PLSC 597: Modern Measurement

Topic Models*

April 5, 2018

*HT to Justin Grimmer's excellent instructional materials on topic models.

Topics in Text

- "Topics" / "themes" / etc.: What the document is about.
- How do we know?
  - Word meanings...
  - Clustering of words
  - Tone (sometimes)
- Complications / challenges...
  - What's a "topic"?
  - (Key)words can be ambiguous ("tennis" vs. "crane")
  - Documents are often about > one topic

Extracting Topics

Dictionary-based / Supervised methods

- A la sentiment analysis...
- Predetermined “topics” (think: dictionaries of keywords)
- Topic$_i$ $\rightsquigarrow$ whatever topic(s) have (proportionally) the most terms

Unsupervised methods

- Extract topics from the corpus itself
- Intuition: *co-occurrence* of terms in documents
- Useful when (a) we don’t know topics *a priori*, and/or (b) term meaning/usage is complex / nonstandard

Latent Dirichlet Allocation

Intuition:

- Start with N documents $i \in \{1...N\}$ in a corpus
  - Each document i has M_i total words
  - The total of all words in the corpus is V
- Each document comprises a mixture of one or more of k topics
- Each topic comprises a mixture of terms
- We observe documents and terms, but not topics; topics are *latent*
- Goals:
  - Infer the latent topic structure of the corpus
  - Assign documents (probabilistically) to topics
- Process:
  - Assign words to topics
  - Assess Pr(topic | document) and Pr(word | topic)
  - Reassign words to topic
  - Repeat...

LDA, continued

For document i with M_i total words $m = \{1 ... M_i\}$, define $\boldsymbol{X}_i$ as an $M_i \times 1$ vector, where X_{im} maps to the mth word used in the document.

The objective function is:

$$\max[f(\boldsymbol{X}, \boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{\alpha})]$$

where:

- $\boldsymbol{X}$ = as above
- $\boldsymbol{\pi} = N \times K$ matrix with row $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iK})$ = the proportion of a document allocated to each topic
- $\boldsymbol{\Theta} = K \times J$ matrix, with row $\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{Jk})$ = the topics
- $\boldsymbol{\alpha} = K$ element vector = population prior for $\boldsymbol{\pi}$.

LDA: The Math

Assume:

$$\begin{aligned}
\boldsymbol{\theta}_k &\sim \text{Dirichlet}(\mathbf{1}) \\
\alpha_k &\sim \text{Gamma}(\alpha, \beta) \\
\boldsymbol{\pi}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{\tau}_{im} | \boldsymbol{\pi}_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}_i) \\
\mathbf{X}_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\theta}_k)
\end{aligned}$$

Implies:

$$\begin{aligned}
\Pr(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) &\propto \Pr(\boldsymbol{\alpha}) \Pr(\boldsymbol{\pi}|\boldsymbol{\alpha}) \Pr(\boldsymbol{T}|\boldsymbol{\pi}) \Pr(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T}) \\
&\propto \Pr(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\Pr(\boldsymbol{\pi}_i|\boldsymbol{\alpha}) \prod_{m=1}^{M_i} \Pr(\boldsymbol{\tau}_{im}|\boldsymbol{\pi}) \Pr(\mathbf{X}_{im}|\boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
&\propto \Pr(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{\mathbf{X}_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}$$

LDA: Estimation

- Variational EM Approximation
  - Intuition: Approximate the posterior via an arbitrary distribution $q(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha})$
  - Via minimizing the Kullback-Leibler divergence between $q(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha})$ and $\Pr(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha}|\boldsymbol{X})$ ("MAP")
  - Simplifying Assumption: $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}) \equiv q(\boldsymbol{\pi})q(\boldsymbol{\theta})q(\boldsymbol{T})q(\boldsymbol{\alpha})$.
  - Via EM-type approach; see (e.g.) Blei et al. for details

- Bayesian / Gibbs Sampling
  - LDA $\equiv$ Bayesian network of documents in a corpus
  - Fitting via iterative sampling from the posterior
  - Standard MCMC... see (e.g.) the Wikipedia page

LDA: Number of Topics

Choosing K:

- Typically try different values of K
- Choose on the basis of model fit, etc.

