Comprehensive Exam: Computational Portion

Computational Science and Informatics PhD

Tommy Jones, 2020-04-17

Background

Review of questions

1. **Exploratory visualization**
   Provide a plot showing the top 20 words in the corpus

2. **Clustering**
   Re-weight term frequencies by TF-IDF
   Calculate cosine similarity between documents
   Cluster using agglomerative hierarchical clustering
   Choose number of clusters using “elbow” method

3. **Topic Modeling**
   Choose number of topics using “elbow” method
   Fit 3 models (chains), with stopping/starting chains
   Compare models across chains (and perhaps within chain)

Tech stack

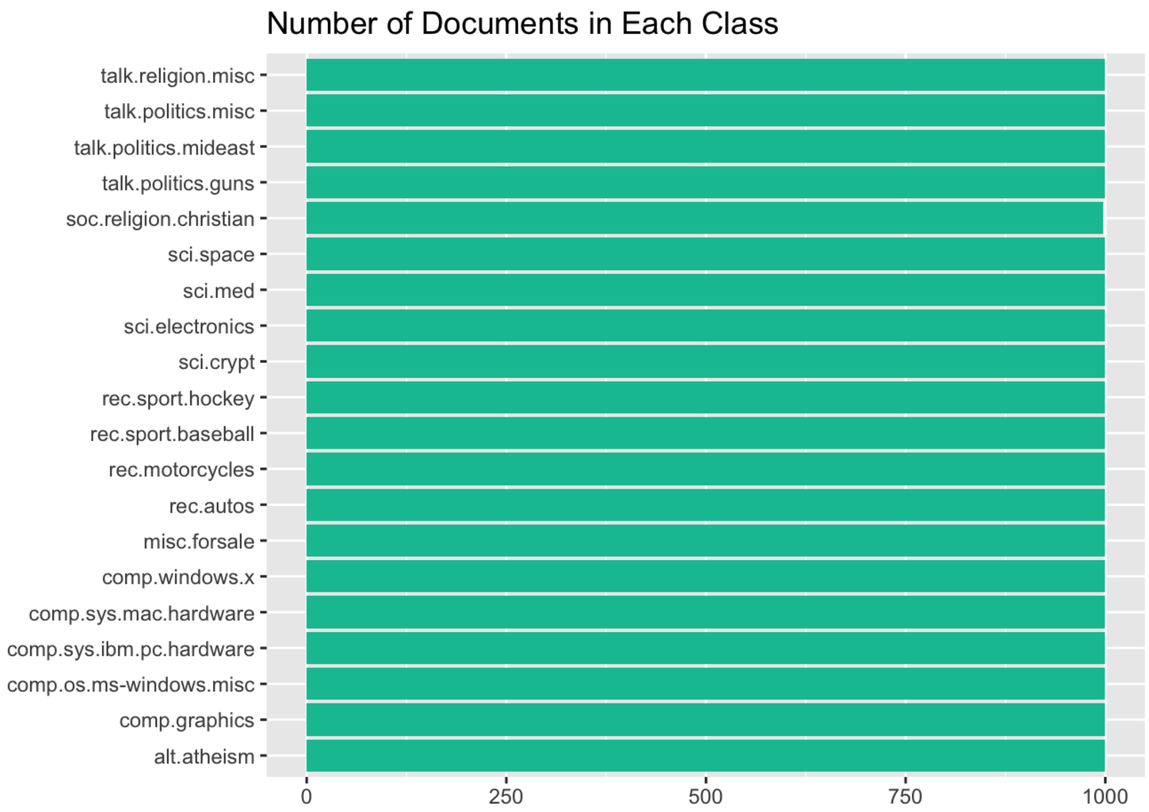
- Programming languages:
  R with extensions in C++

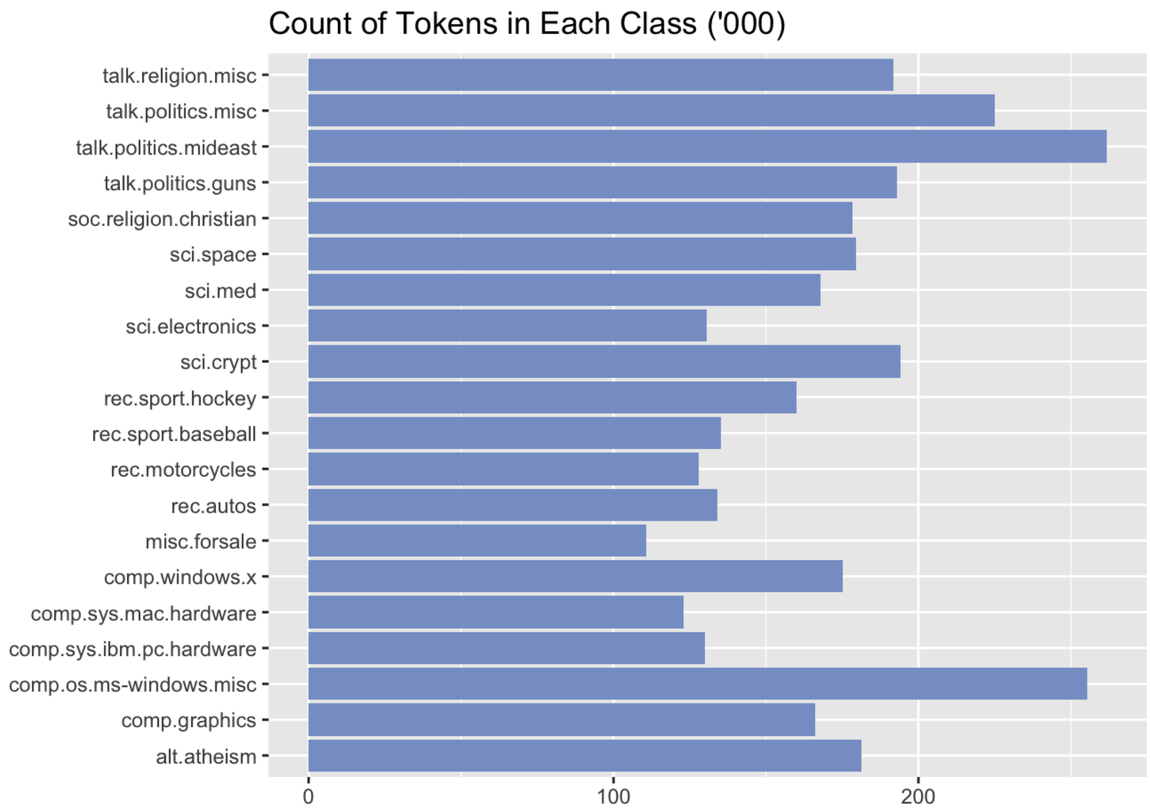
- Packages I wrote:
  textmineR - published on CRAN since 2015
  tidylda - development in process

- Packages I did not write but are worth mentioning:
  stats, cluster, ggplot2, dplyr, coda

- Hardware:
  Late 2019 13” Macbook Pro
  2.8 GHz Quad-Core Intel Core i7
  16 GB RAM

