BristolR Meetup

Not A Gay Cowboy Movie?

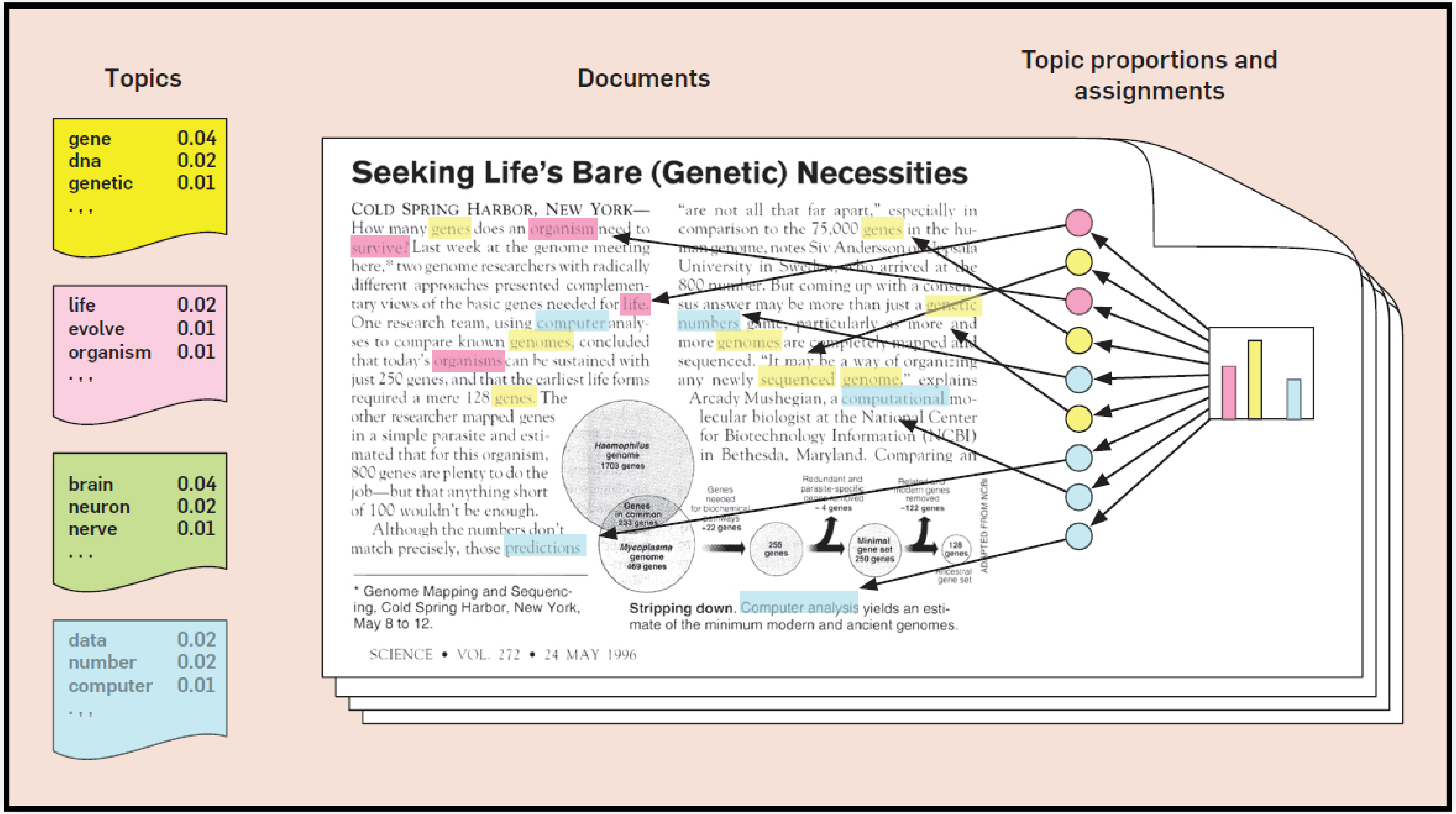
Topic Modelling on Movie Summaries

Paul Matthews

@paulusm

Jun 2019

Definition



Topics

gene	0.04
dna	0.02
genetic	0.01
. , ,	

life	0.02
evolve	0.01
organism	0.01
. , ,	

brain	0.04
neuron	0.02
nerve	0.01
. . .	