Perplexity:

For a (possibly held out) document $\boldsymbol{X}^*_{\text{out}}$

$$\text{Perplexity}_{\text{word}} = \exp\left[-\log \Pr(\boldsymbol{X}^*_{\text{out}} | \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{T}}, \hat{\Theta}, \hat{\boldsymbol{\alpha}}, K)\right]$$

- Perplexity = the geometric mean per-word likelihood
- Declines as $K \rightarrow V$
- Commonly plot perplexity vs. K

Correlated Topic Models ("CTM")

- LDA assumes / requires negative covariance between topics
- The **Logistic Normal Distribution** permits some positive covariance between topics...

$$
\begin{aligned}
\boldsymbol{\theta}_k &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\eta}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim \text{Multivariate Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\boldsymbol{\pi}_i &= \frac{\exp(\boldsymbol{\eta}_i)}{\sum_{k=1}^{K} \exp(\eta_{ik})} \\
\boldsymbol{\tau}_{im} | \boldsymbol{\pi}_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}_i) \\
x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\theta}_k)
\end{aligned}
$$

Structural Topic Models (Roberts et al.)

Intuition: A CTM where topic prevalence (how much of a document is associated with a topic) and/or content (which words go with which topics) varies as a function of document-level metadata predictors.

Some details:

- Predictors enter the MVN via $\boldsymbol{\mu} = \mathbf{Z}_i \gamma$
- No predictors $\equiv$ CTM
- Selection of K is similar to LDA/CTM

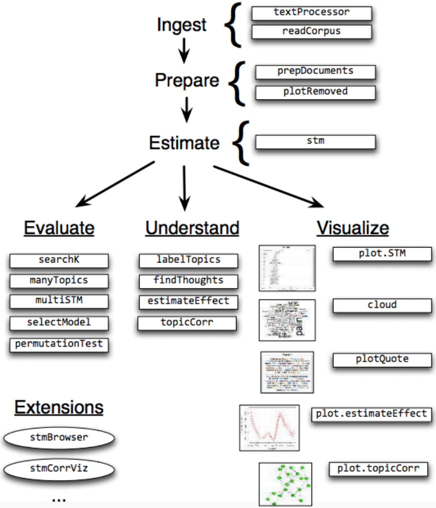
Topic Models in R

- `topicmodels` package
  - Plays well with `tm`
  - LDA and CTM estimation via VEM or Gibbs sampling
  - Some nice graphical tools
  - Is tidy-compatible (see here)
- `stm` package: Structural Topic Models
  - Fits the model in Roberts et al.
  - See the vignette / website
- Others (`quanteda`, `lda`, `text2vec`, `mscstexta4r`)

`topicmodels` Package

- Estimates LDA and CTMs, either via variational approximation (`VEM`, the default) or collapsed Gibbs sampling (`Gibbs`)
- Workhorse functions are `LDA` and `CTM`. Options:
  - `seed` (for replicability)
  - `best` (if `TRUE` (the default), model returns only the model with the highest log-likelihood)
  - Other options related to (VEM or MCMC) optimization...
- Other useful functions:
  - `topics` (extracts most likely topics for each document)
  - `terms` (extracts most likely terms per topic)
  - `posterior` (generates posterior topic probabilities for in- or out-of-sample documents)
  - `perplexity` (calculates model-based perplexity for in- or out-of-sample documents)

stm Package



(from the vignette)

