Literary Data: Some Approaches

Andrew Goldstone
http://www.rci.rutgers.edu/~ag978/litdata

April 16, 2015. Topic modeling (2); being reductive.

```
sidney <- read_mallet_state("mallet-intro/sidney_state.gz")
sidney_lengths_plot <- sidney %>% group_by(doc) %>%
    summarize(length=n()) %>%
    ggplot(aes(length)) +
    geom_bar(binwidth=5, color="gray90") +
    plot_theme()
```

```
sidney_lengths_plot
```

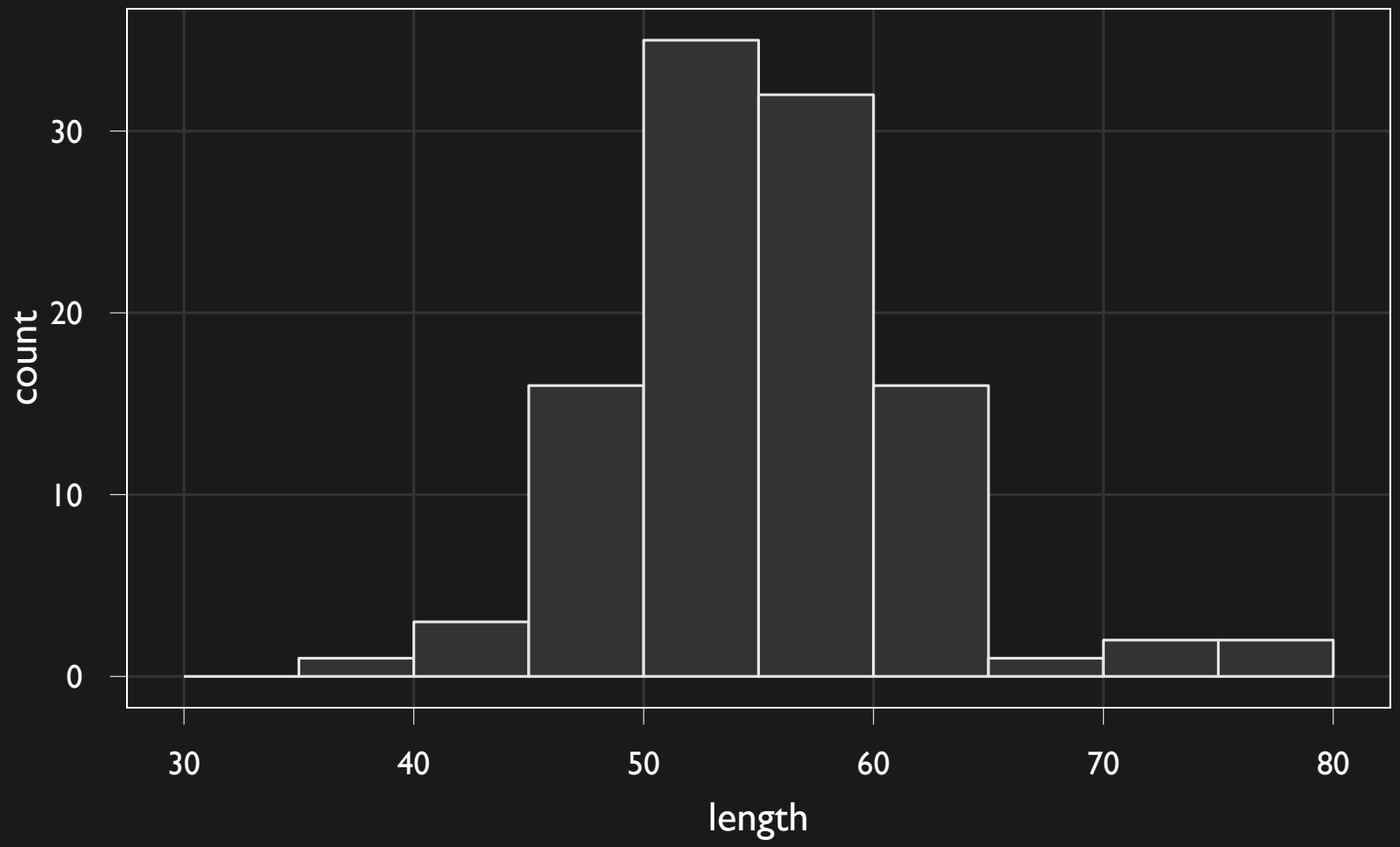

Figure 1: Length distribution of sonnets after stopwording

which words?

```
sidney %>% filter(topic == 1) %>%
    group_by(word) %>%
    summarize(count = n()) %>%
    top_n(8) %>% arrange(desc(count))
```

```
Source: local data frame [11 x 2]

      word count
1    words    15
2   praise    14
3     rich    12
4    write     7
5     fame     6
6   speake     6
7     flow     5
8    farre     4
9  reasons     4
10   skill     4
11   verse     4
```

```r
sidney %>% group_by(doc) %>%
    filter(sum(topic == 1) / n() >= 0.7) %>%
    filter(topic == 1) %>%
    group_by(doc, word) %>%
    summarize(count = n()) %>%
    top_n(3) %>% arrange(desc(count))
```

```
Source: local data frame [9 x 3]
Groups: doc

  doc       word count
1   1     childe     2
2   1 inuentions     2
3   1   pleasure     2
4  35     praise     4
5  35       hope     2
6  35      words     2
7  74     speake     2
8  74      verse     2
9  74        wot     2
```

