08 Case study: mining NASA metadata

H. David Shea

05 Aug 2021

Contents

How data is organized at NASA

```
# pulling NASA metadata takes a long time - saved to .rds file to speed up processing
if(file.exists("data/metadata.rda")) {
    load("data/metadata.rda")
} else {
    metadata <- fromJSON("https://data.nasa.gov/data.json")
    save(metadata, file = "data/metadata.rda")
}

names(metadata$dataset)
#>  [1] "_id"                "@type"              "accessLevel"
#>  [4] "accrualPeriodicity" "bureauCode"         "contactPoint"
#>  [7] "description"        "distribution"       "identifier"
#> [10] "issued"             "keyword"            "landingPage"
#> [13] "language"           "modified"           "programCode"
#> [16] "publisher"          "spatial"            "temporal"
#> [19] "theme"              "title"              "license"
#> [22] "isPartOf"           "references"         "rights"
#> [25] "describedBy"

class(metadata$dataset$title)
#> [1] "character"

class(metadata$dataset$description)
#> [1] "character"

class(metadata$dataset$keyword)
#> [1] "list"
```

Wrangling and tidying the data

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

nasa_title %>%
    slice_head(n = 10)
#> # A tibble: 10 x 2
#>    id                       title
#>    <chr>                    <chr>
#>  1 55942a57c63a7fe59b495a77 15 Minute Stream Flow Data: USGS (FIFE)
#>  2 55942a57c63a7fe59b495a78 15 Minute Stream Flow Data: USGS (FIFE)
#>  3 55942a58c63a7fe59b495a79 15 Minute Stream Flow Data: USGS (FIFE)
#>  4 55942a58c63a7fe59b495a7a 2000 Pilot Environmental Sustainability Index (ESI)
#>  5 55942a58c63a7fe59b495a7b 2000 Pilot Environmental Sustainability Index (ESI)
#>  6 55942a58c63a7fe59b495a7c 2000 Pilot Environmental Sustainability Index (ESI)
#>  7 55942a58c63a7fe59b495a7d 2001 Environmental Sustainability Index (ESI)
#>  8 55942a58c63a7fe59b495a7e 2001 Environmental Sustainability Index (ESI)
#>  9 55942a58c63a7fe59b495a7f 2001 Environmental Sustainability Index (ESI)
#> 10 55942a58c63a7fe59b495a80 2001 Environmental Sustainability Index (ESI)

nasa_desc <- tibble(id = metadata$dataset$`_id`$`$oid`,
                    desc = metadata$dataset$description)

nasa_desc %>%
    select(desc) %>%
    slice_sample(n = 5)
#> # A tibble: 5 x 1
#>   desc
#>   <chr>
#> 1 "The Coastal Zone Color Scanner Experiment (CZCS) was the first instrument de~
#> 2 "The Anthropogenic Sulfur Dioxide Emissions, 1850-2005: National and Regional~
#> 3 "This data set contains mean normalized backscatter coefficient (sigma-naught~
#> 4 "Contains decomposition rates of a standard substrate (wheat straw)."
#> 5 "Energy, carbon dioxide, water vaport, and momentum flux data collected above~

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

nasa_keyword
#> # A tibble: 126,814 x 2
#>    id                       keyword
#>    <chr>                    <chr>
#>  1 55942a57c63a7fe59b495a77 EARTH SCIENCE
#>  2 55942a57c63a7fe59b495a77 HYDROSPHERE
#>  3 55942a57c63a7fe59b495a77 SURFACE WATER
#>  4 55942a57c63a7fe59b495a78 EARTH SCIENCE
#>  5 55942a57c63a7fe59b495a78 HYDROSPHERE
#>  6 55942a57c63a7fe59b495a78 SURFACE WATER
#>  7 55942a58c63a7fe59b495a79 EARTH SCIENCE
#>  8 55942a58c63a7fe59b495a79 HYDROSPHERE
#>  9 55942a58c63a7fe59b495a79 SURFACE WATER
```

```
#> 10 55942a58c63a7fe59b495a7a EARTH SCIENCE
#> # ... with 126,804 more rows

nasa_title <- nasa_title %>%
    unnest_tokens(word, title) %>%
    anti_join(stop_words, by = "word")

nasa_title
#> # A tibble: 210,914 x 2
#>    id                       word
#>    <chr>                    <chr>
#>  1 55942a57c63a7fe59b495a77 15
#>  2 55942a57c63a7fe59b495a77 minute
#>  3 55942a57c63a7fe59b495a77 stream
#>  4 55942a57c63a7fe59b495a77 flow
#>  5 55942a57c63a7fe59b495a77 data
#>  6 55942a57c63a7fe59b495a77 usgs
#>  7 55942a57c63a7fe59b495a77 fife
#>  8 55942a57c63a7fe59b495a78 15
#>  9 55942a57c63a7fe59b495a78 minute
#> 10 55942a57c63a7fe59b495a78 stream
#> # ... with 210,904 more rows

nasa_desc <- nasa_desc %>%
    unnest_tokens(word, desc) %>%
    anti_join(stop_words, by = "word")

nasa_desc
#> # A tibble: 2,677,811 x 2
#>    id                       word
#>    <chr>                    <chr>
#>  1 55942a57c63a7fe59b495a77 usgs
#>  2 55942a57c63a7fe59b495a77 15
#>  3 55942a57c63a7fe59b495a77 minute
#>  4 55942a57c63a7fe59b495a77 stream
#>  5 55942a57c63a7fe59b495a77 flow
#>  6 55942a57c63a7fe59b495a77 data
#>  7 55942a57c63a7fe59b495a77 kings
#>  8 55942a57c63a7fe59b495a77 creek
#>  9 55942a57c63a7fe59b495a77 konza
#> 10 55942a57c63a7fe59b495a77 prairie
#> # ... with 2,677,801 more rows
```

Some initial simple exploration