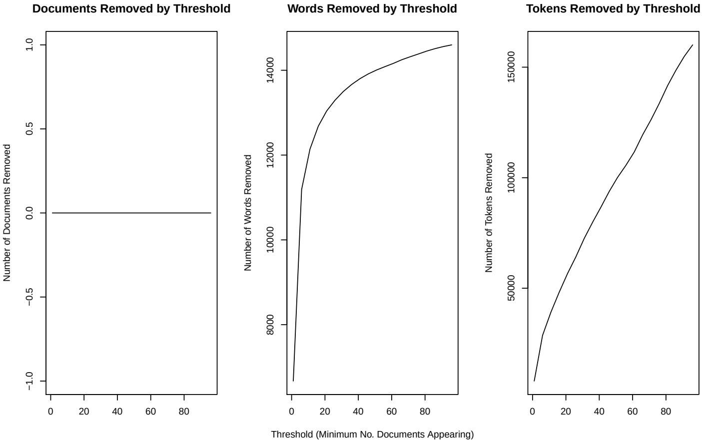
Example, Redux: UNHCR Speeches

- All speeches made by the High Commissioner of the U.N. Refugee Agency, 1970-2016 ($N = 703$)
- Metadata include ID, speaker, title, and date
- Source: https://www.kaggle.com/franciscadias/un-refugee-speech-analysis/

UNHCR Data Prep, Etc.

```
> # Process text (using textProcessor from stm):
> #
> # Note that defaults convert cases, remove stopwords /
> # punctuation / words < 3 characters / extra white space,
> # and stems.
>
> UNHCR <- textProcessor(UN$content, metadata=UN)
Building corpus...
Converting to Lower Case...
Removing punctuation...
Removing stopwords...
Removing numbers...
Stemming...
Creating Output...

> # Create stm corpus. Note that this defaults to dropping
> # words that only appear in one document:
>
> UNCorp <- prepDocuments(UNHCR$documents,UNHCR$vocab,UNHCR$meta)
Removing 6671 of 15742 terms (6671 of 403425 tokens) due to frequency
Your corpus now has 703 documents, 9071 terms and 396754 tokens.>

> # Let's see what happens if we raise that lower threshold:
>
> pdf("Notes and Slides/TopicDocRemoval.pdf",9,6)
> plotRemoved(UNHCR$documents, lower.thresh = seq(1, 100, by = 5))
> dev.off()
```

Effect of `lower.thresh`



Fit a Standard LDA

```
> UN.LDAV.6 <- LDA(UNLDACorp,6,method="VEM"
+                  ,seed=7222009)
> str(UN.LDAV.6)
Formal class 'LDA_VEM' [package "topicmodels"] with 14 slots
  ..@ alpha          : num 0.113
  ..@ call           : language LDA(x = UNLDACorp, k = 6, method = "VEM", seed = 7222009)
  ..@ Dim            : int [1:2] 703 9071
  ..@ control        :Formal class 'LDA_VEMcontrol' [package "topicmodels"] with 13 slots
  .. .. ..@ estimate.alpha: logi TRUE
  .. .. ..@ alpha         : num 8.33
  .. .. ..@ seed          : int 1522857723
  .. .. ..@ verbose       : int 0
  .. .. ..@ prefix        : chr "/var/folders/4p/wkcn3bqs67761813tx05lh9hkvk9km/T//Rtmp8HCEFc/fileba2821eaaa46"
  .. .. ..@ save          : int 0
  .. .. ..@ nstart        : int 1
  .. .. ..@ best          : logi TRUE
  .. .. ..@ keep          : int 0
  .. .. ..@ estimate.beta : logi TRUE
  .. .. ..@ var           :Formal class 'OPTcontrol' [package "topicmodels"] with 2 slots
  .. .. .. .. ..@ iter.max: int 500
  .. .. .. .. ..@ tol     : num 0.000001
  .. .. ..@ em            :Formal class 'OPTcontrol' [package "topicmodels"] with 2 slots
  .. .. .. .. ..@ iter.max: int 1000
  .. .. .. .. ..@ tol     : num 0.0001
  .. .. ..@ initialize    : chr "random"
  ..@ k              : int 6
  ..@ terms          : chr [1:9071] "--camp" "--cuff" "--date" "--job" ...
  ..@ documents      : NULL
  ..@ beta           : num [1:6, 1:9071] -9.34 -225.91 -11.5 -40.89 -26.32 ...
  ..@ gamma          : num [1:703, 1:6] 0.0000786 0.000231 0.0819796 0.0750326 0.0768223 ...
  ..@ wordassignments:List of 5
  .. ..$ i   : int [1:396754] 1 1 1 1 1 1 1 1 1 1 ...
  .. ..$ j   : int [1:396754] 8 48 73 85 107 117 154 174 194 200 ...
  .. ..$ v   : num [1:396754] 6 3 3 6 6 6 6 6 5 6 ...
  .. ..$ nrow: int 703
  .. ..$ ncol: int 9071
  .. ..- attr(*, "class")= chr "simple_triplet_matrix"
  ..@ loglikelihood  : num [1:703] -10851 -3439 -5105 -3402 -4913 ...
  ..@ iter           : int 22
  ..@ logLiks        : num(0)
  ..@ n              : int 906095
```