diagnosis

```
half_words <- sidney %>% mutate(half=doc <= 54) %>%
    group_by(topic, half, word) %>%
    summarize(count=n()) %>%
    filter(count > 1) %>%
    mutate(rank=dense_rank(desc(count))) %>%
    mutate(weight=count / max(count)) %>%
    top_n(2, desc(rank))
half_words %>% filter(topic == 1)
```

```
Source: local data frame [4 x 6]
Groups: topic, half

  topic  half   word count rank    weight
1     1 FALSE praise     7    1 1.0000000
2     1 FALSE  words     7    1 1.0000000
3     1  TRUE   rich    11    1 1.0000000
4     1  TRUE  words     8    2 0.7272727
```

Schmidt-style plot

```
half_plot <- ggplot(half_words,
                    aes(half, weight, label=word)) +
    geom_text(size=2, color="gray90") +
    geom_line(aes(group=word), color="gray90") +
    scale_x_discrete(labels=c("1-54", "55-108")) +
    facet_wrap(~ topic) + plot_theme()
```

```
half_plot
```

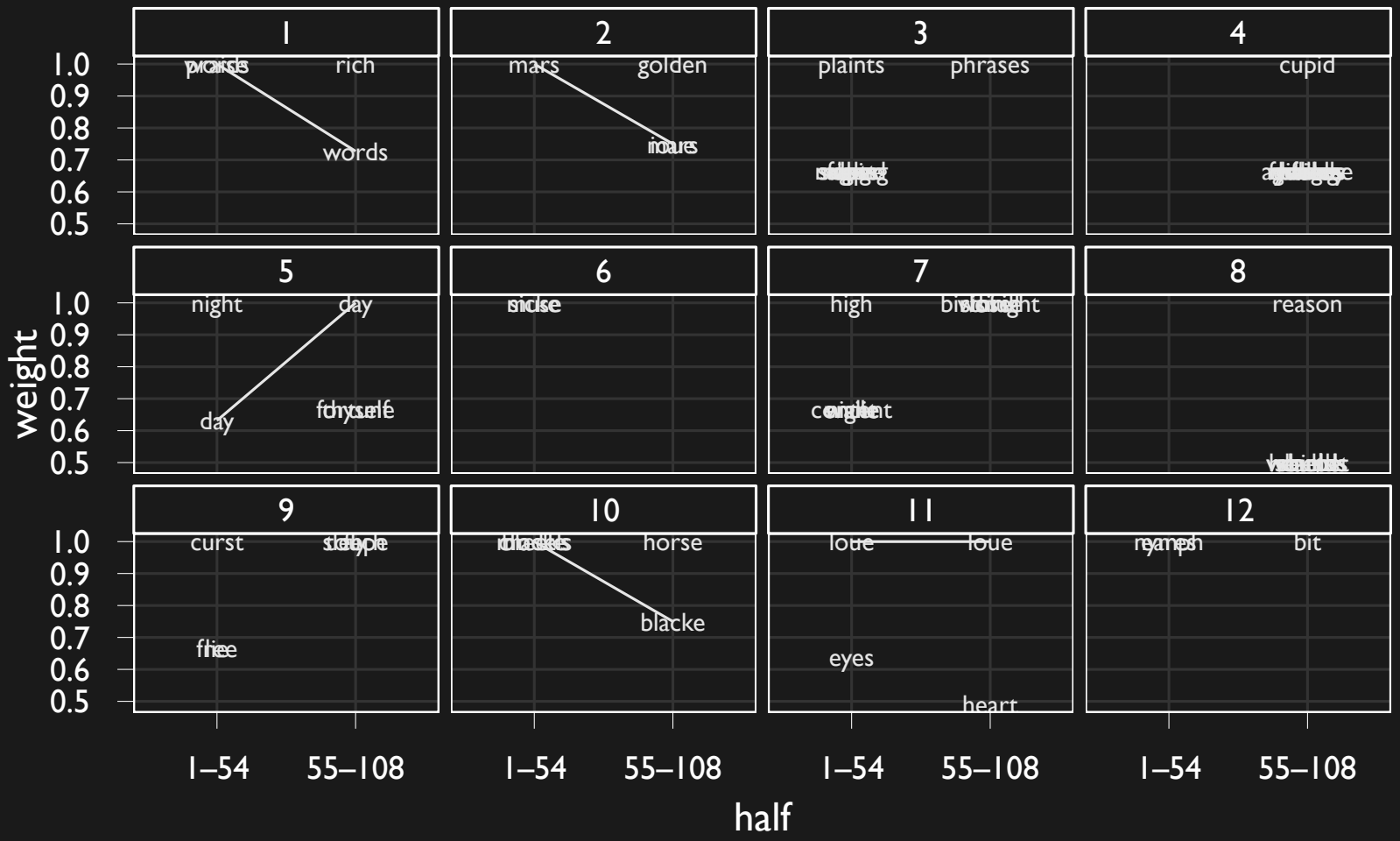

Figure 2: Topic top words can change

a possible shortcut to better featurizing

```
library("SnowballC")
```

```
sidney %>% filter(doc == 1) %>%
    transmute(stemmed=wordStem(word)) %>%
    summarize(str_c(stemmed, collapse="\n")) %>%
    unlist() %>% str_wrap(50) %>% cat()
```

```
lou trueth fayn vers loue show dear som pleasur
pain pleasur read read make knowledg pitti winn
piti grace obtain sought fit word paint blackest
face woe studi inuent fine wit entertain turn
leaue flow fresh fruitful shower sun burnd brain
word halt want inuent stai inuent natur child
fledd step dame studi blow feet seemd stranger
great child speak helpless throw bite trewand pen
beat myself spite fool muse look heart write
```

how about some new data?

in the City. I believe in drawing-rooms for the use of callers, semi-detached villas, nasturtiums in season and dogs with aristocratic, if distant, relatives. I believe that public school-boys, University men (who must not be called Varsity men), and commissioned officers are snobs. I believe that the West End is a gilded haunt of vice. I believe in sober worship once a week, regular payments to the clergy. I believe in temperance, saving an occasional bust, a spree, a night on the tiles (when the wife is in the country), but even then I believe I mustn't go too far. I believe in a bit of fun with a lady now and then, being a dog and all that, so long as there's no harm in it. I believe that I am a gentleman and must be genteel, not too toney though, for it must not be said that I swank. And I believe enough to be saved with. I believe that my wife loves me and that I must reward her by insuring my life; I believe that my sons should be clerks and that my daughters should wait until clerks marry them. I believe that, when I die, the neighbours must approve of my funereal pageant. I believe that I must be honest, that I must not swear in mixed company, that I must visit the upper classes whom I despise. I believe that I am the backbone of England. I am a middle-class man."

* * * *

