Machine Learning, Artificial Intelligence, and Natural Language Processing

Marcelo C. Medeiros

Department of Economics
The University of Illinois at Urbana-Champaign

Natural Language Processing (NLP)

- NLP is a discipline that intersects computer science, linguistics, and artificial intelligence.
- NLP focuses on how computers comprehend, interpret, and handle human languages.
- It can involve things such as interpreting the semantic meaning of a language, translating between human languages, or recognizing patterns in human languages.
- It uses statistical methods, machine learning, and text mining.

"The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades [...]. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.

—Hal Varian, The McKinsey Quarterly, January 2009

Motivation

What is the dataset of interest?

- News from newspapers
- Reports, general documents, ...
- Social media posts
- Central Banks's communications
- Books, papers, etc
- ...

Why should we care about text?

Example: Economic Policy Uncertainty (EPU)

- Index proposed by Baker, Bloom, and Davies (Quarterly Journal of Economics, 2016)
- A measure of policy-related economic uncertainty
- Quantifies newspaper coverage of policy-related economic uncertainty

THE
QUARTERLY JOURNAL
OF ECONOMICS

Vol. 131 November 2016 Issue 4

MEASURING ECONOMIC POLICY UNCERTAINTY*

SCOTT R. BAKER
NICHOLAS BLOOM
STEVEN J. DAVIS

We develop a new index of economic policy uncertainty (EPU) based on newspaper coverage frequency. Several types of evidence—including human readings of 12,000 newspaper articles—indicate that our index proxies for movements in policy-related economic uncertainty. Our U.S. index spikes near tight presidential elections, Gulf Wars I and II, the 9/11 attacks, the failure of Lehman Brothers, the 2011 debt ceiling dispute, and other major battles over fiscal policy. Using firm-level data, we find that policy uncertainty is associated with greater stock price volatility and reduced investment and employment in policy-sensitive sectors like defense, health care, finance, and infrastructure construction. At the macro level, innovations in policy uncertainty foreshadow declines in investment, output, and employment in the United States and, in a panel vector autoregressive setting, for 12 major economies. Extending our U.S. index back to 1900, EPU rose dramatically in the 1930s (from late 1931) and has drifted upward since the 1960s. JEL Codes: D80, E22, E66, G18, L50.

*We thank Adam Jorring, Kyle Kost, Abdulla Al-Kuwari, Sophie Biffar, Jörn Boehnke, Vladimir Dashkeyev, Olga Deriy, Eddie Dirk, Yuto Enzo, Robin Gong, Sonam Jindal, Ruben Kim, Sylvia Klosin, Jessica Koh, Peter Lajewski, David Nekiyu, Rebecca Sacha, Ippei Shibata, Corinne Stephenson, Naoko Takeda, Melissa Tan, Sophie Wang, and Peter Xu for research assistance and the National Science Foundation, MacArthur Foundation, Sloan Foundation, Becker Friedman Institute, Initiative on Global Markets, and Stigler Center at the University of Chicago for financial support. We thank Raedi Bachmann, Sanjai Bhagat, Vincent Bignon, Yongsung Chang, Vladimir Dashkeyev, Jesus Fernandez-Villaverde, Laurent Ferrara, Luis Garicano, Matt Gentzkow, Yuriy Gorodnichenko, Kevin Hassett, Takeo Hoshi, Greg Ip, Anil Kashyap, Patrick Kehoe, John Makin, Johannes Pfeifer, Meijun Qian, Itay Saporta, John Shoven, Sam Schulhofer-Wohl, Jesse Shapiro, Erik Sims, Stephen Terry, Cynthia Wu, and many seminar and conference audiences for comments. We also thank the referees and editors, Robert Barro and Larry Katz, for comments and suggestions.

The Quarterly Journal of Economics (2016), 1593–1636. doi:10.1093/qje/qjw024
Advance Access publication on July 13, 2016.

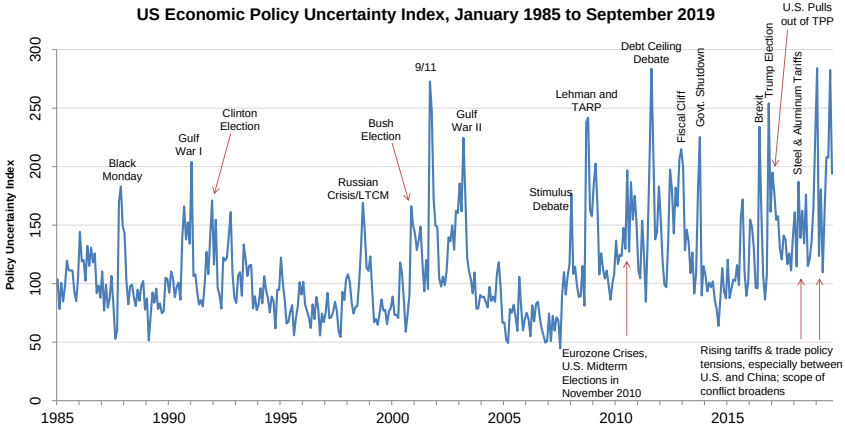
Why should we care about text?

Example: Economic Policy Uncertainty (EPU)

- For the US, the index is mostly based on a normalized index of search results from 10 large newspapers
  - USA Today, the Miami Herald, the Chicago Tribune, the Washington Post, the Los Angeles Times, the Boston Globe, the San Francisco Chronicle, the Dallas Morning News, the New York Times, and the Wall Street Journal
- Extensions to many countries
- Used as input to several econometric/economic models

Why should we care about text?

Example: Economic Policy Uncertainty (EPU)

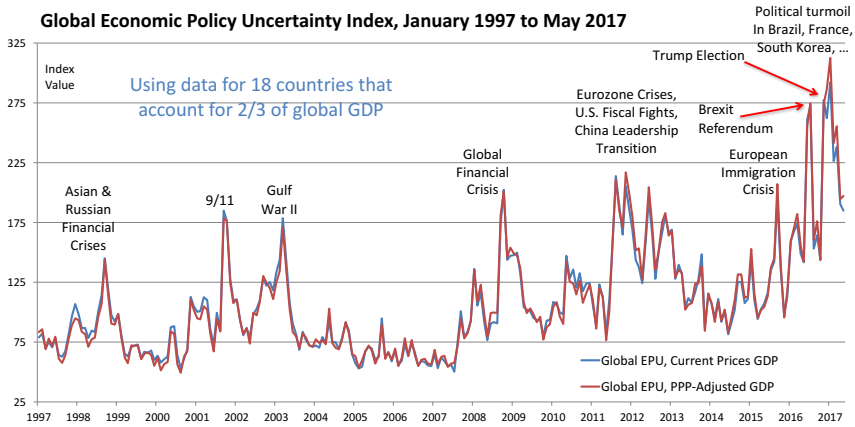


Source: "Measuring Economic Policy Uncertainty" by Scott R. Baker, Nicholas Bloom and Steven J. Davis, as updated at www.policyuncertainty.com. Monthly data normalized to 100 prior to 2010.

Source: http://www.policyuncertainty.com and Baker, Bloom, and Davis (QJE, 2016).

Why should we care about text?

Example: Economic Policy Uncertainty (EPU)

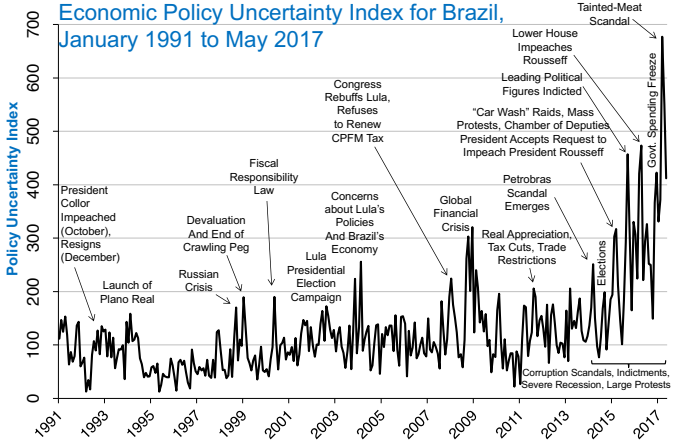


Notes: Global EPU calculated as the GDP-weighted average of monthly EPU index values for US, Canada, Brazil, Chile, UK, Germany, Italy, Spain, France, Netherlands, Russia, India, China, South Korea, Japan, Ireland, Sweden, and Australia, using GDP data from the IMF's World Economic Outlook Database. National EPU index values are from www.PolicyUncertainty.com and Baker, Bloom and Davis (2016). Each national EPU Index is renormalized to a mean of 100 from 1997 to 2015 before calculating the Global EPU Index.

Source: http://www.policyuncertainty.com and Baker, Bloom, and Davis (QJE, 2016).

Why should we care about text?

Example: Economic Policy Uncertainty (EPU)



Notes: Index reflects scaled monthly counts of articles in Folha de São Paulo containing "incerto" or "incerteza", "econômico" or "economia", and one or more policy-relevant terms that include regulação, déficit, orçamento, imposto, "banco central", planalto, congresso, senado, legislação, and tarifa. Normalized to a mean of 100 from 1991 to 2011. Index methods follow "Measuring Economic Policy Uncertainty" by Baker, Bloom and Davis. Data available at www.PolicyUncertainty.com.