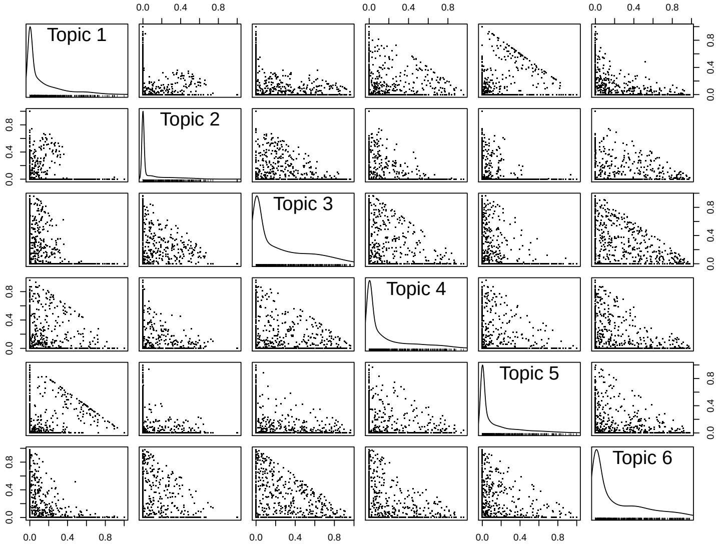
Check Out The Topics

```
> get_terms(UN.LDAV.6,10)

      Topic 1    Topic 2        Topic 3   Topic 4    Topic 5      Topic 6
 [1,] "refuge"   "humanitarian" "refuge"  "unhcr"    "refuge"     "refuge"
 [2,] "countri"  "return"       "unhcr"   "refuge"   "problem"    "intern"
 [3,] "programm" "secur"        "will"    "programm" "work"       "countri"
 [4,] "assist"   "conflict"     "protect" "will"     "nation"     "protect"
 [5,] "govern"   "peac"         "need"    "committe" "commission" "right"
 [6,] "offic"    "displac"      "intern"  "year"     "offic"      "human"
 [7,] "will"     "intern"       "peopl"   "offic"    "high"       "asylum"
 [8,] "problem"  "polit"        "displac" "assist"   "year"       "peopl"
 [9,] "also"     "bosnia"       "countri" "govern"   "unit"       "state"
[10,] "camp"     "forc"         "year"    "continu"  "will"       "nation"
```

