Exploring Data Science Methods in Analyzing Text Information for Datasets

WenWei Zeng

ORISE Data Research Fellow

October 22, 2024

FCSM Research and Policy Conference

U.S. Department of Transportation

Office of the Secretary of Transportation

Bureau of Transportation Statistics

Disclaimer:

- The views represented in this presentation are those of the authors and not necessarily the views of the Bureau of Transportation Statistics (BTS) or the U.S. Department of Transportation (USDOT), or of the U.S. Census Bureau.

Table of contents

Background

Introduction:

- Describe data and research questions

Overview: timeline, workflow

- Text preprocessing, visualization, and analysis

Outputs

- Word frequency graphs
- topic models

Future steps

Background

- **Data analyzed: Commodity Flow Survey (CFS):** a shipper survey conducted by the United States Department of Transportation (USDOT) every five years through a partnership between the Bureau of Transportation Statistics and the U.S. Census Bureau.
  - samples establishments in various industries from manufacturing, wholesale to retail that ship commodities.
  - collected information including **shipment value** and **weight**, **commodity type**, **origin** and **destination** locations of shipments, and mode of transportation
- Available in **2 versions**: publicly available version and a restricted-access version, which the latter is available by request via standard application process (SAP)
- Federal Statistical Research Data Centers (FSRDC): compiles federal datasets, including the CFS dataset, used by government agencies and research institutions
- After approval, data users use FSRDC to get access to the requested data

Introduction to Project

Background: To acquire restricted-access CFS data at a FSRDC, data users need to submit a proposal and receive an approval.

Task: Analyze 44 project proposals with varying subjects (written 1998-2023) from various research institutions for the **research questions:** 1. **which research areas are addressed** and 2. **which aspects of CFS are analyzed**?

Approach: The study used text analysis to review proposals that requested restricted-access CFS data with:

1. Frequency count of word(s)
2. Topic models

Issues that can arise:

- Reading unfamiliar subject matter
- Complex academic language
- Diverse methodologies and findings
- Time constraints
- Difficulty in identifying key themes and connections

What to do?

Potential Resolutions

1. Reorganize thoughts… then set aside time for reviewing,
2. Read through each paper one by one,
3. Try something new… use text analysis

Introduction: what is text analysis?

- In **text analysis**, we treat text as data
  - take text, filter, then count as we do with numeric data analysis,
  - interpret results in context of subject matter
- **Text Mining:**
  - discovers patterns in text, where you find out which certain words or ideas often appear together
  - helps uncover underlying themes or structure
- **Text Models** (one example of text mining)
  - one example is latent Dirichlet allocation (LDA), which takes words and creates clusters of keywords related to a topic
  - groups text from all documents into different number of topics according to their context

timeline

May 2023

Received project proposals

Sep.–Oct. 2023

Familiarized with R program, and tried to extract text

Nov. 2023

Identified keywords, filtered text

Jan.–Mar. 2024

Built text models

Apr. 2024

Finalized text analysis part of the project

May 2024 and beyond

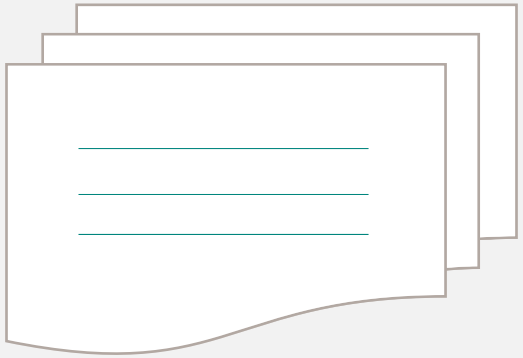
To be determined…

Workflow

- Researchers review 44 papers
- Pull out text and store
- Preprocess (filter text, filter out numbers, equation symbols and punctuation, and other similar filters)
- Reorganize text into frequency count of words
- Place text to a text model
- Receive insights from data exploration and models

Workflow (demo)

Filter



Extract words

organize

Workflow:

1. Pull words out from papers
2. represent words in data matrix*: counting how many words occur per document
3. Take text and graph word frequency plot
4. Build text models