```
nasa_title %>%
    count(word, sort = TRUE) %>%
    slice_max(n, n = 10, with_ties = FALSE) %>%
    kable(caption = "Most common words in titles")
```

Table 1: Most common words in titles

word	n
project	7735
data	3354
1	2841
level	2400
global	1809
v1	1478
daily	1397
3	1364
aura	1363
l2	1311

```
nasa_desc %>%
    count(word, sort = TRUE) %>%
    slice_max(n, n = 10, with_ties = FALSE) %>%
    kable(caption = "Most common words in descriptions")
```

Table 2: Most common words in descriptions

word	n
data	68871
modis	24420
global	23028
2	16599
1	15770
system	15480
product	14780
aqua	14738
earth	14373
resolution	13879

```
my_stopwords <- tibble(word = c(as.character(1:10),
                                "v1", "v1.0", "v03", "l2", "l3", "l4", "v5.2.0", "0.5",
                                "v003", "v004", "v005", "v006", "v7", "ii"))
nasa_title <- nasa_title %>%
    anti_join(my_stopwords, by = "word")

nasa_desc <- nasa_desc %>%
    anti_join(my_stopwords, by = "word")

nasa_keyword %>%
    count(keyword, sort = TRUE) %>%
    slice_max(n, n = 10, with_ties = FALSE) %>%
    kable(caption = "Most common keywords")
```

Table 3: Most common keywords

keyword	n
EARTH SCIENCE	14362
Project	7452
ATMOSPHERE	7321
Ocean Color	7268
Ocean Optics	7268
Oceans	7268
completed	6452
ATMOSPHERIC WATER VAPOR	3142
OCEANS	2765
LAND SURFACE	2720

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

Word co-ocurrences and correlations

We examine which words commonly occur together in the titles, descriptions, and keywords of NASA datasets. Then, we can examine word networks for each showing which datasets might be related.

Networks of Description and Title Words

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

title_word_pairs %>%
    slice_max(n, n = 10, with_ties = FALSE) %>%
    kable(caption = "Most frequent word pairs in titles")
```

Table 4: Most frequent word pairs in titles

item1	item2	n
system	project	796
lba	eco	683
airs	aqua	641
level	aqua	623
level	airs	612
aura	omi	607
global	grid	597
global	daily	574
data	boreas	551
ground	gpm	550

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

desc_word_pairs %>%
    slice_max(n, n = 10, with_ties = FALSE) %>%
    kable(caption = "Most frequent word pairs in descriptions")
```

Table 5: Most frequent word pairs in descriptions

item1	item2	n
data	global	9864
data	resolution	9302
instrument	resolution	8189
data	surface	8180
global	resolution	8139
data	instrument	7994
data	system	7870
resolution	bands	7584
data	earth	7576
orbit	resolution	7462