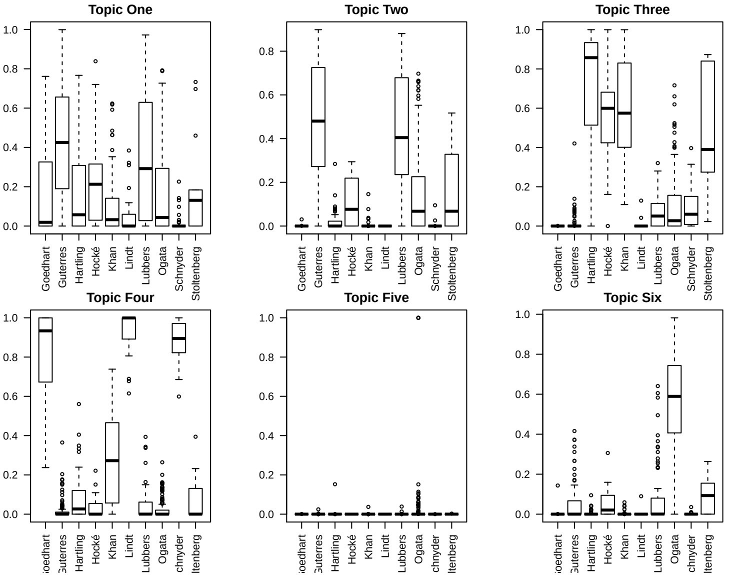
Estimated Pr(Topic | Document)

```
> # Generate posterior probabilities of the topics
> # for each document and the terms for each topic:
>
> V.6.Post <- posterior(UN.LDAV.6)
> cor(V.6.Post$topics)
       1      2      3      4     5      6
1  1.000 -0.138 -0.370 -0.055  0.22 -0.411
2 -0.138  1.000 -0.089 -0.207 -0.24 -0.076
3 -0.370 -0.089  1.000 -0.227 -0.37 -0.189
4 -0.055 -0.207 -0.227  1.000 -0.18 -0.307
5  0.222 -0.245 -0.369 -0.182  1.00 -0.260
6 -0.411 -0.076 -0.189 -0.307 -0.26  1.000
```

Posterior Topic Probabilities

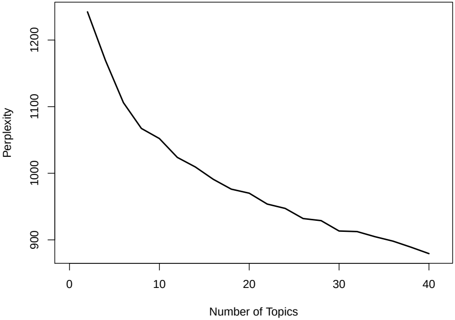


Topic Probabilities by Author



Selecting the Number of Topics

```
MaxTopics <- 40
Seq <- seq(2,MaxTopics,by=2)
Perps <- numeric(MaxTopics/2)
for (i in Seq) {
  foo <- LDA(UNLDACorp,i,method="VEM",
             seed=7222009)
  Perps[i/2] <- perplexity(foo)
}
```



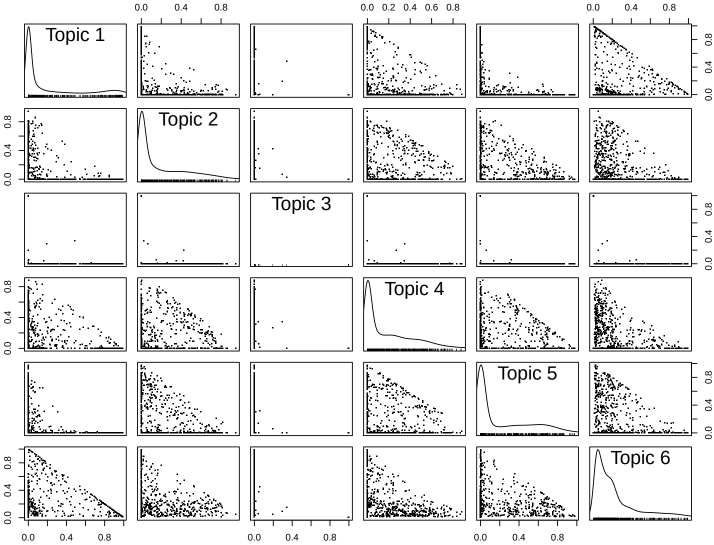
Correlated Topic Model

```
> # Basic CTM:

UN.CTMV.6 <- CTM(UNLDACorp,6,method="VEM",
+                 seed=7222009)

> # Check out topics:
>
> terms(UN.CTMV.6,10)
      Topic 1      Topic 2   Topic 3     Topic 4   Topic 5        Topic 6
 [1,] "refuge"     "refuge"  "los"       "refuge"  "refuge"       "refuge"
 [2,] "problem"    "will"    "que"       "intern"  "humanitarian" "unhcr"
 [3,] "countri"    "unhcr"   "las"       "protect" "intern"       "programm"
 [4,] "offic"      "protect" "refugiado" "countri" "return"       "assist"
 [5,] "will"       "peopl"   "para"      "right"   "displac"      "will"
 [6,] "govern"     "need"    "por"       "human"   "conflict"     "govern"
 [7,] "high"       "year"    "una"       "asylum"  "polit"        "countri"
 [8,] "year"       "also"    "del"       "state"   "peac"         "year"
 [9,] "nation"     "work"    "mas"       "peopl"   "secur"        "continu"
[10,] "commission" "develop" "con"       "nation"  "must"         "need"
```