data	0.02
number	0.02
computer	0.01
. , ,	

Documents

Topic proportions and assignments

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

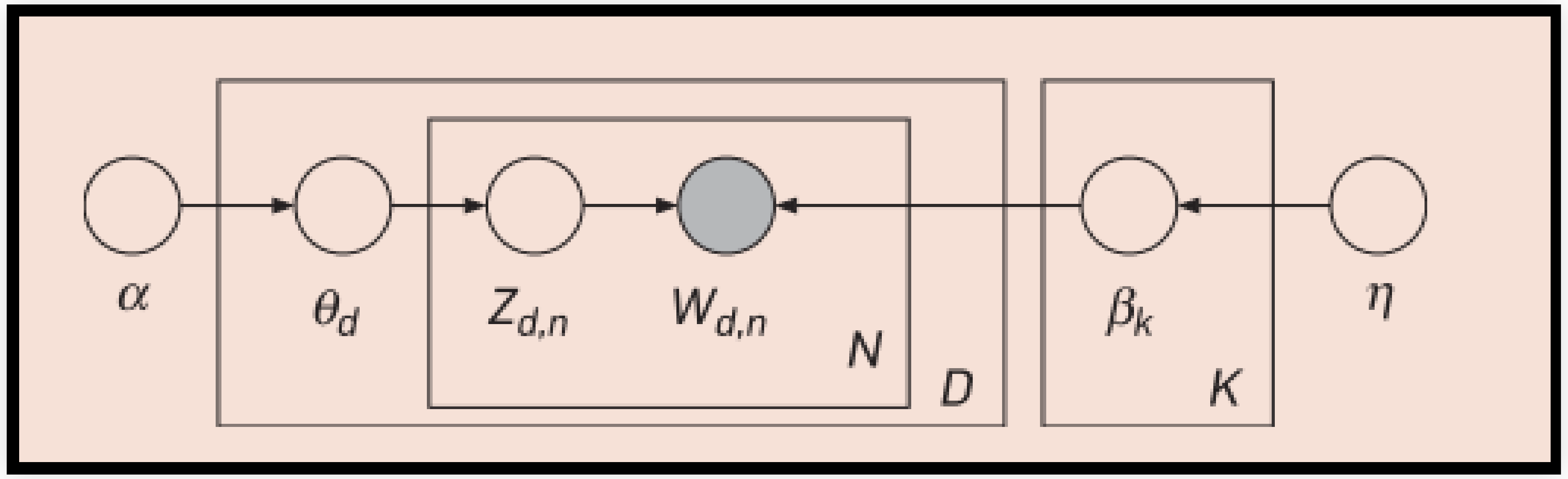
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

