Topic Modeling
RLadies Christmas Meetup

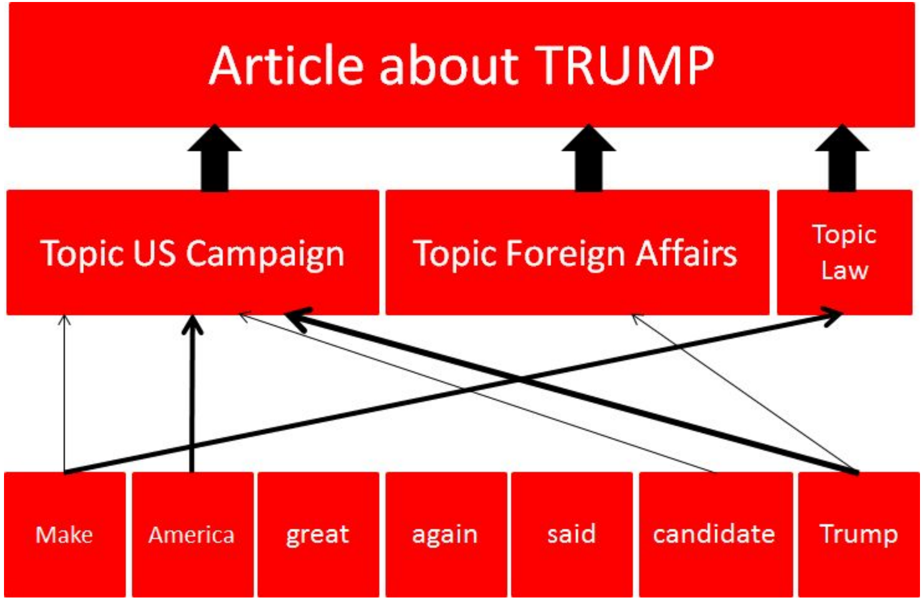
03-12-2018

Agenda

1. What is Topic Models?
2. Practicalities: Training Topic Models in AWS
3. Use cases in VG

Introduction to Topic Models

TOPIC MODELLING - BEYOND TAGGING

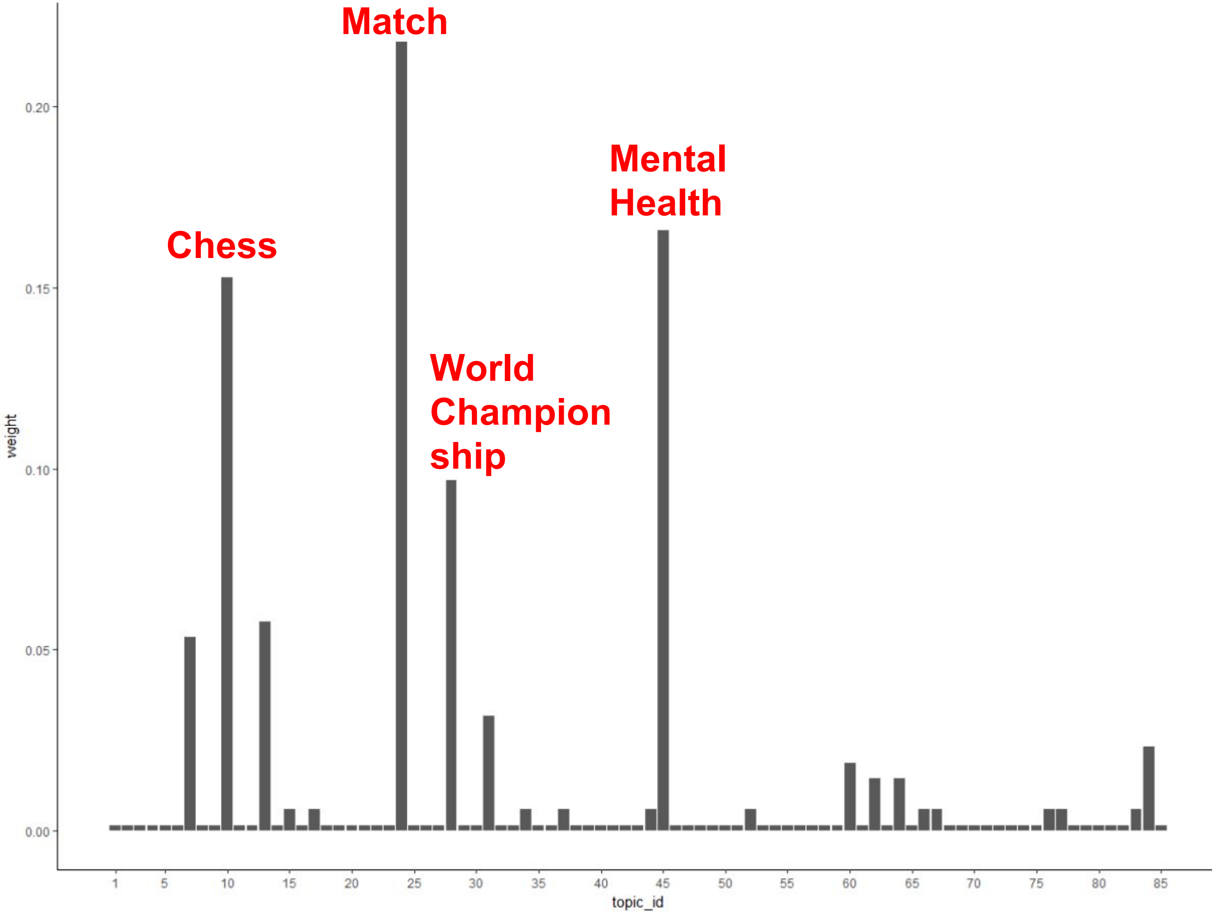


- **Ambiguity:** A single word is related to several topics
- **Content analysis:** A single document may consist of several topics
- **Unsupervised:** Learns from documents and words

Topic Vector

Derfor er Carlsen best når det gjelder

Carlsen hadde blitt nervøs - og at sjakk-lege
nden ikke lenger holdt nordmannen som favoritt
 i omspillet.\n\n- Det psykologiske kan spille
Magnus et puss. Det han gjorde i det 12. part