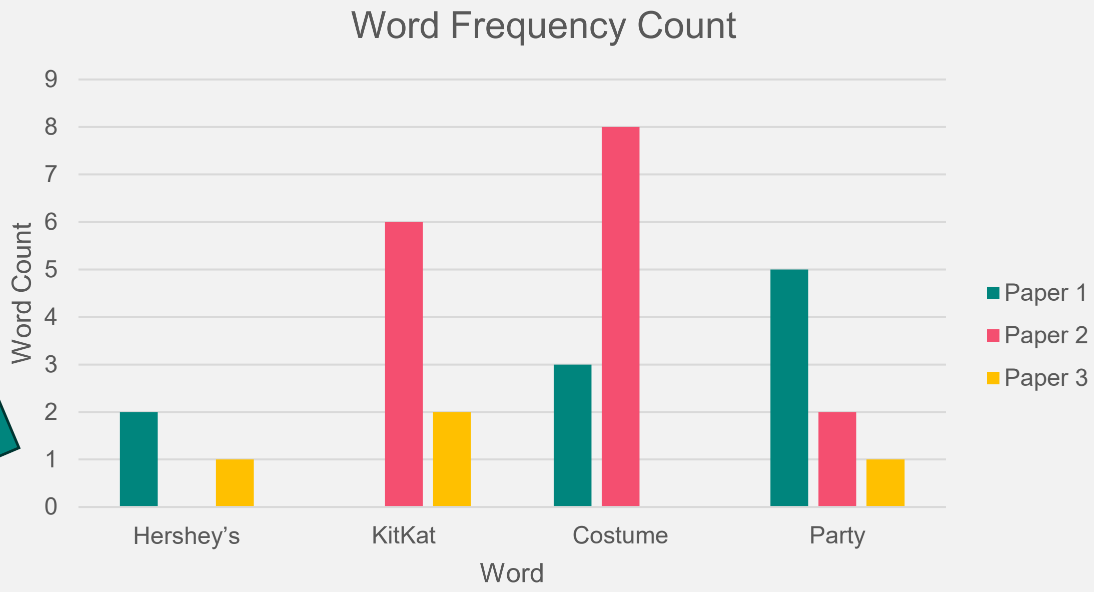
***Term-Document Matrix**

word

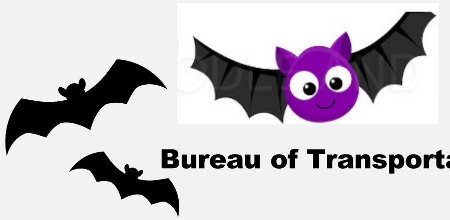
document

	Hershey's	KitKat	Costume	Party
Paper 1	2	0	3	5
Paper 2	0	6	8	2
Paper 3	1	2	0	1

graph



Build text models



Bureau of Transportation Statistics

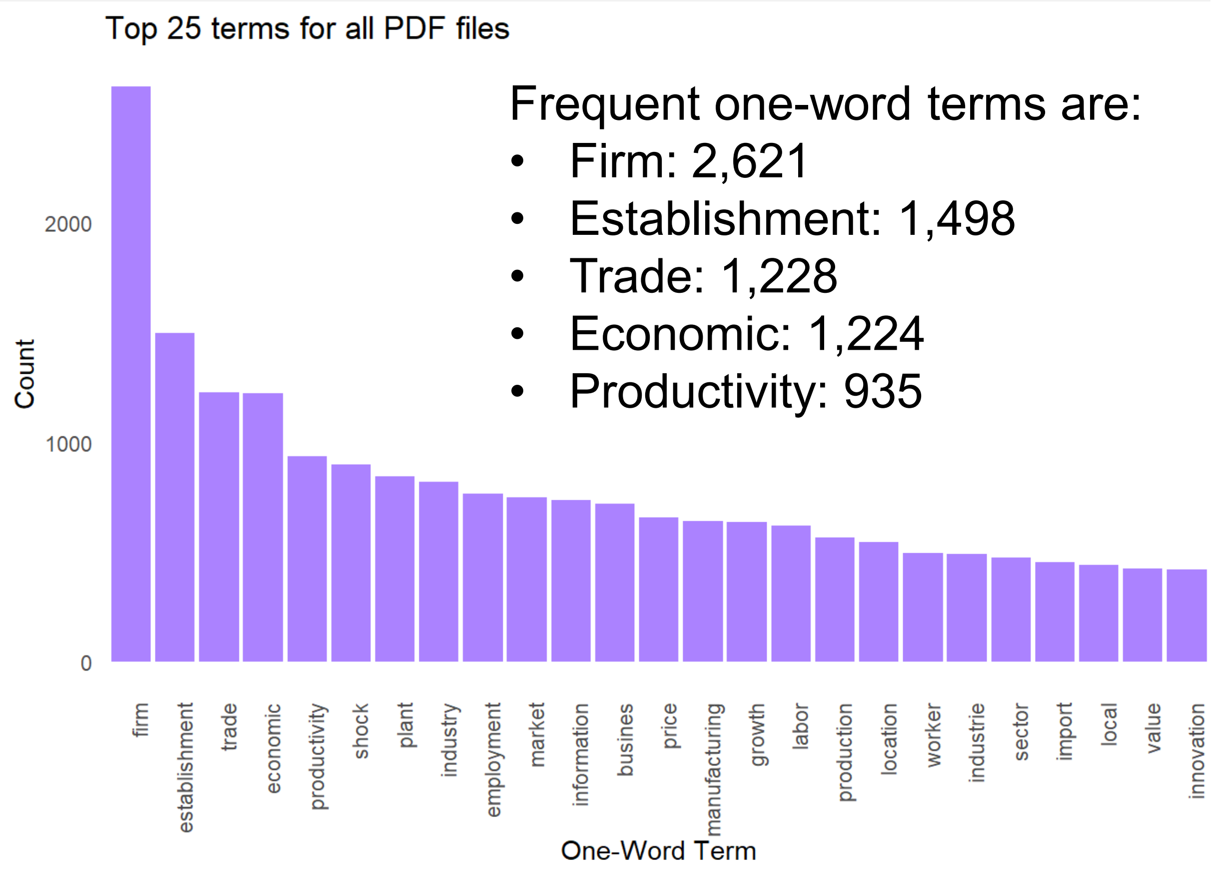
Outputs: word count

- Words found – stored into matrix (sample shown)