I think comment on that valuable document would be supererogatory. Mr. George, having exposed the English in this and two other creeds, became so delighted with them that he wrote a novel to give them a proper setting. Let no one object, but rather sympathise; these syntheses become more effective, if anything, when they are surrounded by particular instances. It was a happy idea to imagine the southern French boy mad on England, mad to be English, and to use him for purposes of indirect satire. And the fact that the exigencies of the market caused all this fine insult and just denunciation to fade away into the sickening strains of Rule Britannia I have already lamented. What more is there to be said? Mr. George had the power of writing a great satire—which would be read rapturously by a few people now, and by a few more in fifty years' time—and he has chosen to mix up his satire with a very clever novel of the ordinary kind. The fastidious of the earth will delight in his satire and ignore the rest; the people who subscribe to circulation libraries will wallow in the bosom and adjacent parts of Edith, and possibly not notice the satire. When Mr. George dies he will go to hell and rest among those who were neither for God nor his enemies.

* * * *

The Effort Libre has taken to Suffragettism, and therefore concerns us not. It contains an article on one Nazi, who is dead, and who, had he lived, would have reformed the world.

* * * *

"Some Ideas on George Brandes," by Henri Albert, Beauduin on the Poesie de l'Epoque, an article on Artificial Cold, and Mlle. Henriette Charasson on Dowson—that is the Mercure de France.

RICHARD ALDINGTON.

A Portrait of the Artist as a Young Man.

BY JAMES JOYCE.

"Et ignotas animum dimittit in artes."
OVID, Metamorphoses, VIII., 18.

I.

ONCE upon a time and a very good time it was, there was a moo-cow coming down along the road, and this moo-cow that was down along the road met a nicens little boy named baby tuckoo.

. . .

His father told him that story: his father looked at him through a glass: he had a hairy face.

He was baby tuckoo. The moo-cow came down the road where Betty Byrne lived: she sold lemon platt.

O, the wild rose blossoms
On the little green place.

He sang that song. That was his song.

O, the green wothe botheth.

When you wet the bed, first it is warm then it gets cold. His mother put on the oilsheet. That had the queer smell.

His mother had a nicer smell than his father. She played on the piano the sailor's hornpipe for him to dance. He danced:

Tralala lala,
Tralala tralaladdy,
Tralala lala,
Tralala lala.

Uncle Charles and Dante clapped. They were older than his father and mother, but Uncle Charles was older than Dante.

Dante had two brushes in her press. The brush with the maroon velvet back was for Michael Davitt and the brush with the green velvet back was for Parnell. Dante gave him a cachou every time he brought her a piece of tissue paper.

The Vances lived in number seven. They had a different father and mother. They were Eileen's father and mother. When they were grown up he was going to marry Eileen.

He hid under the table. His mother said:

— O, Stephen will apologise.