Source: http://www.policyuncertainty.com and Baker, Bloom, and Davis (QJE, 2016).

Inflation Forecasting

Even structured (real-time) data can provide better forecasts…

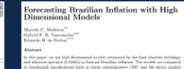
… when carefully estimated ML models are considered


Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods


Real-time inflation forecasting with high-dimensional models: The case of Brazil


Global Inflation Forecasting

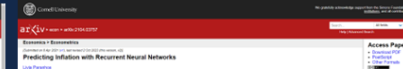

Nowcasting food inflation with a massive amount of online prices


Forecasting Brazilian Inflation with High Dimensional Models


How is machine learning useful for macroeconomic forecasting?


Forecasting CPI inflation components with Hierarchical Recurrent Neural Networks


Forecasting inflation using disaggregates and machine learning


Predicting Inflation with Recurrent Neural Networks


Real-time inflation forecasting using non-linear dimension reduction techniques

Inflation Forecasting

However, news and social-media-based indexes can help further

The art of transforming text into numbers

Inflation forecasting using unstructured data: the benefits of indexes based on Twitter and newspapers

Gilberto Boaretto
Economics, PUC-Rio & Vox Baker
gilbertoboaretto@gmail.com

Marcelo Fernandes
FGV-EPGE
marcelo.fernandes@fgv.br

Marcelo C. Medeiros
Economics, PUC-Rio
mcm@econ.puc-rio.br

Thiago Mitagres
Vox Baker
thiagogaes23agres@gmail.com

This draft: August 14, 2021
Preliminary. Please do not share.

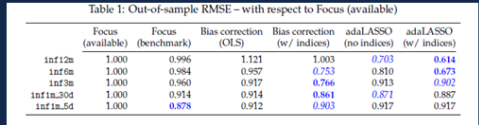
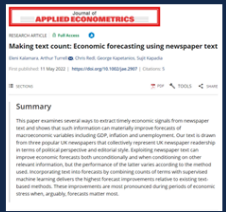


Table 1: Out-of-sample RMSE – with respect to Focus (available)

	Focus (available)	Focus (benchmark)	Bias correction (OLS)	Bias correction (w/ indices)	adaLASSO (no indices)	adaLASSO (w/ indices)
inf12m	1.000	0.996	1.121	1.003	*0.703*	**0.614**
inf6m	1.000	0.984	0.957	*0.753*	0.810	**0.673**
inf3m	1.000	0.960	0.917	**0.766**	0.913	*0.902*
inf1m_30d	1.000	0.914	0.914	**0.861**	*0.871*	0.887
inf1m_5d	1.000	**0.878**	0.912	*0.903*	0.917	0.917



Journal of APPLIED ECONOMETRICS

RESEARCH ARTICLE Full Access

Making text count: Economic forecasting using newspaper text

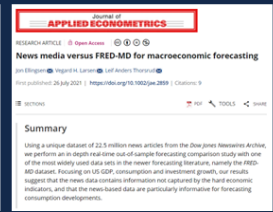
Dave Kazlaowsky, Arthur Turrell, Chris Redl, George Kapetanios, Sujit Kapadia

First published: 11 May 2022 | https://doi.org/10.1002/jae.2907 | Citations: 5

SECTIONS PDF TOOLS SHARE

Summary

This paper examines several ways to extract timely economic signals from newspaper text and shows that such information can materially improve forecasts of macroeconomic variables including GDP, inflation and unemployment. Our test is drawn from three popular UK newspapers that collectively represent UK newspaper readership in terms of political perspective and editorial style. Exploiting newspaper text can improve economic forecasts both unconditionally and when conditioning on other relevant information, but the performance of the latter varies according to the method used. Incorporating text into forecasts by combining counts of terms with supervised machine learning delivers the highest forecast improvements relative to existing text-based methods. These improvements are most pronounced during periods of economic stress when, arguably, forecasts matter most.



Journal of APPLIED ECONOMETRICS

RESEARCH ARTICLE Open Access

News media versus FRED-MD for macroeconomic forecasting

Jon Ellingsen, Vegard H. Larsen, Leif Anders Thorsrud

First published: 26 July 2021 | https://doi.org/10.1002/jae.2859 | Citations: 9

SECTIONS PDF TOOLS SHARE

Summary

Using a unique dataset of 22.5 million news articles from the *Dow Jones Newswires Archive*, we perform an in-depth real-time out-of-sample forecasting comparison study with one of the most widely used data sets in the newer forecasting literature, namely the *FRED-MD* dataset. Focusing on US GDP, consumption and investment growth, our results suggest that the news data contains information not captured by the hard economic indicators, and that the news-based data are particularly informative for forecasting consumption developments.

Text Selection

Bryan Kelly, Asaf Manela & Alan Moreira

Pages 859-879 | Published online: 25 Jun 2021

Cite this article https://doi.org/10.1080/07350015.2021.1947843 Check for updates

Full Article Figures & data References Supplemental Citations Metrics Reprints & Permissions

Read this article

Abstract

Text data is ultra-high dimensional, which makes machine learning techniques indispensable for textual analysis. Text is often selected—journalists, speechwriters, and others craft messages to target their audiences' limited attention. We develop an economically motivated high-dimensional selection model that improves learning from text (and from sparse counts data more generally). Our model is especially useful when the choice to include a phrase is more interesting than the choice of how frequently to repeat it. It allows for parallel estimation, making it computationally scalable. A first application revisits the partisanship of the U.S. congressional speech. We find that earlier spikes in partisanship manifested in increased repetition of different phrases, whereas the upward trend starting in the 1990s is due to distinct phrase selection. Additional applications show how our model can backcast, nowcast, and forecast macroeconomic indicators using newspaper text, and that it substantially improves out-of-sample fit relative to alternative approaches.

Journal of Econometrics

Volume 228, Issue 2, June 2022, Pages 259-277

Can we measure inflation expectations using Twitter?

Cristina Angelico, Jan Marcucci, Marcello Miccoli, Filippo Quarta

Show more

Add to Mendeley Share Cite

https://doi.org/10.1016/j.jeconom.2020.12.008

Get rights and content

Abstract

Drawing on Italian tweets, we employ textual data and machine learning techniques to build new real-time measures of consumers' inflation expectations. First, we select keywords to identify tweets related to prices and expectations thereof. Second, we build a set of daily measures of inflation expectations around the selected tweets, combining the Latent Dirichlet Allocation (LDA) with a dictionary-based approach, using manually labeled bi-grams and tri-grams. Finally, we show that Twitter-based indicators are highly correlated with both monthly survey-based and daily market-based inflation expectations. Our new indicators anticipate consumers' expectations, proving to be a good real-time proxy, and provide additional information beyond market-based expectations, professional forecasts, and realized inflation. The results suggest that Twitter can be a new timely source for eliciting beliefs.

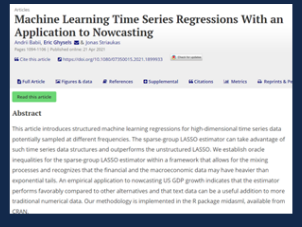
Inflation Forecasting

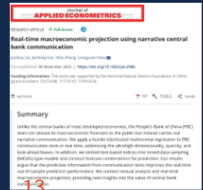
The age of AI

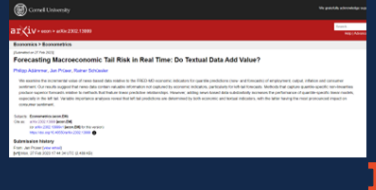
The role of large language models

Beyond Inflation Forecasting

Nowcasting







Transforming text into numbers

- Computers and ML models are able to process numbers
- We must transform text data into numbers

Basic Workflow for Text Analysis

1. Pre-process and represent raw text $\mathcal{D}$ as a numerical array $\boldsymbol{C}$
    * From text to numbers
2. Apply models to map $\boldsymbol{C}$ to predicted outcomes $\widehat{\boldsymbol{V}}$
3. Use $\widehat{\boldsymbol{V}}$ for descriptive or causal inference

- Web pages as HTML or XML files
- Noneditable formats as pdf
- Editable formats as .docx.
- Application Programming Interfaces (API)
- Optical Character Recognition for text extraction
- Manual extraction

- Textual data are inherently high-dimensional
  - Collection $\mathcal{D}$ of D documents, each with N words
  - Each document has the same number of words for simplicity
  - If each word is drawn from a vocabulary of p-possible words, then the unique representation of these documents has dimension p^N
- A simple tweet with $N = 280$ words can have a p^{280}-dimension
  - The length of a basic English speaker is about $10,000$ words

- How do we deal with such huge-dimensional data?
  - Computational capacity quickly escalates without any prior cleaning
- Cleaning is fundamental to organize textual data
  - Assumptions regarding useful or useless words need to be made
  - The order of the cleaning process may differ through different applications

- Divide the raw corpus of text into individual documents
- Level of aggregation depends on the application
  1. If the objective is to nowcast the GDP, and documents are news, we should aggregate by a reference period, e.g. day, week, month
  2. If we are interested in Central Bank Communication, it may be interesting to treat each communication isolated
- Aggregation level may not be straightforward
  - If we analyze news from several newspapers, do we aggregate all articles by day or treat them individually?

