

Introduction to Topic Modeling

Project Mosaic Workshop

Date 2/16/2017

Ryan Wesslen

rwesslen@uncc.edu

Project Mosaic

- ▶ Project Mosaic: What do we do?
 - ▶ Build research methods capability in social sciences
 - ▶ Facilitate research across social science disciplines
 - ▶ Promote social science research
- ▶ Project Mosaic Services
 - ▶ Social sciences research incubator
 - ▶ Facilitate connections
 - ▶ Bring people together to exchange ideas and pursue external funding
 - ▶ Information sharing on research funding opportunities
 - ▶ Consulting
 - ▶ Free to UNC Charlotte faculty, staff and graduate students
 - ▶ Workshops
 - ▶ Open to entire campus community
 - ▶ Provides cutting-edge tools for research and a forum for researchers to network within campus



Workshop Agenda

- ▶ What is topic modeling?
- ▶ Running topic modeling with R
 - ▶ LDA, CTM and STM
- ▶ Advanced Topics (e.g. Validation, Spark)

Learning Objectives

Level	Background	Learning Objective
Beginner	No experience with R or text analysis	<ul style="list-style-type: none">• Learn what problems topic modeling focuses on (text summarization)• Learn why visualizations are important for topic modeling
Intermediate	Experience with R and topic modeling (LDA)	<ul style="list-style-type: none">• Learn how to run topic modeling in R.• Learn the data components of topic modeling (DFM, Output)
Advanced	Experience with R, causal inference (GLM), Spark, Scala, visualizations	<ul style="list-style-type: none">• Learn advanced models like CTM and STM (extensions to LDA)• Learn ways to interpret, analyze and validate topic models.• Learn how to run “big data” LDA on Spark/SOPHI

What is Topic Modeling?

Problem

- ▶ Need to analyze a large collection of text documents.
- ▶ “Text summarization” / information retrieval problem
- ▶ Example: research abstracts, emails, tweets/facebook posts, patents, etc.

Text Analysis

Natural Language
Processing

Bag-of-Words
(Statistical-based)

Supervised

Unsupervised

Topic Modeling

Parsing

Part-of-speech (POS) tagging

Named entity recognition (NER)

Question answering (QA)

Q. How effective is ibuprofen in reducing
fever in patients with acute febrile illness?

Overgeneralization for
illustrative purposes

Topic Modeling

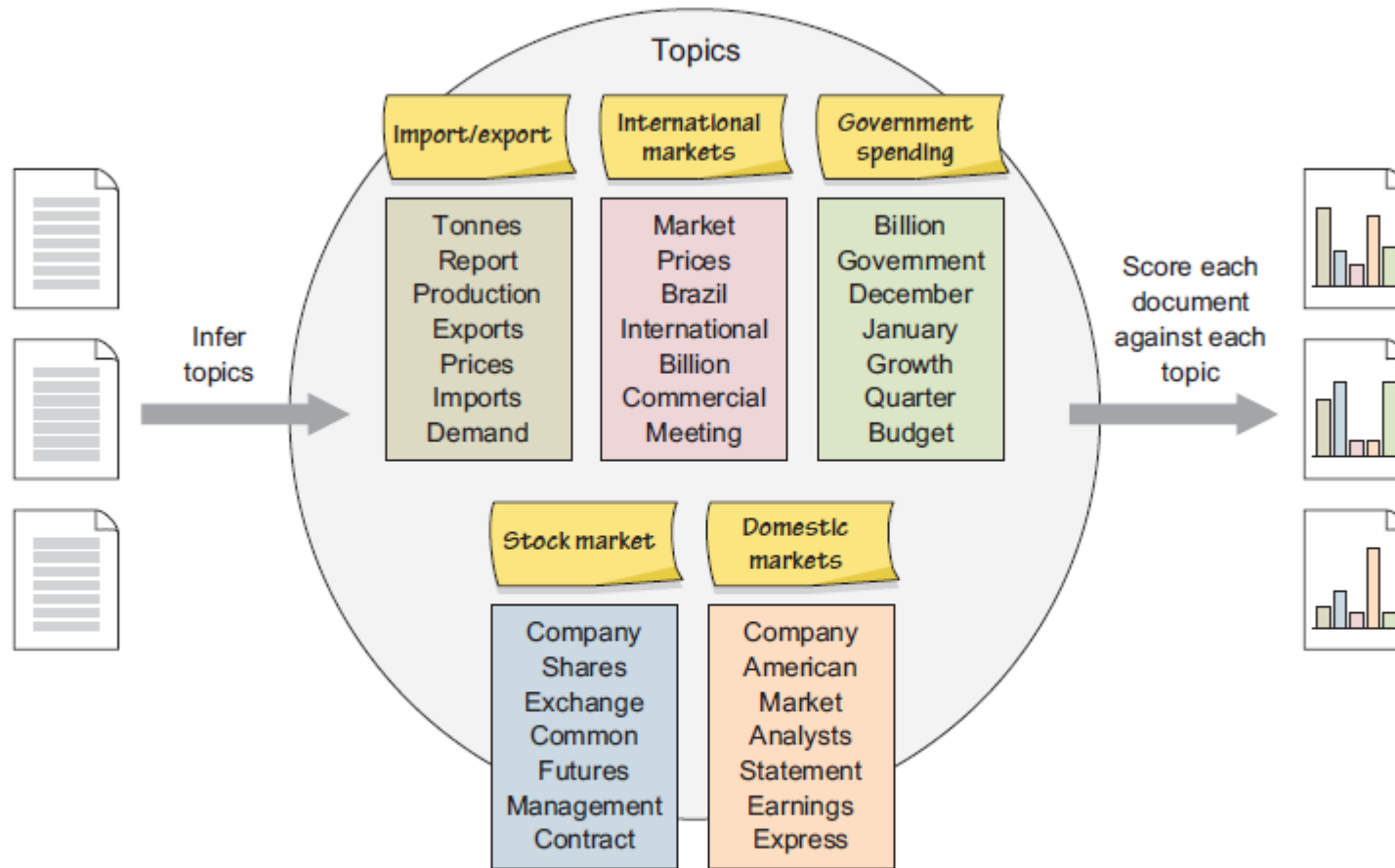


Figure 7.5 Latent Dirichlet Allocation. The topics are the latent variables and are determined automatically by the algorithm. The names of those topics shown in the thin strips are human-inferred and human-applied; the algorithm has no inherent capability to name the topics. Each document expresses each latent variable (topic) to a varying degree.