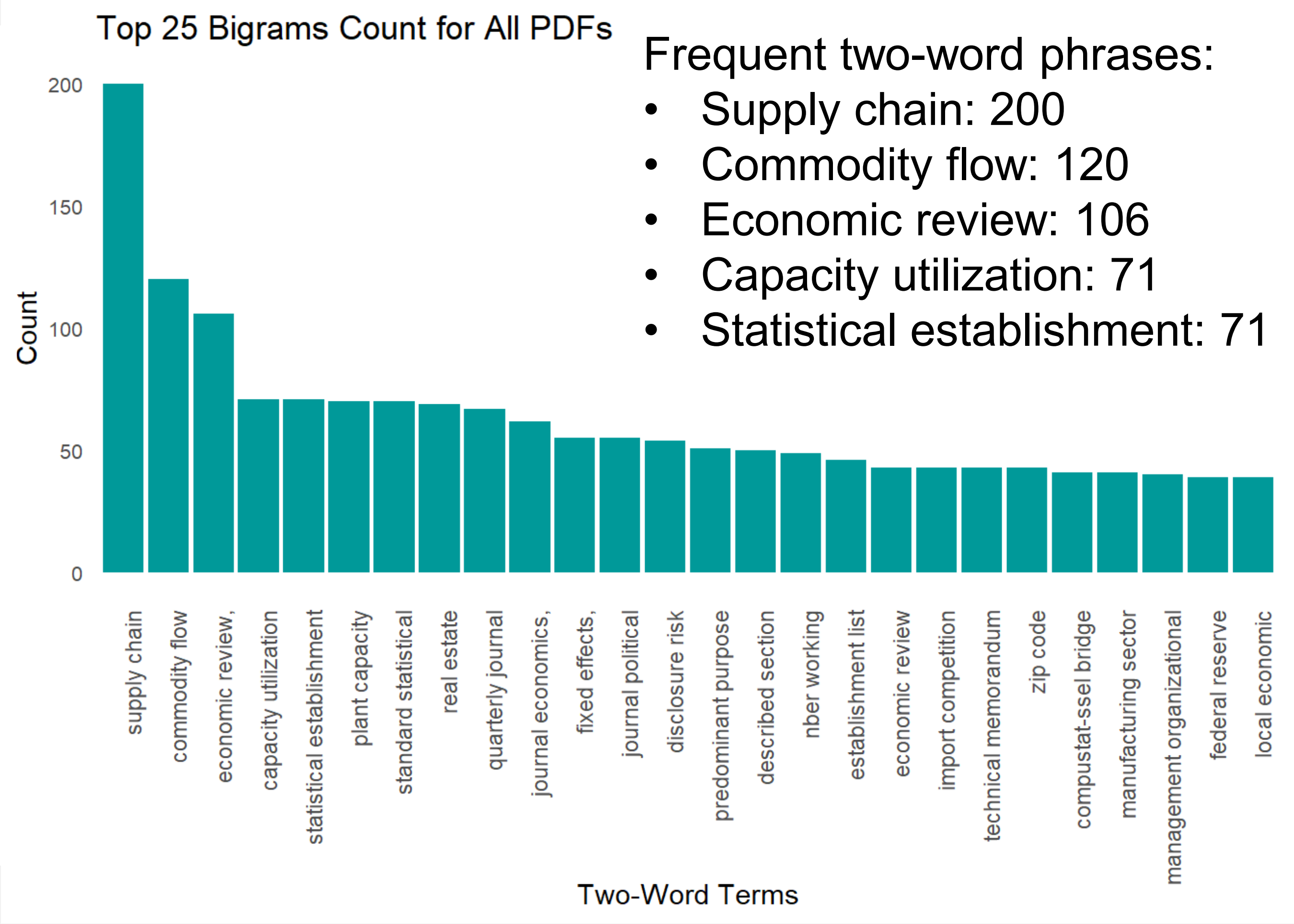
```
## <<TermDocumentMatrix (terms: 13188, documents: 44)>>
```

	term	document	count
1	independent	proposal_1276.pdf	1
2	secondary	proposal_1571.pdf	1
3	rather	proposal_1652.pdf	1
4	hire	proposal_2157.pdf	1
5	product	proposal_1518.pdf	8
6	crt	proposal_2157.pdf	3
7	rdc	proposal_2439.pdf	1
8	secondary	proposal_2427.pdf	2
9	independent	proposal_1287.pdf	1
10	hire	proposal_2439.pdf	1
11	demand	proposal_1975.pdf	7
12	rdc	proposal_2539.pdf	4
13	product	proposal_1499.pdf	1
14	demand	proposal_2389.pdf	2
15	crt	proposal_2396.pdf	3
16	rather	proposal_2210.pdf	4

Outputs: word frequency plots



Frequent one-word terms are:
- Firm: 2,621
- Establishment: 1,498
- Trade: 1,228
- Economic: 1,224
- Productivity: 935