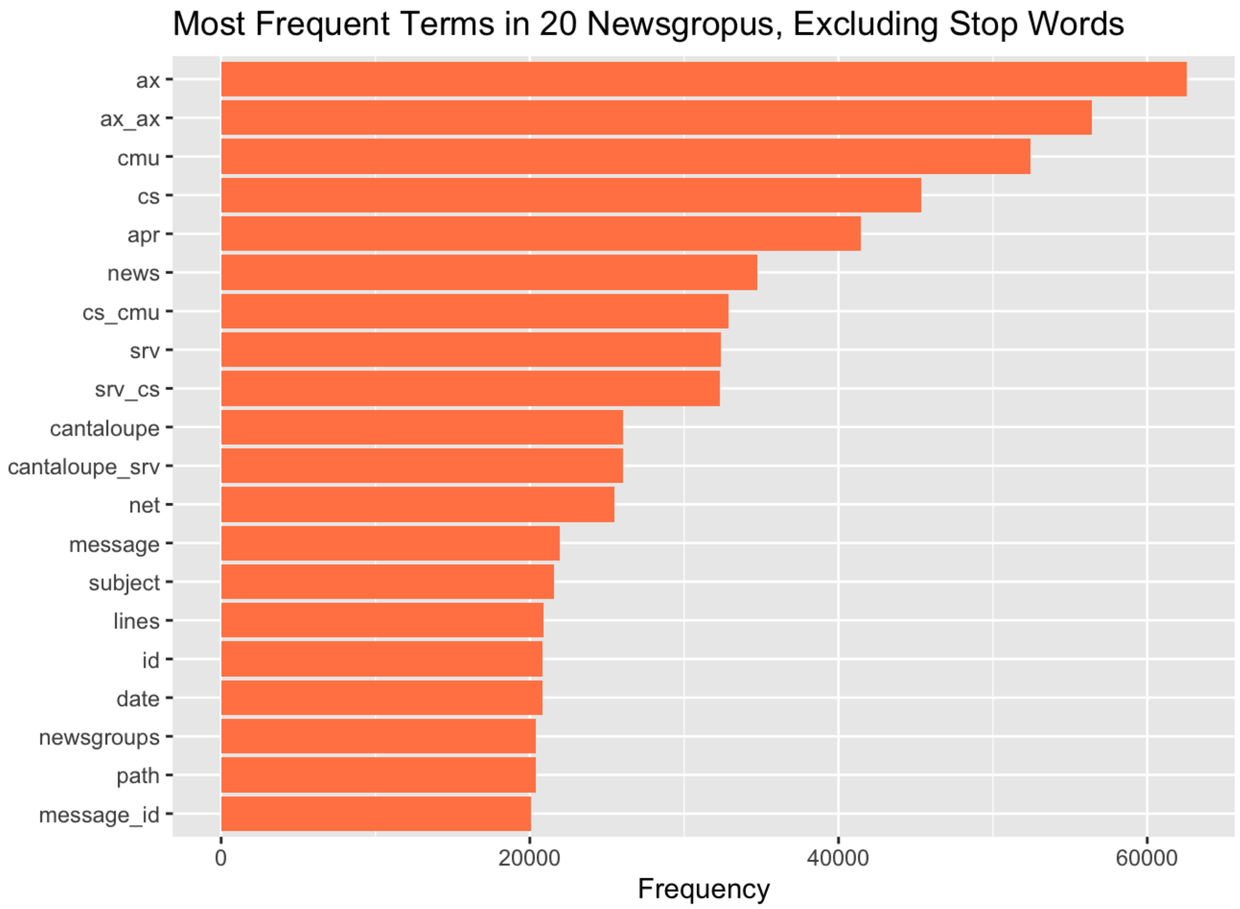
The “20 Newsgroups” Dataset





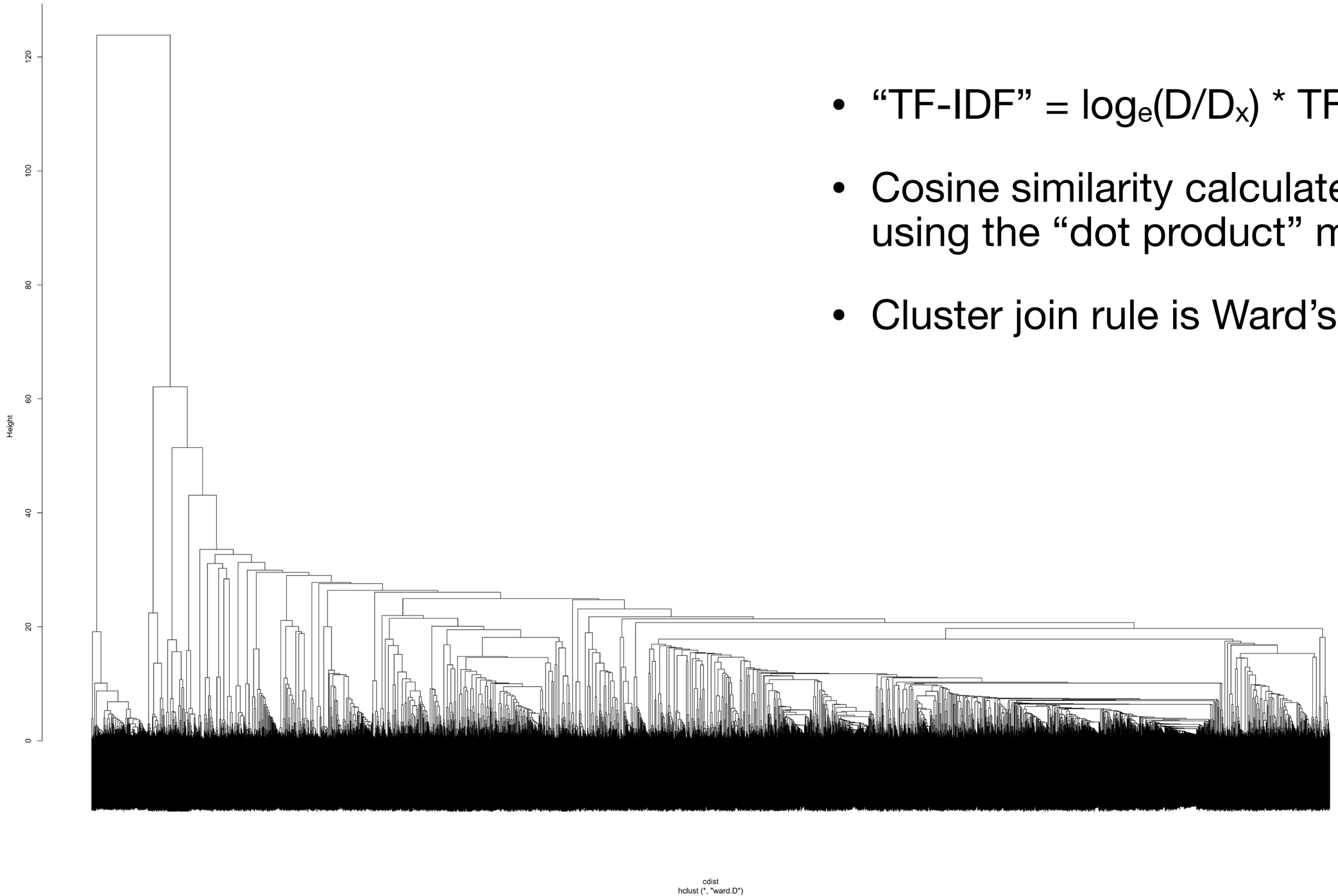
The 20 Newsgroups Dataset

…and data curation decisions



- “token” = unigrams and bigrams
- Non alphabetic characters replaced with a space
- Remove tokens appearing…
  - In all documents
  - In fewer than 5 documents
  - Fewer than 10 times overall
  - In a stop word list of ~ 600 tokens