Bag of Words Assumption

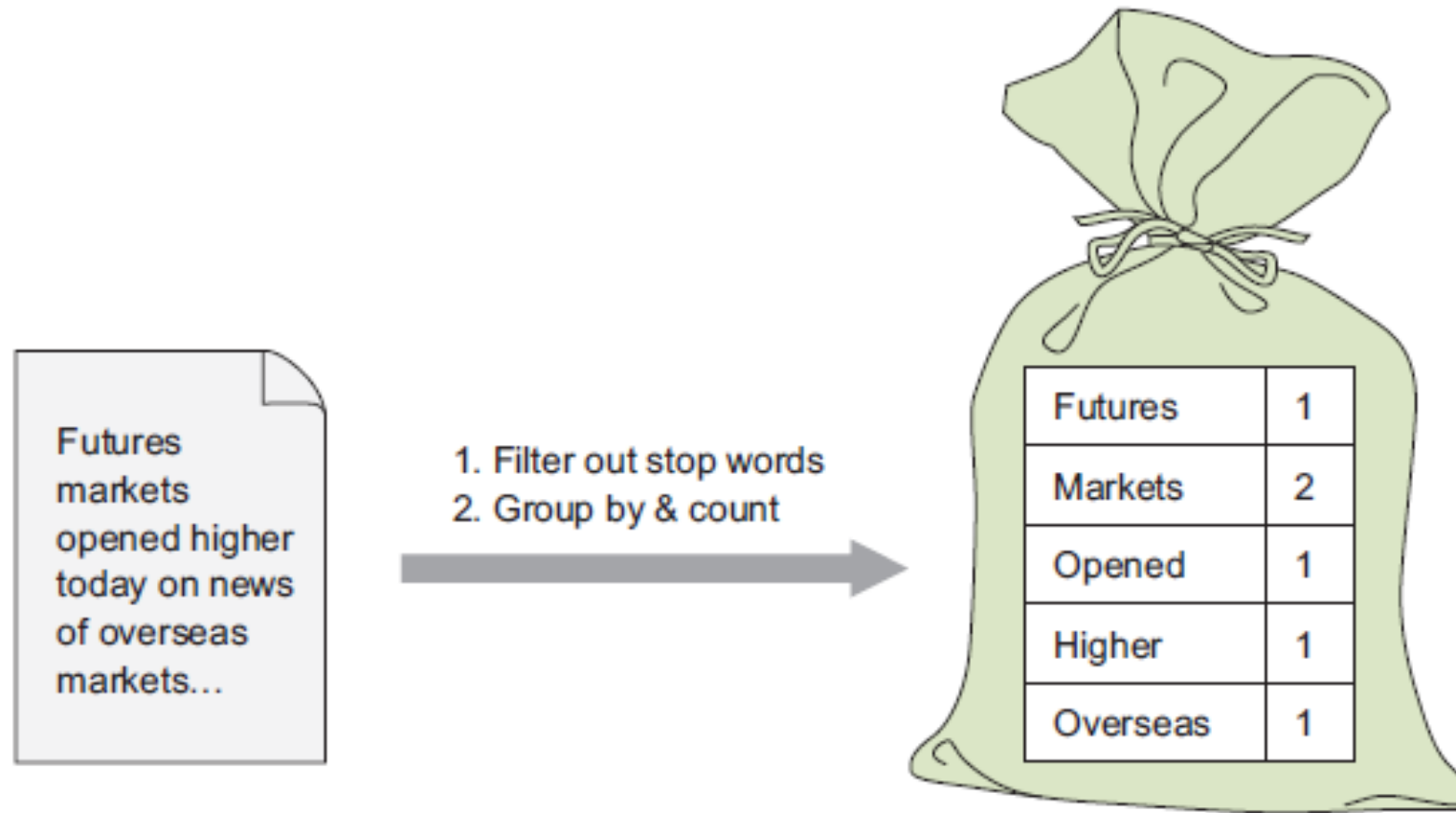
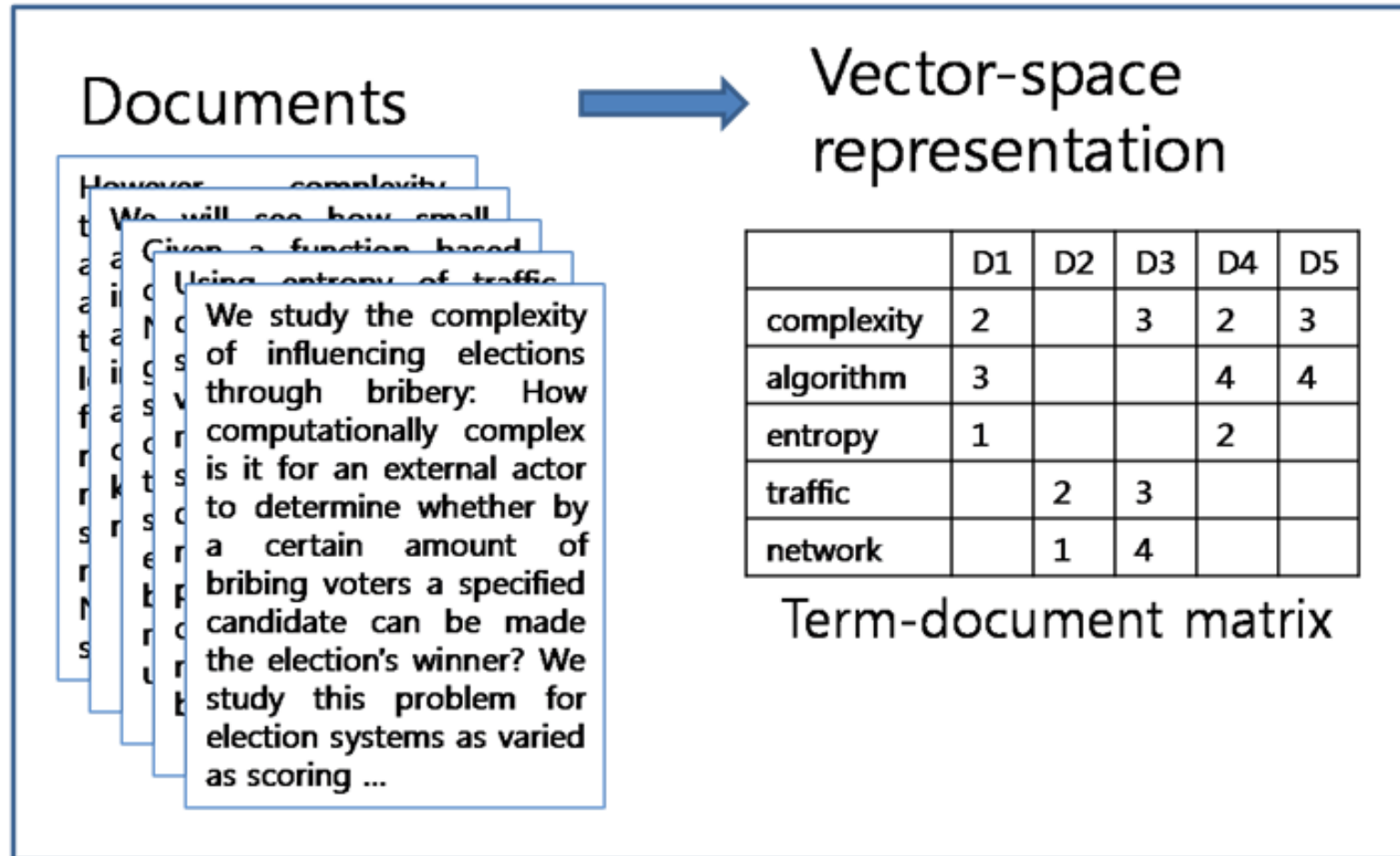


Figure 7.6 Bag of words representation of a document.