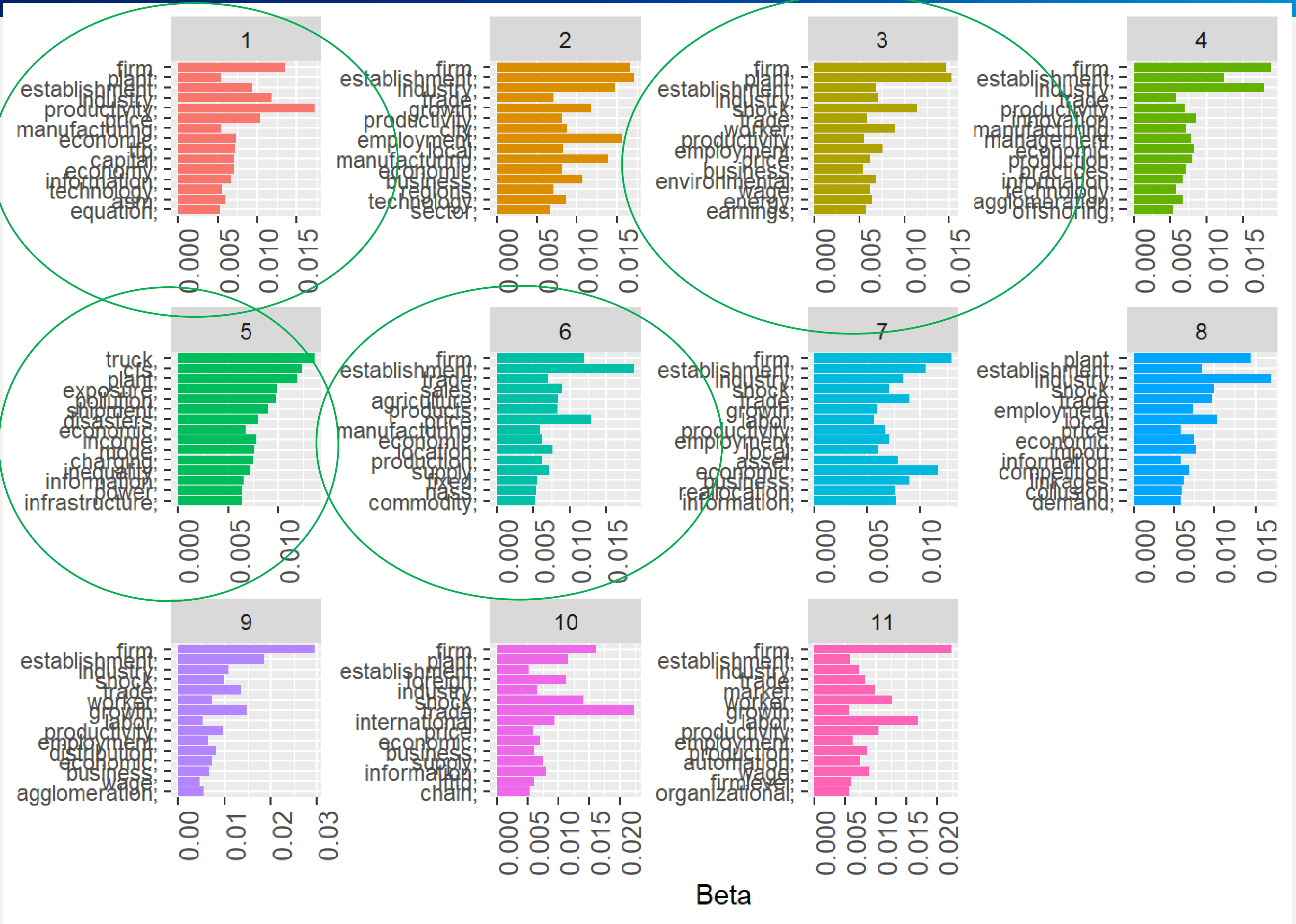


Frequent two-word phrases:
- Supply chain: 200
- Commodity flow: 120
- Economic review: 106
- Capacity utilization: 71
- Statistical establishment: 71

Results from LDA model of 11-topics

This 11-topic LDA model displays different terms per topic with associated beta probabilities of term occurrence within a topic.

We will look at the circled topics.