iet gjør at han tydeligvis ser seg selv som fa
voritt og tvinger seg selv til å levere i omsp
illet. Om dette ekstra presset vil gjøre ham g
odt eller gjøre ham nervøs, gjenstår å se. Det
 kan slå begge veier, sier Dirk Jan ten Geuzen
dam fra New in Chess.\n\n- Etter en tung start
 har Fabiano vist stor psykisk styrke i denne
matchen, sier Elmira Mirzojeva fra Matj TV.\n



Latent Dirichlet Allocation - Latent?

β Word proportion per topic

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

θ Topic proportion per document

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Dirichlet Hyperparameter

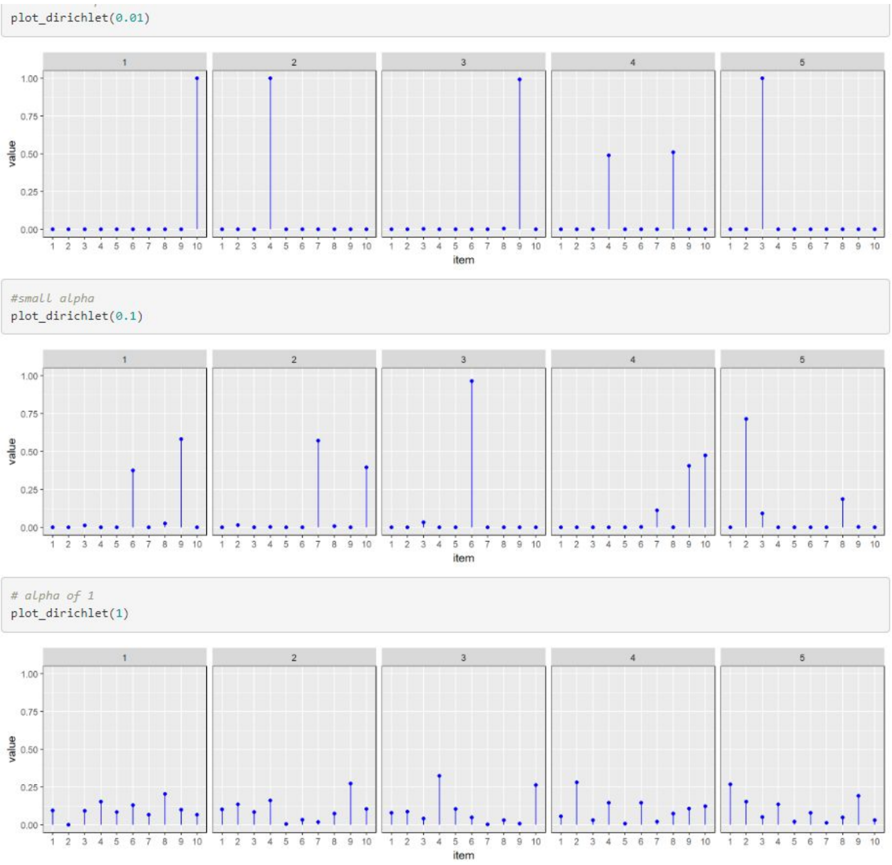
What is a **Dirichlet** distribution?