Term-Document Matrix



- Counting Word-Doc Frequencies yields the TDM (or DTM)

Topic Modeling Intuition: Dimensionality Reduction

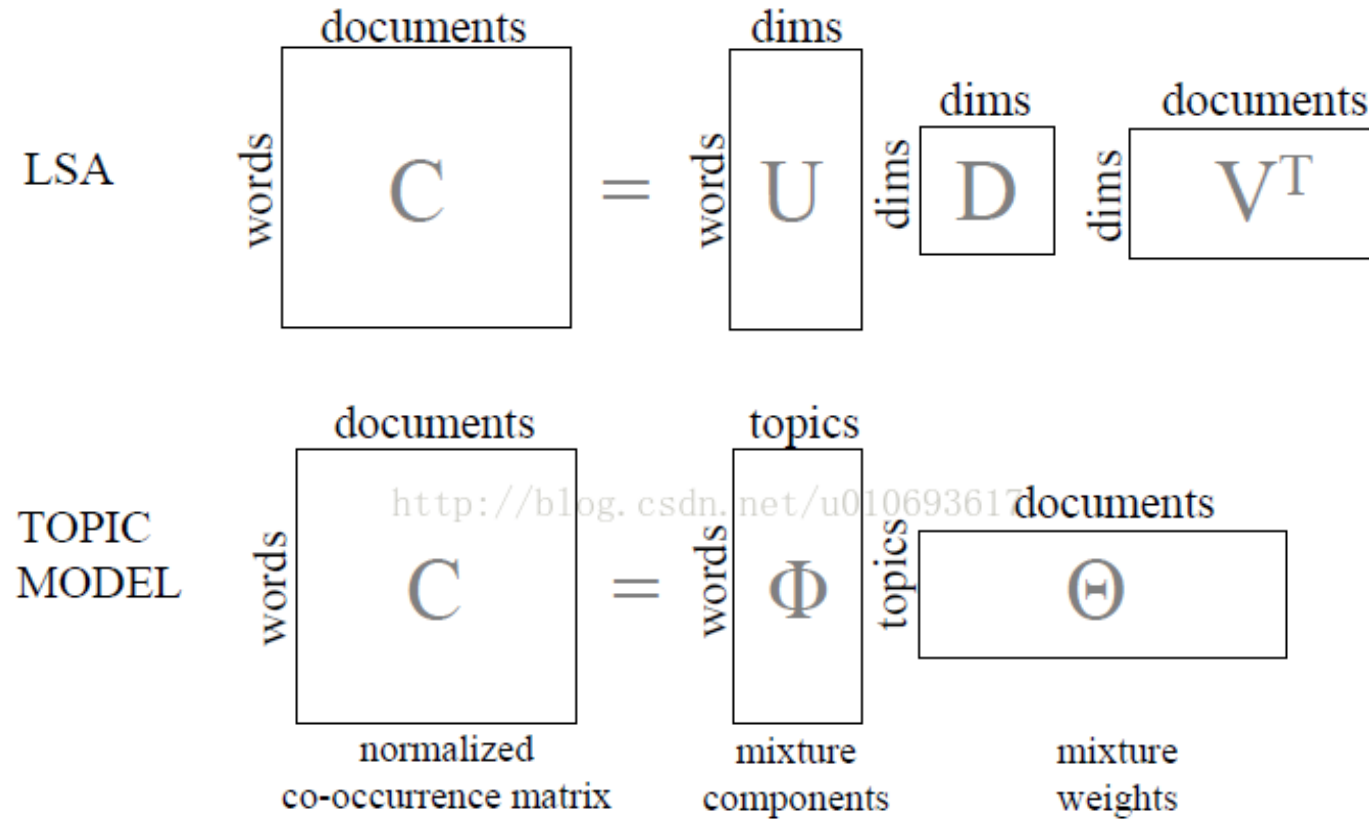
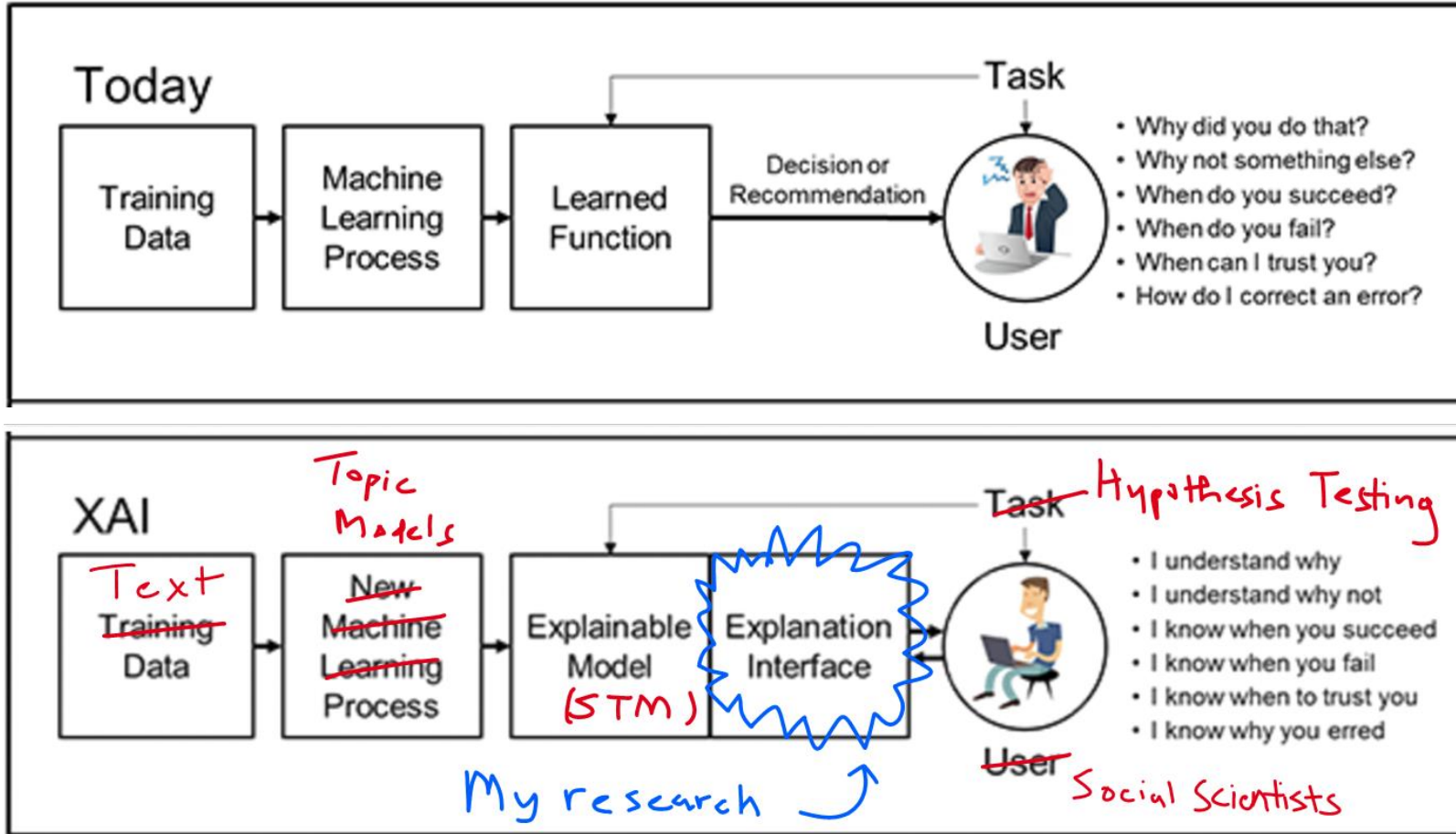


Figure 6. The matrix factorization of the LSA model compared to the matrix factorization of the topic model

Explainable AI and why Visualizations are important



Topic Modeling with R

Case Study

- ▶ We've been asked by the Dean to understand "Computational Social Science" research at UNCC.
- ▶ We have a dataset of nearly 2,500 research abstracts from six years worth of publications by UNCC researchers in Social Science (across dept/college) and Computing & Informatics (the entire college CCI).