CTM Posteriors

```
> CTMV.6.Post <- posterior(UN.CTMV.6)

> cor(CTMV.6.Post$topics)
       1       2      3       4      5      6
1  1.000 -0.3899 -0.044 -0.3087 -0.446 -0.128
2 -0.390  1.0000 -0.056  0.0013 -0.230 -0.190
3 -0.044 -0.0563  1.000 -0.0605 -0.067 -0.077
4 -0.309  0.0013 -0.061  1.0000 -0.138 -0.341
5 -0.446 -0.2303 -0.067 -0.1377  1.000 -0.221
6 -0.128 -0.1898 -0.077 -0.3407 -0.221  1.000
```

Posterior CTM Topic Probabilities



Structural Topic Model

```
> STM.6 <- stm(UNCorp$documents,UNCorp$vocab,6,
             prevalence=~Year+Author,
             data=UNCorp$meta)
```

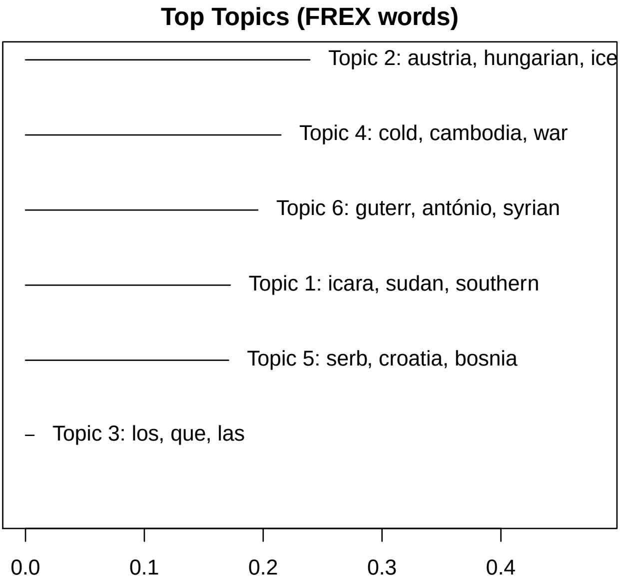
Structural Topic Model

```
> labelTopics(STM.6)
Topic 1 Top Words:
   Highest Prob: refuge, unhcr, programm, assist, will, govern, year
   FREX: icara, sudan, southern, ethiopia, undp, african, execut
   Lift: -site, agronomist, asmara, chi, delicaci, despis, development-rel
   Score: refuge, unhcr, programm, year, assist, chairman, countri
Topic 2 Top Words:
   Highest Prob: refuge, problem, countri, offic, will, govern, high
   FREX: austria, hungarian, icem, iro, connexion, unref, handicap
   Lift: --spot, -employ, -privileg, -root, aac, aacinf, abel
   Score: refuge, countri, problem, offic, year, icem, programm
Topic 3 Top Words:
   Highest Prob: los, que, las, refugiado, para, por, una
   FREX: los, que, las, refugiado, por, del, ms
   Lift: cmo, abandonar, acceso, acontecimiento, actividad, actualment, adecuada
   Score: que, refugiado, las, por, los, ms, como
Topic 4 Top Words:
   Highest Prob: refuge, intern, right, humanitarian, human, protect, countri
   FREX: cold, cambodia, war, violat, right, environment, persecut
   Lift: band-aid, beam, bi-polar, break-, condon, conflagr, creep
   Score: refuge, intern, right, protect, human, countri, war
Topic 5 Top Words:
   Highest Prob: refuge, return, humanitarian, will, displac, unhcr, secur
   FREX: serb, croatia, bosnia, bosnian, herzegovina, osc, sarajevo
   Lift: abkhaz, amunategui, banyamuleng, bijeljina, bosanski, boycott, brahimi
   Score: refuge, bosnia, secur, kosovo, croatia, serb, will
Topic 6 Top Words:
   Highest Prob: refuge, protect, need, will, unhcr, countri, peopl
   FREX: guterr, antnio, syrian, stateless, syria, afghan, reform
   Lift: guterr, -hous, -point, -rs, abidjan, abraham, abu
   Score: refuge, protect, need, countri, intern, unhcr, year
```