```
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()
```

```
set.seed(1234)
desc_word_pairs %>%
    filter(n >= 5000) %>%
    graph_from_data_frame() %>%
    ggraph(layout = "fr") +
    geom_edge_link(aes(edge_alpha = n, edge_width = n), edge_colour = "darkred") +
    geom_node_point(size = 5) +
    geom_node_text(aes(label = name),
                   repel = TRUE,
                   point.padding = unit(0.2, "lines")) +
    theme_void()
```

Networks of Description and Title Words

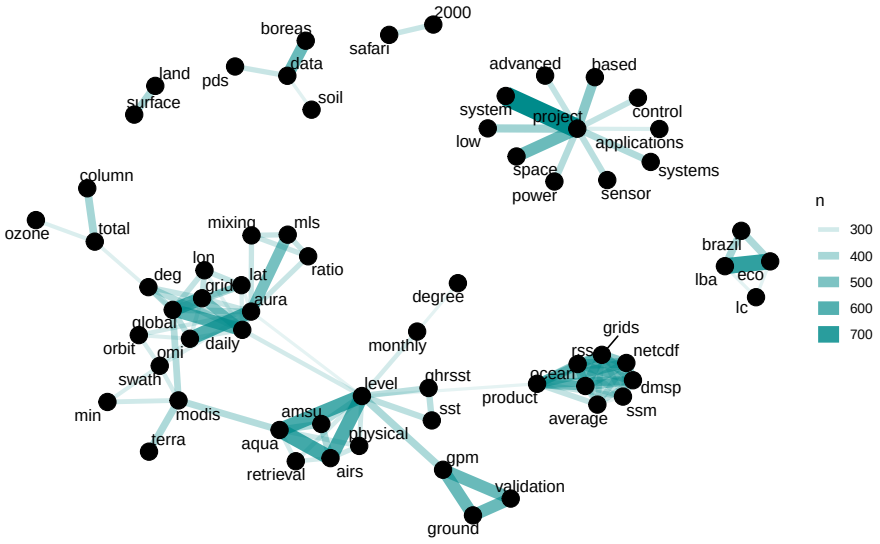

Figure 1: Word network in NASA dataset titles

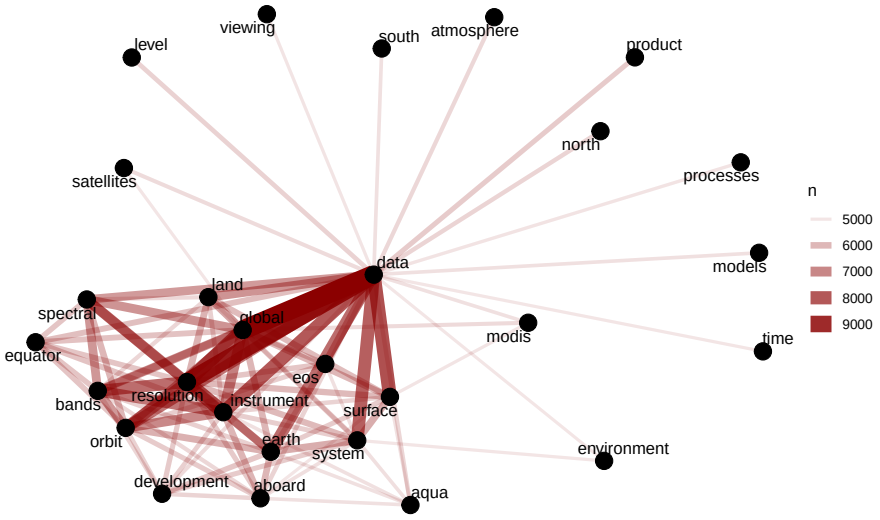

Figure 2: Word network in NASA dataset descriptions

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

keyword_pairs %>%
    slice_max(n, n = 10, with_ties = FALSE) %>%
    kable(caption = "Most frequent keyword pairs")
```

Table 6: Most frequent keyword pairs

item1	item2	n
oceans	ocean optics	7324
earth science	atmosphere	7318
oceans	ocean color	7270
ocean optics	ocean color	7270
project	completed	6450
earth science	atmospheric water vapor	3142
atmosphere	atmospheric water vapor	3142
earth science	oceans	2762
earth science	land surface	2718
earth science	biosphere	2448

```
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()
```

```
keyword_cors <- nasa_keyword %>%
  group_by(keyword) %>%
  filter(n() >= 50) %>%
  pairwise_cor(keyword, id, sort = TRUE, upper = FALSE)

keyword_cors %>%
    slice_max(correlation, n = 10, with_ties = FALSE) %>%
    kable(caption = "Highest correlations in keyword pairs")
```

Table 7: Highest correlations in keyword pairs

item1	item2	correlation
knowledge	sharing	1.0000000
dashlink	ames	1.0000000
schedule	expedition	1.0000000

item1	item2	correlation
turbulence	models	0.9971871
appel	knowledge	0.9967945
appel	sharing	0.9967945
ocean optics	ocean color	0.9952123
atmospheric science	cloud	0.9938681
launch	schedule	0.9837078
launch	expedition	0.9837078

```
set.seed(1234)
keyword_cors %>%
    filter(correlation > .6) %>%
    graph_from_data_frame() %>%
    ggraph(layout = "fr") +
    geom_edge_link(aes(edge_alpha = correlation, edge_width = correlation),
                   edge_colour = "royalblue") +
    geom_node_point(size = 5) +
    geom_node_text(aes(label = name),
                   repel = TRUE,
                   point.padding = unit(0.2, "lines")) +
    theme_void()
```

Calculating tf-idf for the description fields

We apply the tf-idf approach to the description fields of these NASA datasets.

What is tf-idf for the description field words?

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

desc_tf_idf %>%
    arrange(-tf_idf) %>%
    slice_max(tf_idf, n = 10, with_ties = FALSE) %>%
    kable(caption = "Highest tf-idf values for descrition fields")
```

Table 8: Highest tf-idf values for descrition fields

id	word	n	tf	idf	tf_idf
55942a7cc63a7fe59b49774a	rdr	1	1	10.375053	10.375053
55942ac9c63a7fe59b49b688	palsar_radiometric_terrain_corrected_high_res	1	1	10.375053	10.375053
55942ac9c63a7fe59b49b689	palsar_radiometric_terrain_corrected_low_res	1	1	10.375053	10.375053
55942a7bc63a7fe59b4976ca	lgrs	1	1	8.765614	8.765614
55942a7bc63a7fe59b4976d2	lgrs	1	1	8.765614	8.765614
55942a7bc63a7fe59b4976e3	lgrs	1	1	8.765614	8.765614

id	word	n	tf	idf	tf_idf
55942a7dc63a7fe59b497820mri		1	1	8.583293	8.583293
55942ad8c63a7fe59b49cf6c	template_proddescription	1	1	8.295611	8.295611
55942ad8c63a7fe59b49cf6d	template_proddescription	1	1	8.295611	8.295611
55942ad8c63a7fe59b49cf6e	template_proddescription	1	1	8.295611	8.295611

"Notice we have run into an issue here; both n and term frequency are equal to 1 for these terms, meaning that these were description fields that only had a single word in them. If a description field only contains one word, the tf-idf algorithm will think that is a very important word."

"Depending on our analytic goals, it might be a good idea to throw out all description fields that have very few words."