It is a probability distribution of a **probability simplex**. What is a probability simplex? It is a non-negative vector whose values sum up to 1, like so:

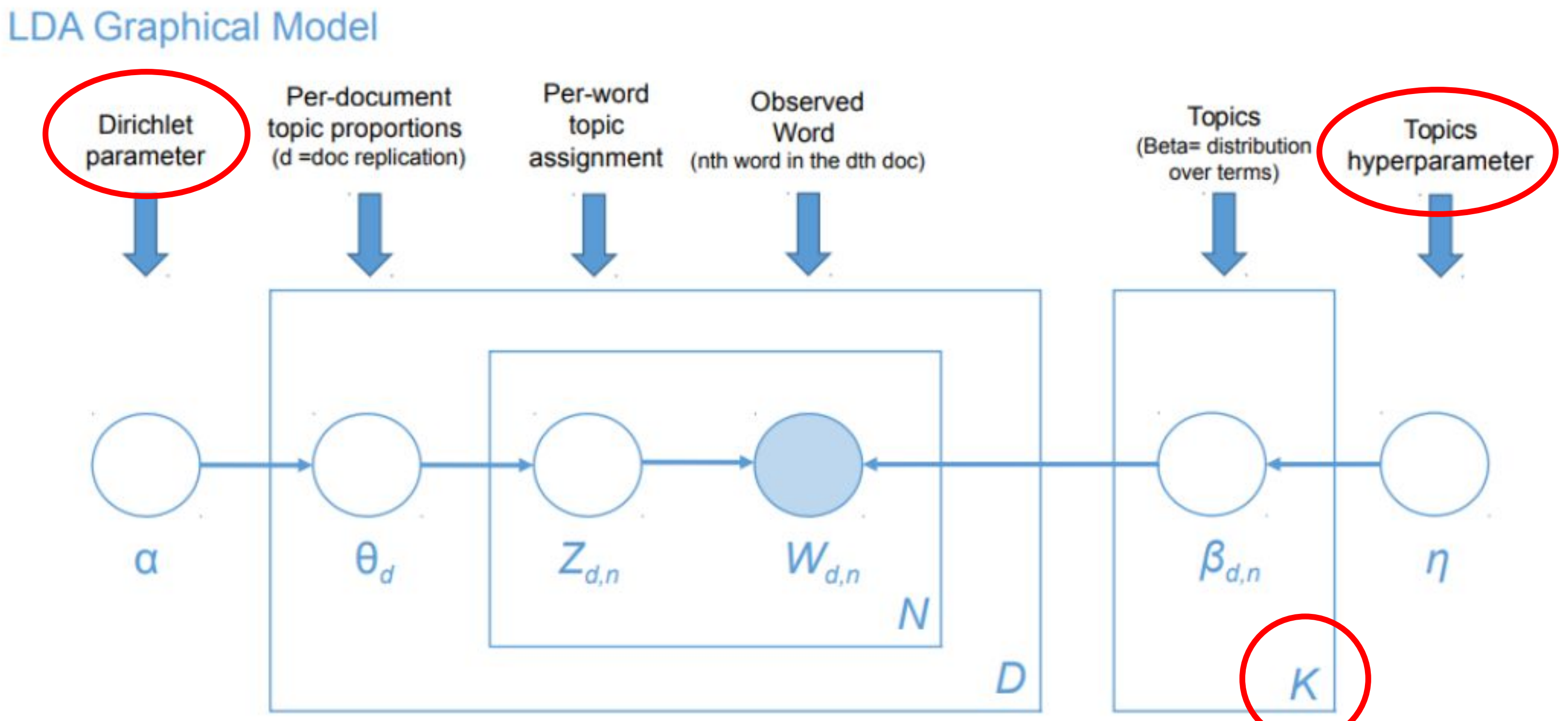
(0.6, 0.4)

(0.2, 0.1, 0.7)

(0.05, 0.1, 0.3, 0.2, 0.2, 0.15)