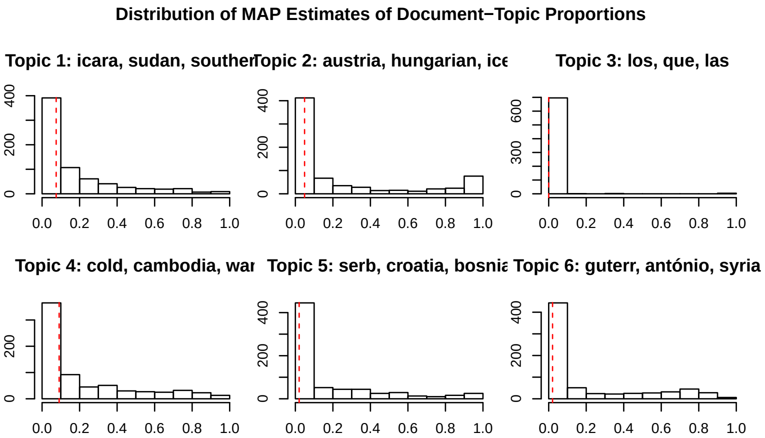
findThoughts: Representative Document(s)

```
> findThoughts(STM.6, UN$content, topic=5)

 Topic 5:

Statement by Mrs. Sadako Ogata, United Nations High Commissioner for Refugees, to the Euro-Atlantic
Partnership Council, Brussels, 18 November 1998 Statements by High Commissioner,
18 November 1998

Deputy Secretary-General Balanzino, Your Excellencies, Ladies and Gentlemen,

I would like to thank you, Deputy Secretary-General, and members of the Council, for this timely
opportunity to address you today. This is a particularly crucial period. The international community
is focusing efforts on addressing the Kosovo crisis within the Federal Republic of Yugoslavia,
while remaining committed to achieving objectives set by the Dayton Peace Accords in
Bosnia and Herzegovina.

Let me start with the crisis in the Yugoslav province of Kosovo...
```