- **Tokenization**: Breaking text into words or phrases
- **Lowercasing, removing punctuation**
- **Stopword removal**: Remove common words (e.g., "the", "and")
- **Stemming/Lemmatization**: Reduce words to base form (e.g., "running" $\rightarrow$ "run")
- **Feature selection**: Keep only informative

Original text:

"Good night, good night! Parting is such sweet sorrow."

After Preprocessing:

- Remove punctuation, stop words
- Apply stemming
- Result: `[good, night, good, night, part, sweet, sorrow]`

- After tokenizing, cleaning and extracting the linguistic roots, we are able to structure the dara
  - The usual representation of text is made in terms of **counts** of sentences, e.g. bag-of-words or dictionaries
  - Any representation of text will throw away information
  - The goal is to keep the relevant part for the application
- Keep in mind: without simplifications, the problem becomes unfeasible

- The decision whether to aggregate or not articles may depend on the application:
  - If we are interested in forecasting using news-analysis, it may be interesting to aggregate articles in daily-terms.
  - If we are interested in Central Bank communication, it may be interesting to treat each document individually.
- Suppose we are interested in forecasting: If we have a set of different newspapers, should we treat them individually, or aggregate the articles as a single representative newspaper?

- After pre-processing, we are left with a finite list of (potentially non-unique) terms
  - Index each unique term in the corpus by $n \in \{1, \cdots, N\}$ where N is the total number of unique terms
  - For each document $d \in \{1, \cdots, D\}$, compute the **count** $x_{d,n}$ of occurrences of term n in document d
  - The $D \times N$ matrix $\boldsymbol{X}_{D,N}$ of all such counts is called the document-term matrix
- Each article treated separately

- Consider the following extension:
  - Index each unique term of the corpus (overall articles) by $n \in \{1, \cdots, N\}$ where N is the number of unique terms
  - For each article $d \in \{1, \cdots, D\}$ compute the count $x_{d,n}$ as the number of occurrences of term n in document D
  - Aggregate the number of occurrences over all documents $x_{d,n}$ for a given period $t \in (1, \cdots, T)$
  - Generate period-counts $x_{n,t}$
  - The $T \times N$ matrix $\boldsymbol{X}_{T,N}$ of all such counts is called the document-term matrix
- This formulation aggregates articles by a period of reference (e.g., days, weeks, months, ...)

- Bag-of-words: the order of words is ignored altogether, and a phrase of length n is called a n-gram:
- Higher order n-grams encode more complex structures
  - Usually, we consider at most 3-grams

- Dictionary: focus on variation across observations along a limited number of words
- Define a dictionary of terms $\mathcal{H} = \{1, \cdots, H\}$:
  - **a) Boolean:** The dictionary-based boolean $b_{d,n}$ of term n in document d is given by: $b_{d,n} = \mathbb{I}(x_{d,n} \in \mathcal{H})$, where $\mathbb{I}$ is the indicator function
  - **b) Counts:** The dictionary-based count $\bar{x}_{d,n}$ of term n in document d is given by: $\bar{x}_{d,n} = \mathbb{I}(x_{d,n} \in \mathcal{H}) x_{d,n} = b_{d,n} x_{d,n}$, where $\mathbb{I}$ is an indicator function.
- Dictionaries are arbitrary

What are Topic Models?

- Topic models are statistical models for discovering abstract topics in a collection of documents
- They help uncover hidden thematic structures in large text corpora
- Useful for document classification, recommendation systems, and information retrieval

What are Topic Models?

Colection of documents

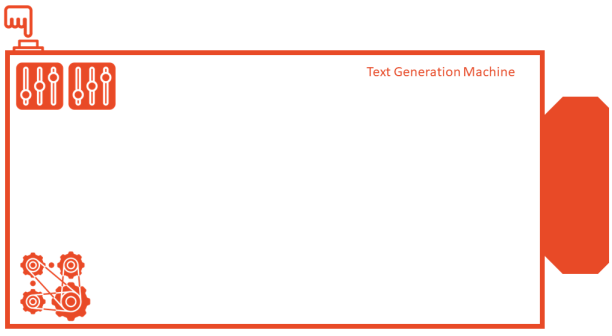
- Documents may consist of one or more topics
- We want the computer to discover the topics of each document

Latent Dirichlet Allocation

- Most popular topic modeling technique
- Assumes each document is a mixture of topics, and each topic is a mixture of words
- Probabilistic generative model

Introduction to Latent Dirichlet Allocation (LDA)

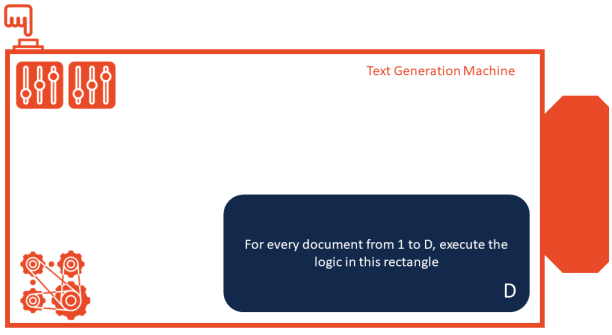
A simple machine to create text



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

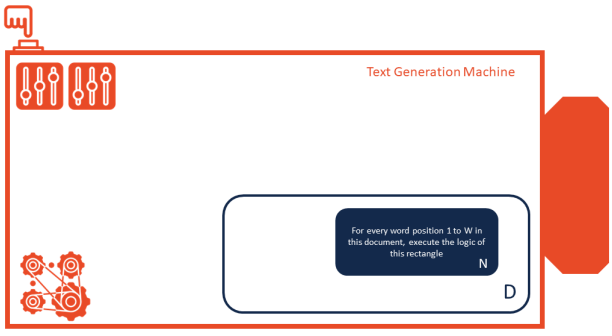
A simple machine to create text



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

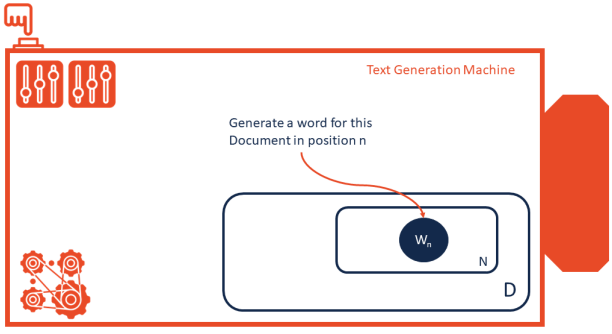
A simple machine to create text



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

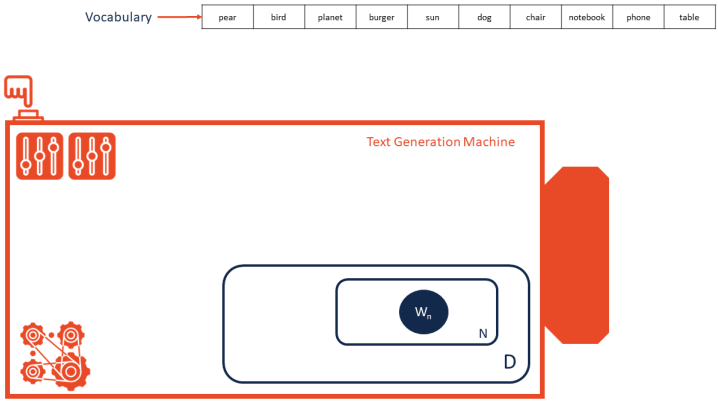
A simple machine to create text – Version 1



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 1



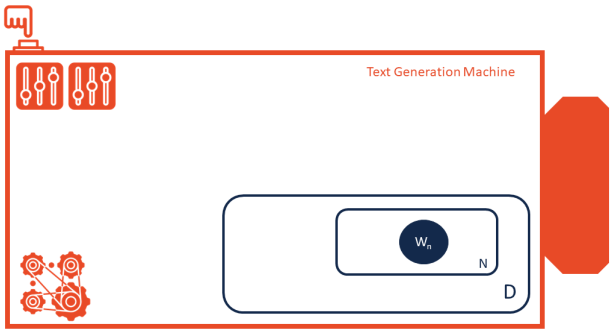
https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 1

Probability distribution

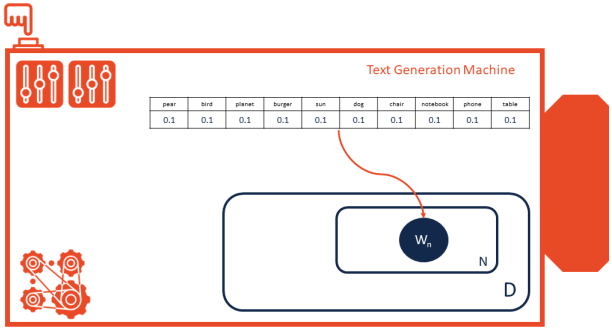
pear	bird	planet	burger	sun	dog	chair	notebook	phone	table
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