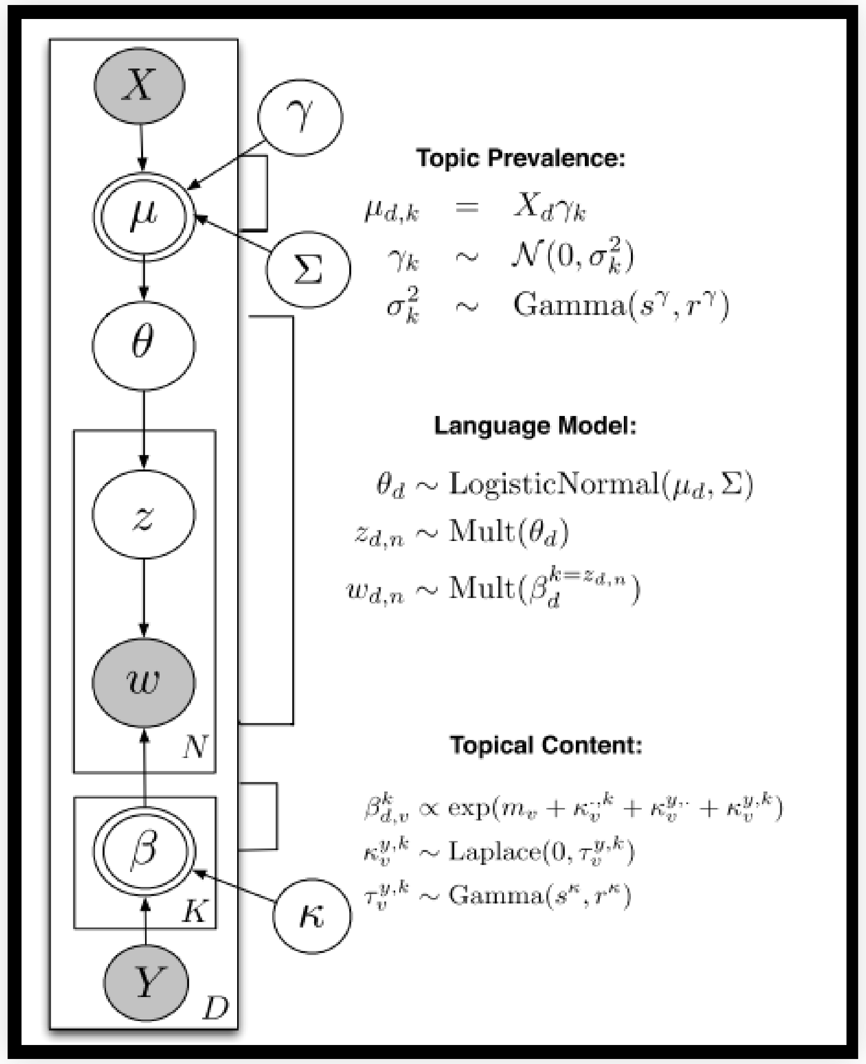
SCIENCE • VOL. 272 • 24 MAY 1996

Variables - The Science Bit

LDA

STM





Topic Prevalence:

$$\mu_{d,k} = X_d\gamma_k$$
$$\gamma_k \sim \mathcal{N}(0, \sigma_k^2)$$
$$\sigma_k^2 \sim \text{Gamma}(s^\gamma, r^\gamma)$$

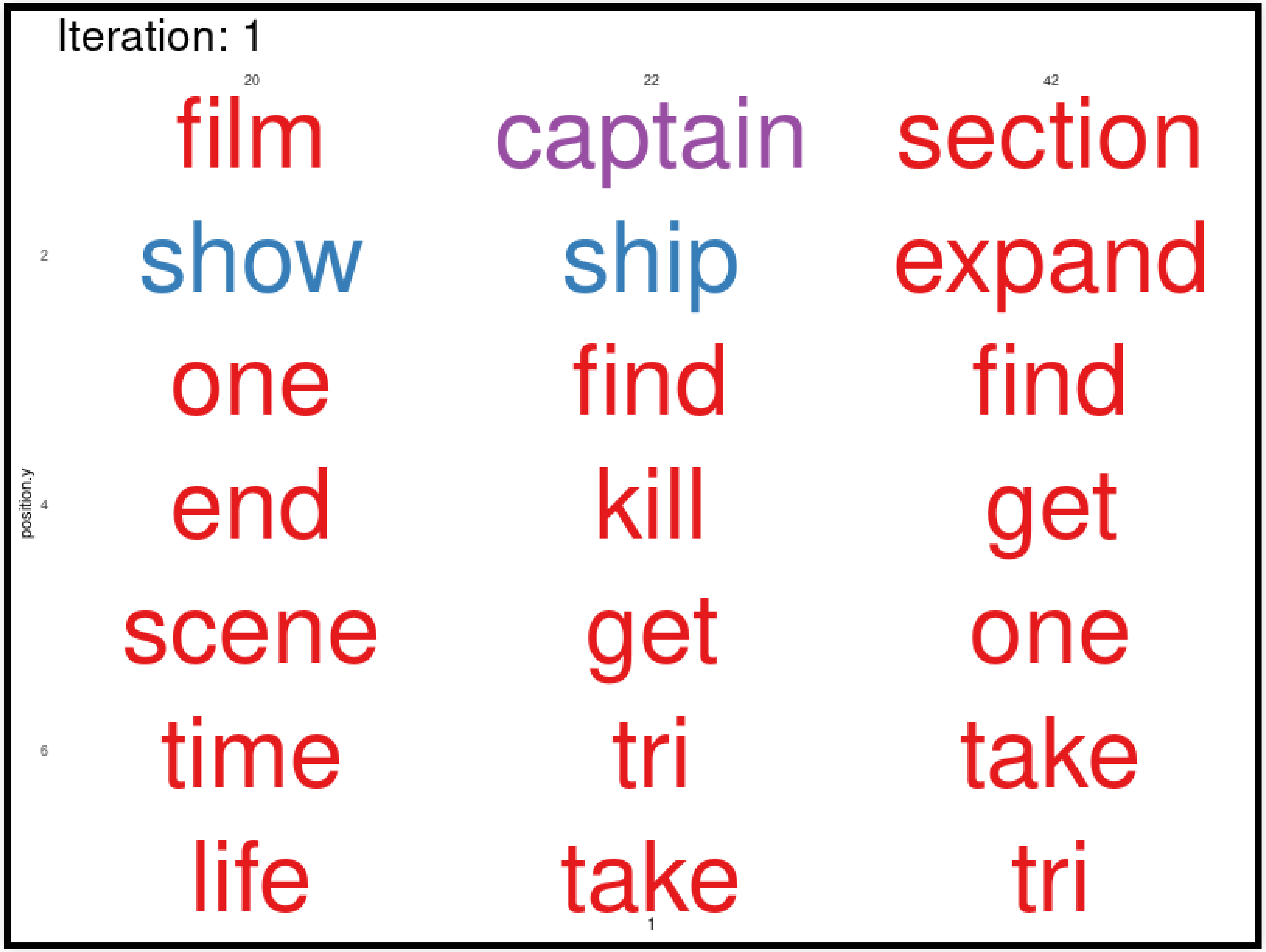
Language Model:

