Class 11: Topic Models

MA5953: Web Scraping and Text Mining

Dr. Miriam Sorace

www.miriamsorace.github.io

25 January 2024

Outline of Today's Class

Topic Models

Topic Modelling in R

Working on the Assignment

Topic Models

Topic Models

What for?

- Introduced by David Blei et al. in 2003
- Aims:
  1. organise vast text collections into *unknown* themes
     - **unsupervised** text mining method
  2. label individual documents in the collection

Topic Models

Logic

- It discovers the latent, unknown themes on the basis of tightly co-occurring words via the known documents: **posterior** inference
  - Latent Dirichlet Allocation/Distribution: probabilistic statistical model of the distribution of topics and words.
  - LDA basically de-composes the document-feature matrix, and is akin to principal component analysis - a dimensionality reduction technique
  - the algorithm 'searches for'/ 'discovers' topics from word usage patterns in the documents

Topic Models
Logic

- In very simple terms, the LDA algorithm follows 2 key rules:
  - Assign tokens of the same word to the same topic
  - Assign tokens found in the same document to the same topic
- This results in words frequently occurring in the same document to form a 'cluster'.

Topic Models

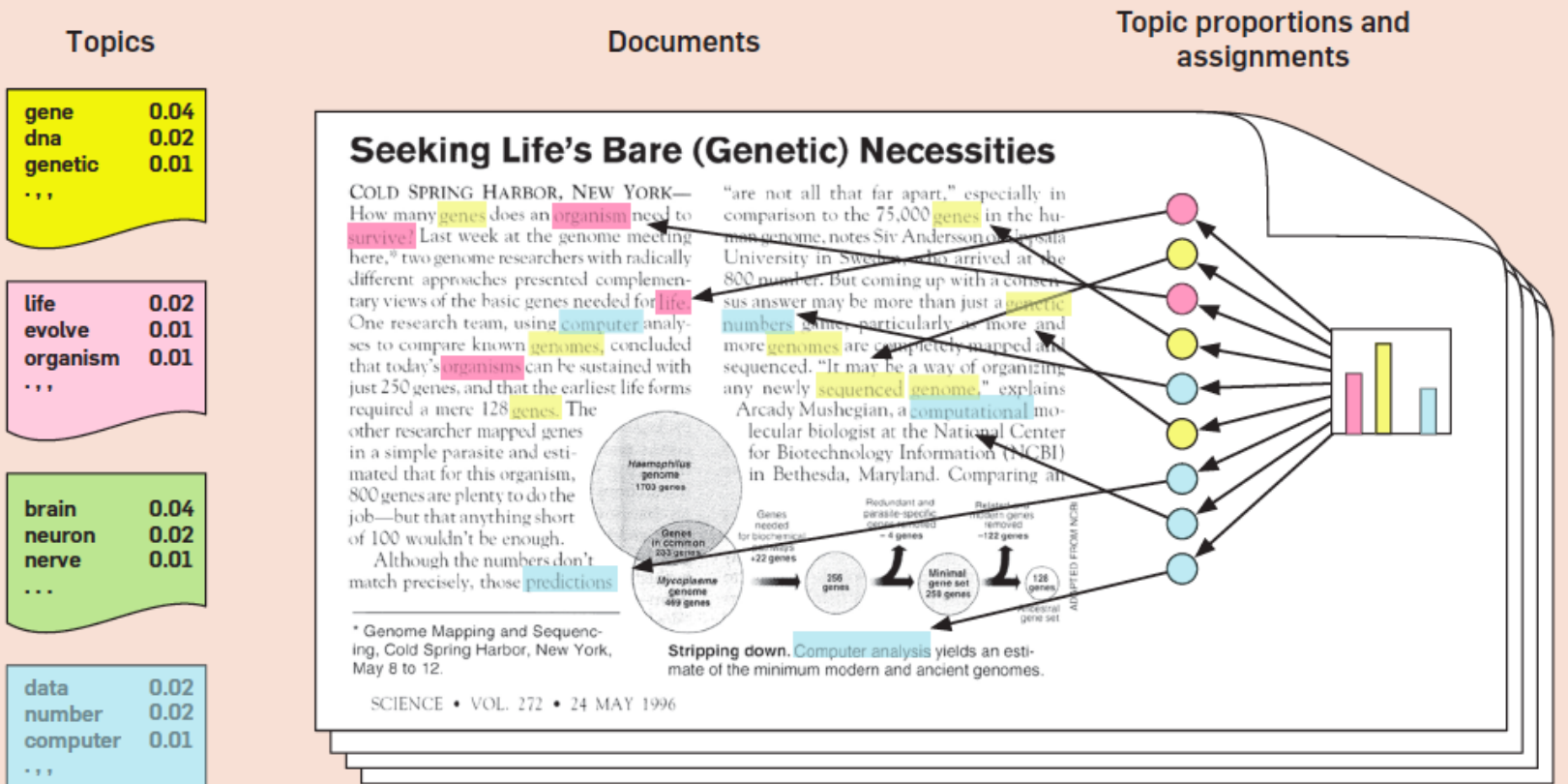
Assumptions

- Topics assumed to be antecedent to document generation & shared across entire set of documents
  - Researcher needs to ask for a pre-determined of topics