Research Questions:

- ▶ What are the topics?
 - ▶ Part 1: Use LDA (Latent Dirichlet Allocation)
- ▶ How are they interrelated?
 - ▶ Part 2: Use CTM (Correlated Topic Model)
- ▶ What is the effect discipline (social sci. vs computing) & year has on the topics?
 - ▶ Part 3: Use STM (Structural Topic Model)

1. Obtain dataset (e.g. webscraping, API, etc.)
2. Pre-processing (Tokenization, stemming, n-grams)
3. Exploratory analysis (Word clouds, clustering)
4. Topic Modeling

We'll do this in three parts (models): LDA, CTM and STM.

Workshop Materials

All workshop materials can be found here:

<https://github.com/wesslen/Topic-Modeling-Workshop-with-R>

Advanced Topics

SYMPOSIUM

We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together

Justin Grimmer, *Stanford University*

Information is being produced and stored at an unprecedented rate. It might come from recording the public's daily life: people express their emotions on Facebook accounts, tweet opinions, call friends on cell phones, make statements on Weibo, post photographs on Instagram, and log locations with GPS on phones. Other information comes from aggregating media news stories through online sources

of statistical techniques. For the analysis of big data to truly yield answers to society's biggest problems, we must recognize that it is as much about social science as it is about computer science.

THE VITAL ROLE OF DESCRIPTION

For the analysis of big data to truly yield answers to society's biggest problems, we must recognize that it is as much about social science as it is about computer science.

stm Package Functions

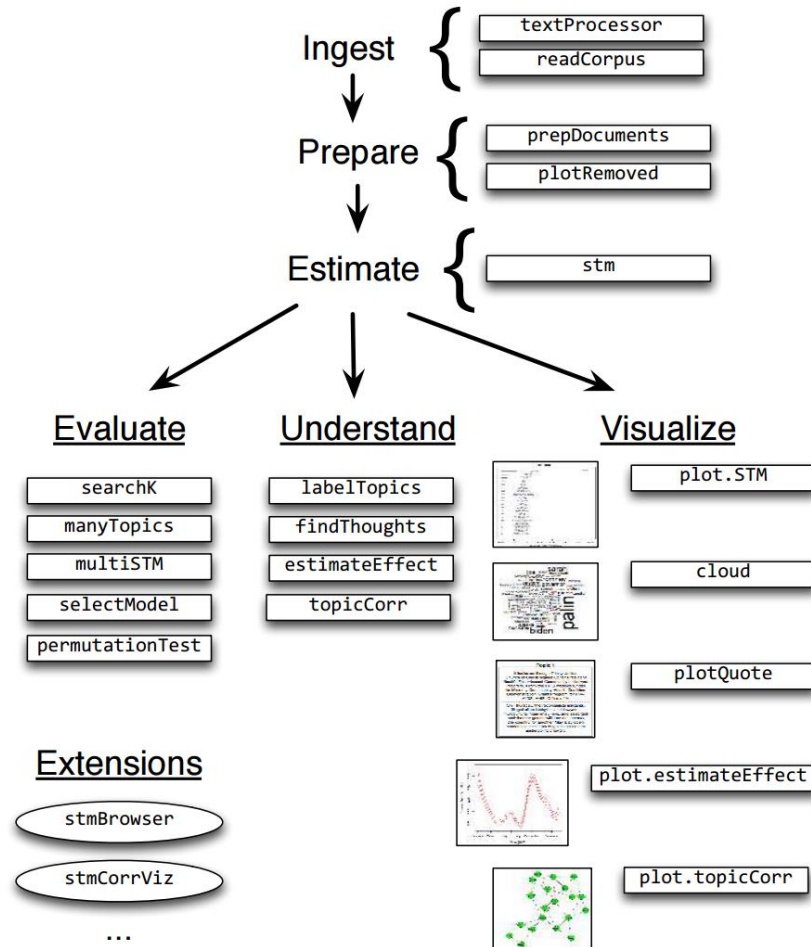
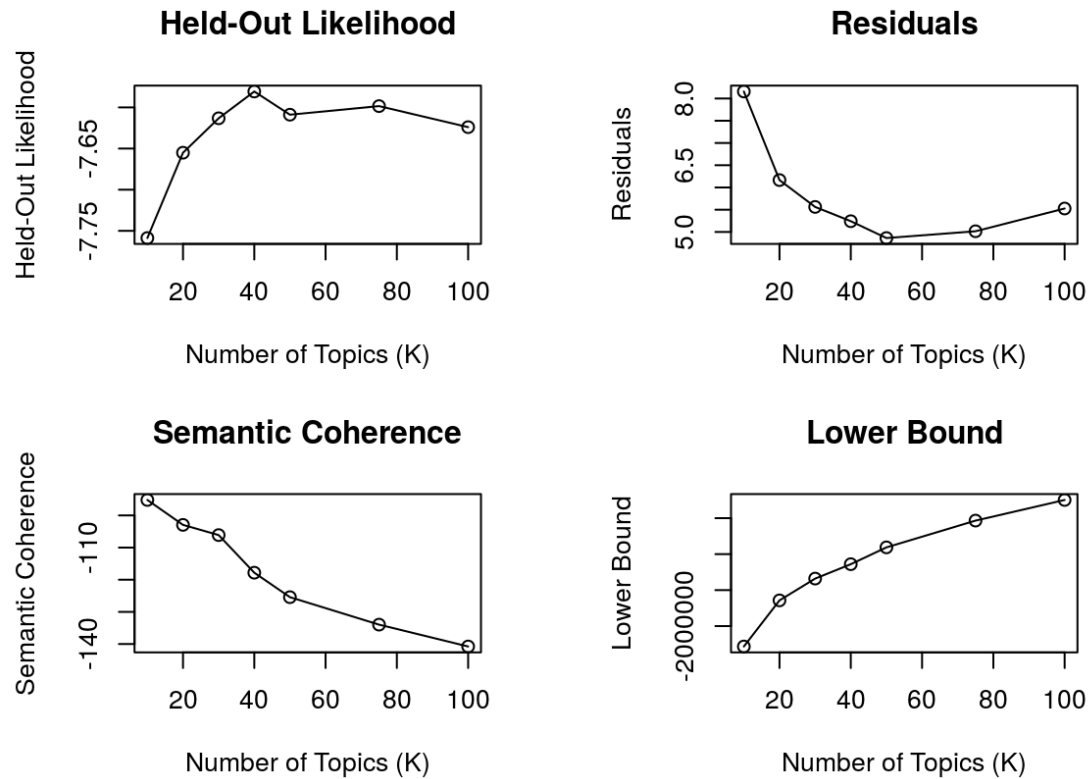


Figure 2: Heuristic description of **stm** package features.

