



# Data Analytics for Social Science

## Topic models

Johan A. Elkink

School of Politics & International Relations  
University College Dublin

28 November 2017



# Outline

---

- 1 Topic models
- 2 Example: Putin & Medvedev
- 3 Afterthoughts
- 4 Project 3



# Topic model

---

A topic model is an **unsupervised** machine learning algorithm that, based on the co-occurrence of words, divides a set of documents in topics.

Words that often occur in the same document, will be in the same topic. If the use of words in two documents is very similar, they are likely to be allocated to the same topics.

All documents will be related to all topics, but to varying degrees.

Unsupervised means that the researcher does not decide on the topics – the computer does.

# Latent Dirichlet Allocation

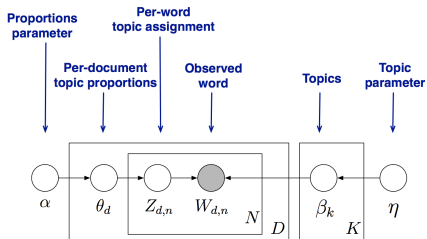


One of the methods for topic modeling is **Latent Dirichlet Allocation (LDA)**, where:

Each **topic** is a distribution over words.

Each **document** is a mixture of topics.

Each **word** is drawn from one of those topics.

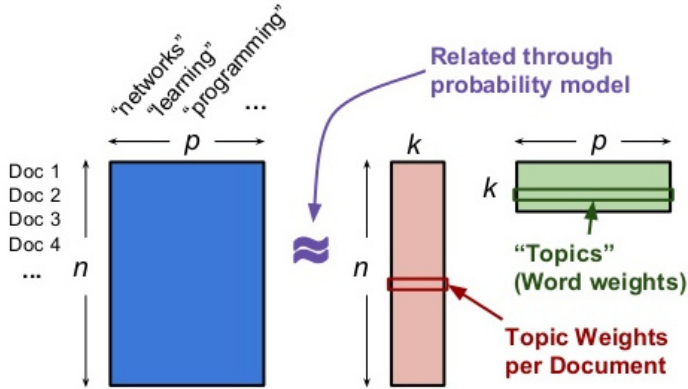




# LDA as matrix decomposition

Digital Representation

Probabilistic Topic Models



- Topic models
- Russia
- Afterthoughts
- Project 3
- References

<http://image.slidesharecdn.com/pres-rpe-apm-2015-150504134806-conversion-gate01/95/admixture-of-poisson-mrfs-a-new-topic-model-with-word-dependencies-13-638.jpg?cb=1430747442>

# LDA optimization

---



There are many possible ways to decompose this matrix. LDA makes the following trade-off:

- 1 For each document, as few topics per word as possible.
- 2 For each topic, as few high probability words as possible.

The trade-off leads to finding the topics where the words are tightly clustered.

<http://yosinski.com/mlss12/media/slides/MLSS-2012-Blei-Probabilistic-Topic-Models.pdf>



# Variations on basic LDA topic modeling

---

Many variations exist, including:

- **Correlated topic models** allows topics co-occurrence to be modeled properly
- **Dynamic topic models** models changing use over time in term usage
- **Supervised topic models** find topics that best predict some observed outcome variable, e.g. reviews
- **Relational topic models** takes into account that linked documents should be similar, e.g. citations or internet links
- **Ideal point topic models** combined document data with voting behaviour to get underlying ideological position
- **Collaborative topic models** find topics that best predict behaviour similar to related individuals, e.g. recommendation software

All much beyond this class. See for more information the slides by David Blei.

# LDA topic model interpretation

---



We can interpret (post-hoc) the output from the topic modeling:

- 1 Identify the meaning of each topic by looking at the terms that have a high probability to be associated with this topic.
- 2 Identify the topic of document by looking at the topics that have highest probability of being associated with this document.
- 3 Look at the use of particular topics over time, by different groups (e.g. parties), etc.
- 4 Identify if there underlying dimensions in the (dis)similarity between topics in terms of term usage.





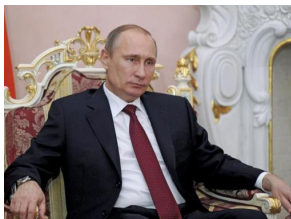
# Outline

---

- 1 Topic models
- 2 Example: Putin & Medvedev
- 3 Afterthoughts
- 4 Project 3

# Russian speeches

---



All speeches of the President of Russia were scraped from the Kremlin website:

<http://www.kremlin.ru/events/president/transcripts/page/>

All speeches of the Prime Minister of Russia for 2008–12 (i.e. V.V. Putin) were scraped from the Premier website:

<http://archive.premier.gov.ru/events/news/>

# Cleaning

---



## After

- removing short words;
- stemming words;
- removing very frequent and very rare words;
- removing stop words;
- removing tiny paragraphs,

we are left with **34,499 paragraphs** in 11,595 documents, with frequencies on 3,282 terms.