```
desc_tf_idf %>%
    filter(n > 3, tf != 1) %>%
    arrange(-tf_idf) %>%
    slice_max(tf_idf, n = 10, with_ties = FALSE) %>%
    kable(caption = "Highest tf-idf values for descrition fields (n > 3 and tf != 1)")
```

Table 9: Highest tf-idf values for descrition fields (n > 3 and tf != 1)

id	word	n	tf	idf	tf_idf
56cf5b00a759fdadc44e56d4	ug3	10	0.2000000	8.583293	1.716659
56cf5b00a759fdadc44e56d6	ug3	10	0.2000000	8.583293	1.716659
56cf5b00a759fdadc44e56d0	td	128	0.2184300	7.735995	1.689774
56cf5b00a759fdadc44e56d7	ug3	10	0.1960784	8.583293	1.682999
56cf5b00a759fdadc44e56d5	ug3	10	0.1886792	8.583293	1.619489
55942a88c63a7fe59b498280	nbsp	655	0.3825935	4.205442	1.608974
56cf5b00a759fdadc44e56d2	ug3	10	0.1851852	8.583293	1.589499
56cf5b00a759fdadc44e56d3	ug3	10	0.1818182	8.583293	1.560599
55942a86c63a7fe59b49803b	nbsp	204	0.3682310	4.205442	1.548574
55942a5cc63a7fe59b495e15	nsa	5	0.2500000	5.554771	1.388693

Connecting description fields to keywords

```
desc_tf_idf <- full_join(desc_tf_idf, nasa_keyword, by = "id")

desc_tf_idf %>%
    filter(!near(tf, 1)) %>%
    filter(
        keyword %in% c(
            "solar activity",
            "clouds",
            "seismology",
            "astrophysics",
            "human health",
            "budget"
        )
```

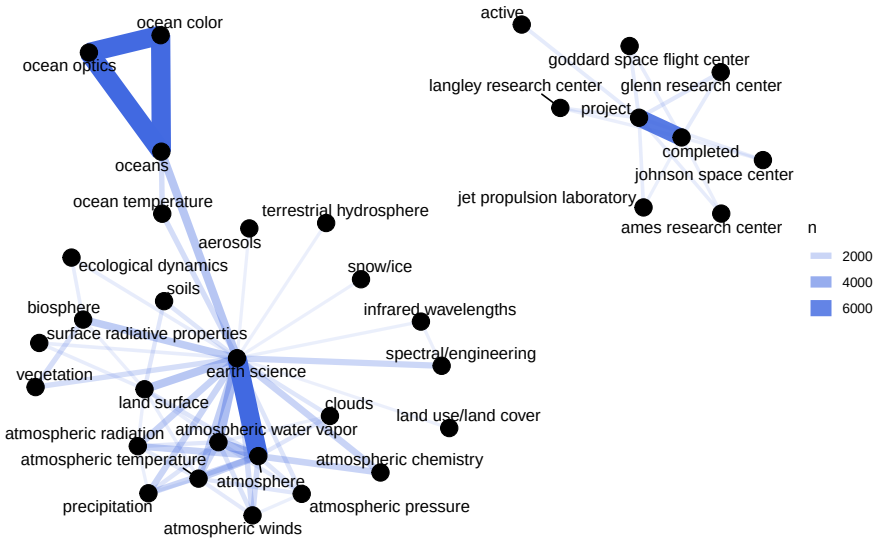

Figure 3: Co-occurrence network in NASA dataset keywords

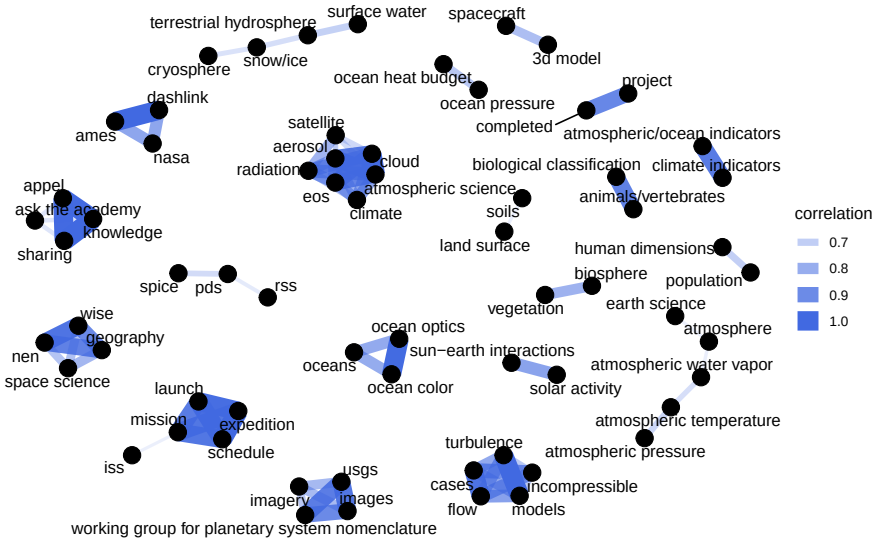

Figure 4: Correlation network in NASA dataset keywords