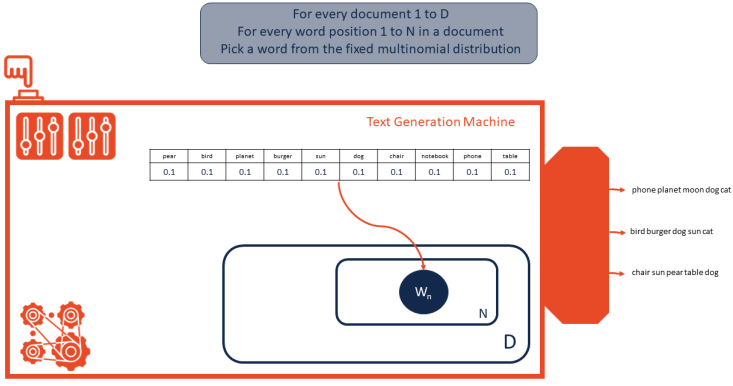
A simple machine to create text – Version 1



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 1

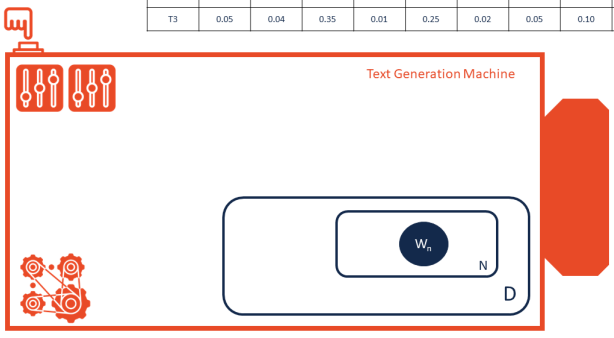


https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 2

Topics	pear	bird	planet	burger	sun	dog	chair	notebook	phone	table
T1	0.07	0.30	0.03	0.01	0.05	0.40	0.04	0.05	0.04	0.01
T2	0.25	0.10	0.01	0.31	0.01	0.15	0.05	0.01	0.02	0.10
T3	0.05	0.04	0.35	0.01	0.25	0.02	0.05	0.10	0.03	0.10

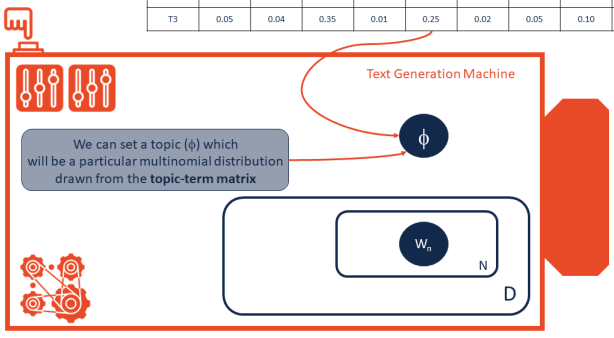


https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 2

Topics	pear	bird	planet	burger	sun	dog	chair	notebook	phone	table
T1	0.07	0.30	0.03	0.01	0.05	0.40	0.04	0.05	0.04	0.01
T2	0.25	0.10	0.01	0.31	0.01	0.15	0.05	0.01	0.02	0.10
T3	0.05	0.04	0.35	0.01	0.25	0.02	0.05	0.10	0.03	0.10

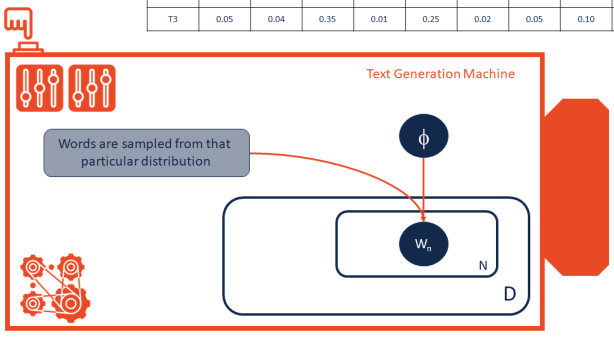


https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 2

Topics	pear	bird	planet	burger	sun	dog	chair	notebook	phone	table
T1	0.07	0.30	0.03	0.01	0.05	0.40	0.04	0.05	0.04	0.01
T2	0.25	0.10	0.01	0.31	0.01	0.15	0.05	0.01	0.02	0.10
T3	0.05	0.04	0.35	0.01	0.25	0.02	0.05	0.10	0.03	0.10

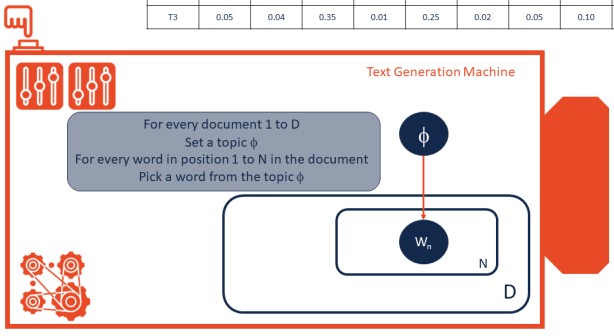


https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 2

Topics	pear	bird	planet	burger	sun	dog	chair	notebook	phone	table
T1	0.07	0.30	0.03	0.01	0.05	0.40	0.04	0.05	0.04	0.01
T2	0.25	0.10	0.01	0.31	0.01	0.15	0.05	0.01	0.02	0.10
T3	0.05	0.04	0.35	0.01	0.25	0.02	0.05	0.10	0.03	0.10

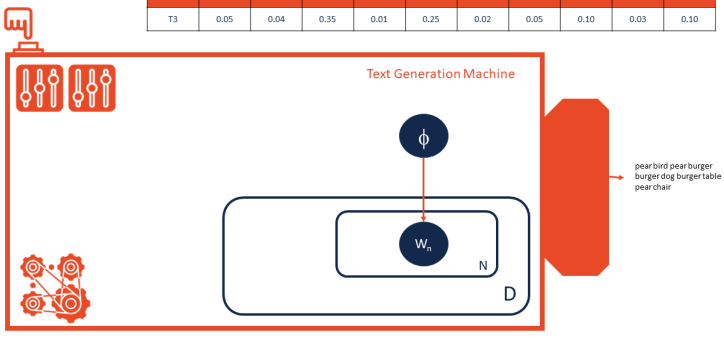


https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 2

Topics	pear	bird	planet	burger	sun	dog	chair	notebook	phone	table
T1	0.07	0.30	0.03	0.01	0.05	0.40	0.04	0.05	0.04	0.01
T2	0.25	0.10	0.01	0.31	0.01	0.15	0.05	0.01	0.02	0.10
T3	0.05	0.04	0.35	0.01	0.25	0.02	0.05	0.10	0.03	0.10

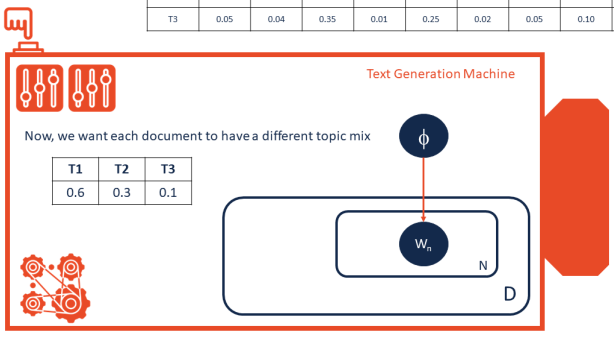


https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 3

Topics	pear	bird	planet	burger	sun	dog	chair	notebook	phone	table
T1	0.07	0.30	0.03	0.01	0.05	0.40	0.04	0.05	0.04	0.01
T2	0.25	0.10	0.01	0.31	0.01	0.15	0.05	0.01	0.02	0.10
T3	0.05	0.04	0.35	0.01	0.25	0.02	0.05	0.10	0.03	0.10

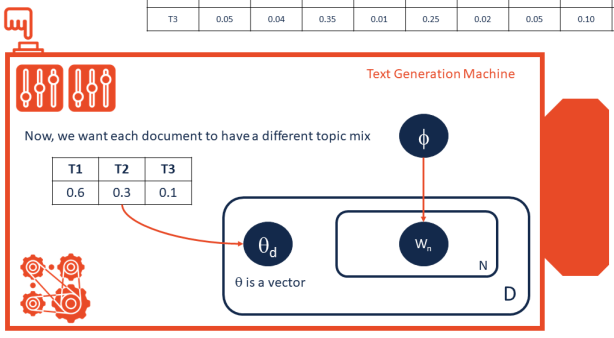


https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 3

Topics	pear	bird	planet	burger	sun	dog	chair	notebook	phone	table
T1	0.07	0.30	0.03	0.01	0.05	0.40	0.04	0.05	0.04	0.01
T2	0.25	0.10	0.01	0.31	0.01	0.15	0.05	0.01	0.02	0.10
T3	0.05	0.04	0.35	0.01	0.25	0.02	0.05	0.10	0.03	0.10