Determining # of Topics

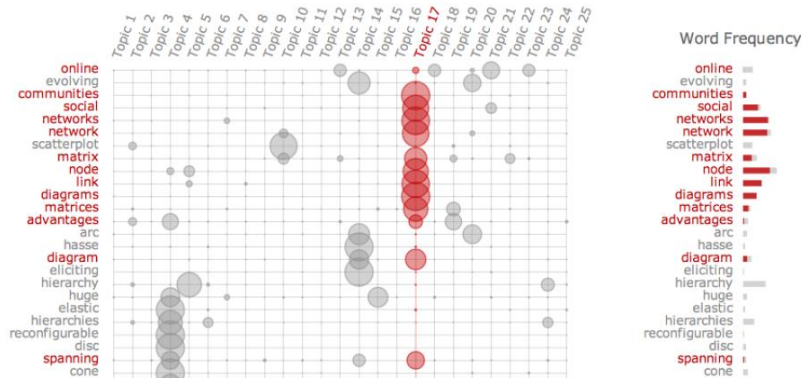
- The stm package has a helpful function (searchK) to help determine the number of topics.

Diagnostic Values by Number of Topics



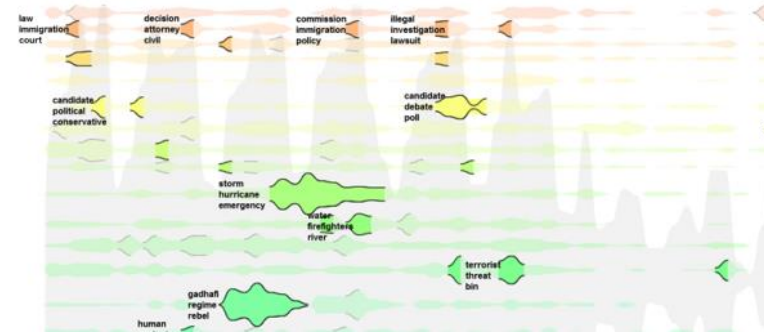
Visualizing Topic Models

Topic-Oriented



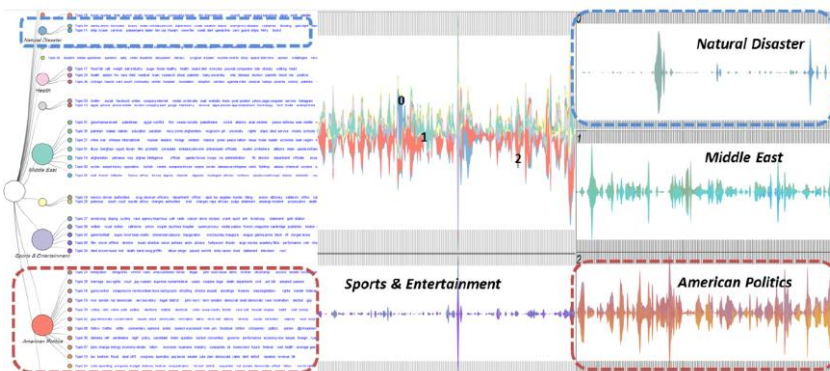
Termite (Chuang et al 2012)

Time-Oriented



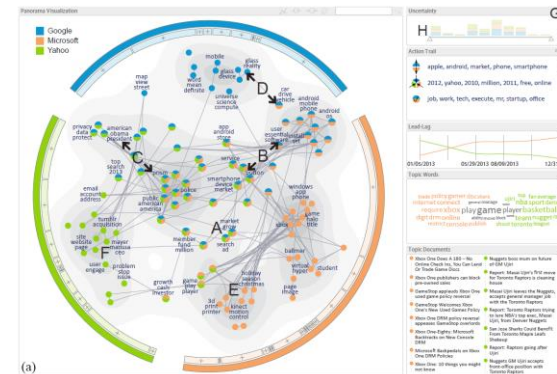
LeadLine (Dou et al 2012)

Hierarchical



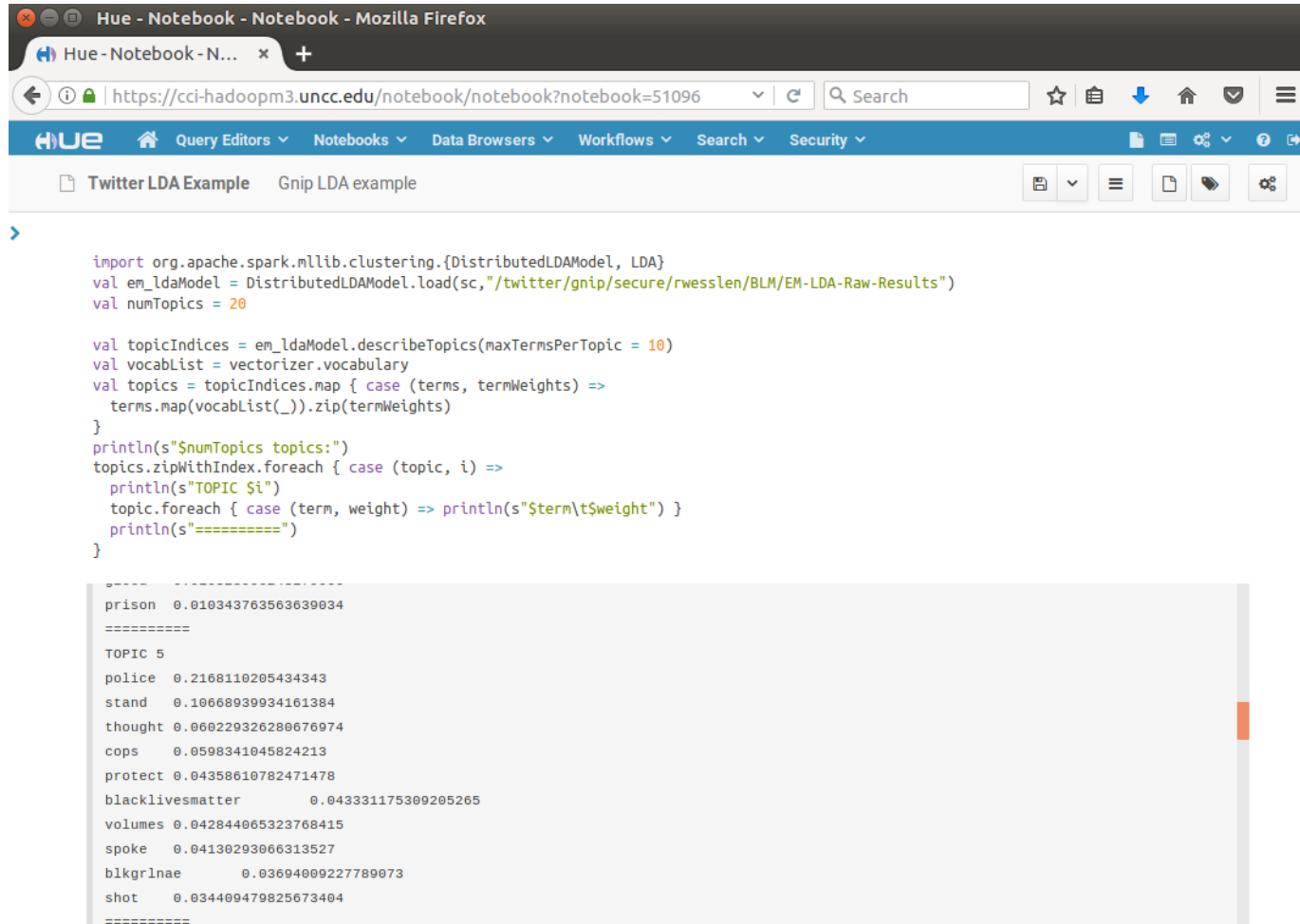
HierarchicalTopics (Dou et al 2013)