```
    ) %>%
    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
    ) +
    theme_light()
```

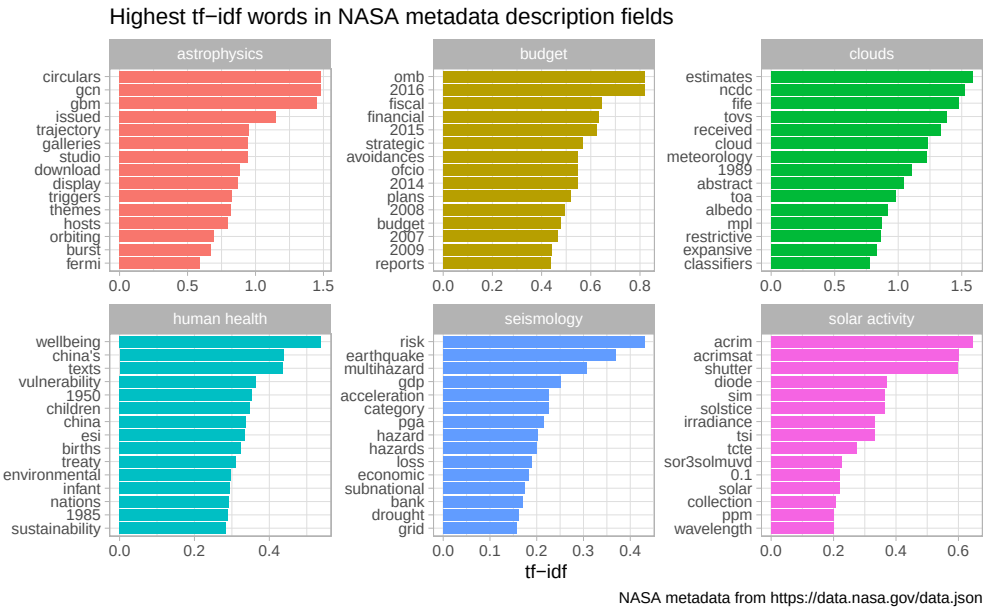

Figure 5: Distribution of tf-idf for words from datasets labeled with selected keywords

Topic modeling

We use topic modeling to model each document description field as a mixture of topics and each topic as a mixture of words. We will use LDA for our topic modeling.

Casting to a document-term matrix

```
my_stop_words <- bind_rows(stop_words,
                           tibble(word = c("nbsp", "amp", "gt", "lt",
```

```
                                          "timesnewromanpsmt", "font",
                                          "td", "li", "br", "tr", "quot",
                                          "st", "img", "src", "strong",
                                          "http", "file", "files",
                                          as.character(1:12)),
                                 lexicon = rep("custom", 30)))

word_counts <- nasa_desc %>%
  anti_join(my_stop_words, by = "word") %>%
  count(id, word, sort = TRUE) %>%
  ungroup()

word_counts %>%
    slice_max(n, n = 10, with_ties = FALSE) %>%
    kable(caption = "Highest word count in decsriptions - stop words removed")
```

Table 10: Highest word count in decsriptions - stop words removed

id	word	n
55942a8ec63a7fe59b4986ef	suit	82
55942a8ec63a7fe59b4986ef	space	69
56cf5b00a759fdadc44e564a	data	41
56cf5b00a759fdadc44e564a	leak	40
56cf5b00a759fdadc44e564a	tree	39
55942a8ec63a7fe59b4986ef	pressure	34
55942a8ec63a7fe59b4986ef	system	34
55942a89c63a7fe59b4982d9	em	32
55942a8ec63a7fe59b4986ef	al	32
55942a8ec63a7fe59b4986ef	human	31

```
desc_dtm <- word_counts %>%
  cast_dtm(id, word, n)

desc_dtm
#> <<DocumentTermMatrix (documents: 32003, terms: 35898)>>
#> Non-/sparse entries: 1892658/1146951036
#> Sparsity           : 100%
#> Maximal term length: 166
#> Weighting          : term frequency (tf)
```

Ready for topic modeling

To determine the number of topics to use, the authors tested increments of 8 from 8 to 64. They found that at 24, documents were still getting sorted into topics cleanly. Higher numbers produced flatter, less discerning distributions of gamma.