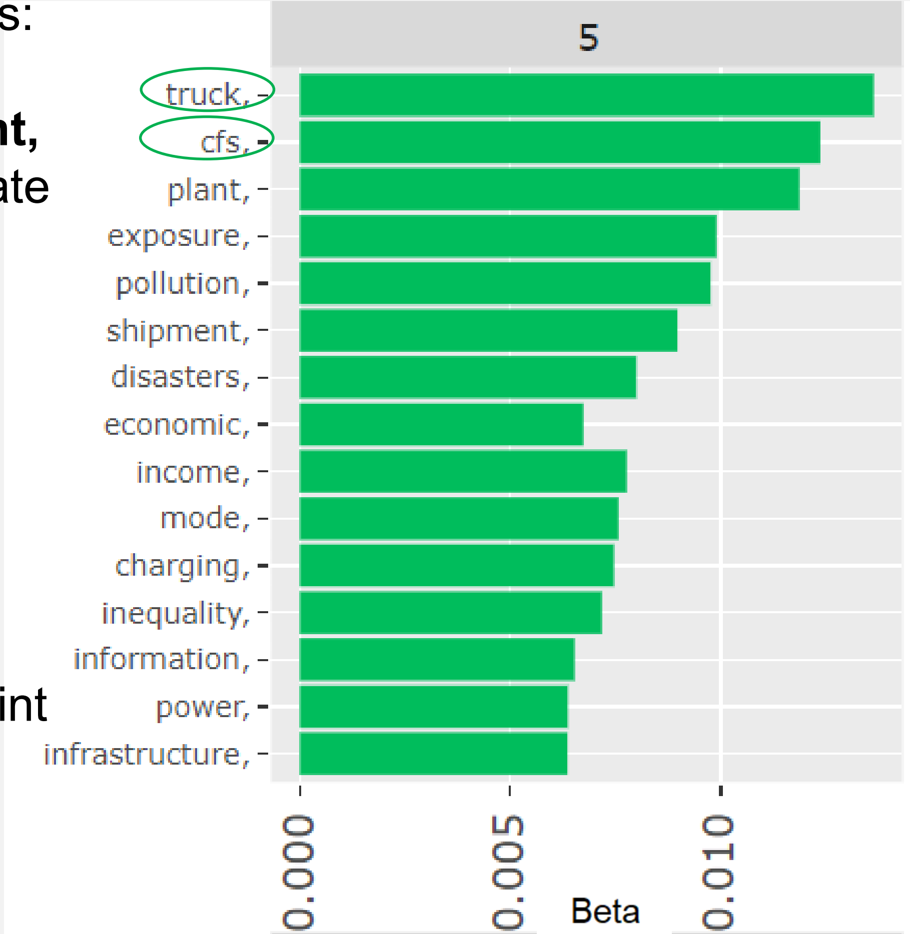


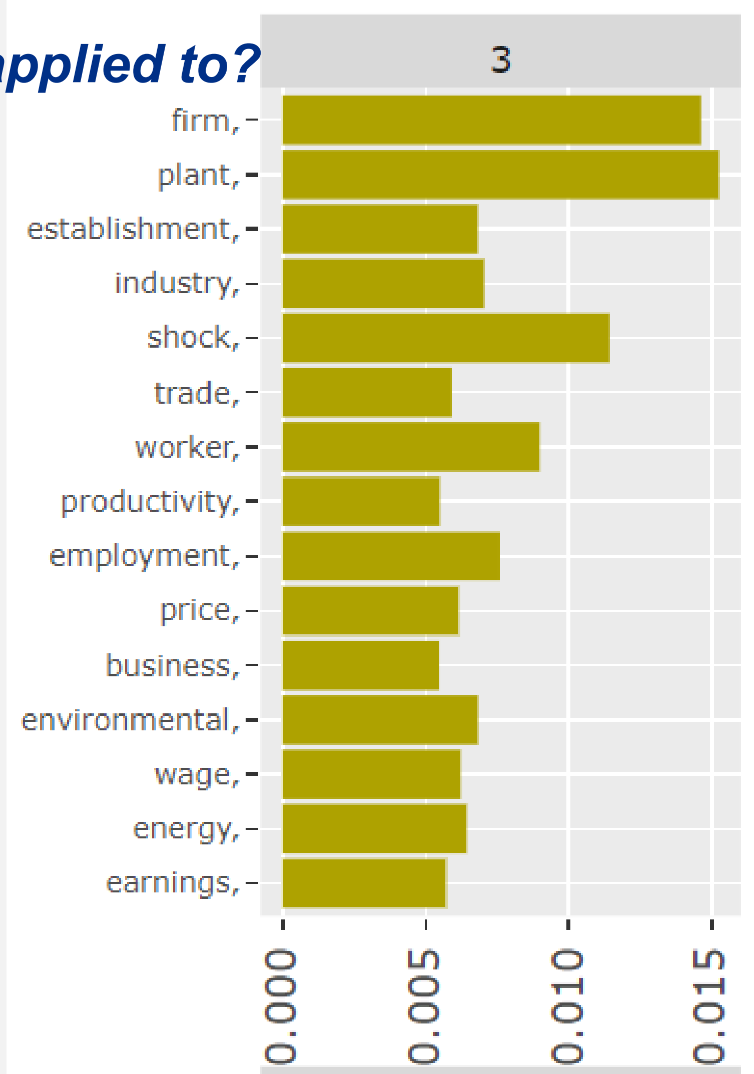
Results from 11-topic LDA model:

Addressed question: For which research areas will the papers be applied to?

- Topic 5 covers the terms: **truck (beta= 0.0136), CFS (0.0124), shipment, mode, information** relate to transportation, while terms like **economic, income, plant, infrastructure** suggest economics terms,

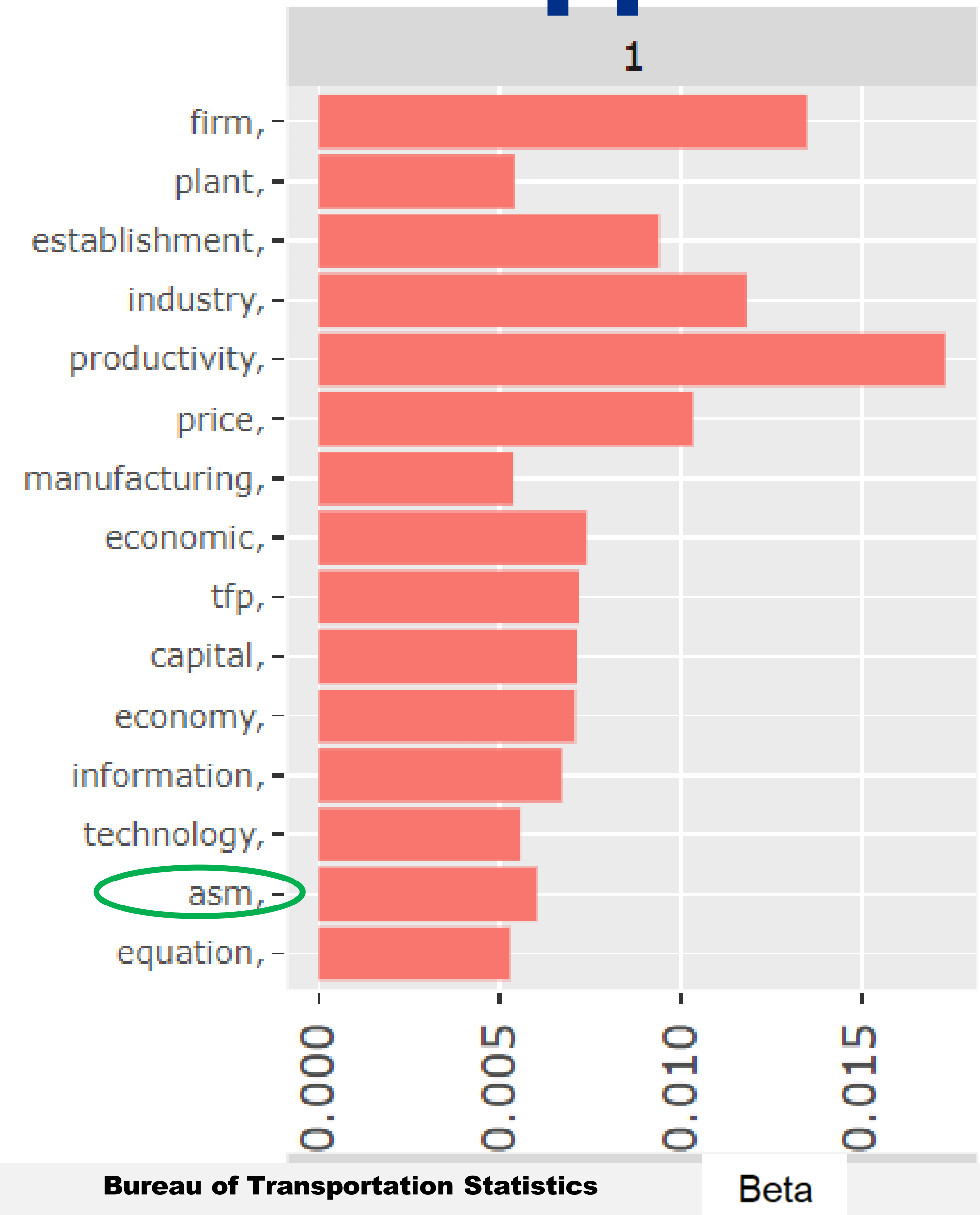
- Topic 5 also included words such as **power, charging, pollution** point to environmental concerns, indicating a possible topic: transportation-economics-environmental as a subject of discussion