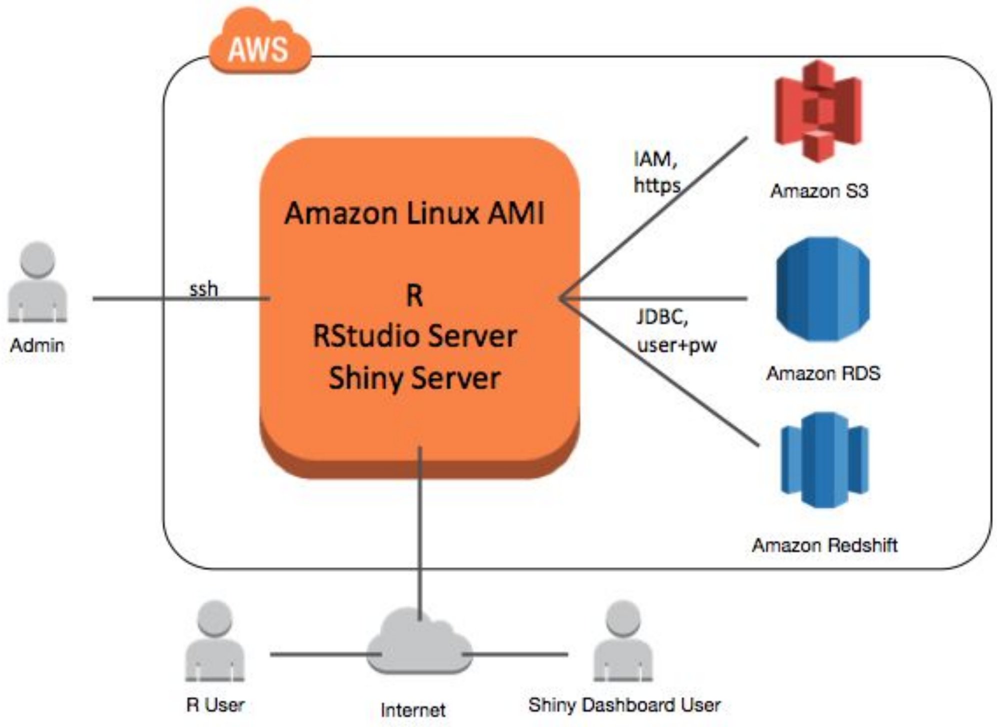


LDA - The graphical model



Practicalities

AWS Setup



- I ♥ R
- But...what about Scaling?
- Main Challenge with TM: we need a lot of brute force computational power
- AWS is one solution to solve this

How to debug AWS?

- AWS and R do not play well together out of the box
- You need extra development packages that do not come with standard installation.
- Depending on what packages you need, edit what default gcc compiler R should use
- Tutorial/blog in RMarkdown

```
# updated the instance:
sudo yum update -y

# checked if there is any version of GCC installed in the instance:
sudo yum list installed gcc*

# remove all gcc instances older than 48, like the following
sudo yum remove gcc72-c++.x86_64 libgcc72.x86_64
sudo yum remove gcc64-gfortran.x86_64
sudo yum remove gcc64.x86_64
# remove other gcc version should you have any
# since the blog post recommended to install GCC version 4.8, so did I:
sudo yum install -y gcc48

# needed for stm package
sudo yum install R-devel

cd /usr/lib64/R/etc
sudo vi Makeconf
# insert the following
#   CC = gcc64 back to CC = gcc
# then save and exit

# start R
sudo R
# once in R install stm package with dependencies
install.packages("stm", dependencies = T)
library(stm)
```

Preprocessing Text

1. Annotate Text: tokenization, lemmatization and pos tagging
2. Filter:
    a. Keep Noun, Verb and Prop Noun
    b. Remove common words: tell, say, come, go, etc.
    c. Remove word with frequency of 1 in vocabulary
3. Concatinate Prop Noun:
   "Manchester", "United" => "Manchester United"
   "Magnus", "Carlsen" => "Magnus Carlsen"

The magic of "udpipe" library in R

install.packages(“udpipe”)

```r
# Load data to annotate into R on EC2
dt <- readRDS("df_to_featureEngineering.rds")
dir()

#text annotation for VG in 30 datasets because udpipe limits annot
sequence <- seq(0, 300000, by = 10000)

for(i in 1:length(sequence)-1){

print(paste("session", i, Sys.time()))
start <- sequence[i] + 1
end <- sequence[i + 1]
to_feature <- dt[start:end,]
model <- udpipe_load_model("ud_norwegian.udpipe")

annotated_dt <- udpipe_annotate(model, x = to_feature$text)
annotated_dt <- as.data.frame(annotated_dt)

temp_out <- paste0("featured", i, ".rds")
saveRDS(annotated_dt,file=temp_out)
print(paste("done with session", i, "starting with row" , start))
Sys.time()

}
```