Dante said:

— O, if not, the eagles will come and pull out his eyes.—

Pull out his eyes,
Apologise,
Apologise,
Pull out his eyes.

Apologise,
Pull out his eyes,
Pull out his eyes,
Apologise.

* * * *

The wide playgrounds were swarming with boys. All were shouting and the prefects urged them on with strong cries. The evening air was pale and chilly, and after every charge and thud of the footballers the greasy leather orb flew like a heavy bird through the grey light. He kept on the fringe of his line, out of sight of his prefect, out of the reach of the rude feet, feigning to run now and then. He felt his body small and weak amid the throng of players,

- *Egoist* TEI from MJP Lab
- Processed into text files with `XML` functions

```
egoist_texts <- read.table("egoist_texts.tsv", sep="\t",
                           as.is=T, header=T, quote="",
                           comment.char="") %>%
    mutate(issue=str_replace_all(issue,
                                 fixed("."), "_")) %>%
    group_by(issue) %>%
    mutate(item_id=str_c(issue, "_", 1:n())) %>% ungroup()
```

sort out the mess a little

```
issues_meta <- read.table("egoist_meta.tsv", sep="\t",
                          as.is=T, header=T, quote="",
                          comment.char="") %>%
    mutate(issue_id=sprintf("Egoist%03d_%d_%02d",
        seq_along(pubdate), volume, issue))
egoist_meta <- egoist_texts %>%
    select(item_id, issue_id=issue, type) %>%
    inner_join(issues_meta, by="issue_id")
egoist_texts <- egoist_texts %>%
    select(item_id, text) %>%
    inner_join(egoist_meta, by="item_id") %>%
    # prose only, please
    filter(type %in% c("articles", "fiction"))
```

some featurization refinements

- start with the basic one-row-per-feature frame:

```
egoist_features <- egoist_texts %>%
    group_by(item_id) %>%
    do({
        data_frame(feature=featurize(.$text), # well...
                   item_id=.$item_id)
    })
```

- then produce a list of features to *include:*

```
stoplist <- readLines("stoplist_default.txt")
keep_feats <- egoist_features %>%
    group_by(feature) %>%
    summarize(count=n()) %>%
    filter(!(feature %in% stoplist)) %>% # stopword filter
    filter(str_detect(feature, "\\D")) %>% # digits-b-gone
    mutate(rank=min_rank(desc(count))) %>%
    filter(rank < 10000) # rank filter
```

- avoid the sonnet trap by keeping longer items only:

```
egoist_features <- egoist_features %>%
    filter(feature %in% keep_feats$feature) %>%
    filter(n() > 500)
```

- then keep `egoist_meta` for matching items only (convenient later):

```
# assuming we haven't reordered rows, only deleted some!
egoist_meta <- egoist_meta %>%
    filter(item_id %in% egoist_features$item_id)
```

```
dtm <- egoist_features %>%
    group_by(item_id, feature) %>%
    summarize(weight=n()) %>%
    mutate(weight=weight / sum(weight)) %>%
    spread(feature, weight, fill=0) %>%
    select(-item_id)
```

- rows of `dtm` are vectors in `ncol(dtm)` dimensions
- what can we learn from the distribution of points in space?
- (especially:) what can we learn from nearness?

dimensionality reduction (1)

```
top2 <- keep_feats %>%
    filter(rank %in% 1:2) %>%
    arrange(rank) %>%
    select(feature) %>% unlist()
top2_plot <- dtm[ , colnames(dtm) %in% top2] %>%
    cbind(egoist_meta) %>%
    ggplot(aes_string(top2[1], top2[2])) +
        geom_point(aes(color=type)) + plot_theme() +
        scale_color_brewer(type="qual")
```

```
top2_plot
```