```
# running a 24 topic LDA on this data takes a long time - saved to .rds file to speed up processing
if(file.exists("data/desc_lda.rda")) {
    load(file = "data/desc_lda.rda")
} else {
```

```
    desc_lda <- LDA(desc_dtm, k = 24, control = list(seed = 1234))
    save(desc_lda, file = "data/desc_lda.rda")
}

desc_lda
#> A LDA_VEM topic model with 24 topics.
```

Interpreting the topic model

```
tidy_lda <- tidy(desc_lda, matrix = "beta")

tidy_lda %>%
    slice_head(n = 10) %>%
    kable()
```

topic	term	beta
1	suit	0.0000000
2	suit	0.0000000
3	suit	0.0000000
4	suit	0.0000000
5	suit	0.0000000
6	suit	0.0000000
7	suit	0.0003284
8	suit	0.0000000
9	suit	0.0000000
10	suit	0.0000000

```
top_terms <- tidy_lda %>%
  group_by(topic) %>%
  slice_max(beta, n = 10, with_ties = FALSE) %>%
  ungroup() %>%
  arrange(topic, -beta)

top_terms %>%
    slice_head(n = 10) %>%
    kable(caption = "Top terms by beta")
```

Table 12: Top terms by beta

topic	term	beta
1	data	0.0448896
1	soil	0.0367620
1	moisture	0.0295456
1	amsr	0.0243775
1	sst	0.0168400
1	validation	0.0132246
1	temperature	0.0131707

topic	term	beta
1	surface	0.0129005
1	accuracy	0.0122513
1	set	0.0115537

```
top_terms %>%
    mutate(term = reorder_within(term, beta, topic)) %>%
    group_by(topic, term) %>%
    arrange(desc(beta)) %>%
    ungroup() %>%
    ggplot(aes(beta, term, fill = as.factor(topic))) +
    geom_col(show.legend = FALSE) +
    scale_y_reordered() +
    labs(title = "Top 10 terms in each LDA topic",
         x = expression(beta), y = NULL) +
    facet_wrap( ~ topic, ncol = 4, scales = "free")
```

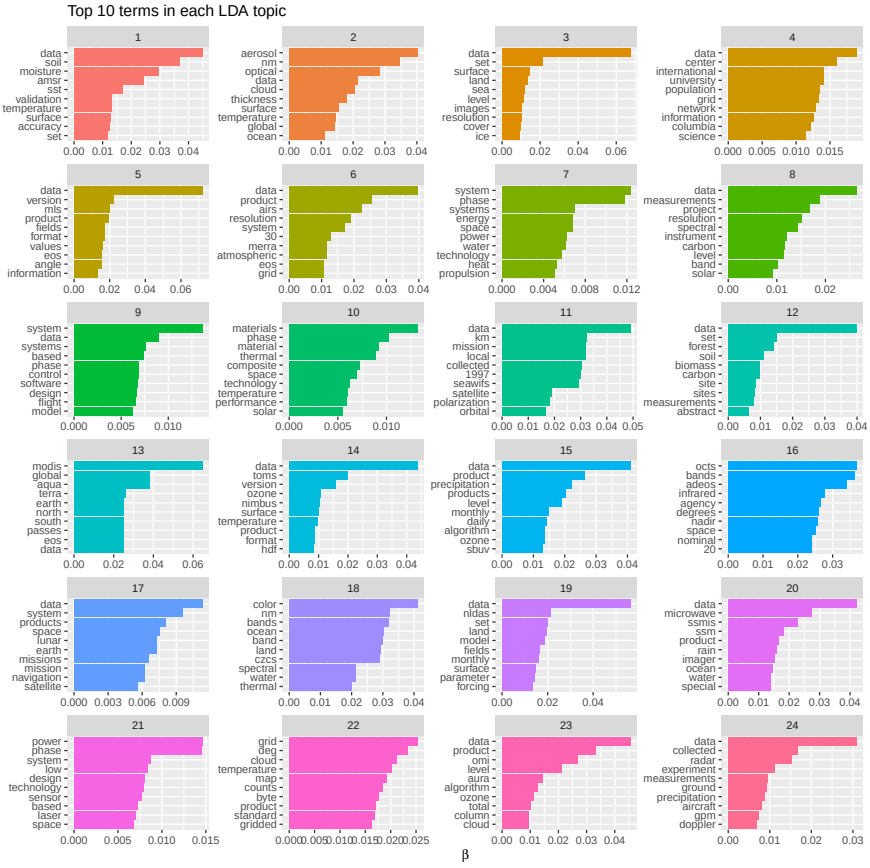

Figure 6: Top terms in topic modeling of NASA metadata description field texts

```
lda_gamma <- tidy(desc_lda, matrix = "gamma")

lda_gamma %>%
```

```
    slice_head(n = 10) %>%
    kable()
```

document	topic	gamma
55942a8ec63a7fe59b4986ef	1	0.0000065
56cf5b00a759fdadc44e564a	1	0.0000116
55942a89c63a7fe59b4982d9	1	0.0491744
56cf5b00a759fdadc44e55cd	1	0.0000225
55942a89c63a7fe59b4982c6	1	0.0000661
55942a86c63a7fe59b498077	1	0.0000567
56cf5b00a759fdadc44e56f8	1	0.0000475
55942a8bc63a7fe59b4984b5	1	0.0000431
55942a6ec63a7fe59b496bf7	1	0.0000441
55942a8ec63a7fe59b4986f6	1	0.0000288