Untitled1 * | to_LDA | test

Filter

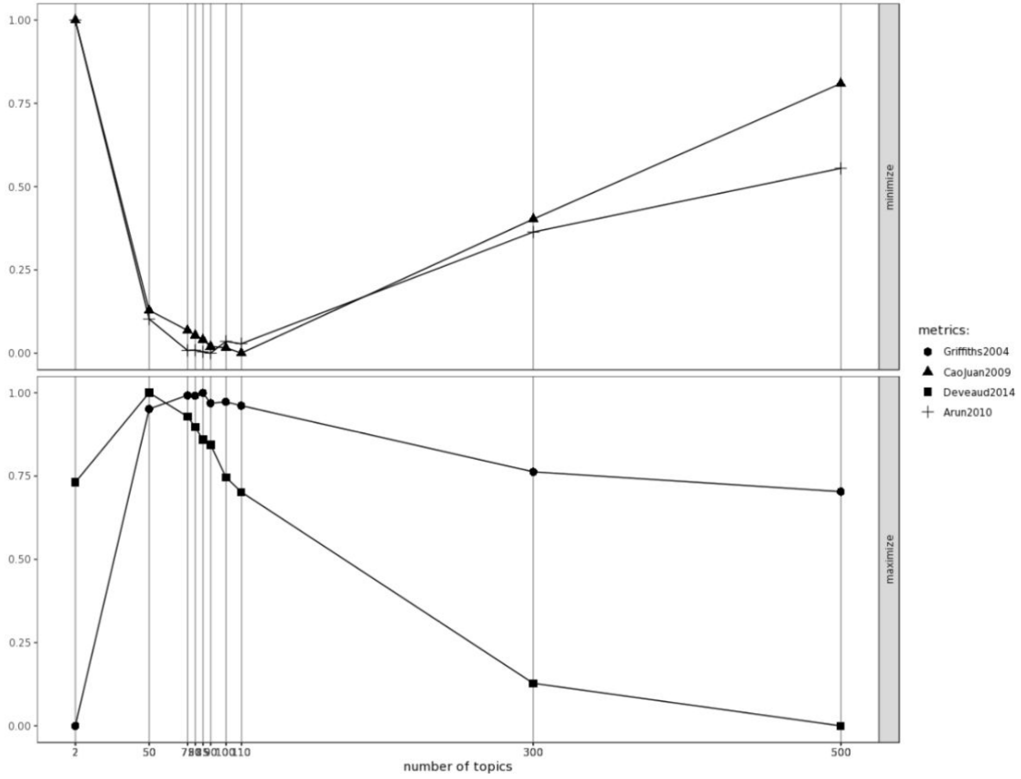
	doc_id	paragraph_id	sentence_id	sentence	token_id	token	lemma	upos
27	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	1	Alt	alt	PRON
28	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	2	tyder	tyde	VERB
29	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	3	på	på	ADP
30	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	4	at	at	SCONJ
31	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	5	du	du	PRON
32	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	6	ikke	ikke	ADV
33	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	7	får	få	AUX
34	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	8	se	se	VERB
35	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	9	norsk	norsk	ADJ
36	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	10	fotball	fotball	NOUN
37	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	11	på	på	ADP
38	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	12	riksdekkende	riksdekkende	ADJ
39	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	13	TV	TV	NOUN
40	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	14	da	da	SCONJ
41	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	15	sesongen	sesong	NOUN
42	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	16	starter	starte	VERB
43	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	17	igjen	igjen	ADV
44	doc50257	3	4	Alt tyder på at du ikke får se norsk fotball på riksdekkende T...	18	i	i	ADP

Document term matrix

```
library(udpipe)
x <- document_term_frequencies(filtered_df,
                               document = "doc_id",
                               term = "lemma")

dtm <- document_term_matrix(x)
```

How to choose the K - number of topics?

```r
library(ldatuning)

result <- FindTopicsNumber(
  dtm,
  topics = c(2, 50, 75, 80, 85, 90,
             100, 110, 300, 500),
  metrics = c("Griffiths2004", "CaoJuan2009",
              "Deveaud2014", "Arun2010"),
  method = "Gibbs",
  control = list(seed = 77),
  mc.cores = 50L,
  verbose = TRUE
)

# Plot result
FindTopicsNumber_plot(result)
```



Parameters

library(topicmodels)

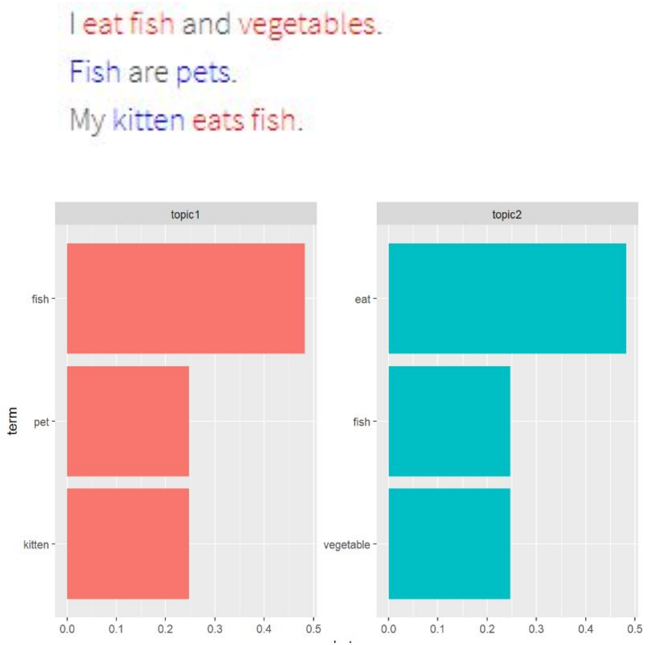
- Number of Topic K
- Dirichlet hyperparameter θ and β