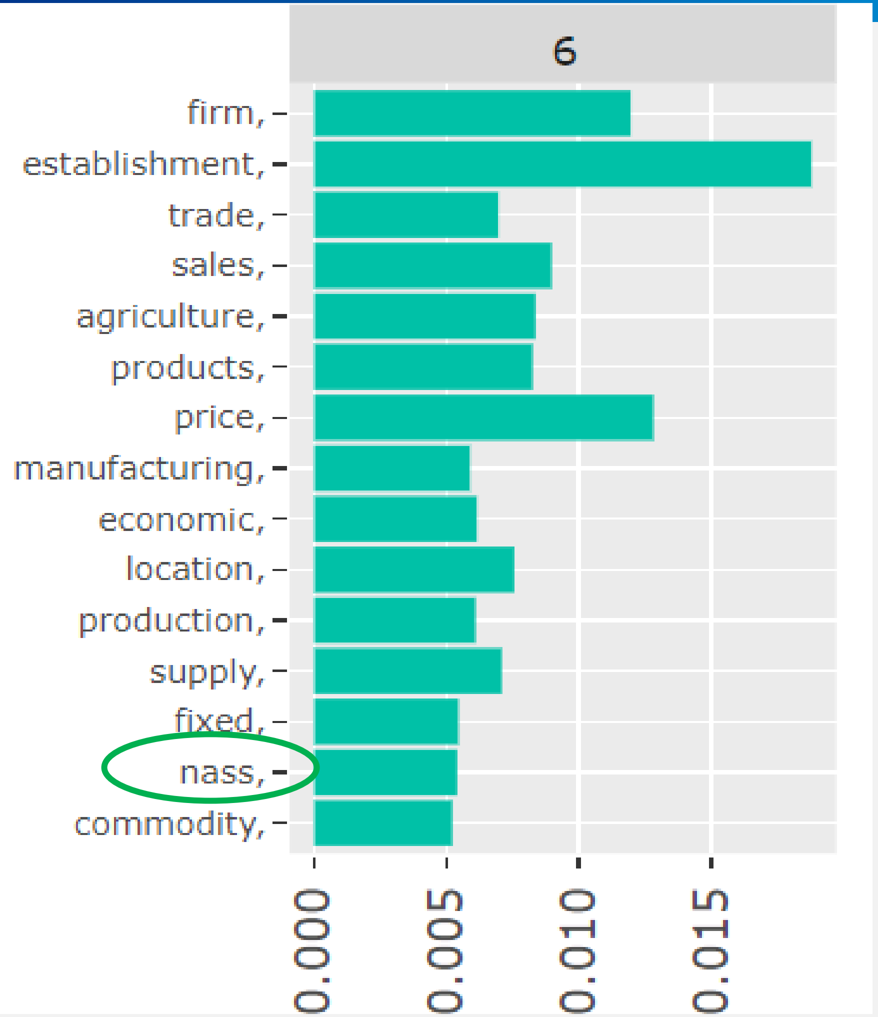




- Topic 3 grouped words like **trade, worker, productivity, employment, wage, earnings**, indicating *financial or business operations*, alongside terms such as **firm, plant, establishment, industry**, related to *where economic activity occurs*, indicating financial-business-economic topic