"Path: cantaloupe.srv.cs.cmu.edu!crabapple.srv.cs.cmu.edu!fs7.ece.cmu.edu!europa.eng.gtefsd.com!emory!sol.ctr.columbia.ed
u!destroyer!newsrelay.iastate.edu!iscsvax.uni.edu!sunfish!pwiseman\nNewsgroups: comp.os.ms-windows.misc\nSubject: roman.bm
p 12/14 \nMessage-ID: <C65Gv3.1uz@sunfish.usd.edu>\nFrom: pwiseman@salmon.usd.edu (Cliff)\nDate: Tue, 27 Apr 1993 16:09:01
 GMT\nReply-To: pwiseman@salmon.usd.edu (Cliff)\nSender: pwiseman@sunfish.usd.edu (Cliff)\nDistribution: usa\nOrganizatio
n: University of South Dakota\nLines: 956\n------------ Part 12 of 14 ------------\nMAX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'A
X>'AX>'AX>'AX>'AX>'AX>'\nMAX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'\nMAX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'A
X>'AX>'AX>'AX>'AX>'AX>'AX>'\nMAX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'\nMAX>',<3$9@L+I:5'W]_?W]_?W]_?>
GIZ*BJ[N[M>7EY>7EY>7@,#`P,#`P->\nM`P,#`P->*BHJ*KN[N[M>`P,#`P->NRHJ*BIZ1PMF,8>'AX>'AX>'AX>'AX>'\nMAX>'AX>'AX>'AX>'AX>'AX>'A
X>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'\nMAX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'\nMAX>'AX>'AX>'AX>'AX>'A
X>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'\nMAX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'\nMAX>'AX>'AX>'AX>'A
X>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'\nMAX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'\nMAX>'AX>'AX>'A
X>'AX>'AX>'AX>'AS'$Q,1F\"PL+I:6EI:6EI:5'1T='1T??\nMW]_?>M_?W]_?W]_?W]_?WT='1Z4+\"V;$,8>'AX>'AX>'AX>'AX>'AX>'AX>'\nMAX>'AX
>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'\nMAX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'\nMAX
>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'\nMAX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX
>'\nMAX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'\nMAX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX
>'AX>'\nMAX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'\nMAX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX
>'AX>'AX>'\nMAX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'\nMAX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX>'AX



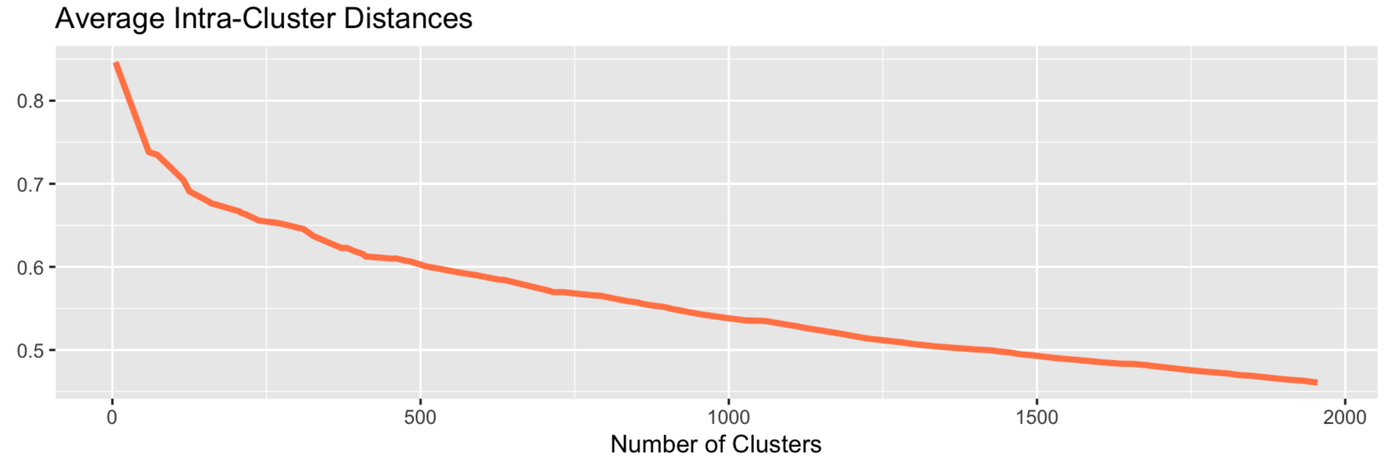
- “TF-IDF” = $\log_e(D/D_x) * TF(x)$
- Cosine similarity calculated using the “dot product” method
- Cluster join rule is Ward’s method

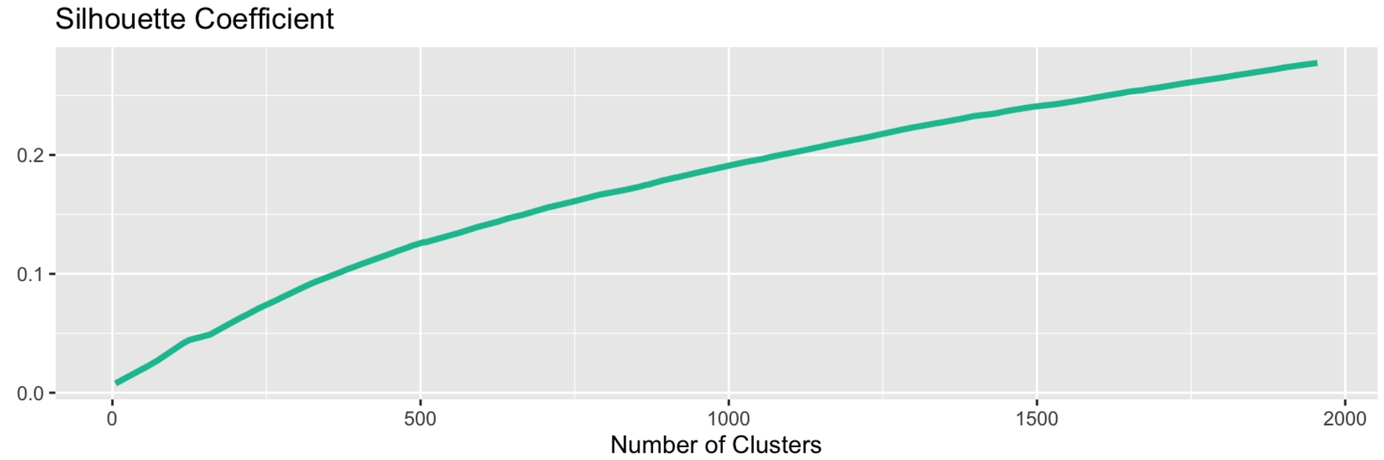
Choosing the number of topics for hierarchical clustering