  α = 50/K
- Number of iterations
- Burn-in

```
control_LDA_Gibbs <- list(alpha = 50/85,
                          estimate.beta = TRUE,
                          verbose = 0,
                          prefix = tempfile(),
                          save = 0,
                          keep = 0,
                          seed = as.integer(848),
                          nstart = 1,
                          best = TRUE,
                          delta = 0.1,
                          iter = 10000,
                          burnin = 1000,
                          thin = 2000)


Sys.time()
LDAmodel_vgpluss_final <- LDA(dtm, k = ,
                              method = "Gibbs",
                              control = control_LDA_Gibbs)
Sys.time()
```

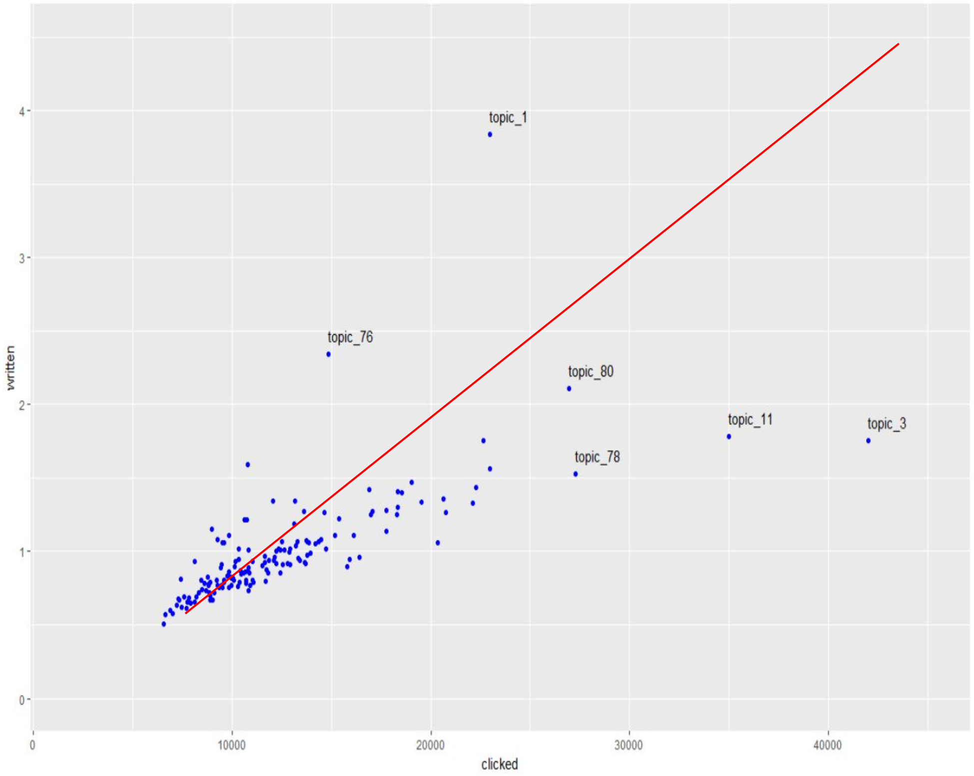
How the algorithm works?

1. Parameterisation
2. Initialisation
3. Topic Allocation
4. Count Matrix
5. Iterations
6. Visualizing output



Use Cases in VG

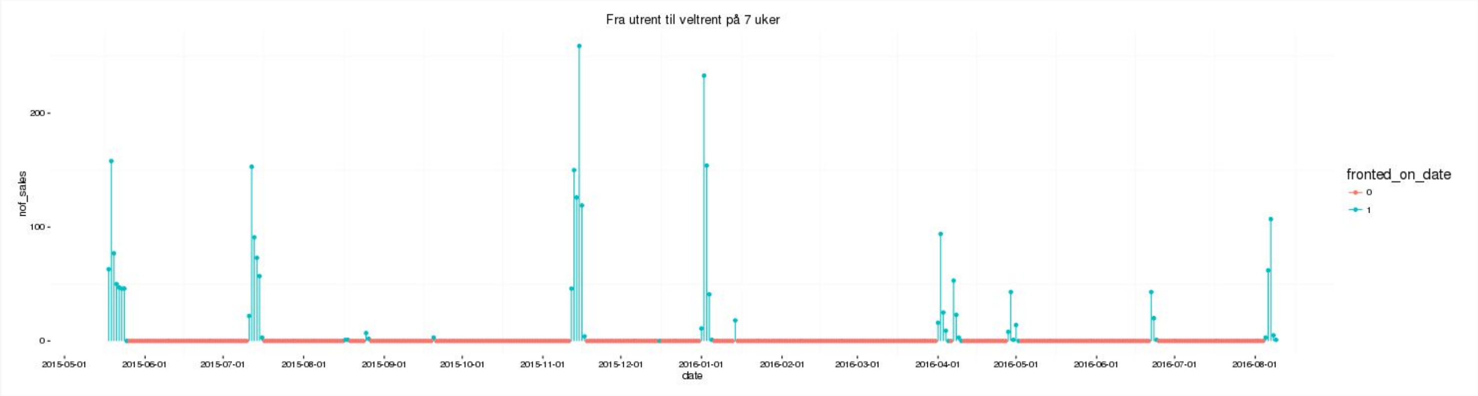
Analysing article production



- Finding what we should write more about
- Finding what we write a lot about that is not that popular
- What are the least popular articles and why?