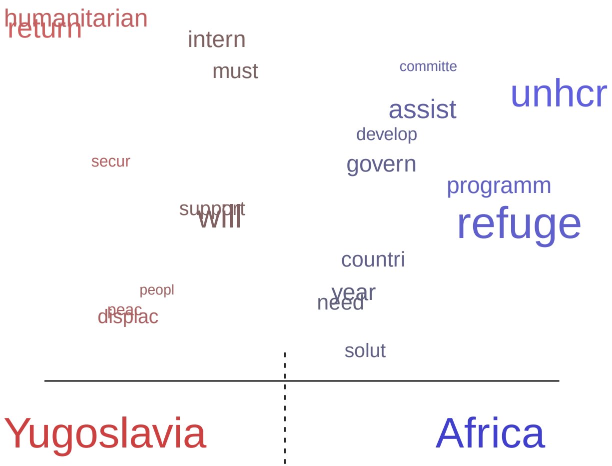
STM Plots: Summary



STM Plots: MAP Histograms



STM Plots: Labels



STM Plots: WORD CLOUDS LOL

Covariate Effects

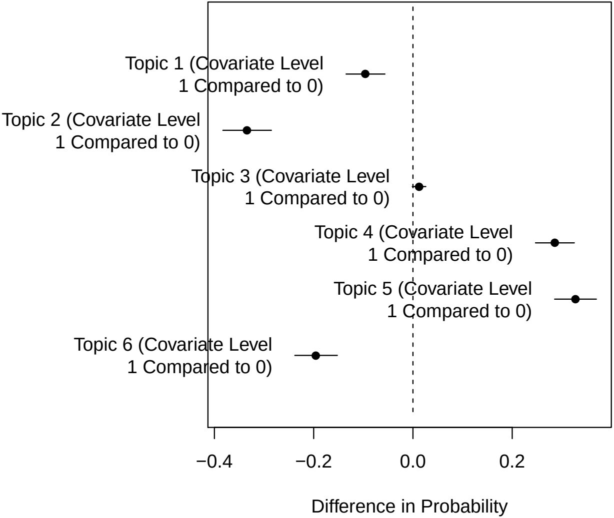
```
> UN$Ogata <- ifelse(UN$Author=="Ogata",1,0)
> STM.Ogata<- estimateEffect(1:6~Ogata,STM.6,metadata=UN)
> summary(STM.Ogata)

Call:
estimateEffect(formula = 1:6 ~ Ogata, stmobj = STM.6, metadata = UN)

Topic 1:

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   0.2111     0.0123   17.14   < 2e-16 ***
Ogata        -0.0961     0.0199   -4.83 0.0000017 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Topic 2:

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3675     0.0153    24.0   <2e-16 ***
Ogata        -0.3336     0.0239   -13.9   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Topic 3:

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.00497    0.00436    1.14    0.255
Ogata        0.01257    0.00689    1.83    0.068 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Covariate Effects (continued)

```
.
.
.
Topic 4:

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1047     0.0126    8.31  4.9e-16 ***
Ogata         0.2856     0.0202   14.14  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Topic 5:

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0496     0.0116    4.27 0.000022 ***
Ogata         0.3273     0.0215   15.26  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Topic 6:

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2621     0.0138   19.05   <2e-16 ***
Ogata        -0.1955     0.0219   -8.94   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

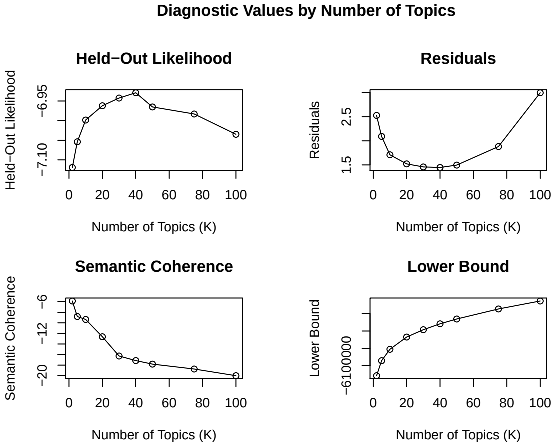
STM: More Covariate Effects

```
> plot(STM.Ogata,"Ogata",method="difference",
     cov.value1=0,cov.value2=1)
```



Selection of K

```
STM.searchK <- searchK(UNCorp$documents,UNCorp$vocab,
        c(2,5,seq(10,50,by=10),75,100))
```



Things to Think About: How Many Topics?

From the `stm` documentation:

> *"The most important user input in parametric topic models is the number of topics. There is no right answer to the appropriate number of topics.* **More topics will give more fine-grained representations of the data at the potential cost of being less precisely estimated.** *The number must be at least 2 which is equivalent to a unidimensional scaling model. For short corpora focused on very specific subject matter (such as survey experiments) 3-10 topics is a useful starting range. For small corpora (a few hundred to a few thousand) 5-50 topics is a good place to start. Beyond these rough guidelines it is application specific. Previous applications in political science with medium sized corpora (10k to 100k documents) have found 60-100 topics to work well. For larger corpora 100 topics is a useful default size. Of course, your mileage may vary." (emphasis added)*

More Things...

- STM integrates measurement and model fitting...
- For STM: Covariates $\rightarrow$ topic *prevalence* or topical *content*?
  - MC region $\rightarrow$ (e.g.) more likely to discuss agriculture, less mass transit
  - MC ideology $\rightarrow$ talk about foreign policy as "humanitarian" vs. "nuclear threat"
- As always, validation is useful...