

Class 11: Topic Models

MA5953: Web Scraping and Text Mining

Dr. Miriam Sorace

www.miriamsorace.net

25 February 2022

Outline of Today's Class

Topic Models

Topic Modelling in R

Learning Outcomes

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

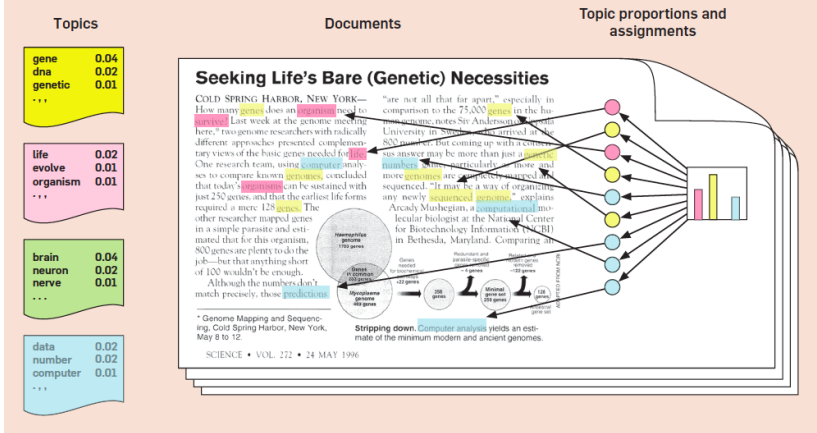
Assumptions

- ▶ 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.

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

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 1985 2015	department government state policy number development staff body public programme fund department government state programme development number agency body public.service staff policy government new programme also work support service department need ensure year
2	Tax Administration	0.022	1975 1985 2015	tax income.tax revenue.commissioners year rate relief capital.gains pay taxation income finance tax year revenue.commissioners finance taxpayer amount income.tax respect income cost vat tax ireland revenue finance rate credit.unions central_bank fund credit_union irish revenue.commissioners
3	Banking / Financial Crisis	0.017	1975 1985 2015	company loan bank money state building.societies investment society share interest finance company bank state investment loan money fund interest account corporation irish bank irish.water government water nama mortgage debt people central_bank loan water.charges
4	Defense	0.009	1975 1985 2015	army defence defence.forces officer vessel service force men personnel member duty limerick west defence army defence.forces irish.shipping service number force aer.lingus airport defence.forces defence aer.lingus personnel member force shannon.airport mission shannon airport military
5	Procedural / Points of Order	0.022	1975 1985 2015	deputy question chair matter order must house amendment minister time bill deputy question chair order matter house minister time amendment debate would deputy time question agreed bill please house ask proposal need member
6	Farming & Agricultural Trade	0.016	1975 1985 2015	price irish industry country farmer increase market export import industry.and.commerce eec industry price irish market trade agriculture product country commerce.and.tourism export import farmer industry market sector agriculture price farm ireland beef food irish
7	Healthcare	0.021	1975 1985 2015	board hospital health health.boards service health.board patient person number medical available board hospital health service health.boards health.boards patient number area available nurse patient hospital health treatment doctor service medical group consultant clinical national
8	Education – Third Level & Vocational	0.009	1975 1985 2015	parliamentary.secretary dublin central parliamentary.secretarys university education student council course college institution education dún.laoghaire student college teacher course third.level fee educational examination council
9	European Union & British-Irish Relations	0.024	1975 1985 2015	education teacher student school college education.and.skills system parent third.level course minister community eec country ireland agreement council irish policy meeting european british community ireland ec agreement irish country european northern.ireland meeting europe would
10	Foreign Affairs & Human Rights	0.022	1975 1985 2015	ireland eu agreement european.union government country northern.ireland europe european irish people government foreign.affairs country irish convention ireland british.government united.nations matter visit conference government foreign.affairs country ireland irish matter visit united.states state united.nations convention
11	Social Welfare Benefits	0.013	1975 1985 2015	ireland government eu human.rights irish country also people state international issue claim payment appeal decision made paid person.concerned due benefit case unemployment.assistance person.concerned mean claim unemployment.assistance case benefit made payment payable entitlement paid 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

Learning Outcomes

What did we learn ...

- ▶ The goal and logic of topic modelling
- ▶ How to perform a topic modelling analysis in R

What did we learn - overall

- ▶ Web scraping from HTML pages vs. using Twitter APIs
- ▶ How to transform text into numbers
- ▶ How to do sentiment analysis and topic categorisation in R

Working on the Assignment

Assignment

- ▶ Due March 11th, 2022 **at 4pm!** - 1,000 words report
 1. Scrape tweets from 2 politicians/political actors of your choice (ideally from ideologically opposed camps)
 2. Transform the tweets' texts into a dfm and explore/summarise the corpus
 3. Choose a text analysis method (sentiment analysis or 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 11th, 2022 **at 4pm!** - 1,000 words report
- ▶ Try to submit in advance as experiencing computer issues close to the deadline can mean you have to request a formal extension to CEMS, Moodle won't allow submissions after the deadline.
- ▶ If glitches arise, submit to the late submission box then immediately alert CEMS to explain the late submission.
- ▶ For issues with submission, and to request extensions/mitigation contact CEMS Student Support: cemssupport@kent.ac.uk

Tips

Common Mistakes in 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](#))!
- ▶ 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 [Course GitHub 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!