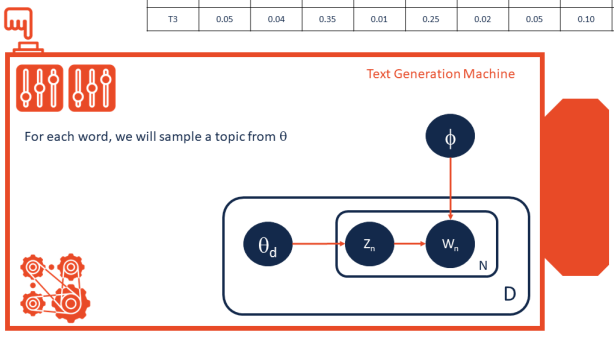


https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 3

Topics	pear	bird	planet	burger	sun	dog	chair	notebook	phone	table
T1	0.07	0.30	0.03	0.01	0.05	0.40	0.04	0.05	0.04	0.01
T2	0.25	0.10	0.01	0.31	0.01	0.15	0.05	0.01	0.02	0.10
T3	0.05	0.04	0.35	0.01	0.25	0.02	0.05	0.10	0.03	0.10

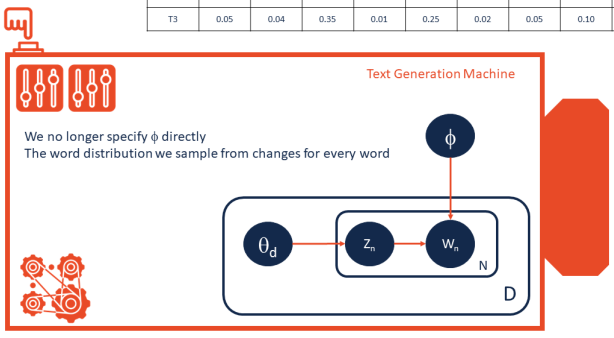


https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 3

Topics	pear	bird	planet	burger	sun	dog	chair	notebook	phone	table
T1	0.07	0.30	0.03	0.01	0.05	0.40	0.04	0.05	0.04	0.01
T2	0.25	0.10	0.01	0.31	0.01	0.15	0.05	0.01	0.02	0.10
T3	0.05	0.04	0.35	0.01	0.25	0.02	0.05	0.10	0.03	0.10

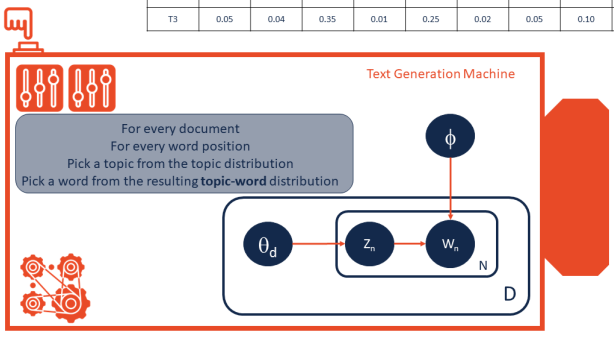


https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 3

Topics	pear	bird	planet	burger	sun	dog	chair	notebook	phone	table
T1	0.07	0.30	0.03	0.01	0.05	0.40	0.04	0.05	0.04	0.01
T2	0.25	0.10	0.01	0.31	0.01	0.15	0.05	0.01	0.02	0.10
T3	0.05	0.04	0.35	0.01	0.25	0.02	0.05	0.10	0.03	0.10

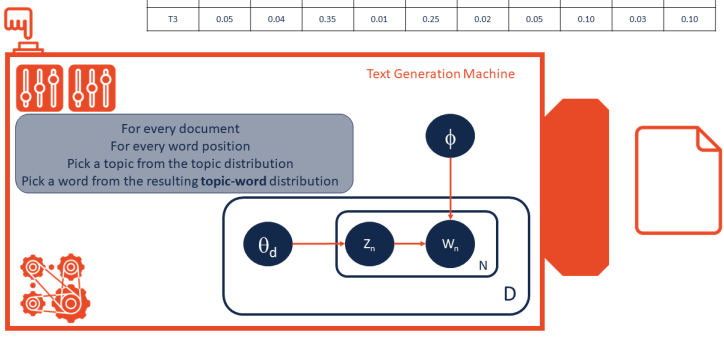


https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 3

Topics	pear	bird	planet	burger	sun	dog	chair	notebook	phone	table
T1	0.07	0.30	0.03	0.01	0.05	0.40	0.04	0.05	0.04	0.01
T2	0.25	0.10	0.01	0.31	0.01	0.15	0.05	0.01	0.02	0.10
T3	0.05	0.04	0.35	0.01	0.25	0.02	0.05	0.10	0.03	0.10

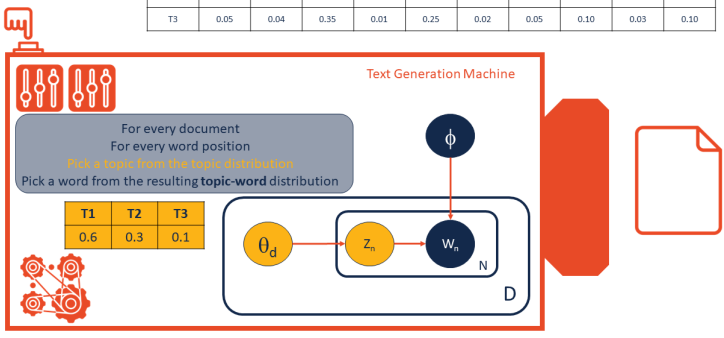


https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 3

Topics	pear	bird	planet	burger	sun	dog	chair	notebook	phone	table
T1	0.07	0.30	0.03	0.01	0.05	0.40	0.04	0.05	0.04	0.01
T2	0.25	0.10	0.01	0.31	0.01	0.15	0.05	0.01	0.02	0.10
T3	0.05	0.04	0.35	0.01	0.25	0.02	0.05	0.10	0.03	0.10

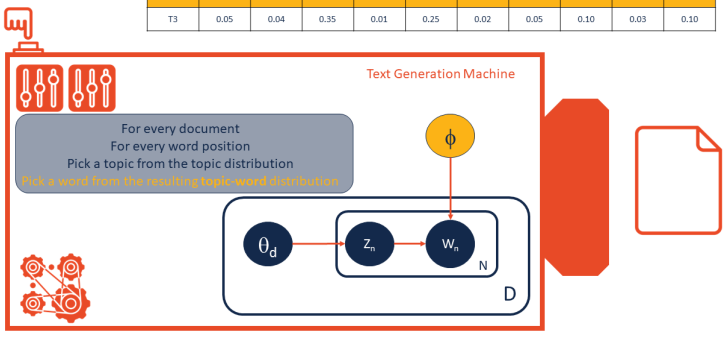


https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 3

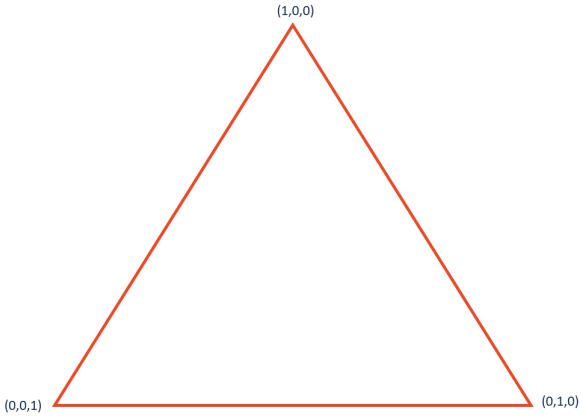
Topics	pear	bird	planet	burger	sun	dog	chair	notebook	phone	table
T1	0.07	0.30	0.03	0.01	0.05	0.40	0.04	0.05	0.04	0.01
T2	0.25	0.10	0.01	0.31	0.01	0.15	0.05	0.01	0.02	0.10
T3	0.05	0.04	0.35	0.01	0.25	0.02	0.05	0.10	0.03	0.10