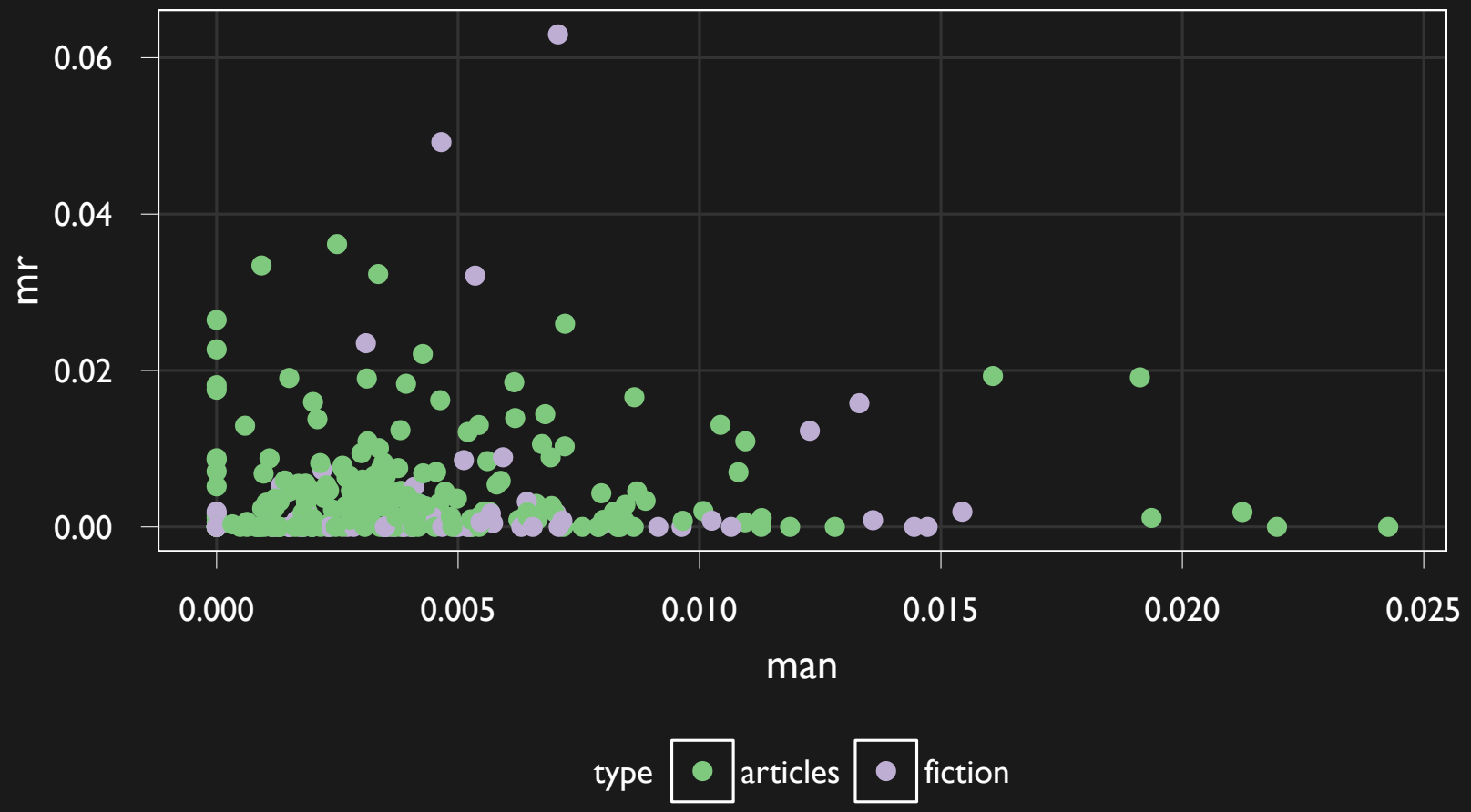

Figure 3: *Egoist* prose in mr-man space

a better angle of vision: PCA

```
set.seed(293) # prcomp can be randomly flipped
dtm_pca <- prcomp(dtm, scale.=T)
```

- PCA: rotate coordinates so that variance of 1st dimension is maximized, variance of 2nd dimension maximizes variance in orthogonal subspace, …
- `dtm_pca$x`: rotated `dtm`
- `dtm_pca$rotation`: "loadings"

dimensionality reduction (2)

```
# extract first two principal components
pca2d <- data.frame(pc1=dtm_pca$x[, 1],
                    pc2=dtm_pca$x[, 2],
                    type=egoist_meta$type,
                    item_id=egoist_meta$item_id)
pca2_plot <- ggplot(pca2d, aes(pc1, pc2, color=type)) +
        geom_point() + plot_theme() +
        scale_color_brewer(type="qual")
```

```
pca2_plot
```