$$\theta_d \sim \text{LogisticNormal}(\mu_d, \Sigma)$$
$$z_{d,n} \sim \text{Mult}(\theta_d)$$
$$w_{d,n} \sim \text{Mult}(\beta_d^{k=z_{d,n}})$$

Topical Content:

$$\beta_{d,v}^k \propto \exp(m_v + \kappa_v^{\cdot,k} + \kappa_v^{y,\cdot} + \kappa_v^{y,k})$$
$$\kappa_v^{y,k} \sim \text{Laplace}(0, \tau_v^{y,k})$$
$$\tau_v^{y,k} \sim \text{Gamma}(s^\kappa, r^\kappa)$$

Model estimation

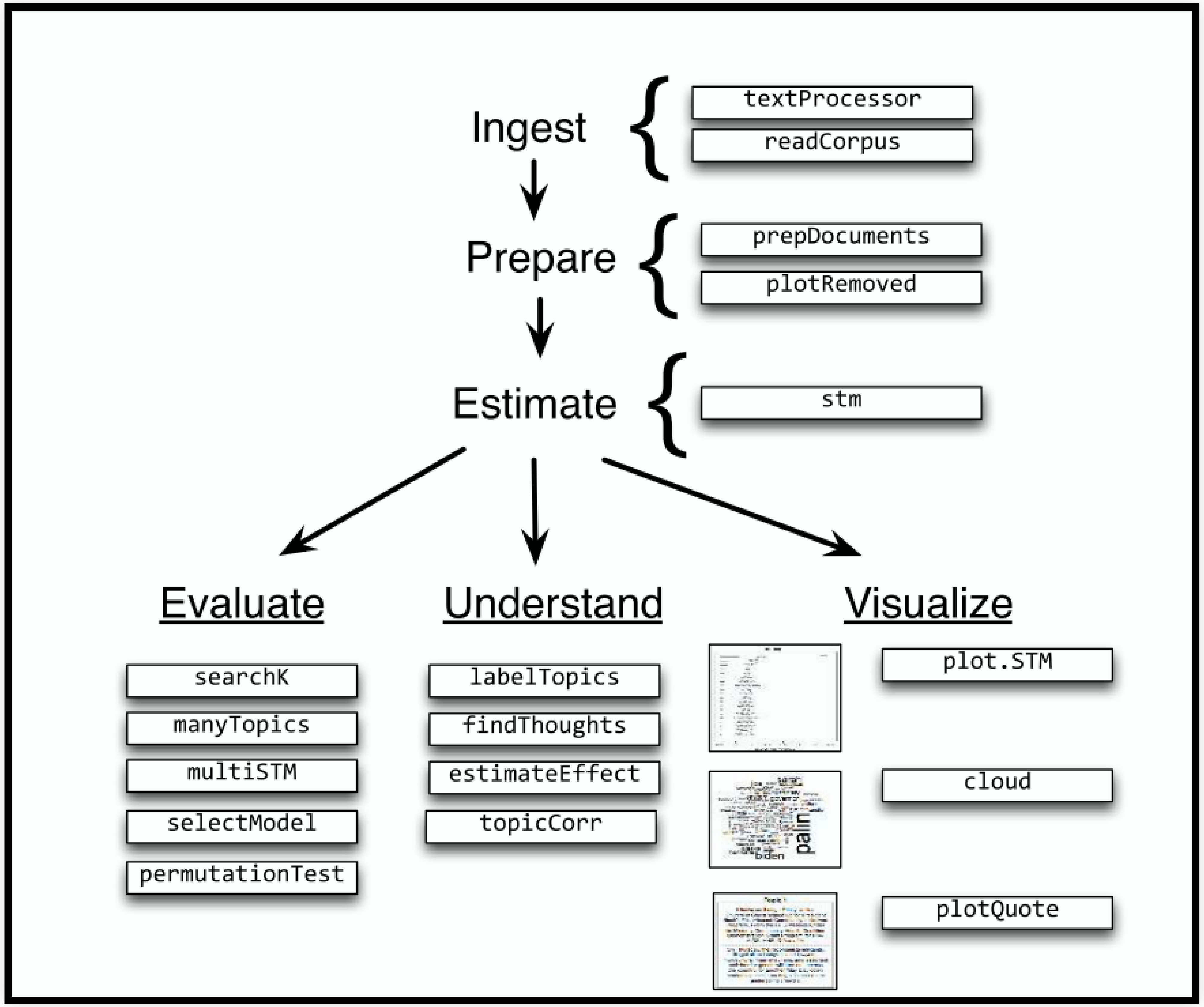


(Animation Aside - ggplot + gganimate)

```
anim <- finalwords.df %>% filter(topic %in% c(22,20,42) )  %>%
  ggplot(aes(x=1, y=position.y, label=token, colour=as.factor(position.x) )) +
  scale_color_brewer(palette="Set1") +
  geom_text(size=22) +
  theme_minimal() +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        line = element_blank(),legend.position="none", plot.title = element_text(size=28)) +
  scale_x_discrete(1) +
  scale_y_reverse()  +
  facet_grid(~topic) +
  transition_states(iteration.y , transition_length = 4, state_length = 2) +