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

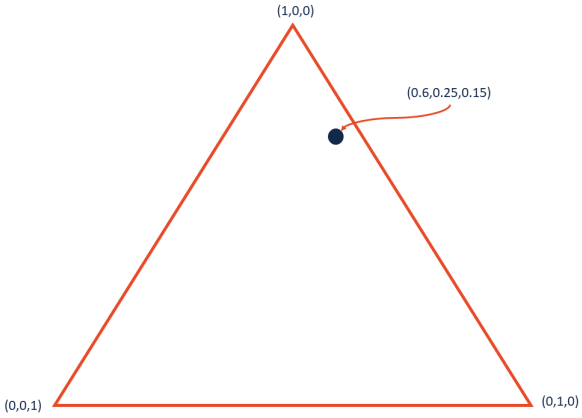
Generating multinomial distributions



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

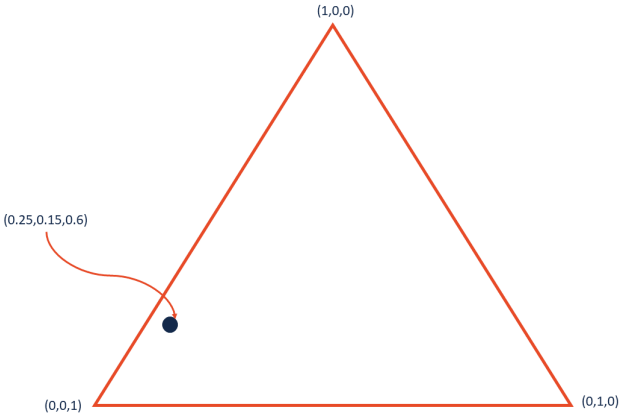
Generating multinomial distributions



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

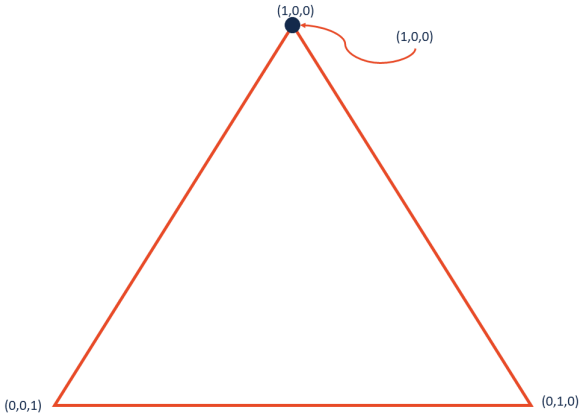
Generating multinomial distributions



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

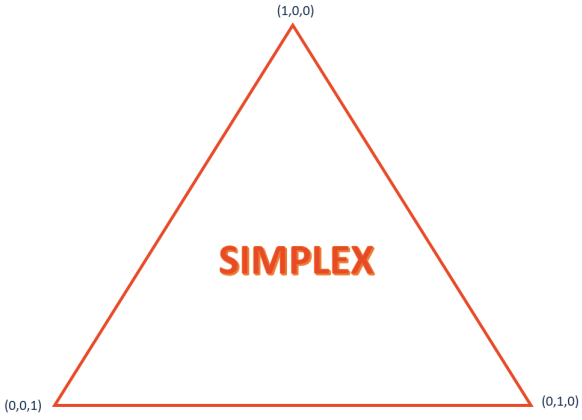
Generating multinomial distributions



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

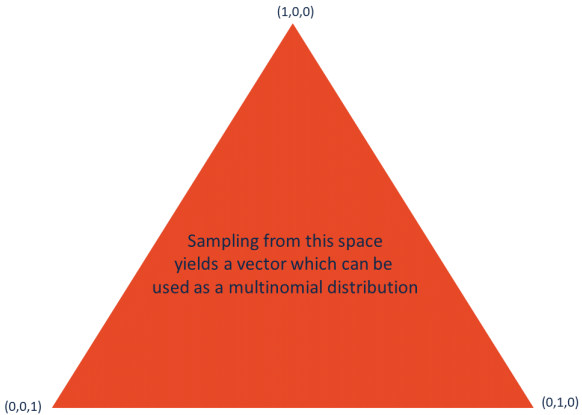
Generating multinomial distributions



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

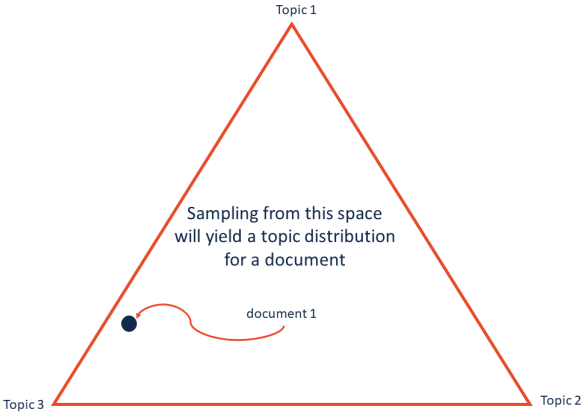
Generating multinomial distributions



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

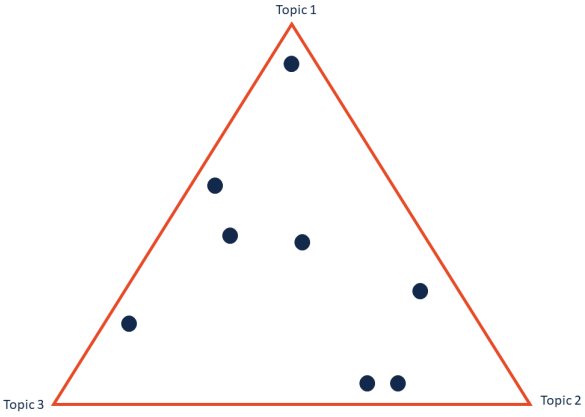
Generating multinomial distributions



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

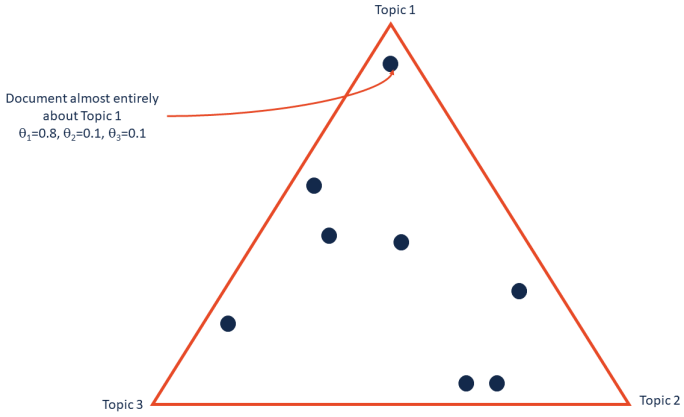
Generating multinomial distributions



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

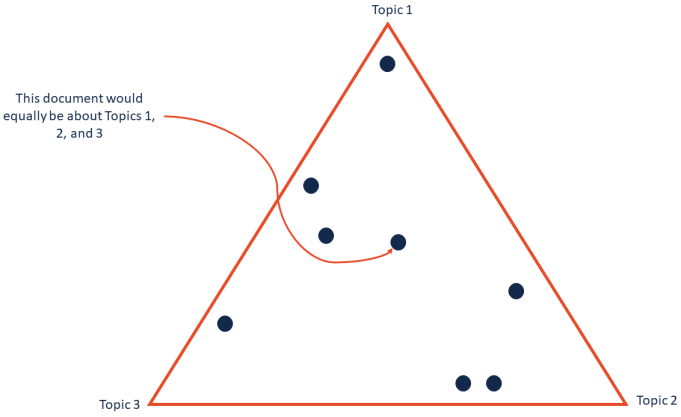
Generating multinomial distributions



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

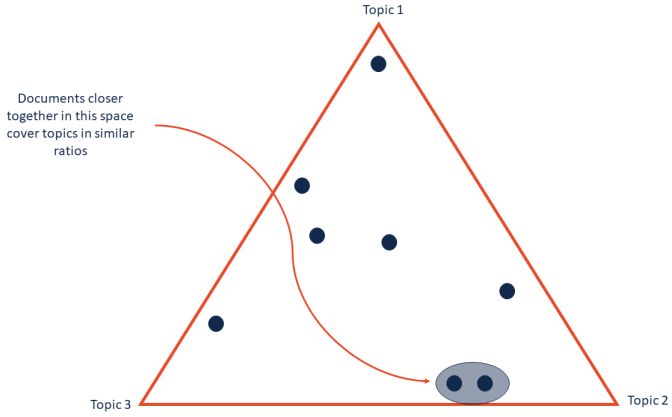
Generating multinomial distributions



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

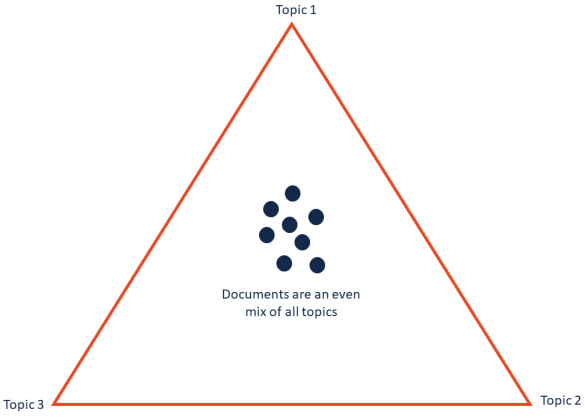
Generating multinomial distributions



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

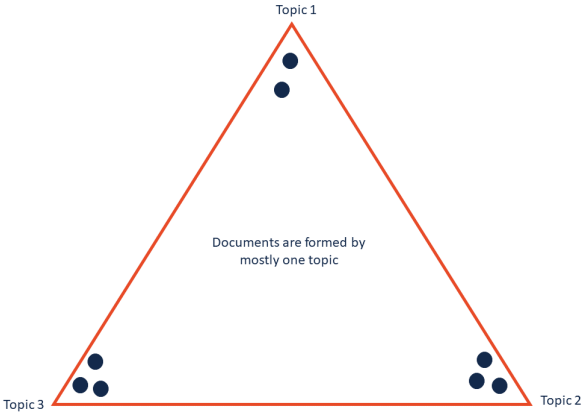
Generating multinomial distributions