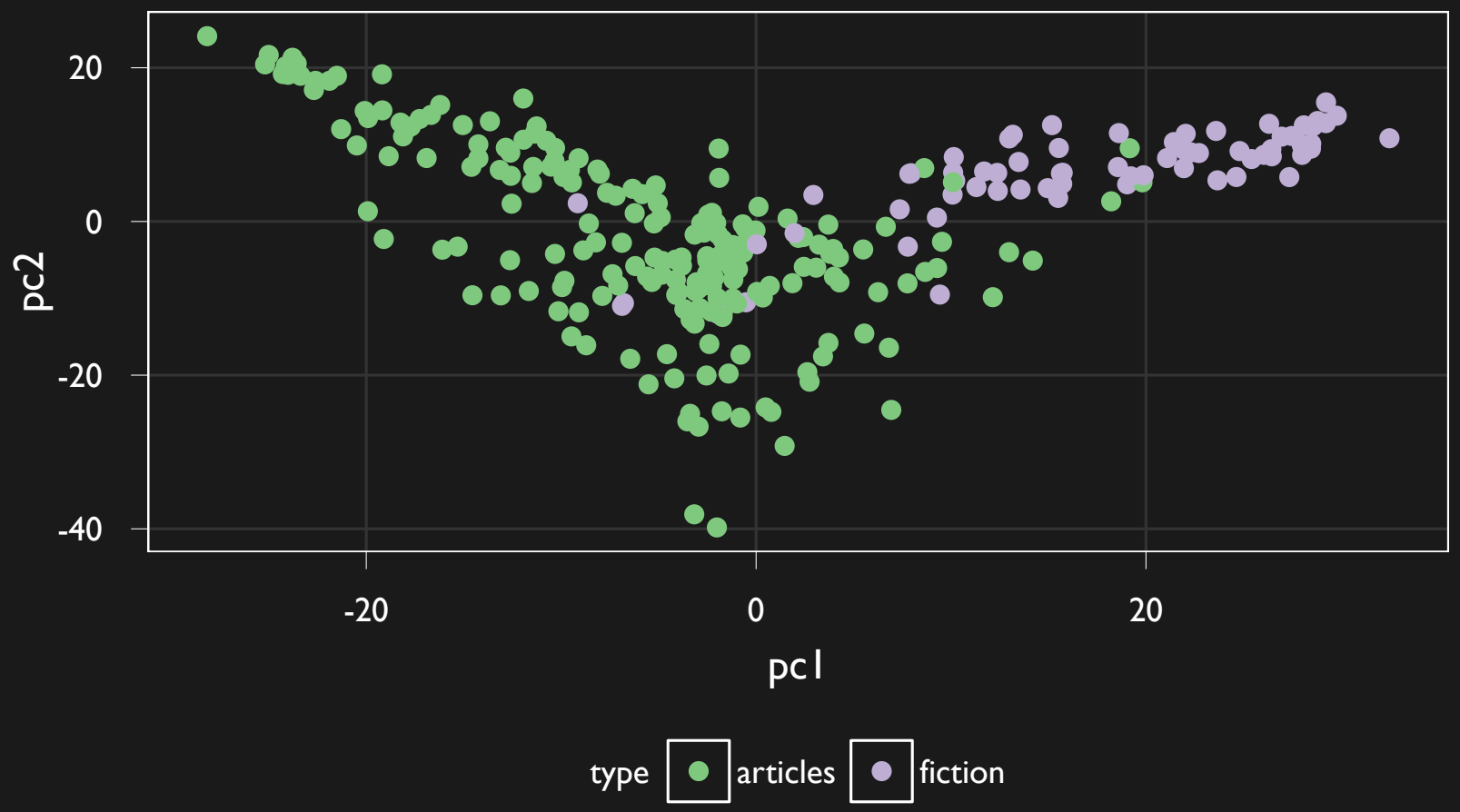

Figure 4: *Egoist* prose, first two principal components

“loadings”

```r
load1 <- dtm_pca$rotation[, 1]
signif(sort(load1, decreasing=T)[1:20], 2)
```

```
   eyes    face   heard   stood    back 
  0.052   0.048   0.044   0.044   0.043 
   head   asked  slowly   night  walked 
  0.042   0.041   0.041   0.039   0.038 
 looked    dark  turned  passed morning 
  0.038   0.038   0.038   0.038   0.037 
  round evening    door    fell     air 
  0.037   0.036   0.036   0.035   0.035 
```

go negative

```r
signif(sort(load1)[1:10], 2)
```

```
     fact   effects     forms     terms
   -0.041    -0.037    -0.035    -0.035
character     means      form     sense
   -0.034    -0.034    -0.034    -0.033
 instance   feature
   -0.033    -0.033
```

```r
load2 <- dtm_pca$rotation[, 2]
signif(sort(load2, decreasing=T)[1:10], 2)
```

```
     entire    specific    organism
      0.035       0.034       0.034
      power     feature        fact
      0.033       0.032       0.032
 constitute constitutes      powers
      0.031       0.031       0.031
      total
      0.031
```

oddballs might be interesting

```
pca2d %>%
    filter(type == "fiction") %>%
    top_n(4, desc(pc1)) %>% arrange(desc(pc1)) %>%
    inner_join(egoist_texts, by="item_id") %>%
    select(text) %>%
    mutate(text=str_sub(text, 1, 54))
```

```
                                                        text
1 UNE FEMME EST UN ÉTAT DE NOTRE AME Peace WHAT is her l
2 A DRAMA Translated from the Russian of A. P. Chekhov b
3 DIALOGUES OF FONTENELLE Translated by Ezra Pound VI CH
4 TARR By Wyndham Lewis PART V A MEGRIM OF HUMOUR CHAPTE
```

dimensionality reduction (3): from LSA to LDA

```
# dumb but easier than alternatives
egoist_pseudotexts <- egoist_features %>%
    group_by(item_id) %>%
    summarize(text=str_c(feature, collapse=" "))
instances <- mallet.import(egoist_pseudotexts$item_id,
                           egoist_pseudotexts$text,
    preserve.case=T, stoplist.file="stoplist_empty.txt",
    token.regexp="\\S+")
```

```
# normally...
write_mallet_instances(instances, "egoist.mallet")
```

model

```
n_topics <- 18
egoist_model_statefile <- "egoist_model_state.gz"
```

```
model <- MalletLDA(n_topics)
model$model$setRandomSeed(as.integer(42))
model$loadDocuments(instances)
model$setAlphaOptimization(20, 50)
model$train(500)
model$maximize(10)
write_mallet_state(model, egoist_model_statefile) # etc.
```