  ease_aes('linear') +
  labs(title = 'Iteration: {closest_state}')

animate(anim, nframes=118, width=800, height=600, end_pause=0, fps=9)
anim_save("animation/iterations.gif")
```

STM Workflow



Ingest

- textProcessor
- readCorpus

Prepare

- prepDocuments
- plotRemoved

Estimate

- stm

Evaluate

- searchK
- manyTopics
- multiSTM
- selectModel
- permutationTest

Understand

- labelTopics
- findThoughts
- estimateEffect
- topicCorr

Visualize

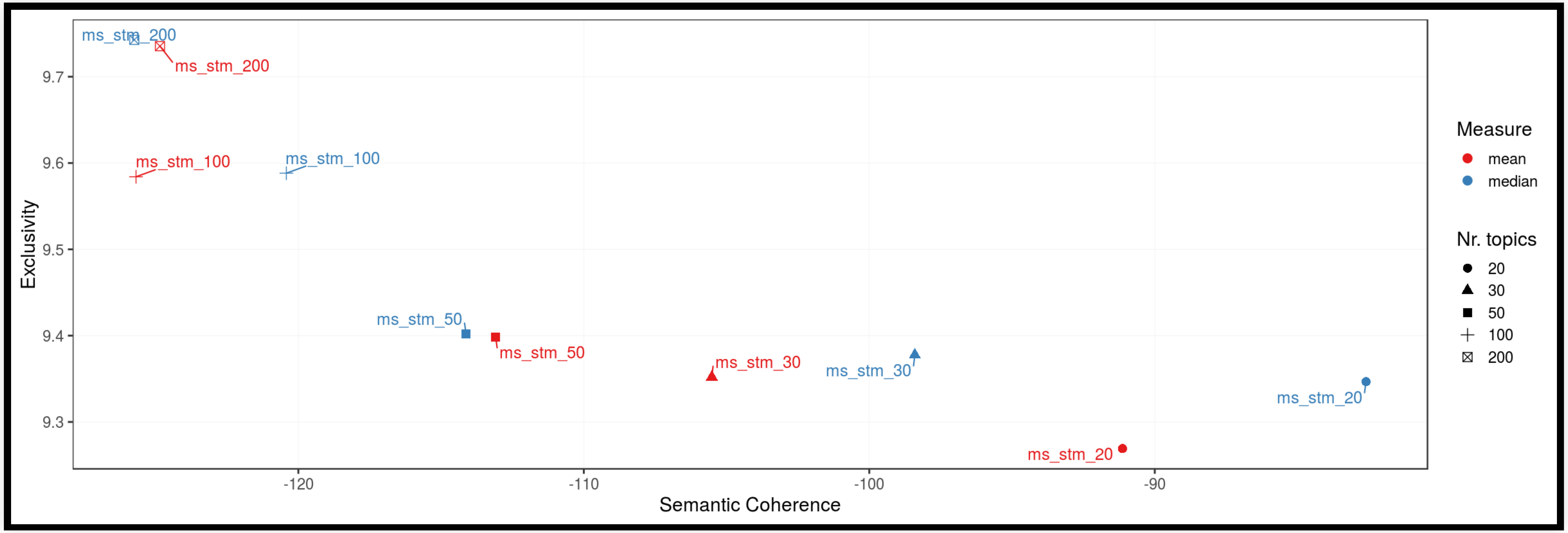
- plot.STM
- cloud
- plotQuote

Data

CMU Movie Summary Corpus, 42k movie plot summaries extracted from Wikipedia + aligned metadata from Freebase (Wikidata)
http://www.cs.cmu.edu/~ark/personas/

title	summary	release	scifi	bollywood
The Hunger Games	The nation of Panem consists of a wealthy Capitol ...	2012	1	0
Narasimham	Poovalli Induchoodan is sentenced for six years p...	2000	0	1
The Lemon Drop Kid	The Lemon Drop Kid , a New York City swindler, is ...	1951	0	0

How many topics?