Graph (Relational) Based



TopicPanorama (Wang et al 2016)

Running Large-Scale LDA



```
import org.apache.spark.mllib.clustering.{DistributedLDAModel, LDA}
val em_ldaModel = DistributedLDAModel.load(sc,"/twitter/gnip/secure/rwesslen/BLM/EM-LDA-Raw-Results")
val numTopics = 20

val topicIndices = em_ldaModel.describeTopics(maxTermsPerTopic = 10)
val vocabList = vectorizer.vocabulary
val topics = topicIndices.map { case (terms, termWeights) =>
  terms.map(vocabList(_)).zip(termWeights)
}
println(s"$numTopics topics:")
topics.zipWithIndex.foreach { case (topic, i) =>
  println(s"TOPIC $i")
  topic.foreach { case (term, weight) => println(s"$term\t$weight") }
  println(s"=====")
}

-----
prison 0.010343763563639034
=====
TOPIC 5
police 0.2168110205434343
stand 0.10608939934161384
thought 0.060229326280676974
cops 0.0598341045824213
protect 0.04358610782471478
blacklivesmatter 0.043331175309205265
volumes 0.042844065323768415
spoke 0.04130293066313527
blkgrlnae 0.03694009227789073
shot 0.034409479825673404
=====
```

Need to know either Java, Scala or Python (and Spark).

<https://github.com/wesslen/Code-Tutorials-for-SOPHI>

More About Services

▶ Research Incubator

▶ Affiliates Program

- ▶ Faculty Affiliates are hand picked for their research expertise
- ▶ Our affiliates leverage the core functionality and expertise of Project Mosaic

▶ Seed Grants Program

- ▶ Geared towards the formation of new teams of researchers in the social, behavior and economic sciences
- ▶ Aim is to pursue external funding

▶ Consulting

▶ Project Mosaic offers three types of consulting:

- ▶ Software-centric
- ▶ Dissertation/thesis assistance
- ▶ Research collaboration

Make an appointment on
our website!

▶ Workshops

- ▶ Our workshops fulfill a commitment to enhance data literacy and analytical capabilities of UNC Charlotte researchers

Find workshops online on
our Events List.

Contact Project Mosaic



- ▶ Jean-Claude Thill is the director of Project Mosaic. A broadly trained geographer, he is a 'Knight' Distinguished Professor of Public Policy at UNC Charlotte.
- ▶ Contact Jean-Claude:
 - ▶ Email: Jean-Claude.Thill@uncc.edu
 - ▶ Phone: 704-687-5931 ext. 75909



- ▶ Leonora is the Administrative Support for Project Mosaic. She manages our not-so-massive paperwork, coordinates meetings and assists with administrative functions.
- ▶ Contact Leonora:
 - ▶ Email: projectmosaic@uncc.edu
 - ▶ Phone: 704-687-5931

Visit our website!
Projectmosaic.uncc.edu

Additional Resources: Consultants



- ▶ Shaoyu Li is the head consultant in the Center of Statistics and Applied Mathematics Consulting Center (CSAMC) and works with Project Mosaic to coordinate consulting requests for statistical and mathematical expertise.
- ▶ Contact Shaoyu:
 - ▶ Email: shaoyu.li@uncc.edu



- ▶ Kailas Venkitasubramanian is a research methodologist and manages the consulting service and the workshop program of Project Mosaic. Kailas is experienced in a variety of applied statistical techniques and works fluently on multiple software platforms.
- ▶ Contact Kailas:
 - ▶ Email: kvenkita@uncc.edu

Questions?

STM: Graphical Model

Correlated Topic Model (Blei & Lafferty, 2007)

- Allow topic correlation structure

SAGE (Eisenstein et al., 2011)

- Topic-word computational approximation
- Covariates that impact “how” topics (content)

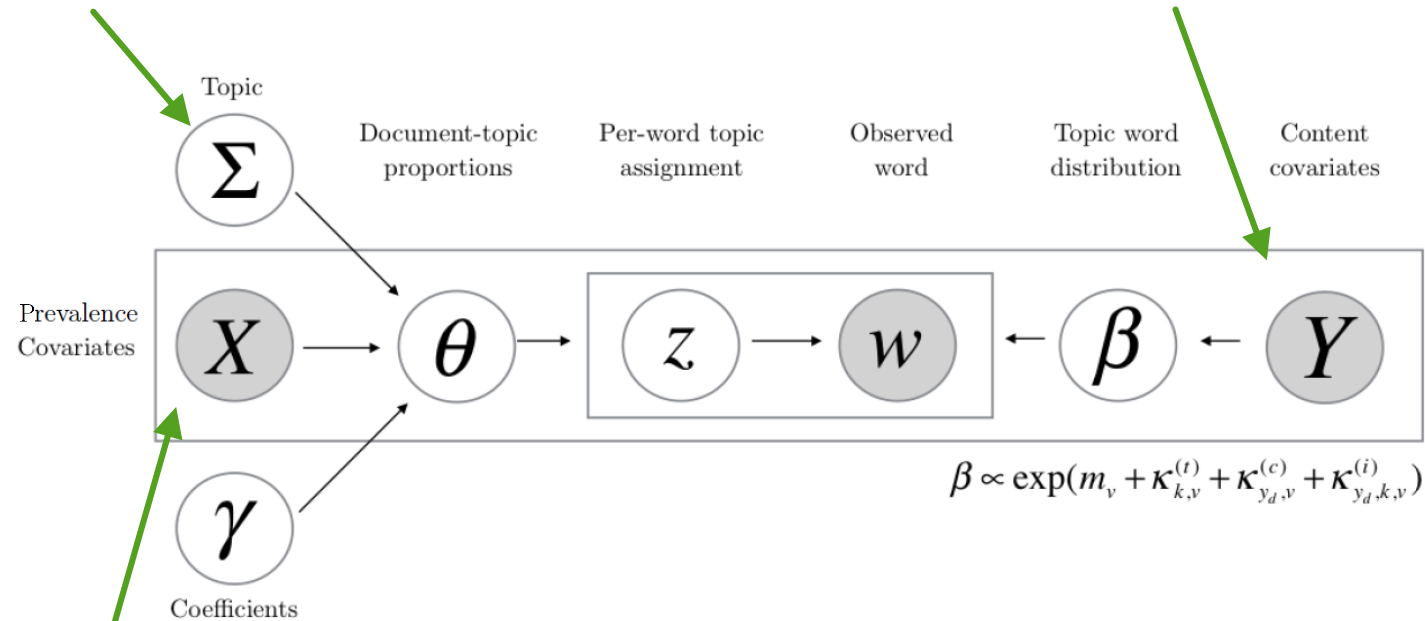


Figure 1: A graphical illustration of the structural topic model.