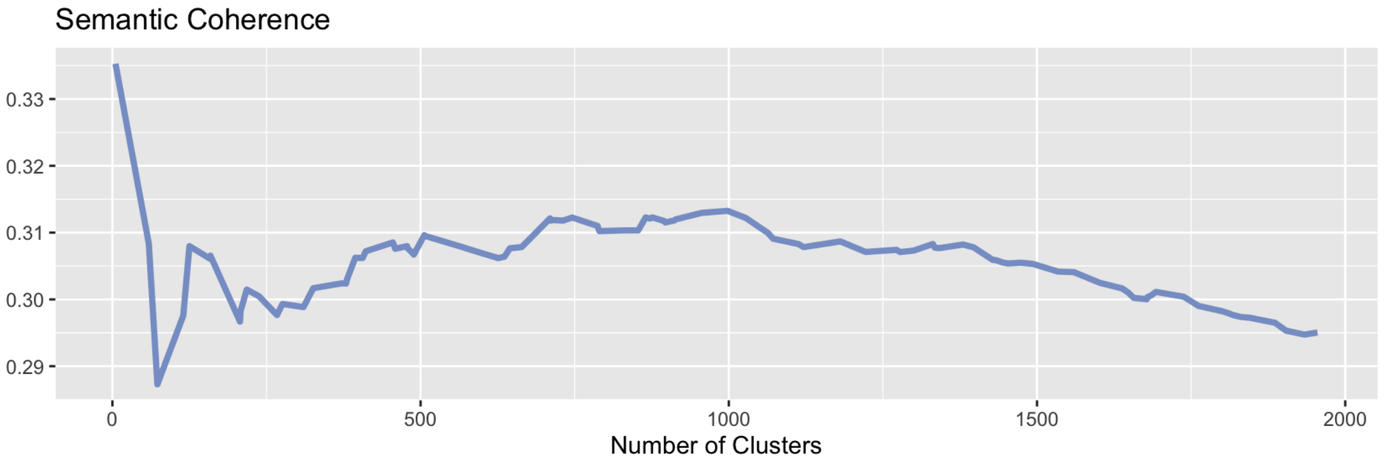
Average total inverse cosine similarity of each cluster

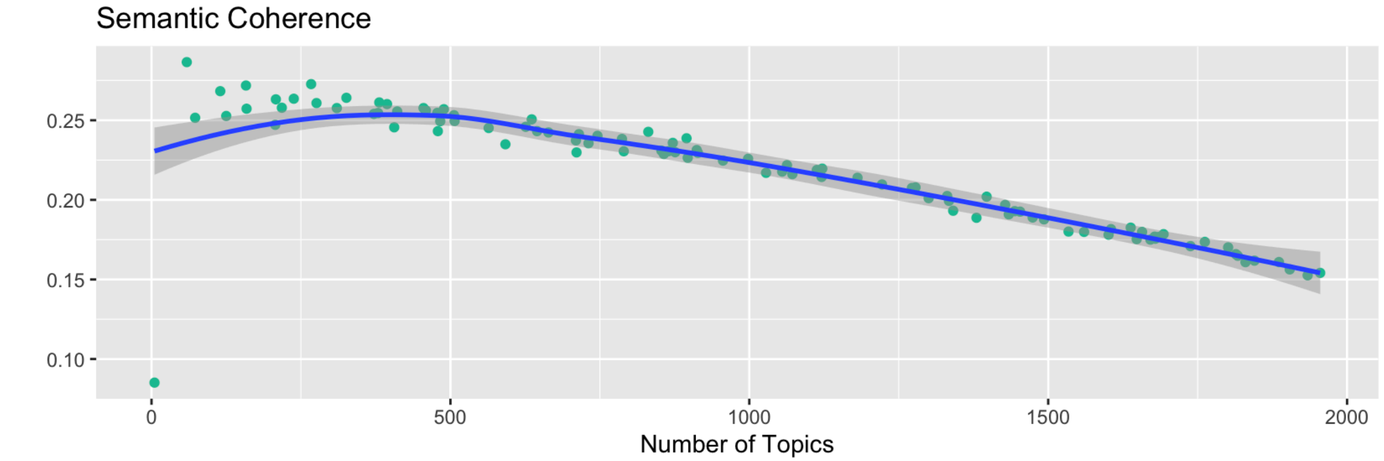
Silhouette averaged across all documents

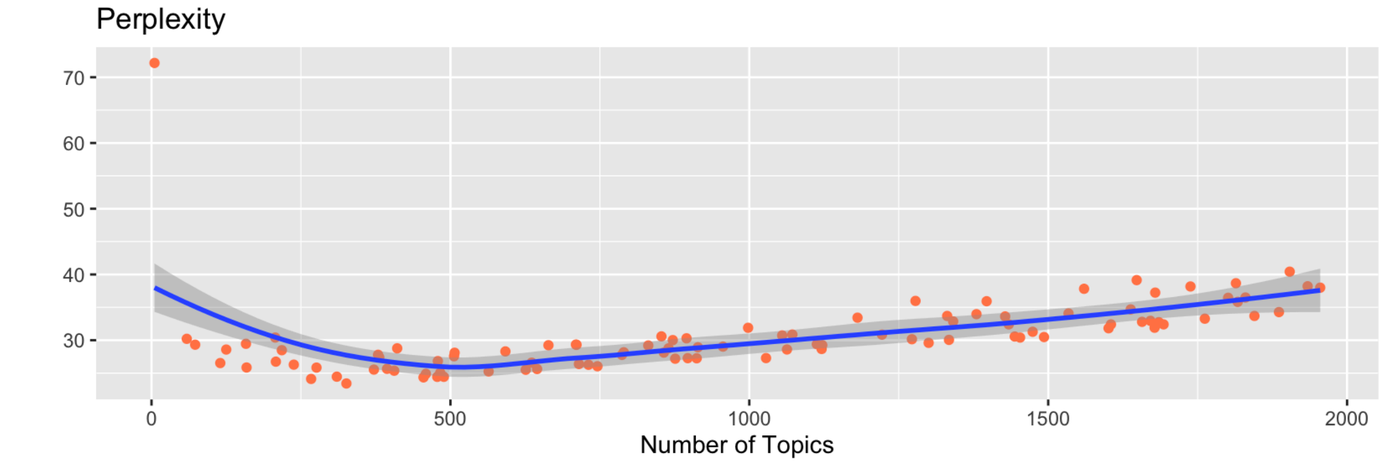
What if document clustering is just topic modeling where each document has only one topic?