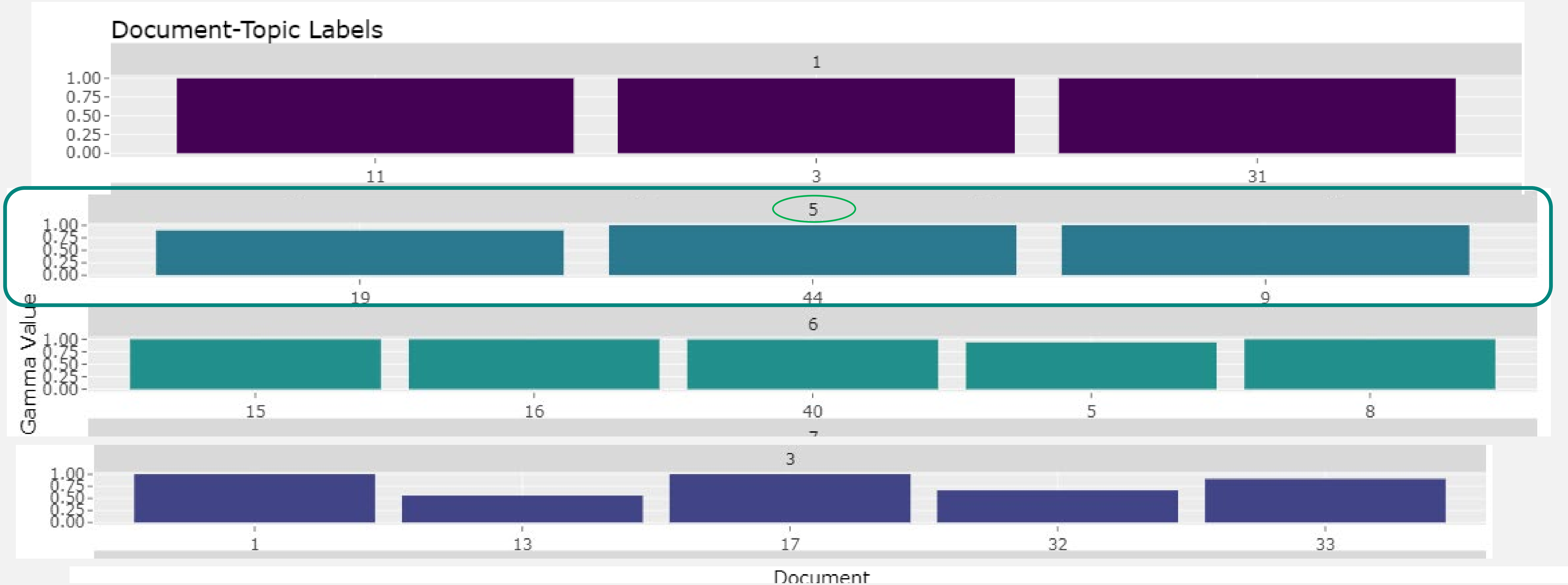
Results: applications





- Topic models also identified other data sources:
  - **ASM** (Annual Survey of Manufactures) for manufacturing and economics research,
  - **NASS** (National Agricultural Statistics Service) for agriculture trade and related subjects