Modelling

```
# ms_data is the movie summary dataset
firstnamefilter <- unique(firstnames[firstnames$percent > 0.0005,]$word)
ms_processed <- textProcessor(ms_data$summary, ms_data, customstopwords = firstnamefilter)

# Increasing lower.thresh gives performance enhancement on modelling
ms_prepped <- prepDocuments(ms_processed$documents, ms_processed$vocab , ms_processed$meta, lower.thresh = 10)

# 1) Unlabelled standard CTM (Blei & Lafferty, 2007)
# Around 60-90 seconds per iteration (Intel Core i7-7500, 32GB), so 1.5 hours with d = 44,000
ms_stm_50 <- stm(prepped$documents,ms_prepped$vocab,
            data=prepped$meta, K=50, max.em.its=75)

# 2) Labelled
ms_stm_50_labelled <- stm(ms_prepped$documents,ms_prepped$vocab,
              data=ms_prepped$meta, K=50, max.em.its=75, content=~genre, prevalence=~genre+release_year)

# Useful to save model
save(ms_stm_50, file="models/ms_stm_50.RData")
```

Genres

Linking topic distribution to document level classification

Stagecoach (1939)

Brokeback Mountain (2006)

Westerns - Topic Signature

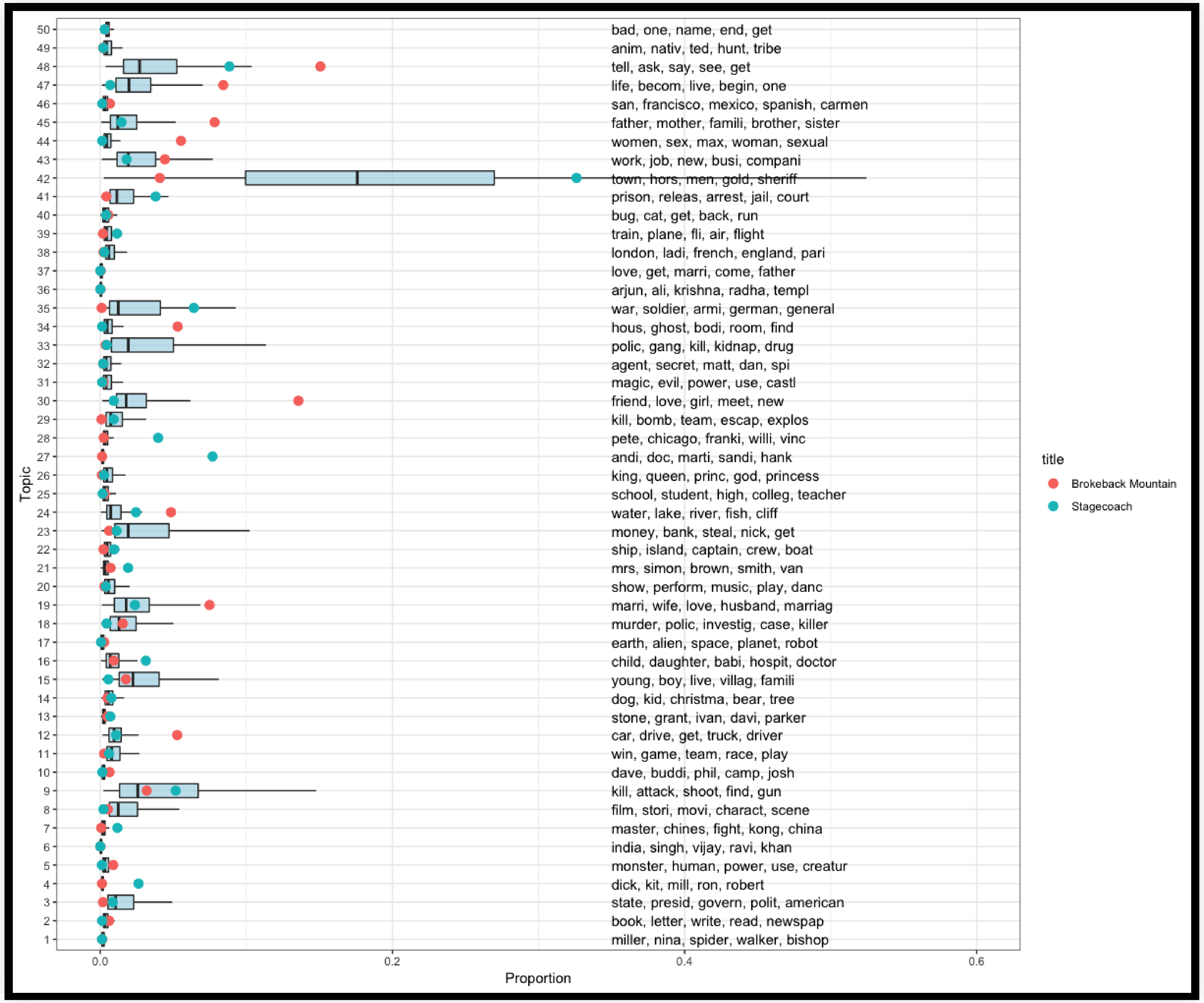
```
genre_data <- cbind(ms_prepped$meta, ms_stm_50$theta) %>%
    filter("genre.Western") == TRUE) %>%
    melt(id.vars=c("id", "title"), measure.vars=c(374:424)))

genre_data_hl <- genre_data %>% filter(id %in% hl)

topic_keywords <- labelTopics(ms_stm_50)

genre_data %>%
  ggplot(aes(x=variable, y=value)) +
    geom_boxplot(fill = 'lightblue', alpha=0.7, outlier.alpha = 0) +
    geom_point(data=genre_data_hl, size=3, aes(x=variable, y=value, color=title)) +
    theme_bw() + scale_y_continuous("Proportion", limits=c(0,0.6)) +
    scale_x_discrete("Topic") +
    coord_flip() +
    annotate("text", x = t, y = 0.35, hjust=0,
                      label=paste(topic_keywords$prob[t,1:5], collapse=", "))
```