Dirichlet-Multinomial Regression (Minmo & McCallum, 2008)

- Covariates that impact “what” topics (prevalence)

On average good; but high variance Better captures covariate relationship

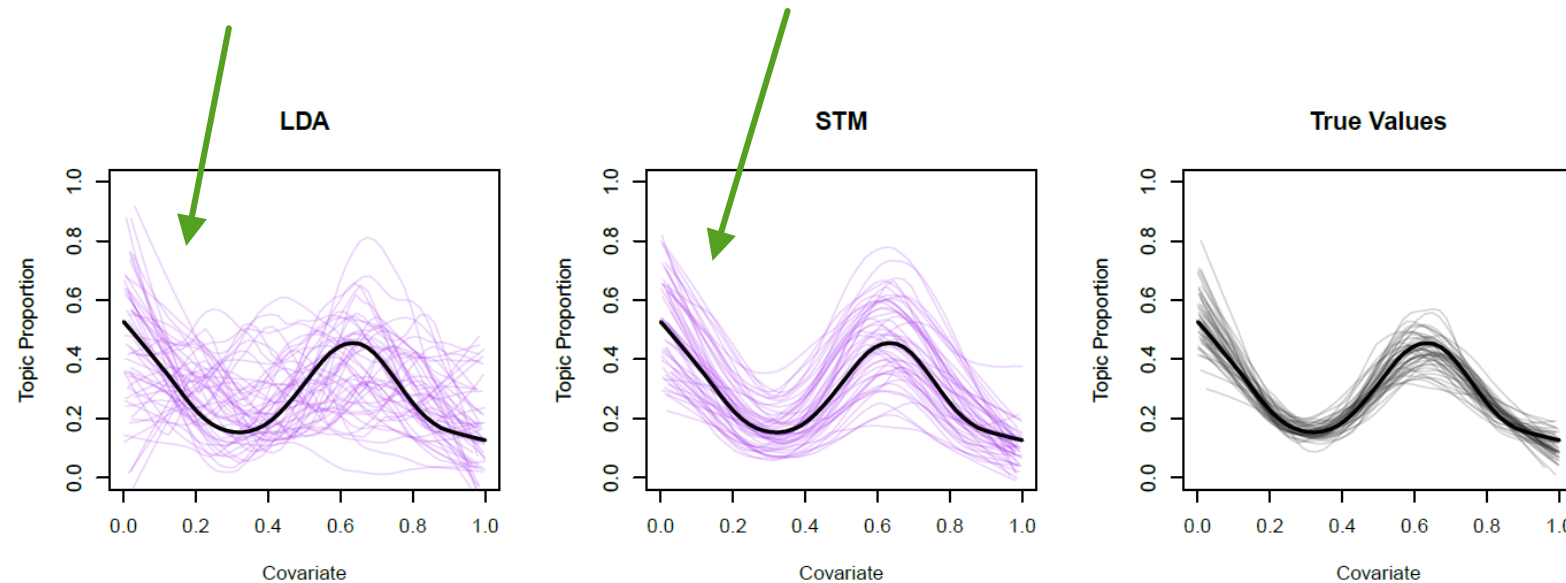


Figure 2: Plot of fitted covariate-topic relationships from 50 simulated datasets using LDA and the proposed structural topic model of text. The third panel shows the estimated relationship using the true values of the topic and thus only reflects sampling variability in the data generating process.

Performance on Holdout Sample

Higher Values => Better holdout prediction

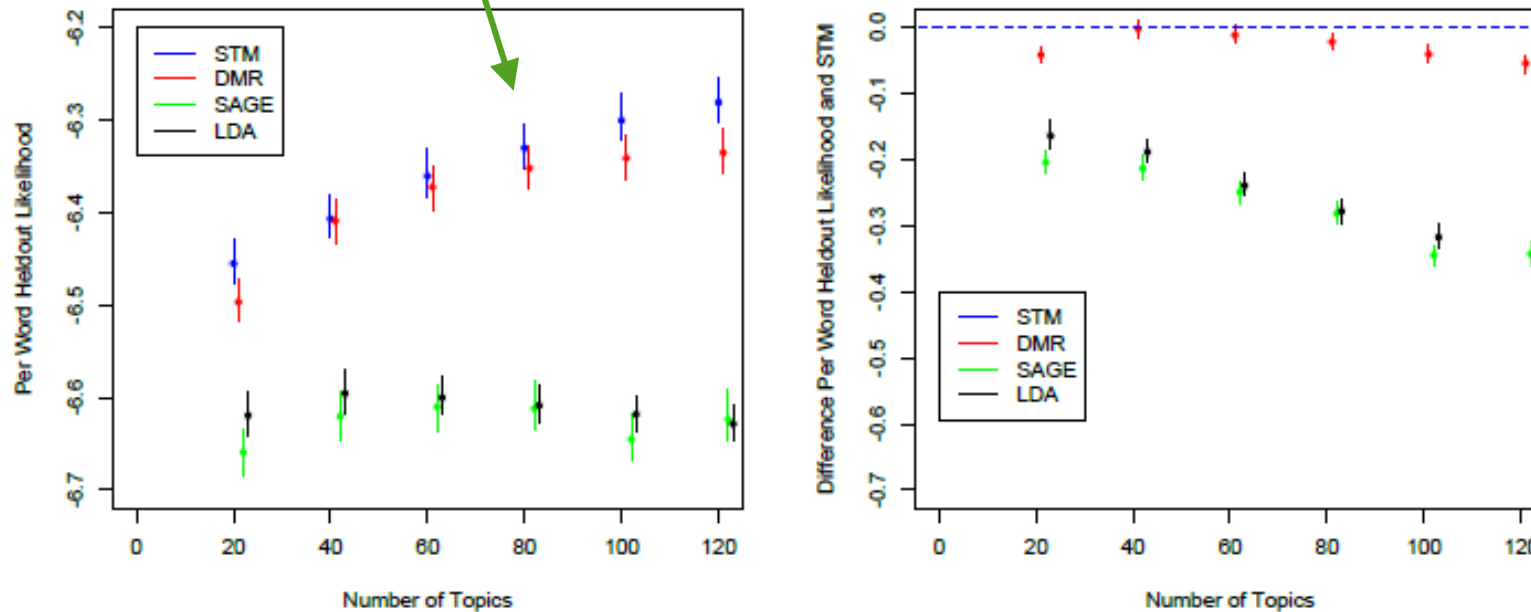


Figure 4: STM vs. SAGE, LDA and DMR Heldout Likelihood Comparison . On the left is the mean heldout likelihood and 95% quantiles. On the right is the mean paired difference between the three comparison models and STM.

Example: Media Slant on China

A Model of Text for Experimentation in the Social Sciences

Margaret E. Roberts^a, Brandon M. Stewart^b, and Edoardo M. Airolidi^c

^aDepartment of Political Science, University of California San Diego, San Diego, CA, USA; ^bDepartment of Sociology, Princeton University, Princeton, NJ, USA; ^cDepartment of Statistics, Harvard University, Cambridge, MA, USA

ABSTRACT

Statistical models of text have become increasingly popular in statistics and computer science as a method of exploring large document collections. Social scientists often want to move beyond exploration, to measurement and experimentation, and make inference about social and political processes that drive discourse and content. In this article, we develop a model of text data that supports this type of substantive research. Our approach is to posit a hierarchical mixed membership model for analyzing topical content of documents, in which mixing weights are parameterized by observed covariates. In this model, topical *prevalence* and topical *content* are specified as a simple generalized linear model on an arbitrary number of document-level covariates, such as news source and time of release, enabling researchers to introduce elements of the experimental design that informed document collection into the model, within a generally applicable framework. We demonstrate the proposed methodology by analyzing a collection of news reports about China, where we allow the prevalence of topics to evolve over time and vary across newswire services. Our methods quantify the effect of news wire source on both the frequency and nature of topic coverage. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received March 2014
Revised October 2015