23323739

2016-08-09



Slide meg for å velge en topic:

0 77 149

0 15 30 45 60 75 90 105 120 135 149

Ordskyen for:
Samliv

Topic score:

Topp Salg for topic
Samliv

Show 10 entries Search:

	id	title	score	salg
1	23473652	Evig singel?	32.70	1678
2	23456976	Elskerinner og elskere forteller	59.87	959
3	23465997	Derfor er kvinner utro	41.45	818
4	23644653	Tegnene på at dere ikke passer sammen	30.60	623
5	23501571	Sex pluss én	46.33	449
6	23490411	Derfor faller vi for andre	18.52	440
7	23658358	Han avslører sjekketriksene: Slik får de jenter på kroken	26.04	351
8	23432792	Gjør slutt på dårlige vennskap	19.14	305
9	23602637	Tisser du foran kjæresten?	31.67	300

Topp scorede topic

Show 10 entries Search:

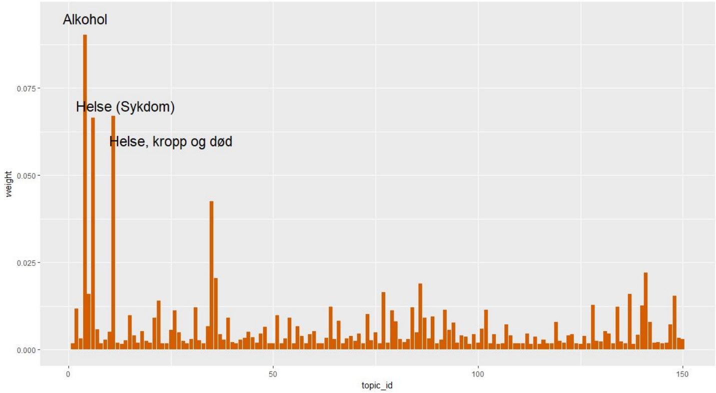
	name	salg	score	antall	Topic
1	Samliv	10207	16.587069	58	Topic 77
2	Sex	9882	17.400556	55	Topic 2
3	Ernring	6912	16.023469	49	Topic 78
4	Mental Helse	6471	10.701310	85	Topic 9
5	Sykdom	5339	10.875915	72	Topic 10
6	Historie	4251	11.159036	83	Topic 23
7	Kroppen	3868	10.928400	25	Topic 86
8	Bil/Motor	3733	8.105775	71	Topic 79
9	Ting du ikke visste om utlandet	3278	11.546389	36	Topic 139
10	Familie & Barn	2897	12.691538	27	Topic 29

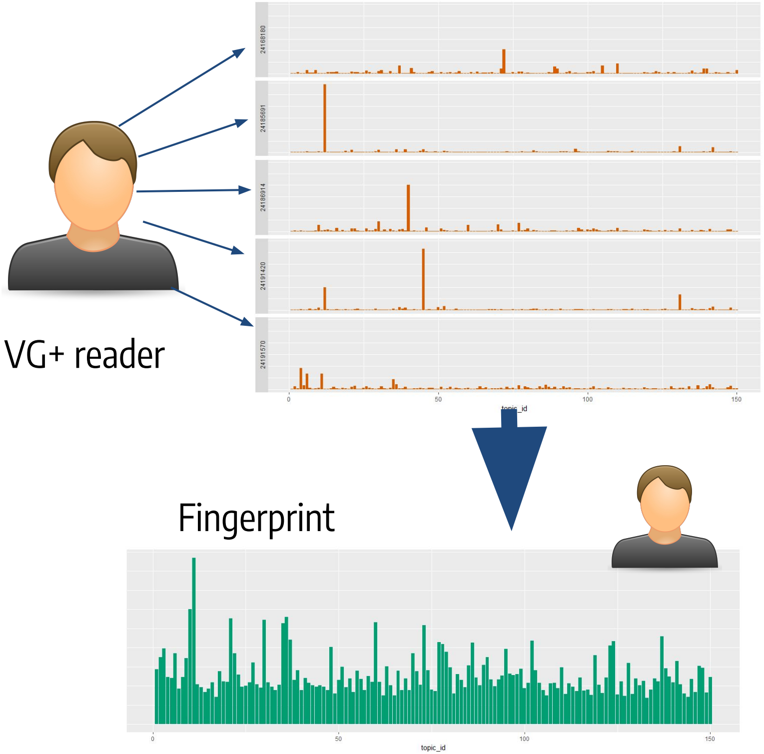
HUNTING RELEVANT PREMIUM ARTICLES

Insert an open VG article

Finds the most relevant premium articles based on their textual content

GENERATING A USER "FINGERPRINT"