    - this is an arbitrary decision, a model selection research problem to be solved with trial & error
    - check if topics discovered are coherent enough
    - formal model diagnostics also exist (held-out methods, topic coherence/exclusivity measures).
- Documents assumed to contain several topics in differing proportions

Topic Models

From: Blei (2012) Probabilistic Topic Models *Communications of the ACM*, 55(4)

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



Topic Models

Assigning a topic to a document - From: Blei et al. (2003). Latent dirichlet allocation.*Journal of machine Learning research*,3

"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

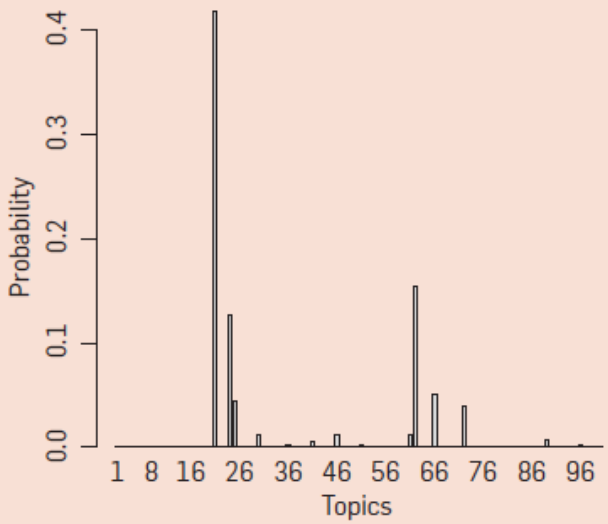
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our 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.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Topic Models

From: Blei (2012) Probabilistic Topic Models *Communications of the ACM*, 55(4)

Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Topic Models

Steps

- Run the algorithm on the dfm
- Interpret the results:
  - Top words
  - Top documents
- Validate
  - Correlational validity: compare human-coding with topic model results
  - Predictive validity: check association between topics and relevant events/meta-data
  - Summarize the results

Topic Models

Labelling Step - From: Boussalis, McElroy & Sorace (*working paper*) Exploring the Conditional Nature of the Descriptive-Substantive Link in the Representation of Women