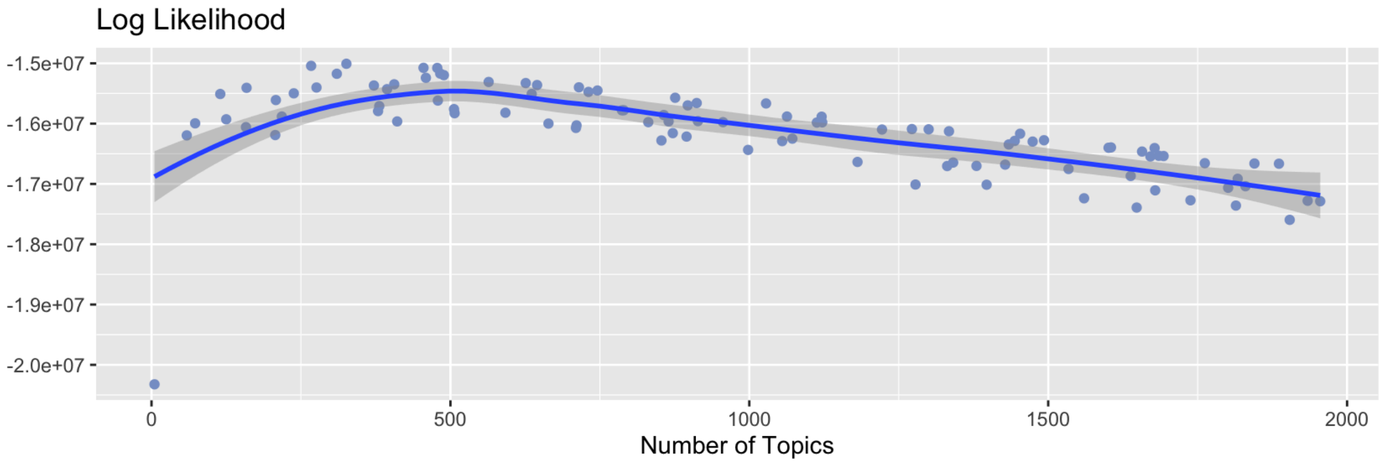










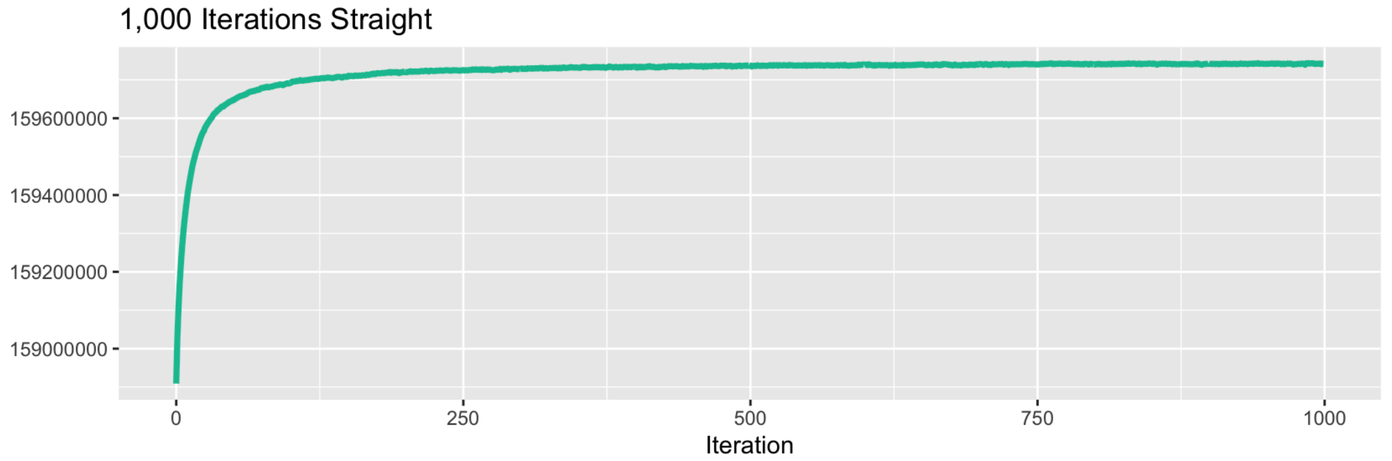


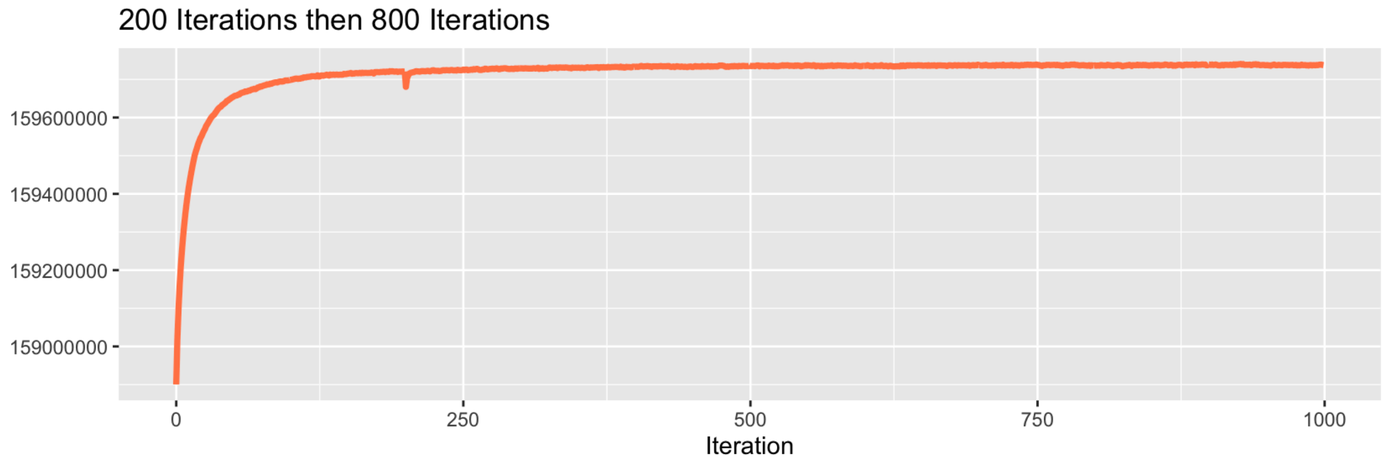
Choosing the number of topics for LDA

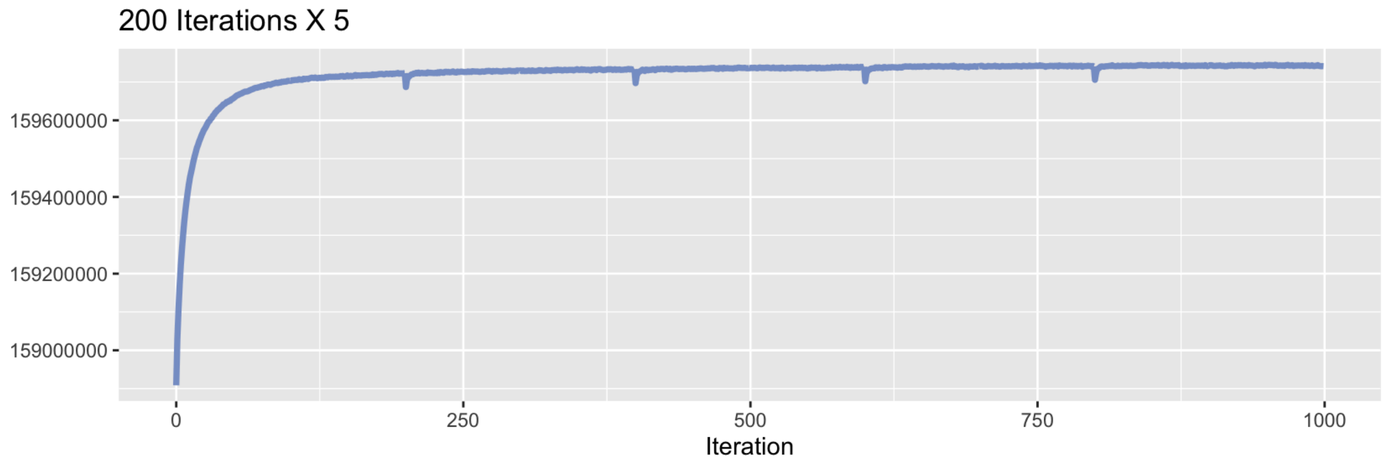
- Sample 100 “k” between 5 and 2,000
- For each “k” sample 1,000 documents
- Run LDA Gibbs sampler for 200 iterations, averaging over the last 50 iterations
  (no chain converged)
- Calculate loess curve
- k = 394 {coherence}
- k = 507 {perplexity, likelihood}