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

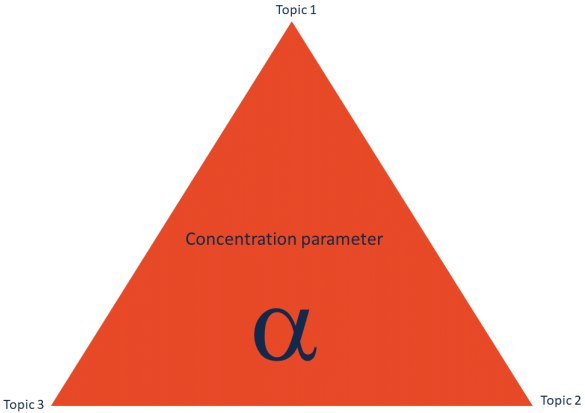
Generating multinomial distributions



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

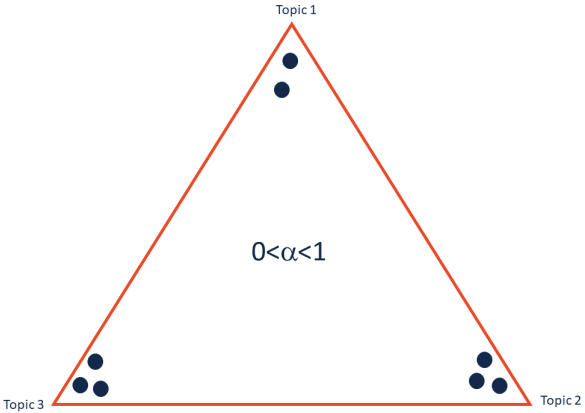
Generating multinomial distributions



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

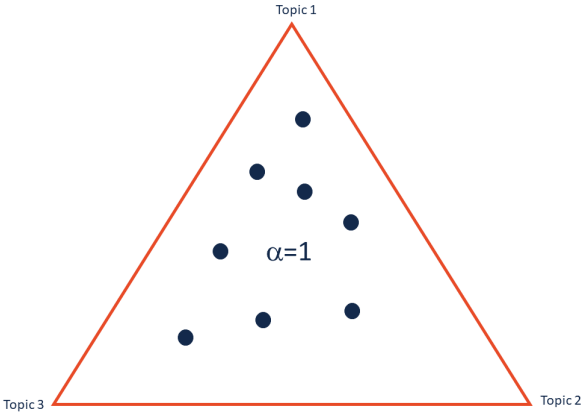
Generating multinomial distributions



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

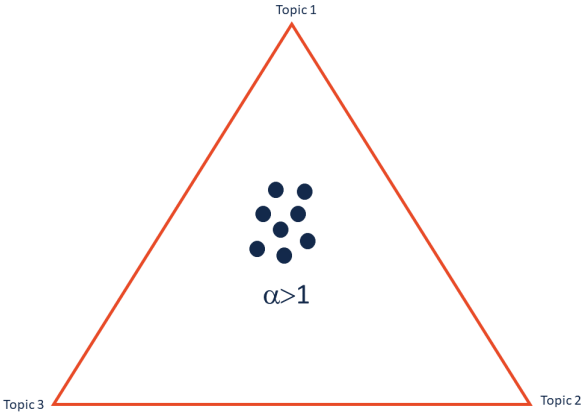
Generating multinomial distributions



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

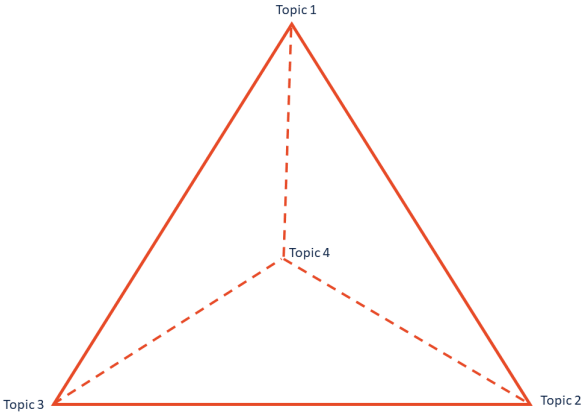
Generating multinomial distributions



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

Generating multinomial distributions

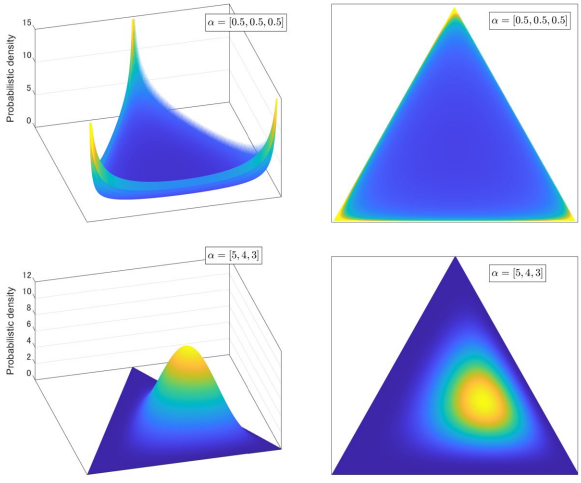


https://www.youtube.com/watch?v=9mNV4AwA9QI

This is the Dirichlet distribution

Introduction to Latent Dirichlet Allocation (LDA)

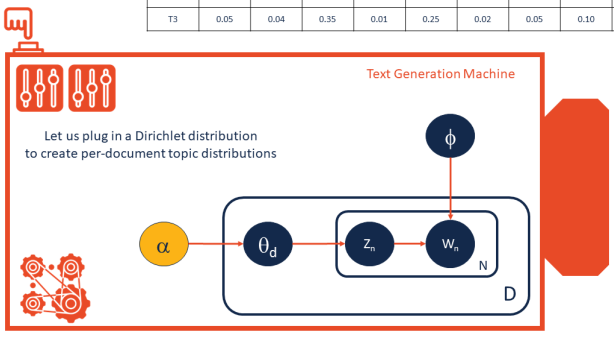
The Dirichlet distribution



Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 4

Topics	pear	bird	planet	burger	sun	dog	chair	notebook	phone	table
T1	0.07	0.30	0.03	0.01	0.05	0.40	0.04	0.05	0.04	0.01
T2	0.25	0.10	0.01	0.31	0.01	0.15	0.05	0.01	0.02	0.10
T3	0.05	0.04	0.35	0.01	0.25	0.02	0.05	0.10	0.03	0.10

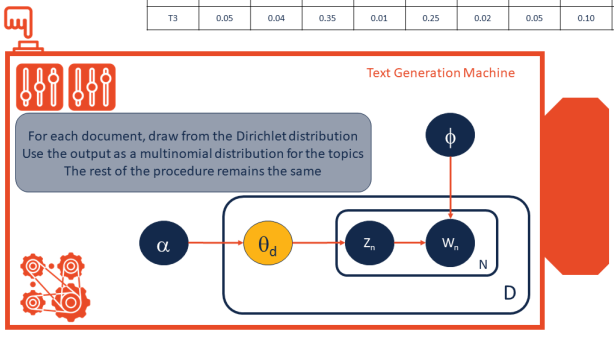


https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 4

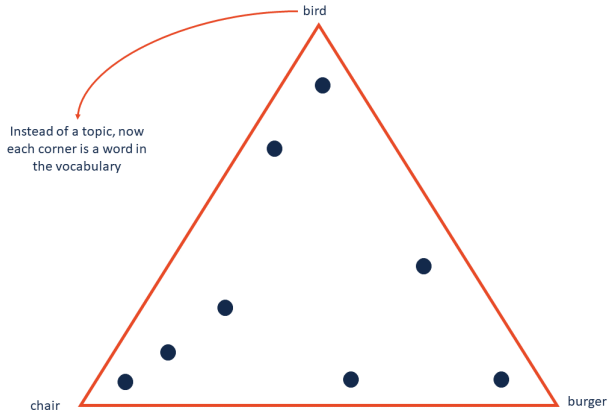
Topics	pear	bird	planet	burger	sun	dog	chair	notebook	phone	table
T1	0.07	0.30	0.03	0.01	0.05	0.40	0.04	0.05	0.04	0.01
T2	0.25	0.10	0.01	0.31	0.01	0.15	0.05	0.01	0.02	0.10
T3	0.05	0.04	0.35	0.01	0.25	0.02	0.05	0.10	0.03	0.10