Topic ID	Topic Label	Prevalence	Time Slice	Top 10 Terms
1	Policy Review & Government Programmes	0.073	1975	department government state policy number development staff body public programme fund
			1985	department government state programme development number agency body public_service staff policy
			2015	government new programme also work support service department need ensure year
2	Tax Administration	0.022	1975	tax income_tax revenue_commissioners year rate relief capital_gains pay taxation income finance
			1985	tax year revenue_commissioners finance taxpayer amount income_tax respect income cost vat
			2015	tax ireland revenue finance rate credit_unions central_bank fund credit_union irish revenue_commissioners
3	**Banking / Financial Crisis**	0.017	1975	company loan bank money state building_societies investment society share interest finance
			1985	company bank state investment loan money fund interest account corporation irish
			2015	bank irish_water government water nama mortgage debt people central_bank loan water_charges
4	**Defense**	0.009	1975	army defence defence_forces officer vessel service force men personnel member duty
			1985	limerick west defence army defence_forces irish_shipping service number force aer_lingus airport
			2015	defence_forces defence aer_lingus personnel member force shannon_airport mission shannon airport military
5	Procedural / Points of Order	0.022	1975	deputy question chair matter order must house amendment minister time bill
			1985	deputy question chair order matter house minister time amendment debate would
			2015	deputy time question agreed bill please house ask proposal need member
6	Farming & Agricultural Trade	0.016	1975	price irish industry country farmer increase market export import industry_and_commerce eec
			1985	industry price irish market trade agriculture product country commerce_and_tourism export import
			2015	farmer industry market sector agriculture price farm ireland beef food irish
7	**Healthcare**	0.021	1975	board hospital health health_boards service health_board patient person number medical available
			1985	board hospital health service health_board health_boards patient number area available nurse
			2015	patient hospital health treatment doctor service medical group consultant clinical national
8	**Education – Third Level & Vocational**	0.009	1975	parliamentary_secretary dublin central parliamentary_secretarys university education student council course college institution
			1985	education dún_laoghaire student college teacher course third_level fee educational examination council
			2015	education teacher student school college education_and_skills system parent third_level course minister
9	**European Union & British-Irish Relations**	0.024	1975	community eec country ireland agreement council irish policy meeting european british
			1985	community ireland ec agreement irish country european northern_ireland meeting europe would
			2015	ireland eu agreement european_union government country northern_ireland europe european irish people
10	**Foreign Affairs & Human Rights**	0.022	1975	government foreign_affairs country irish convention ireland british_government united_nations matter visit conference
			1985	government foreign_affairs country ireland irish matter visit united_states state united_nations convention
			2015	ireland government eu human_rights irish country also people state international issue
11	**Social Welfare Benefits**	0.013	1975	claim payment appeal decision made paid person_concerned due benefit case unemployment_assistance
			1985	person_concerned mean claim unemployment_assistance case benefit made payment payable entitlement paid
			2015	appeal department social_protection social_welfare application review jobbridge person_concerned decision case scheme

Topic Modelling in R

Core R Functions

- library: `seededlda`
  - `textmodel_lda`
  - `terms`
  - `topics`
- check out:
  - `https://tutorials.quanteda.io/machine-learning/topicmodel/`

R Functions: Structural Topic Model

- Variation that allows to include covariates/meta-data in the estimation step. This is what we'll be using in this module.
- library: `stm`
  - `stm`
  - `labelTopics`
  - `top_docs`
- check out:
  - `https://cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf`
- R Code

Working on the Assignment

Assignment 2

- Due March 18th, 2024 **at 4pm!** - 1,000 words report
  1. Scrape speeches (from Hansard website) from 2 politicians/political actors of your choice (from ideologically opposed camps)
  2. Transform the texts into a dfm and explore/summarise the corpus using data visualisations as well! Give technical interpretations of the results.
  3. Use Topic Models, explain in the report what the method does/achieves!
  4. Perform all the necessary steps for the text analysis method chosen and write-up the results - use numerical summaries, tables, graphs ...

Assignment

- Due March 18th, 2024 **at 4pm!** - 1,000 words report
- You can submit in advance if you'd like!
- For issues with submission, and to request extensions/mitigation contact CEMS Student Support: cemssupport@kent.ac.uk

Tips

Common Mistakes from Assessment 1

- Not following the lines of code presented in class - and/or not troubleshooting R errors correctly (StackOverflow is a life-saver + do use the Discussion Forum on the Course Moodle Page!
- Failing to respond/understand the assessment brief: do what is required of you.
  - Read the assessment prompt carefully, there is a lot of detail on the Couse Moodle Page!
- Uncritical usage/application of the R code learnt. **NB:** aim is NOT to see whether you can copy/paste lines of code exactly in the order I presented them! Relevant lines of codes are presented in different classes, some lines of code are duplicated from one class to the next! Show that you've understood what each line of code does and can apply it when it suits you.
  - **Study the code and the notes I provide on it in the .Rmd files or go back to the lecture recordings**.
  - Practice every week what we've learnt in class. Do not leave it to the week your assignment is due!

Lab Work

Preparing the Report

- Let's work together on your R code!
- Feel free to ask any question!