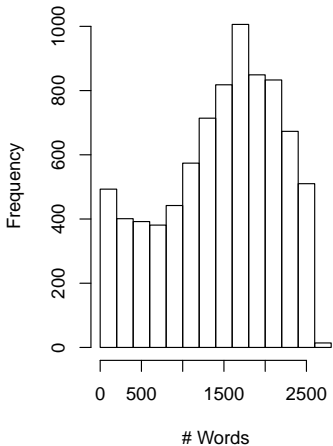
Of the 7,587 documents, 3,015 are from the Prime Minister data, 4,572 from the President data.

Of the 34,499 paragraphs, 9,430 are attributed to Putin, 2,375 to Medvedev; 18,873 are from the presidential data, 15,626 from the premier data.

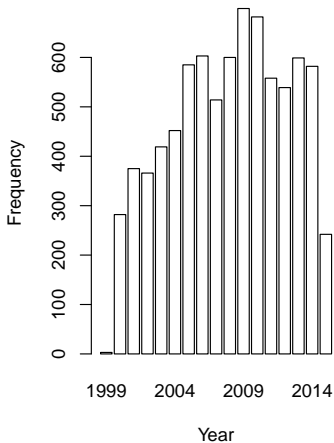
# Texts distribution



### Length of speech in words



### Speeches by year



# Topic modeling output

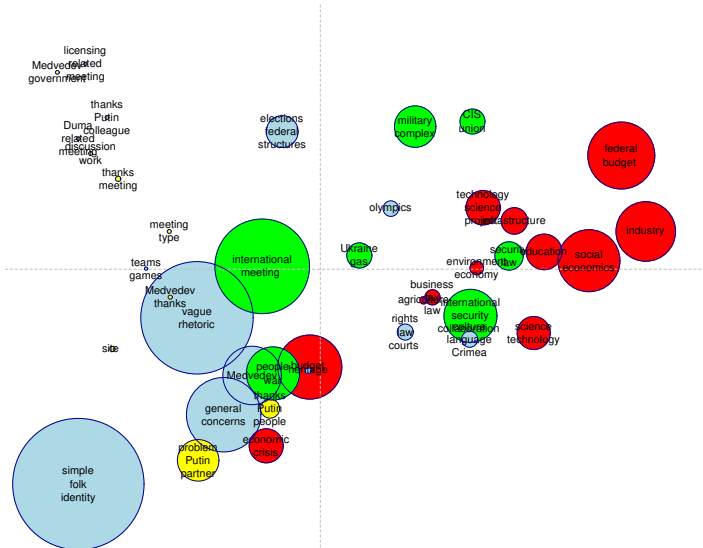
---

The highest probability terms for the first 10 of 40 topics are as follows:

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
defense military forces armed technician	states people w its solved	economies develop tasks social growth	company invest productions market project	investments investor zones investment environmental
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
farms earth land state council rural	negotiations worker problems meetings affordable	waspi vladimir vladimir finish rescue	developed roads project east transport	vladimirovich vladimir colleagues rescue groups