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

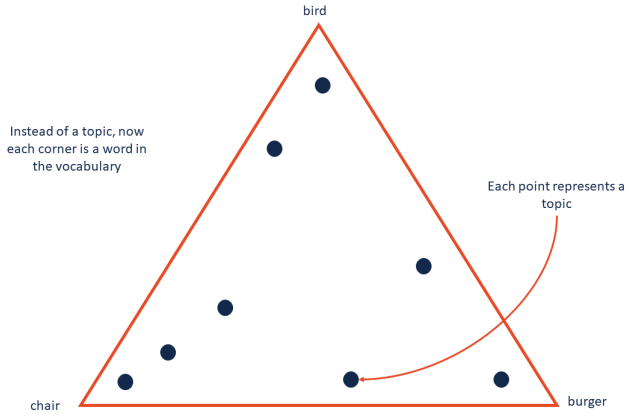
Another Dirichlet distribution



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

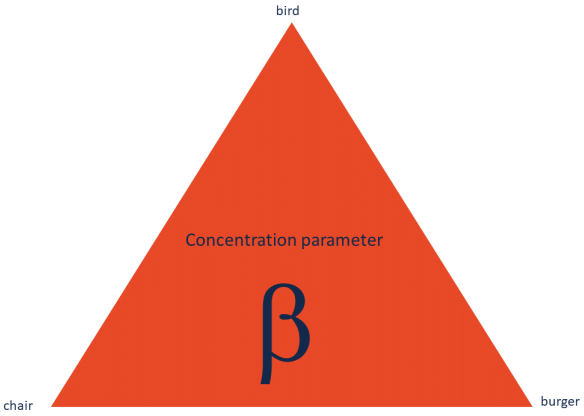
Another Dirichlet distribution



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

Another Dirichlet distribution

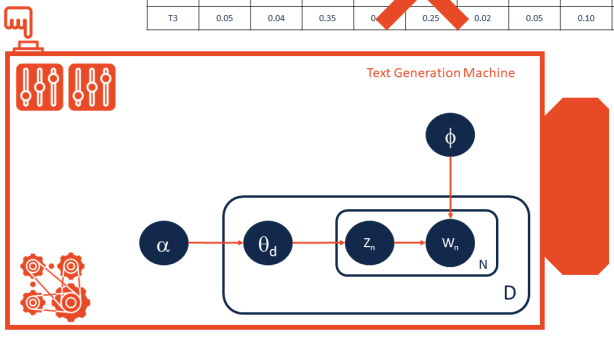


https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 5

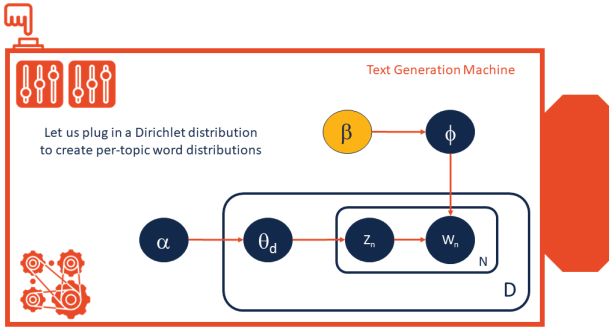
Topics	pear	bird	planet	bu	sun	dog	chair	notebook	phone	table
T1	0.07	0.30	0.03	0.01		0.40	0.04	0.05	0.04	0.01
T2	0.25	0.10	0.01	0.31		0.15	0.05	0.01	0.02	0.10
T3	0.05	0.04	0.35	0	0.25	0.02	0.05	0.10	0.03	0.10



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

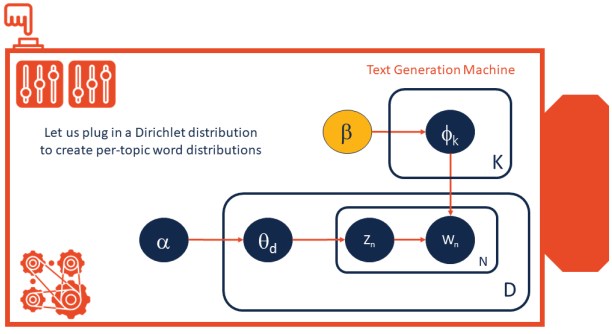
A simple machine to create text – Version 5



https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 5

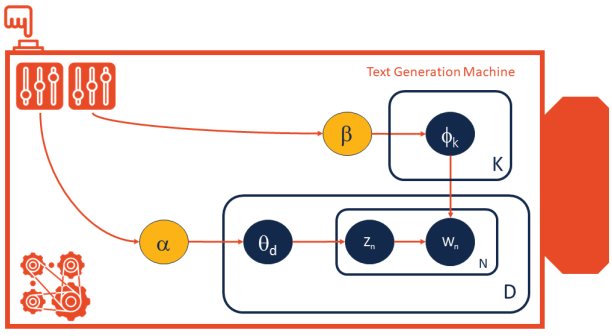


https://www.youtube.com/watch?v=9mNV4AwA9QI

Introduction to Latent Dirichlet Allocation (LDA)

A simple machine to create text – Version 5

For every document d = 1,..., D, draw a multinomial distribution θ_d from the Dirichlet distribution with parameter α
For every topic k = 1,..., K, draw a multinomial distribution ϕ_k from the Dirichlet distribution with parameter β
For every word position n = 1,..., N in the document d, assign a topic Z_n from θ_d and draw a work from ϕ_k with $k = Z_n$



https://www.youtube.com/watch?v=9mNV4AwA9QI

- Algorithm:
  1. Draw β_k independently for $k = \{1, \cdots, K\}$ from Dir($\boldsymbol{\nu}$).
  2. Draw θ_d independently for $d = \{1, \cdots, D\}$ from Dir($\boldsymbol{\alpha}$).
  3. Each word $w_{d,n}$ in document d is generated from a two-step process:
     - 3.1 Draw a topic assignment $z_{d,n}$ from θ_d.
     - 3.2 Draw $w_{d,n}$ from $\boldsymbol{\beta}_{\boldsymbol{z_{d,n}}}$
- Estimate the hyperparameters $\boldsymbol{\nu}$ and $\boldsymbol{\alpha}$ from the Dirichlet distributions.

Example

Newspaper Data Collection

- Scrape online news from major newspapers
  - News from January, 2009 up to December, 2021 from Folha de São Paulo, Estadão, and Valor Econômico
- Daily collected and weekly aggregated.

- **Bag-of-words:** for each individual article count the number of occurrences of a given term (or a composition of terms) on the cleaned text.
  - Generate a matrix of all such counts
  - Approach: 1-grams up to 3-grams counting
    - e.g. 'Covid-19', 'Presidente.Bolsonaro' and 'Supremo.Tribunal.Federal'
  - Aggregate the term-counts to generate the **Corpus**.
- **Generate the Corpus:** a matrix $\boldsymbol{X}_{T,V}$ where rows denote months and columns denotes terms

Textual Data

Number of news collected

	2009	2010	2011	2012	2013	2014	2015
Folha	598	5329	6012	5235	15476	30628	24048
Estadão	55298	94182	90182	92639	61469	54209	31085
Valor	-	-	-	65545	51057	45509	45048
Sum	55896	99511	96194	163419	128002	130346	100181

	2016	2017	2018	2019	2020	2021	Total
Folha	22073	18479	19853	16375	15278	13988	194944
Estadão	35758	40258	34957	32405	36337	31459	692866
Valor	47010	46156	47001	46059	60773	64969	525309
Sum	104841	104893	101811	94839	112388	110416	1413119

Textual Data

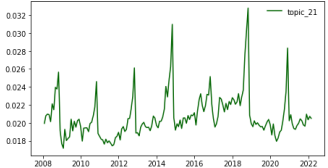
Pre-processing

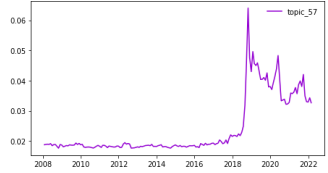
grafton , wisconsin - quinhentas doses da vacina contra a covid-19 tiveram que ser descartadas porque não foram devidamente refrigeradas em wisconsin , nos estados unidos . de acordo com a rede de hospitais aurora medical center , os imunizantes foram aparentemente estragadas de forma deliberada por um funcionário . no sábado , o hospital havia revelado que as doses foram acidentalmente deixadas em temperatura ambiente durante a noite por um funcionário da unidade de grafton . no entanto , nesta quarta-feira , 30 , o aurora divulgou uma nota afirmando que o funcionário envolvido `` reconheceu que as vacinas foram retiradas intencionalmente da geladeira '' . o comunicado diz ainda que o funcionário foi demitido e o assunto foi entregue às autoridades para uma investigação mais aprofundada . o depoimento não menciona o possível motivo dessa ação , e os executivos do sistema de saúde não responderam no momento às mensagens que lhes foram enviadas em busca de mais informações . " continuamos a acreditar que a vacinação é a nossa saída para a pandemia . estamos mais do que frustrados com o fato de o comportamento desse indivíduo atrasar a vacinação de mais de 500 pessoas " , disse a nota . o aurora medical center se recusou a fornecer informações adicionais , mas disse que daria mais detalhes na quinta-feira./ap

Cleaning and pre-processing of textual data from a particular article from Estadão. We mark the punctuation extraction in blue together with the number and single-letter removal. We highlight the stopwords, rare words, and temporal markers removed in red. In green, we point out the words that should be lemma-reduced. In black, we highlight words that should not be ex-ante modified.

Textual Data

Attention to topics





Obrigado