Running 3 LDA chains

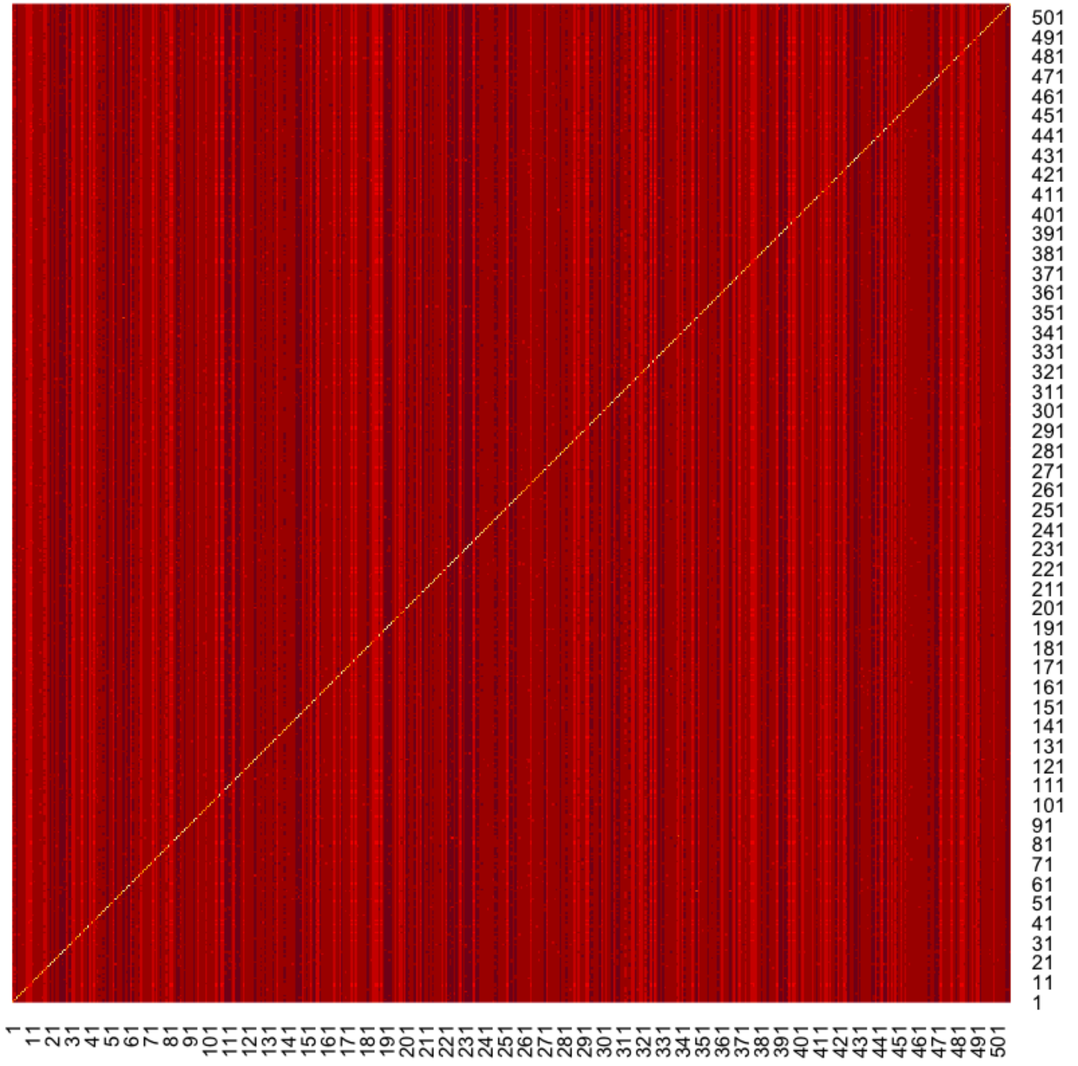
- “Blips” are artifacts of how I re-initialize the chain in tidylda
- Magnitude of log-likelihood calculation is clearly incorrect (because it’s positive) but still useful for assessing convergence
- All chains showed signs of convergence according to Geweke stat between 800 and 1,000 iterations
- Parameters averaged across last 100 samples of each chain