Topic Distribution - Westerns



Road Movies

Topic Distribution - Road Movies

Road Movie Identification?

```
# Get medians
rm_median <- cbind(ms_prepped$meta, ms_stm_50$theta) %>% filter(genre.Road.movie == TRUE) %>%
  melt(id.vars=c("id", "title"), measure.vars=c(374:(373 + 50))) %>% group_by(variable) %>% summarise(med_proporti

# Calc distances for Road Movies
rm_dist <- cbind(ms_prepped$meta, ms_stm_50$theta) %>% filter(genre.Road.movie == TRUE & nchar(summary) > 2000) %>
  melt(id.vars=c("id", "title"), measure.vars=c(374:(373 + 50))) %>% arrange(id, variable) %>%
  cbind(rm_median[,2]) %>% mutate(dist=((med_proportion-value)^2) * med_proportion) %>%
  group_by(id, title) %>% summarise(distance=sum(dist)) %>% arrange(distance) %>% head(20)

# Distances for all others
non_rmdist <- cbind(ms_prepped$meta, ms_stm_50$theta) %>% filter(genre.Road.movie == FALSE & nchar(summary) > 2000
  melt(id.vars=c("id", "title"), measure.vars=c(374:(373 + 50))) %>% arrange(id, variable) %>%
  cbind(rm_median[,2]) %>% mutate(dist=((med_proportion-value)^2) * med_proportion) %>%
  group_by(id, title) %>% summarise(distance=sum(dist)) %>% arrange(distance) %>% head(20)

rm_table <- cbind(rm_dist, non_rmdist) %>% select(title, distance, title1, distance1) %>%
  rename("In road movie genre"=title, "Distance from median"=distance, "Not in genre"=title1, "Distance from media
```