```
ggplot(lda_gamma, aes(gamma)) +
    geom_histogram(alpha = 0.8) +
    scale_y_log10() +
    labs(title = "Distribution of probabilities for all topics",
         y = "Number of documents",
         x = expression(gamma)) +
    theme_light()
```

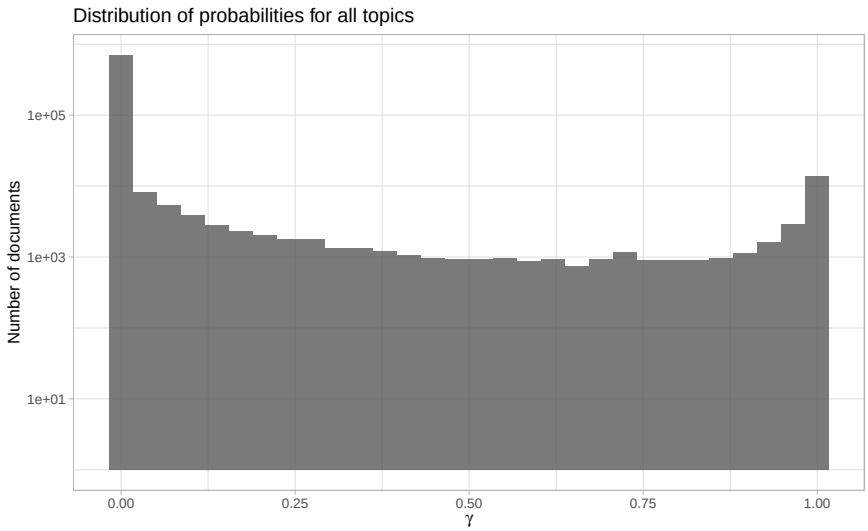

Figure 7: Probability distribution in topic modeling of NASA metadata description field texts

```
ggplot(lda_gamma, aes(gamma, fill = as.factor(topic))) +
    geom_histogram(alpha = 0.8, show.legend = FALSE) +
    facet_wrap( ~ topic, ncol = 4) +
    scale_y_log10() +
    labs(title = "Distribution of probability for each topic",
         y = "Number of documents",
         x = expression(gamma)) +
```

```
    theme_light()
#> Warning: Transformation introduced infinite values in continuous y-axis
#> Warning: Removed 67 rows containing missing values (geom_bar).
```

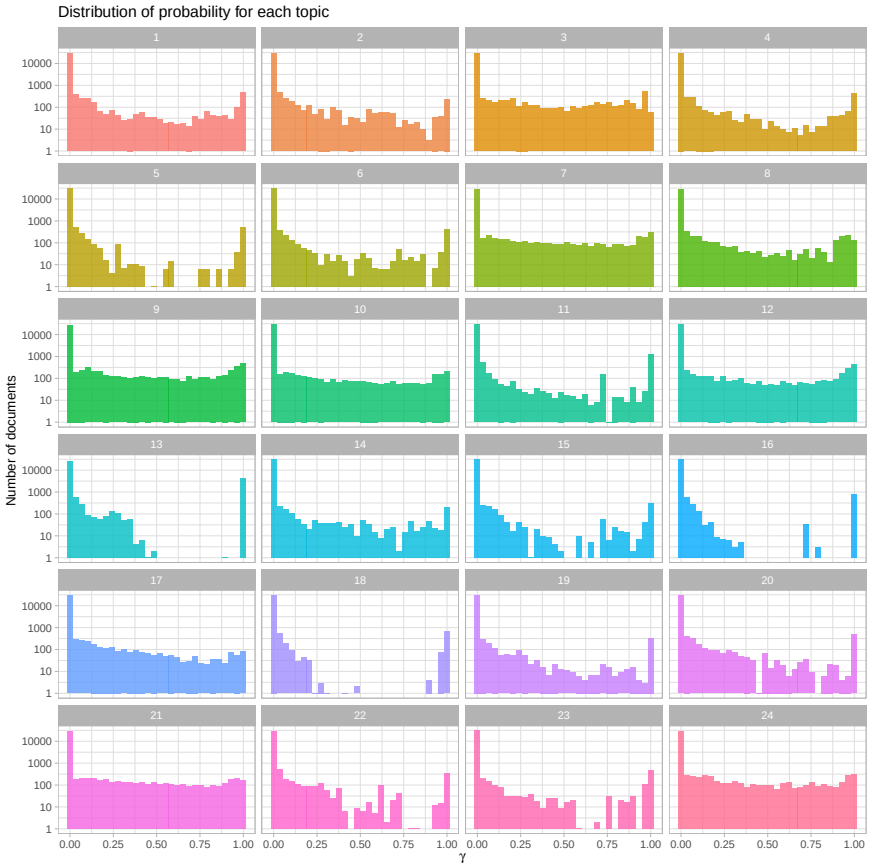

Figure 8: Probability distribution for each topic in topic modeling of NASA metadata description field texts

A "good" distribution for all topics and individual topics will show a clustering near gamma = 0 - documents that **do not** belong to the topic - and a clustering near gamma = 1 - documents that **do** belong to the topic.

Looking at topic gamma distributions can help in determining the number of topics to model. Flat distributions with little or no clustering at the extremes indicate that documents are not getting sorted into topics very well. A lower number might be better.

Connecting topic modeling with keywords

The topic model data combined with the human-tagged keywords may provide a solid way to identify or gategorize the different topics selected by the model.