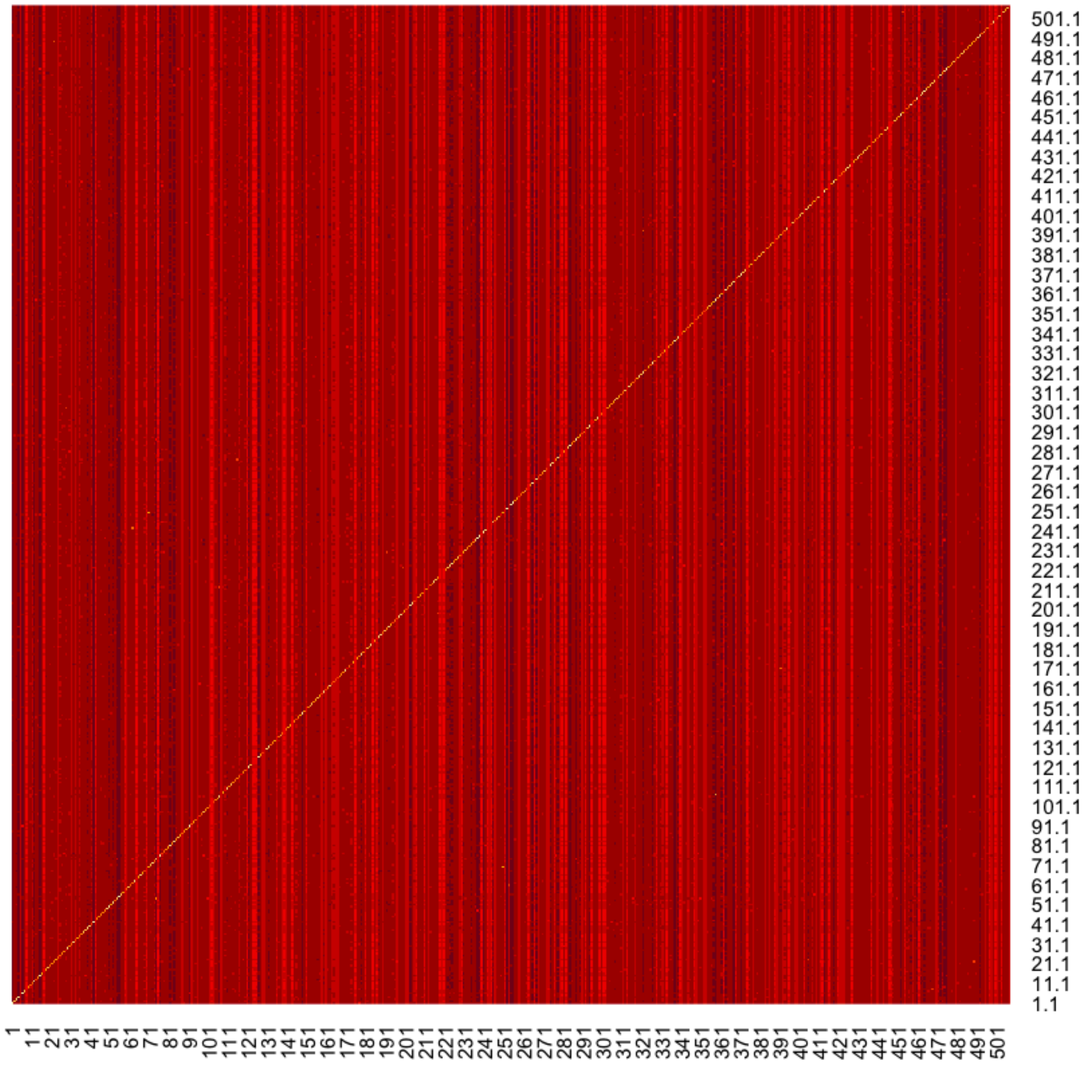




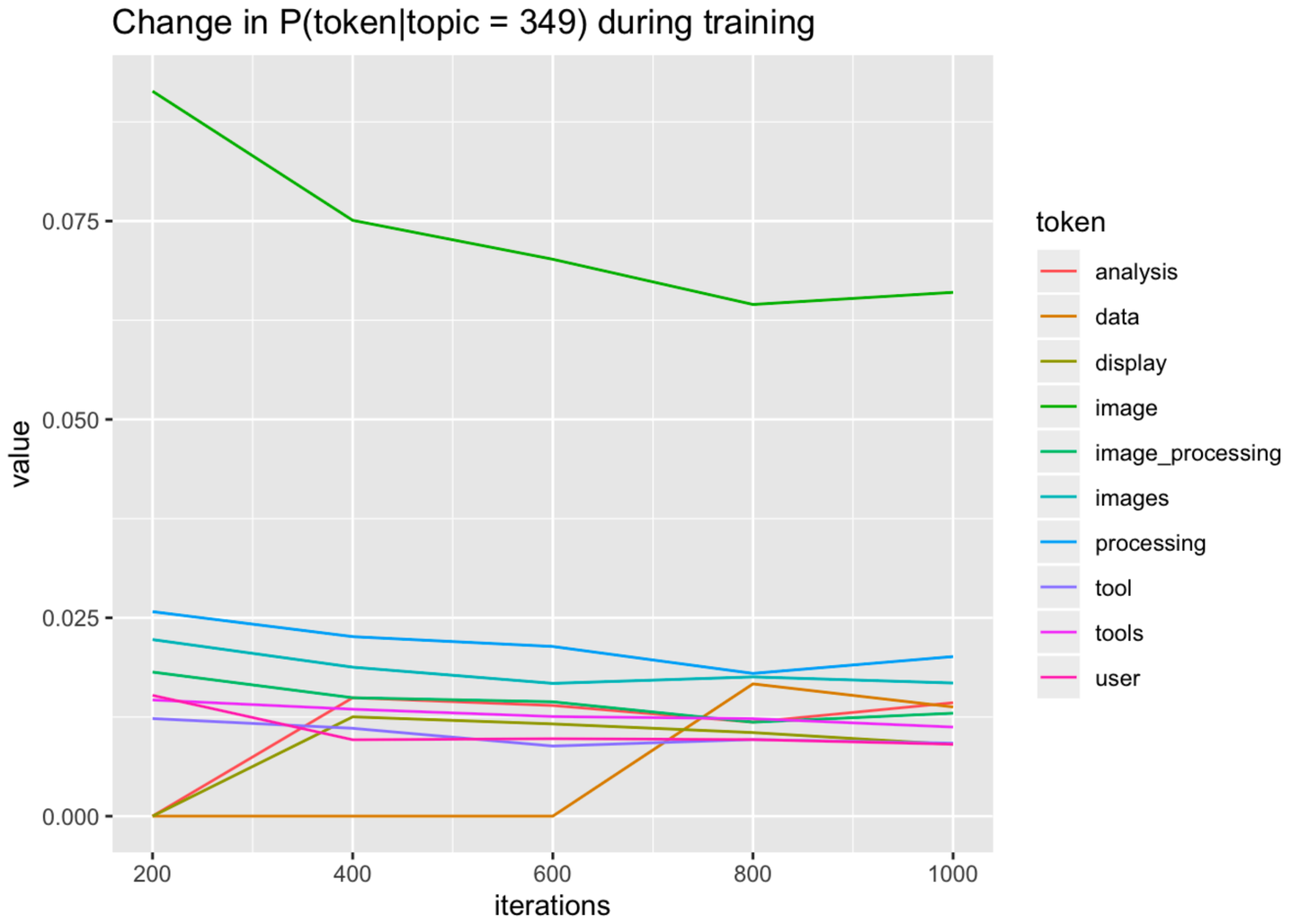
LDA baseline: model 1 to itself



Distance from 200 iterations to 1000 iterations



How do parameter estimates change as we iterate?



Chain 1

```
The  10  most prevalent topics are:
# A tibble: 507 x 4
   topic prevalence coherence top_terms
   <dbl>      <dbl>     <dbl> <chr>
 1    85       5.49   0.489   ax, ax_ax, max, ...
 2   170       3.2    0.0938  time, good, years, ...
 3   453       3.11   0.413   apr, eng, gtefsd, ...
 4   504       2.76   0.0991  news, apr, harvard, ...
 5   166       2.59   0.166   state, apr, ohio, ...
 6   290       2.3    0.428   dos, dos_dos, windows, ...
 7   310       2.27   0.0835  people, make, true, ...
 8   372       1.35  -0.00522 apr, news, rochester, ...
 9   399       1.3    0.350   net, ans_net, ans, ...
10   111       1.15   0.277   disk, drive, drives, ...
# … with 497 more rows
```