Predicting gender with user fingerprint

- Predict gender, age
- Prediction on gender with 76% accuracy
- Age did not predict as well
- Not in production

Learning Resources for LDA

- https://github.com/trinker/topicmodels_learning

- Dirichlet function in R: https://www.rdocumentation.org/packages/DirichletReg/versions/0.6-3/topics/Dirichlet

- Dirichlet wikipedia page: https://en.wikipedia.org/wiki/Dirichlet_distribution

- Professor Blei KDD Tutorial:
http://www.ccs.neu.edu/home/jwvdm/teaching/cs6220/fall2016/assets/pdf/blei-kdd-tutorial.pdf

- Professor Blei lectures on Topic models at Machine Learning Summer School (MLSS), Cambridge 2009 part 1 & 2 with slides: http://videolectures.net/mlss09uk_blei_tm/

- Introduction into Latent Dirichlet Allocation by Professor Bobby B. Lyle at SMU School of Engineering URL:
https://pdfs.semanticscholar.org/presentation/7f54/8af3930a4f10a012a46bc7956ac6da8c38e3.pdf

- Introduction to Markov Chain Monte Carlo: https://nicercode.github.io/guides/mcmc/

Learning resources for udpipe

- udpipe wedsite: http://ufal.mff.cuni.cz/udpipe

- udpipe on github: https://github.com/ufal/udpipe

- vignett for udpipe:
https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-usecase-topicmodelling.html

- R-Bloggers on udpipe:
https://www.r-bloggers.com/is-udpipe-your-new-nlp-processor-for-tokenization-parts-of-speech-tagging-lemmatization-and-dependency-parsing/

References

- Blei DM, Ng AY, Jordan MI (2003b). “Latent Dirichlet Allocation.” Journal of Machine Learning Research, 3, 993–1022, page 1009. URL http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

- Griffiths TL, Steyvers M (2004). “Finding Scientific Topics.” Proceedings of the National Academy of Sciences of the United States of America, 101, 5228–5235. URL http://psiexp.ss.uci.edu/research/papers/sciencetopics.pdf

- Grün, B. & Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models.. Journal of Statistical Software, 40(13), 1-30.

- Ponweiser M., "Latent Dirichlet Allocation in R", Diploma Thesis, Institute for Statistics and Mathematics, 2012. URL http://epub.wu.ac.at/3558/1/main.pdf

Isabelle Valette
Data Scientist - VG

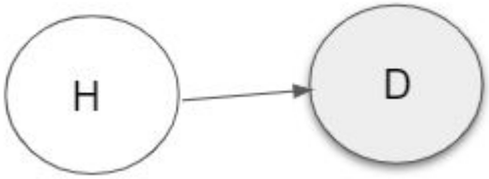
@ZazzValette

Appendix

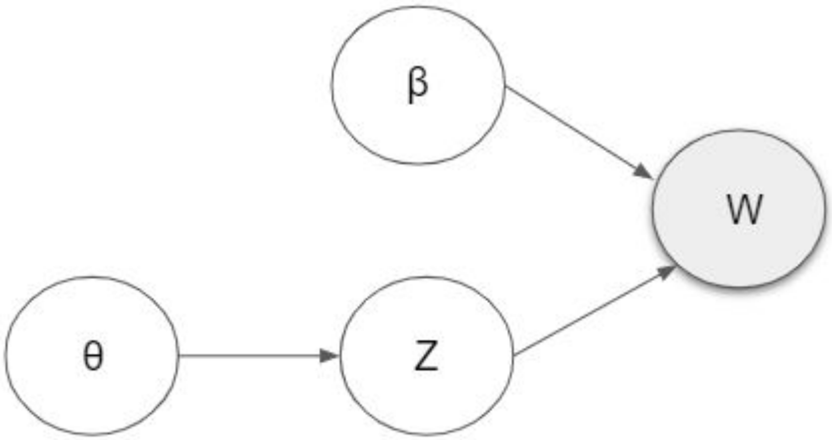
Bayesian Problem

$$PosteriorProbabilityOfAnEvent = \frac{PriorKnowledge * Likelihood}{Evidence(MarginalLikelihood)}$$

$$P(H|D) = \frac{P(H) * P(D|H)}{P(D)}$$



Bayesian Problem to solve in TM



$$P(\theta, z, \beta | w, \eta, \alpha) = \frac{\Pi P(\beta|\eta) * \Pi P(\theta|\alpha) * \Pi P(z|\theta) P(w|z, \beta)}{P(w, \eta, \alpha)}$$

Gibbs sampler

Why Gibbs? Most popular Monte Carlo sampling algorithm - Unbiased, easy to implement, computationally simple, requires little memory and is competitive in speed and performance