KEYWORDS

Causal inference;
Experimentation;
High-dimensional inference;
Social sciences; Text analysis;
Variational approximation

[Link for PDF: Highly Recommend!](#)

- ▶ Western and Chinese media slant on China's Rise
 - ▶ Sample of 11,980 (English) news articles containing the term “China” between 1997 and 2006
 - ▶ Five different international news sources
 - ▶ AFP (France), AP (USA), BBC (British), JEN (Japan), XIN (China)
- ▶ Run 100 topic STM model
 - ▶ Prevalence covariates: Time (month) and News Source
 - ▶ Content covariate: News Source

Findings: Content & Prevalence

“Demonstration”

AP / AFP covered
topic more often
than Xinhua

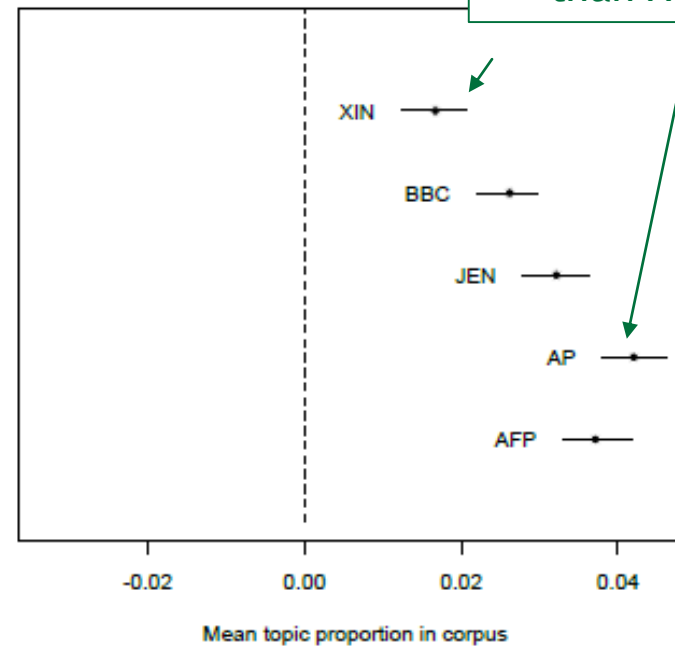
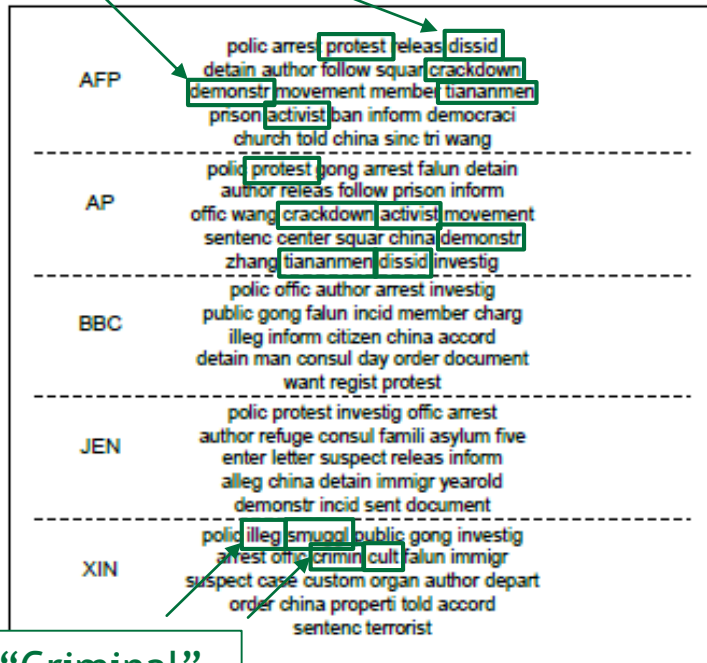


Figure 7: Falungong Topic. Each group of words are the highest probability words for the news source (left panel). Mean prevalence of Falungong topic within each news source corpus (right panel).

Example: Time as Covariate

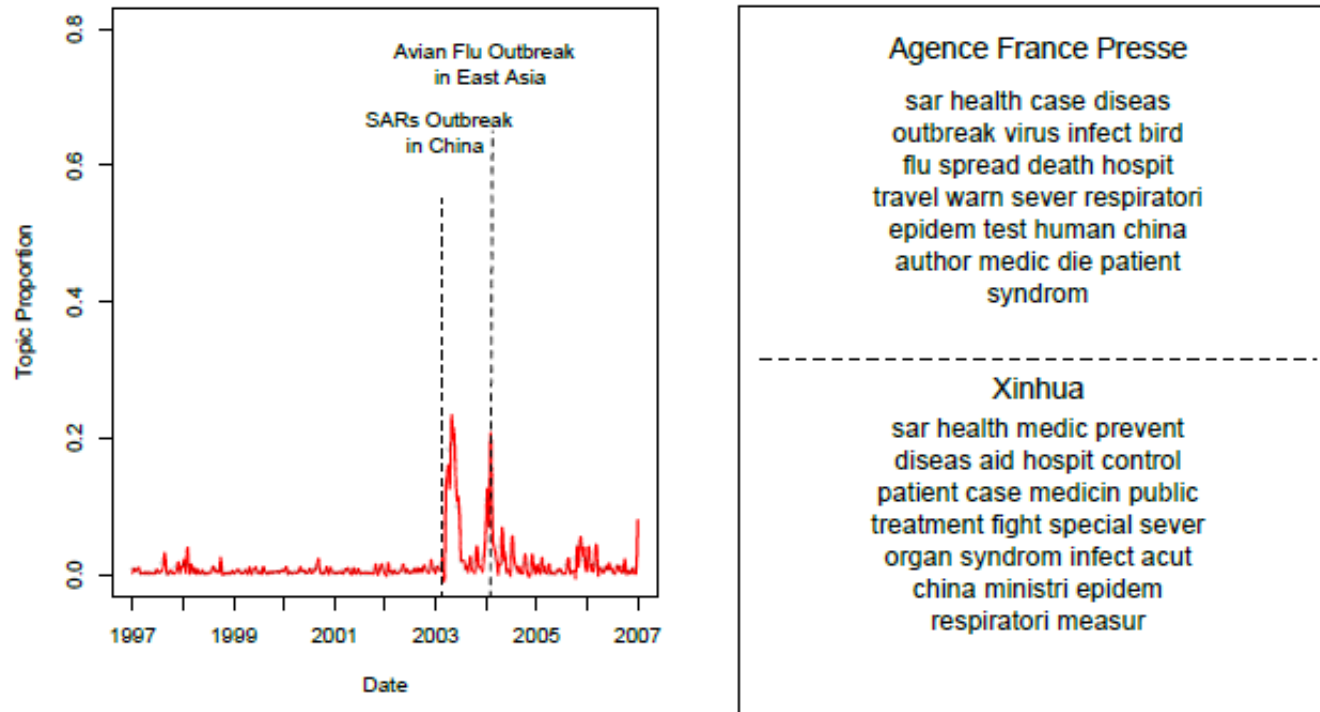


Figure 8: SARS and Avian Flu. Each dot represents the average topic proportion in a document in that month and the line is a smoothed average across time (left panel). Comparisons between news sources (right panel).