node_point(size = 5) +
  geom_node_text(aes(label = name), repel = TRUE,
    point.padding = unit(0.2, "lines")) +
  theme_void()
```



```
desc_tf_idf <- nasa_desc %>%
  count(id, word, sort = TRUE) %>%
  bind_tf_idf(word, id, n)
desc_tf_idf %>%
  arrange(-tf_idf)
```

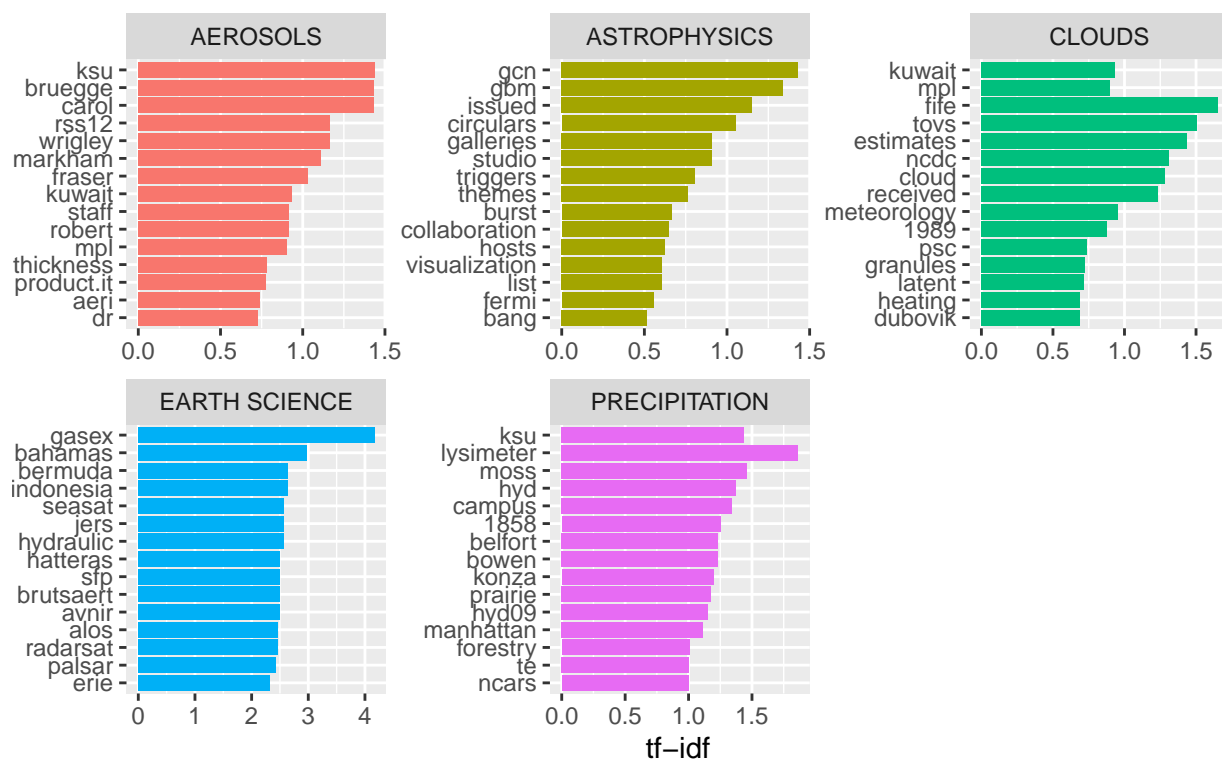
```
## # A tibble: 1,123,591 x 6
```

##	id	word	n	tf	idf	tf_idf
##	<chr>	<chr>	<int>	<dbl>	<dbl>	<dbl>
##	1	C1206487217-ASF	pals~	1	10.0	10.0
##	2	C1206487504-ASF	pals~	1	10.0	10.0

```
## 3 C1633360161-OB_DAAC      bio_~      1      1 10.0   10.0
## 4 urn:nasa:pds:context_pds3:data~ unk      1      1  9.31   9.31
## 5 urn:nasa:pds:context_pds3:data~ unk      1      1  9.31   9.31
## 6 urn:nasa:pds:lab.hydrocarbon_s~ ____~      1      1  9.31   9.31
## 7 urn:nasa:pds:mgs_tes_recalib_a~ ____~      1      1  9.31   9.31
## 8 C2263929260-OB_DAAC      temp~      1      1  7.93   7.93
## 9 C2263929262-OB_DAAC      temp~      1      1  7.93   7.93
## 10 C2263929265-OB_DAAC      temp~      1      1  7.93   7.93
## # i 1,123,581 more rows
```

```
desc_tf_idf <- full_join(desc_tf_idf, nasa_keyword, by = "id")
desc_tf_idf %>%
  filter(!near(tf, 1)) %>%
  filter(keyword %in% c( "CLOUDS",
                        "AEROSOLS", "ASTROPHYSICS",
                        "PRECIPITATION", "EARTH SCIENCE")) %>%
  arrange(desc(tf_idf)) %>%
  group_by(keyword) %>%
  distinct(word, keyword, .keep_all = TRUE) %>%
  slice_max(tf_idf, n = 15, with_ties = FALSE) %>%
  ungroup() %>%
  mutate(word = factor(word, levels = rev(unique(word)))) %>%
  ggplot(aes(tf_idf, word, fill = keyword)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~keyword, ncol = 3, scales = "free") +
  labs(title = "Highest tf-idf words in NASA metadata description fields",
       caption = "NASA metadata from https://data.nasa.gov/data.json",
       x = "tf-idf", y = NULL)
```

Highest tf-idf words in NASA metadata description fields



NASA metadata from <https://data.nasa.gov/data.json>