Chain 2

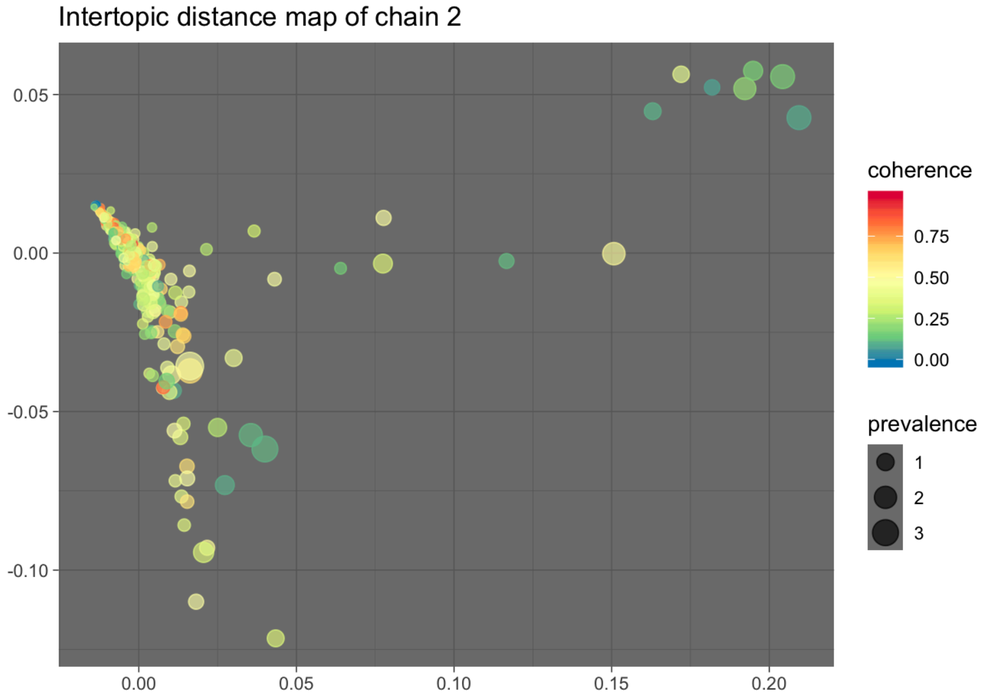
```
The  10  most prevalent topics are:
# A tibble: 507 x 4
   topic prevalence coherence top_terms
   <dbl>      <dbl>     <dbl> <chr>
 1   225       3.52    0.481  ax, max, ax_max, ...
 2   281       2.97    0.109  people, make, good, ...
 3    82       2.53    0.620  ax_ax, ah, uj, ...
 4    67       2.43    0.166  state, news, ohio, ...
 5   205       2.42    0.0892 news, apr, net, ...
 6   104       2.22    0.115  time, back, years, ...
 7    28       2.03    0.514  nntp, host, posting, ...
 8    10       1.96    0.211  net, apr, eng, ...
 9   500       1.54    0.344  dos, windows, microsoft, ...
10   248       1.32    0.171  state, ohio, ohio_state, ...
# … with 497 more rows
```

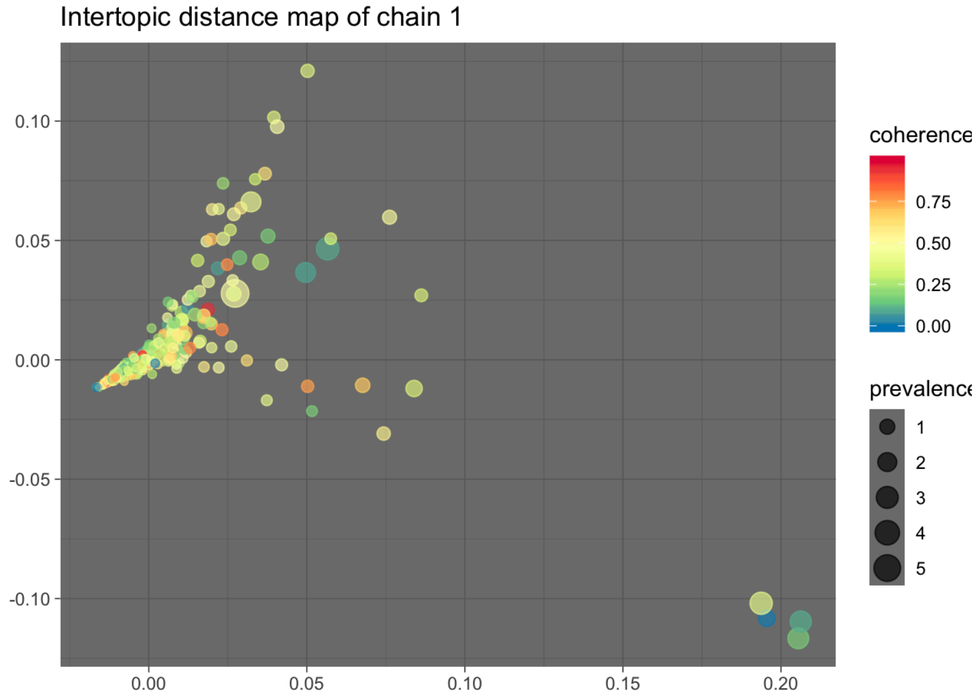
Chain 3

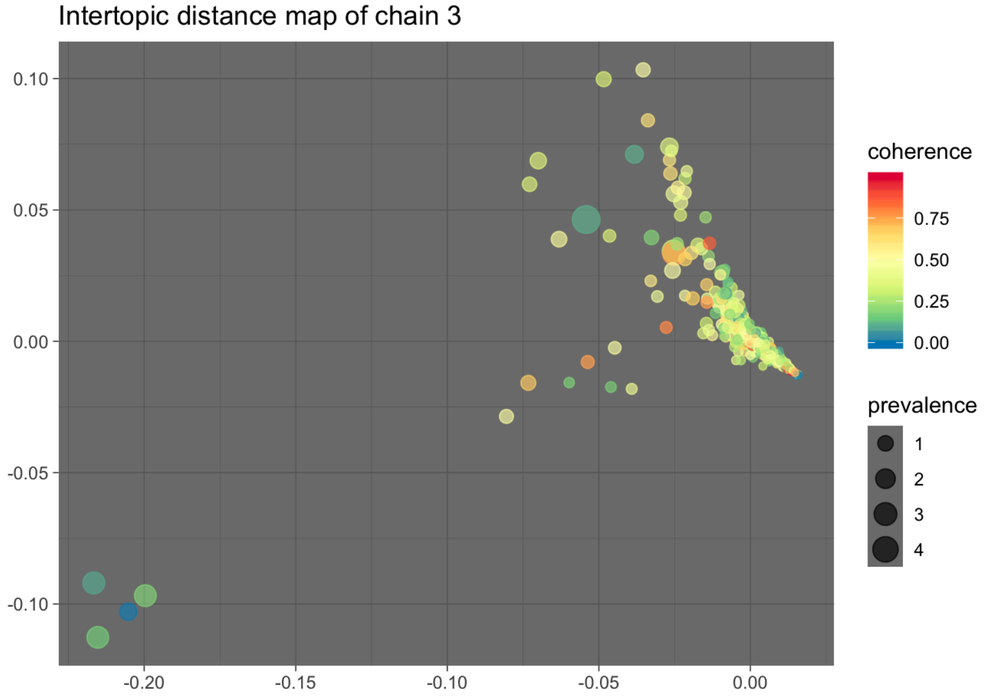
```
The  10  most prevalent topics are:
# A tibble: 507 x 4
   topic prevalence coherence top_terms
   <dbl>      <dbl>     <dbl> <chr>
 1   473       4.98   0.108   people, good, time, ...
 2   250       3.15   0.748   ax_ax, max, ax_max, ...
 3    70       2.9    0.313   ax, sl, ei, ...
 4   147       2.73   0.0892  news, apr, net, ...
 5   502       2.66   0.183   apr, eng, news, ...
 6   453       2.62   0.166   state, apr, ohio, ...
 7    77       1.53   0.102   software, system, systems, ...
 8   489       1.42   0.381   dos, windows, microsoft_windows, ...
 9   297       1.38  -0.00522 apr, news, rochester, ...
10    68       1.17   0.350   net, ans_net, ans, ...
# … with 497 more rows
```

Plotting Intertopic Distance by MDS

- Hellinger distance of the “phi” matrix
- Conventional multidimensional scaling for the layout







LDAvis?

One more thing…