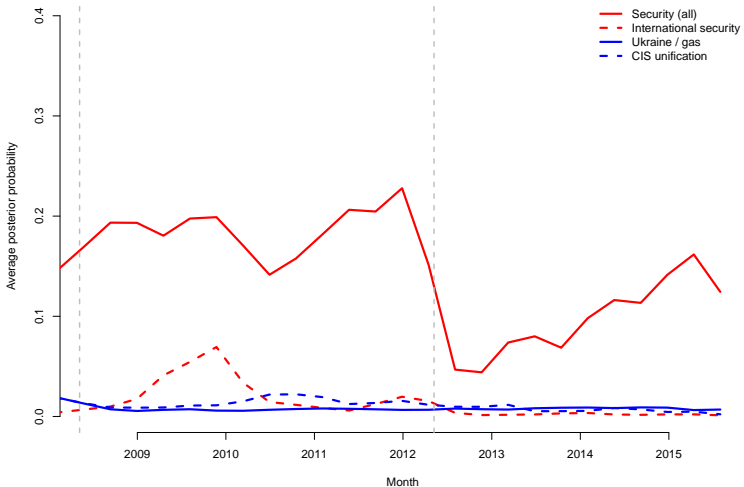
# MDS on topic-term matrix





# Putin on high politics

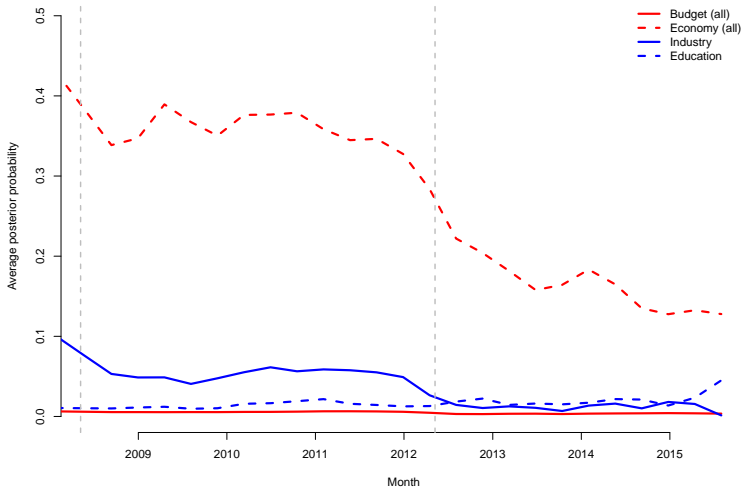
Relative prominence of Putin on some high politics topics





# Putin on economics

Relative prominence of Putin on some low politics topics







# Outline

---

- 1 Topic models
- 2 Example: Putin & Medvedev
- 3 Afterthoughts**
- 4 Project 3



# What have we learned?

---

## A lot!

- Basic skills in R and Markdown.
- Visualisation of statistical data.
- Basic regression, classification, and dimension reduction techniques.
- Both supervised and unsupervised learning methods.
- Using survey data, country data, and text as data.

# Issues to watch out for

---



- Models based on theoretical expectations (usually supervised) versus post-hoc interpretation of model output (usually unsupervised).
- Models that report uncertainty and allow for statistical inference versus models that are specific to the data set.
- Prediction versus description versus causal inference.





# What have we not really covered?

---

- Conventional econometric theory, the fundamentals behind ordinary least squares and maximum likelihood.
- Bayesian statistical inference.
- Methods for causal inference.
- Database access and management, such as SQL.
- General programming languages, such as Python.
- Big data and parallel computation.

We also did not cover anything related to “domain knowledge,” but this is what you get in all other modules!

We have roughly covered what a typical **multivariate analysis** textbook will cover, ignoring the technicalities of an **econometrics** textbook and the computer science technicalities of **data science**.



# MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21st century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

## MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

## DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

## PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

## COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau





# Outline

---

- 1 Topic models
- 2 Example: Putin & Medvedev
- 3 Afterthoughts
- 4 Project 3**

# Project 3: assignment

---



Based on Labs 9–12:

- Choose one of the following two to focus your essay:
  - ① What do presidents talk about in speeches to joint sessions of Congress since 1945? Are there any noticeable patterns?
  - ② Can we use the statistical text data to detect the ideological positioning of presidents, which can include left-right, liberal-conservative, nativist-globalist, etc.?
- Choose one or two approaches from: cluster analysis, principal components analysis, factor analysis, wordscores, wordfish, or topic modelling.
- Word clouds and other visualisations can be used in addition to, but not instead of, the above.



## Project 3: essay structure

---

Write a 2,000-2,500 word essay (excluding captions, tables and bibliography from the count), including the graphs and tables.

Remember that this is an essay about the ideological positioning of State of the Union addresses and presidents, not about the methods used.

Each essay should consist of

- 1 a short introduction,
- 2 a description and motivation of the data and methods used (approximately 25% of the essay),
- 3 the analysis including necessary graphs and tables (approximately 35%), and
- 4 an interpretation and conclusion (approximately 40%).

Everything needs to be properly referenced.





# Project 3: practicalities

---

## Literature

A good starting point for references is probably to search for terms like “state of the union”, “ideological space”, “ideal point” or “spatial model of politics” in Google Scholar (<https://scholar.google.com/>).

Note that you significantly strengthen your essay by referencing two or three (at least) academic references that discuss the role of ideological spaces in politics and/or the estimation of ideological spaces, typically in the section where you describe your model (before the analysis).

## Markdown

Best is to make a separate Markdown file for the analysis required for the assignments, which can be copy/pasted and then edited from the lab sheets. You can potentially open the HTML file in your browser and save as, or print to, PDF.

## Project 3: submission

---



All assignments should be submitted electronically to `jos.elkink@ucd.ie`, consisting of either:

- a PDF file containing the essay and an R- or Rmd-file with all the commands necessary for the analysis, or
- a PDF file generated using RMarkdown with the essay and all necessary commands and results.

Recommendation: the deadline is **12 December, 1 pm**.



Feinerer, I and K Hornik. 2015. "Package tm: A framework for text mining applications within R." *R Foundation for Statistical Computing, Vienna* .

Feinerer, Ingo, Kurt Hornik and David Meyer. 2008. "Text mining infrastructure in R." *Journal of statistical software* 25(5):1–54.

Grün, Bettina and Kurt Hornik. 2011. "topicmodels: An R package for fitting topic models." *Journal of Statistical Software* 40(13).