topics and metadata

```
model_state <- read_mallet_state(egoist_model_statefile) %>%
    mutate(item_id=egoist_pseudotexts$item_id[doc]) %>%
    # works here, but don't try it on a huge state
    inner_join(egoist_meta, by="item_id")
```

topic	label
1	called human interest make men nature thing
2	de des est la le les à
3	author book de france french paris war
4	egoist fact made man things time work
5	china country government great literary people times
6	berkeley ego image images language mind thing
7	asked cranly dedalus father man mr stephen
8	book english great mr poems poet poetry
9	expression modern music musical spirit work works
10	day good life long woman women world
11	back dark eyes hand heard night white
12	anastasya back bertha don felt kreisler tarr
13	day french german germans paris soldiers war
14	appearance philosophy real reality relation term terms
15	art artist drama form life mr theatre
16	good means people power state war world
17	form forms life organism power sense world
18	death god life love sin soul yang

topics over time (once more)

```
library("lubridate")  # useful date functions
```

```
sample_topics <- c(10, 11)
sample_labels <- str_c("t", sample_topics)
topic_time_series <- model_state %>%
    mutate(year=year(pubdate)) %>% # lubridate::year
    group_by(year, topic) %>%
    summarize(count=n()) %>%
    mutate(weight=count / sum(count)) %>%
    filter(topic %in% sample_topics) %>%
    ggplot(aes(year, weight)) +
        geom_bar(stat="identity", color="white") +
        facet_wrap(~ topic) + plot_theme()
```

```
topic_time_series
```