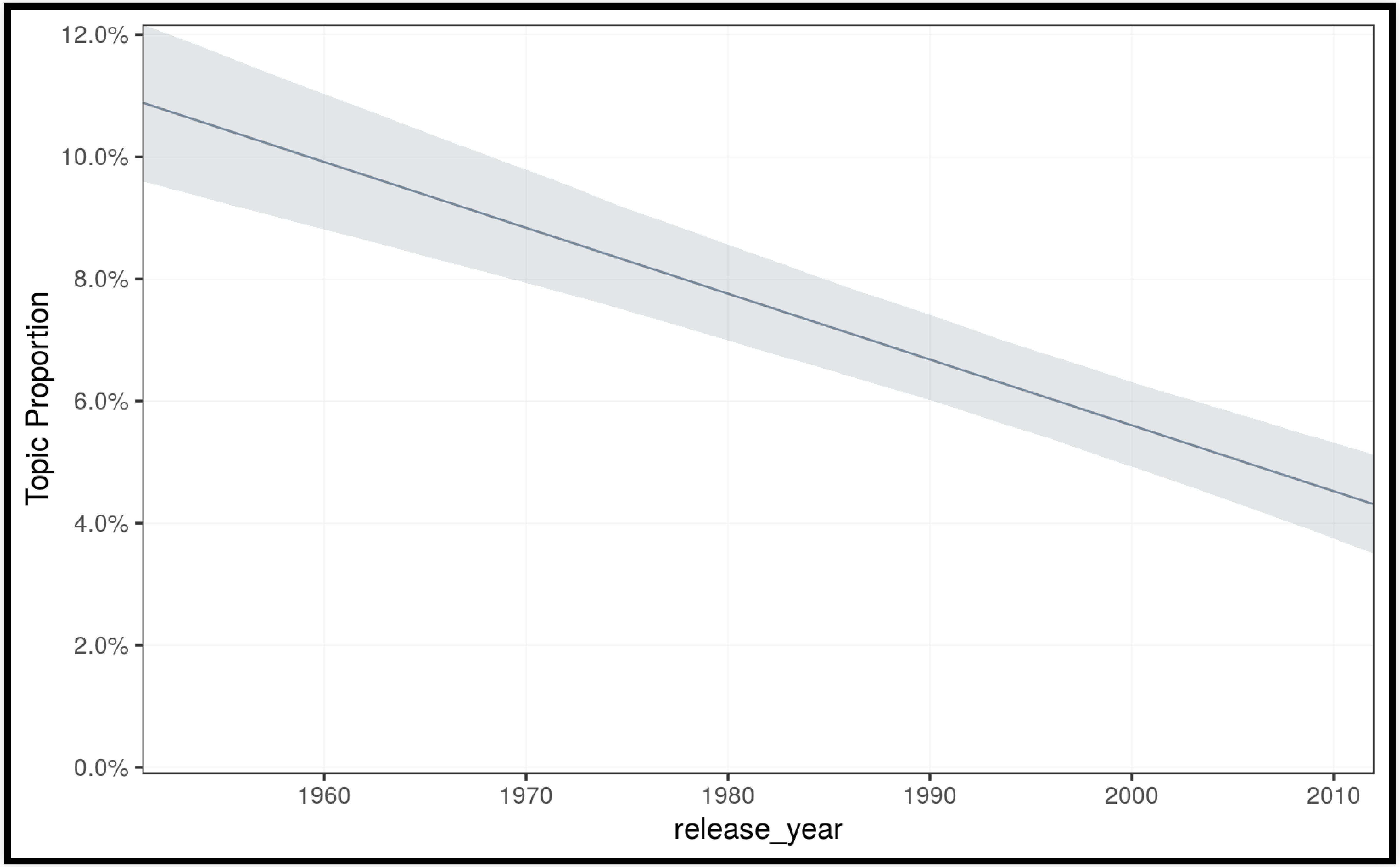
Road Movie Identification?

id	In road movie genre	Distance from median	Not in genre	Distance
20927392	My Name is Khan	0.0002774	Gadar: Ek Prem Katha	0.0001569
22216	O Brother, Where Art Thou?	0.0004099	Alien from L.A.	0.0002011
167303	A Canterbury Tale	0.0004282	The Stepford Children	0.0002023
7047921	Little Miss Sunshine	0.0004370	Arthur and the Vengeance of Maltazard	0.0002227
238906	Blues Brothers 2000	0.0004389	Wake in Fright	0.0002234
13473149	Singh Is Kinng	0.0004441	Hum Ek Hain	0.0002239
163457	The Last Detail	0.0004475	Kim Possible: A Sitch in Time	0.0002369
4198757	Space Truckers	0.0004552	Cat's Eye	0.0002408
73402	Sullivan's Travels	0.0004688	Just Imagine	0.0002422
17786854	Five Dollars a Day	0.0005223	The Petrified Forest	0.0002457
6238106	The Darjeeling Limited	0.0005426	Stand by Me	0.0002603
583758	Cannonball Run II	0.0005559	The Business	0.0002629
873029	To Wong Foo, Thanks for Everything! Julie Newmar	0.0005862	Role Models	0.0002631

Wake in Fright (1971)

Topics over time (K=20) - Party

Topics over time (K=20) - War



Finale: Little Miss Sunshine (2006)