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()
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]

<table>
<thead>
<tr>
<th>word</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>words</td>
<td>15</td>
</tr>
<tr>
<td>praise</td>
<td>14</td>
</tr>
<tr>
<td>rich</td>
<td>12</td>
</tr>
<tr>
<td>write</td>
<td>7</td>
</tr>
<tr>
<td>fame</td>
<td>6</td>
</tr>
<tr>
<td>speake</td>
<td>6</td>
</tr>
<tr>
<td>flow</td>
<td>5</td>
</tr>
<tr>
<td>farre</td>
<td>4</td>
</tr>
<tr>
<td>reasons</td>
<td>4</td>
</tr>
<tr>
<td>skill</td>
<td>4</td>
</tr>
<tr>
<td>verse</td>
<td>4</td>
</tr>
</tbody>
</table>
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
```r
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

<table>
<thead>
<tr>
<th>topic</th>
<th>half</th>
<th>word</th>
<th>count</th>
<th>rank</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FALSE</td>
<td>praise</td>
<td>7</td>
<td>1</td>
<td>1.000000000</td>
</tr>
<tr>
<td>2</td>
<td>FALSE</td>
<td>words</td>
<td>7</td>
<td>1</td>
<td>1.000000000</td>
</tr>
<tr>
<td>3</td>
<td>TRUE</td>
<td>rich</td>
<td>11</td>
<td>1</td>
<td>1.000000000</td>
</tr>
<tr>
<td>4</td>
<td>TRUE</td>
<td>words</td>
<td>8</td>
<td>2</td>
<td>0.727272727</td>
</tr>
</tbody>
</table>
Schmidt-style plot

```r
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()
```
Figure 2: Topic top words can change
a possible shortcut to better featurizing

```r
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 knowldg 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
```
A Portrait of the Artist as a Young Man.

By James Joyce.

"El es un animal doméstico en el más alto sentido." - Virtus, Metamorphoses, VIII, 10.

Once upon a time and a very good time it was, there was a moocow coming along the road, and this moocow, that was down along the road met a nieue little boy named baby tuckoo.

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

He was baby tuckoo. The moocow came down the road where Betty Byrne lived: she sold lemon plott.

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

He sang that song. That was his song.

O, the green-sleeved lady.

When you met the bed, first it was warm then it gets cold. His mother put on the shoonet. That had the queer smell.

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

Tedala lala. Tedala ladladlulu. Tedala lala.

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

Dante had two brushes in his press. The brush with the moocow velvet butt was for Michael Derrit and the brush with the green velvet butt was for Parrell. Dante gave him a card every time he brought her a piece of tissue paper.

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

He sat under the table. His mother said:

— O, Stephen will apolgise.

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.

— O, the wide playgrounds were swarming with boys. All were shouting and the pretzels urged them on with strong cries. The evening air was pale and chilly, and after every change and thrust of the foot-ballers the grey and leather ball flew like a hairy bird through the sky light. He kept on the fringe of his line, out of sight of his protect, out of the reach of the rude foot, begging 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

```r
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") %>%
  filter(type %in% c("articles", "fiction"))
some featurization refinements

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

```r
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*:

```r
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:

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

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

```r
# 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?
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")
Figure 3: *Egoist* prose in mr-man space
a better angle of vision: PCA

```r
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”
# 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")
Figure 4: *Egoist* prose, first two principal components
“loadings”

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

<table>
<thead>
<tr>
<th></th>
<th>eyes</th>
<th>face</th>
<th>heard</th>
<th>stood</th>
<th>back</th>
</tr>
</thead>
<tbody>
<tr>
<td>val</td>
<td>0.052</td>
<td>0.048</td>
<td>0.044</td>
<td>0.044</td>
<td>0.043</td>
</tr>
<tr>
<td></td>
<td>head</td>
<td>asked</td>
<td>slowly</td>
<td>night</td>
<td>walked</td>
</tr>
<tr>
<td>val</td>
<td>0.042</td>
<td>0.041</td>
<td>0.041</td>
<td>0.039</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td>looked</td>
<td>dark</td>
<td>turned</td>
<td>passed</td>
<td>morning</td>
</tr>
<tr>
<td>val</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0.037</td>
</tr>
<tr>
<td></td>
<td>round</td>
<td>evening</td>
<td>door</td>
<td>fell</td>
<td>air</td>
</tr>
<tr>
<td>val</td>
<td>0.037</td>
<td>0.036</td>
<td>0.036</td>
<td>0.035</td>
<td>0.035</td>
</tr>
</tbody>
</table>
go negative

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

<table>
<thead>
<tr>
<th>fact</th>
<th>effects</th>
<th>forms</th>
<th>terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.041</td>
<td>-0.037</td>
<td>-0.035</td>
<td>-0.035</td>
</tr>
<tr>
<td>character</td>
<td>means</td>
<td>form</td>
<td>sense</td>
</tr>
<tr>
<td>-0.034</td>
<td>-0.034</td>
<td>-0.034</td>
<td>-0.033</td>
</tr>
<tr>
<td>instance</td>
<td>feature</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.033</td>
<td>-0.033</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```r
load2 <- dtm_pca$rotation[, 2]
signif(sort(load2, decreasing=T)[1:10], 2)

<table>
<thead>
<tr>
<th>entire</th>
<th>specific</th>
<th>organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.035</td>
<td>0.034</td>
<td>0.034</td>
</tr>
<tr>
<td>power</td>
<td>feature</td>
<td>fact</td>
</tr>
<tr>
<td>0.033</td>
<td>0.032</td>
<td>0.032</td>
</tr>
<tr>
<td>constitute</td>
<td>constitutes</td>
<td>powers</td>
</tr>
<tr>
<td>0.031</td>
<td>0.031</td>
<td>0.031</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.031</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
oddballs might be interesting

```r
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

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