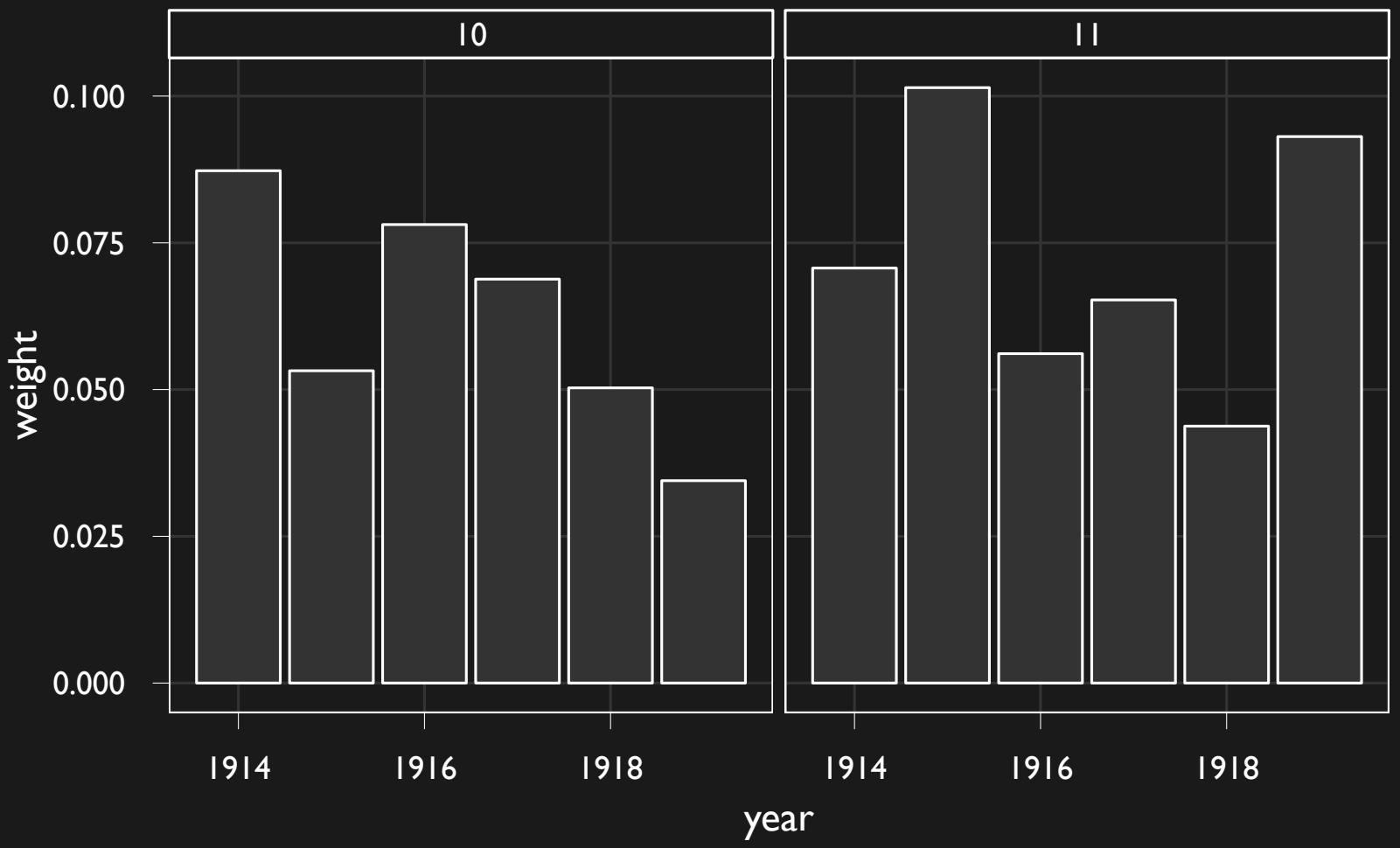

Figure 5: Two *Egoist* topics over time

documents IN SPACE

```
docs_space <- model_state %>% group_by(doc, topic) %>%
    summarize(count=n(), type=first(type)) %>%
    mutate(weight=count / sum(count)) %>%
    mutate(topic=str_c("t", topic)) %>%
    select(-count) %>%
    spread(topic, weight, fill=0) %>%
    ggplot(aes_string(sample_labels[1], sample_labels[2],
                      color="type")) +
        geom_point() + scale_color_brewer(type="qual") +
        plot_theme()
```

```
docs_space
```

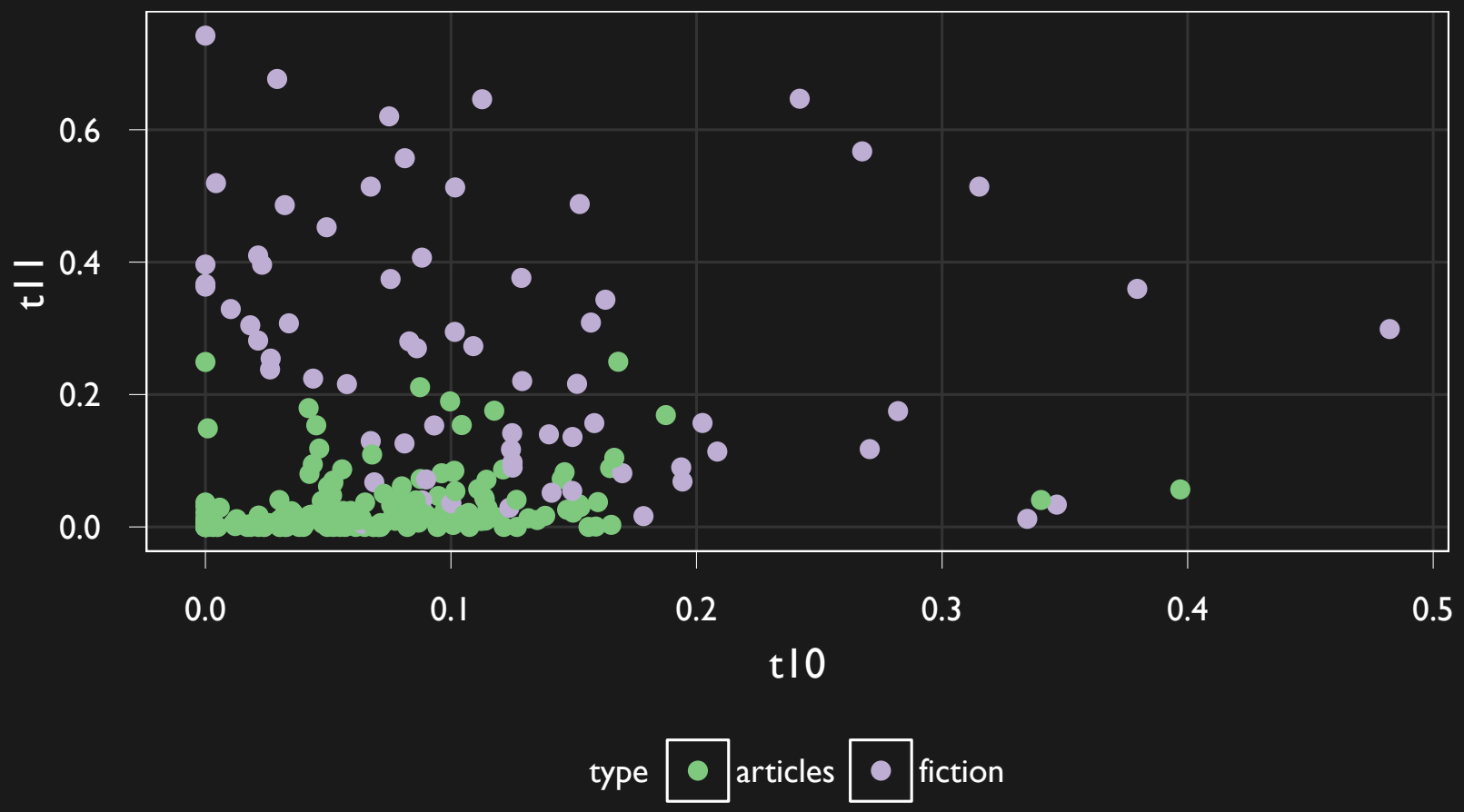

Figure 6: Documents in the space of those same topics