```
lda_gamma <- full_join(lda_gamma, nasa_keyword, by = c("document" = "id"))

lda_gamma %>%
    slice_head(n = 10) %>%
    kable(caption = "Gamma data - the probability that each document belongs in each topic - joined wit
```

Table 14: Gamma data - the probability that each document belongs in each topic - joined with keywords

document	topic	gamma	keyword
55942a8ec63a7fe59b4986ef	1	0.0000065	johnson space center
55942a8ec63a7fe59b4986ef	1	0.0000065	project
55942a8ec63a7fe59b4986ef	1	0.0000065	completed
56cf5b00a759fdadc44e564a	1	0.0000116	dashlink
56cf5b00a759fdadc44e564a	1	0.0000116	ames
56cf5b00a759fdadc44e564a	1	0.0000116	nasa
55942a89c63a7fe59b4982d9	1	0.0491744	goddard space flight center
55942a89c63a7fe59b4982d9	1	0.0491744	project
55942a89c63a7fe59b4982d9	1	0.0491744	completed
56cf5b00a759fdadc44e55cd	1	0.0000225	dashlink

```
top_keywords <- lda_gamma %>%
  filter(gamma > 0.9) %>%
  count(topic, keyword, sort = TRUE)

top_keywords %>%
    slice_max(n, n = 10, with_ties = FALSE) %>%
    kable(caption = "Gamma > 0.9 with keywords")
```

Table 15: Gamma > 0.9 with keywords

topic	keyword	n
13	ocean color	4480
13	ocean optics	4480
13	oceans	4480
11	ocean color	1216
11	ocean optics	1216
11	oceans	1216
9	project	926
12	earth science	909
9	completed	834
16	ocean color	768

```
top_keywords %>%
    group_by(topic) %>%
    slice_max(n, n = 5, with_ties = FALSE) %>%
    ungroup %>%
    mutate(keyword = reorder_within(keyword, n, topic)) %>%
    ggplot(aes(n, keyword, fill = as.factor(topic))) +
    geom_col(show.legend = FALSE) +
    labs(title = "Top keywords for each LDA topic",
         x = "Number of documents", y = NULL) +
    scale_y_reordered() +
    facet_wrap( ~ topic, ncol = 4, scales = "free")
```

“By using a combination of network analysis, tf-idf, and topic modeling, we have come to a greater understanding of how datasets are related at NASA. Specifically, we have more information now about how

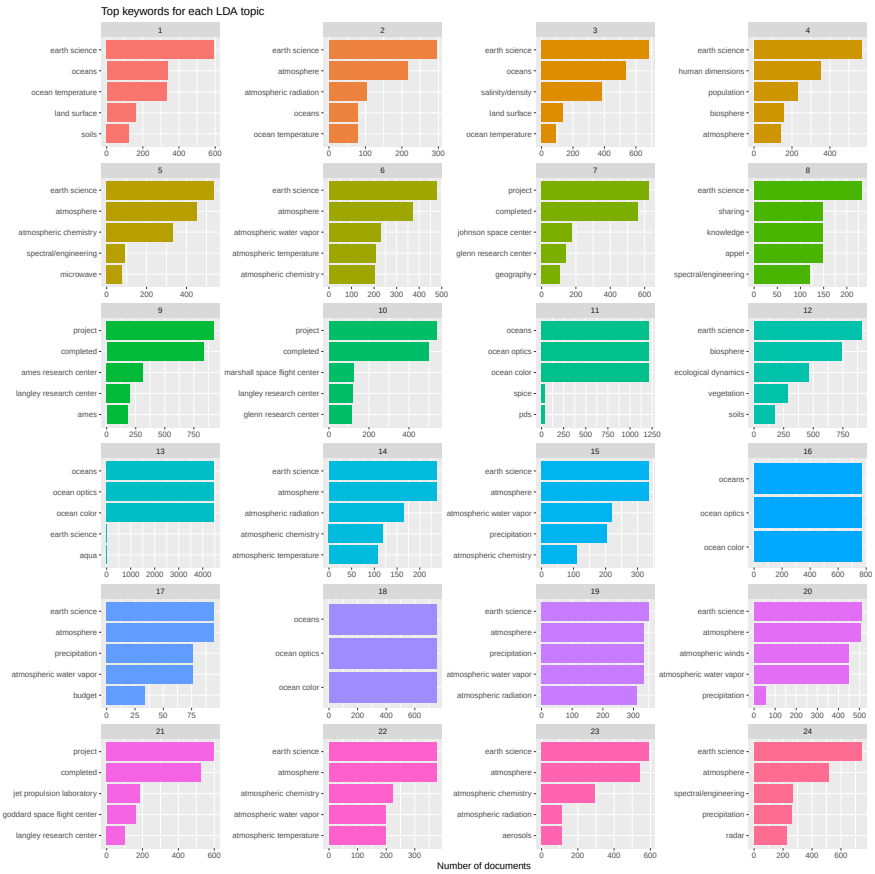

Figure 9: Top keywords in topic modeling of NASA metadata description field texts

keywords are connected to each other and which datasets are likely to be related. The topic model could be used to suggest keywords based on the words in the description field, or the work on the keywords could suggest the most important combination of keywords for certain areas of study."