Outputs: how topic models connect to papers

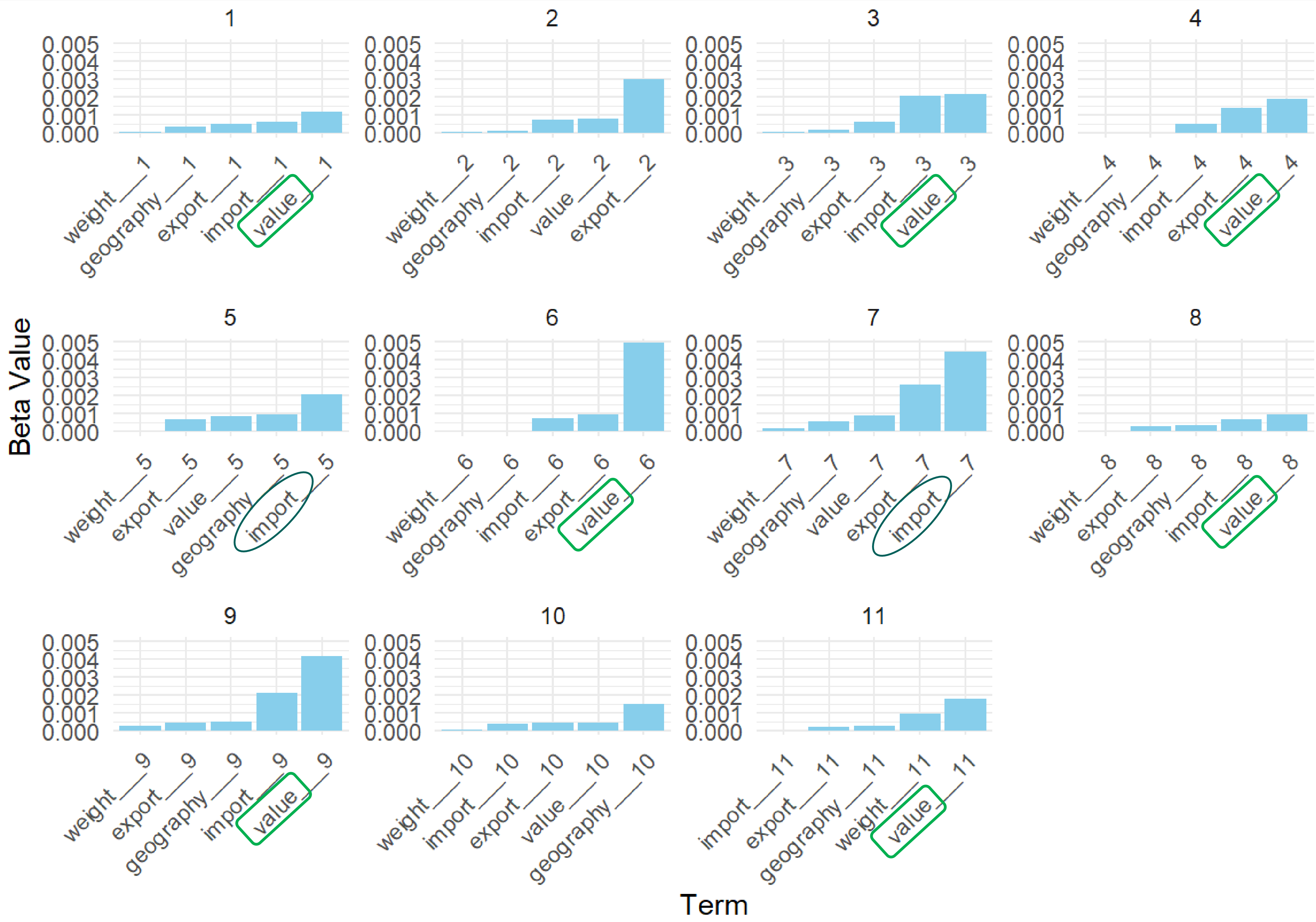


- Gamma value represents the document-topic distribution, which indicates how much each topic matches a specific document,
- Here, papers 9, 19, and 44 frequently discussed about topic 5 *transportation-economics-environmental topic,* which includes CFS

Results: Topic models of customized variables of CFS

Answers question: Which aspects of the CFS data are analyzed?

Results found: **shipment value** in 7 out of 11 topics as the most frequent variable, the **import** variable was found in 2 out of 11 topics



Findings:

- From reviewing 44 FSRDC proposals, we identified and sorted 13,188 unique terms:

1. Frequently occurring words were detected from the fields of **transportation** with *CFS* and *commodity flow***,** along with **economics, agriculture, environment, finance**, and **business**
2. Identified frequent CFS variables across proposals, with **<u>shipment value</u>** variable appearing in 7 of 11 topics as the most frequent, and the **<u>import</u>** variable found in 2 out of 11 topics

- This approach provides a supplementary method to:
  - Visualize how **CFS** and other **federal or Census datasets** are used by data requesters
  - Explore key subjects and keywords, offering insights into areas of research interest
  - Help analysts focus on relevant topics, connect the topics found to specific papers, improving the efficiency of the review process

Future considerations:

- **Search for additional natural language processing (NLP) models**
  - Explore other natural language processing (NLP) models to represent subjects from large text datasets
  - Look into models in addition to Latent Dirichlet Allocation (LDA)
- **Build visualizations**
  - Develop visualizations, such as counting words found per paper to enhance data interpretation.
- **Explore applications in transportation datasets and other data sources**
  - Example: Use text summarization techniques to condense long articles into shorter summaries and keywords
- **Search for validation techniques and metrics for model fit**
  - Identify and apply methods to validate the accuracy and effectiveness of text models

References

- Bureau of Transportation Statistics. Commodity Flow Survey (CFS). U.S. Department of Transportation, 2024. https://www.bts.gov/cfs. Accessed July 30, 2024.
- U.S. Census Bureau. Research Data Centers. U.S. Department of Commerce, 2022. https://www.census.gov/about/adrm/fsrdc/locations.html. Accessed July 30, 2024.
- U.S. Census Bureau. Federal Statistical Research Data Centers. U.S. Department of Commerce, 2023. https://www.census.gov/about/adrm/fsrdc.html. Accessed July 30, 2024.
- Bureau of Transportation Statistics and U.S. Census Bureau. 2017 Commodity Flow Survey (CFS) Public Use File (PUF) Data Users Guide. U.S. Department of Transportation and U.S. Department of Commerce, Washington, DC, 2020.
- U.S. Census Bureau. Projects. U.S. Department of Commerce, July 7, 2023. https://www.census.gov/about/adrm/fsrdc/about/ongoing-projects.html. Accessed October 17, 2023.
- Ooms, J. pdftools: Text Extraction, Rendering and Converting of PDF. R package version 3.4.9, 2023. https://cran.r-project.org/web/packages/pdftools/.
- Gupta, R.K., Agarwalla, R., Naik, B. H., Evuri, J.R., Thapa, A., and Singh, T.D. Prediction of Research Trends using LDA based Topic Modeling. Global Transitions Proceedings, 2022. 3(1): 298-304.
- Silge, J., and Robinson, D. Text mining with R. O'Reilly Media, Inc., 2017.
- Feinerer, I. and Hornik, K. tm: Text Mining Package. R package version 0.7-13, 2024. https://CRAN.R-project.org/package=tm.
- Asmussen, C.B., and Møller, C. Smart Literature Review: A practical topic modelling approach to exploratory literature review. Journal of Big Data, 2019. 6:93. https://doi.org/10.1186/s40537-019-0255-7.
- Weston S.J., Shryock I., Light R., and Fisher P.A. Selecting the Number and Labels of Topics in Topic Modeling: A Tutorial. Advances in Methods and Practices in Psychological Science, 2023. 6(2). https://doi.org/10.1177/25152459231160105.
- Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. Optimizing Semantic Coherence in Topic Models. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011. pp. 262-272. Edinburgh, Scotland, UK. Association for Computational Linguistics.
- The National Academies of Sciences, Engineering, and Medicine. TRID. https://trid.trb.org/. Accessed November 16, 2023.
- Ponweiser, M. Latent Dirichlet Allocation in R (Diploma Thesis). Vienna University of Economics and Business, Vienna, 2012. https://doi.org/10.57938/533618e5-dcd9-4c8f-913a2339fa145c71.
- Sievert, C., and Shirley, K. LDAvis: A Method for Visualizing and Interpreting Topics. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, 2014. pp. 63-70. Baltimore, Maryland. Association for Computational Linguistic.
- Chuang, J., Ramage, D., Manning, C.D., and Heer. J. Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis. In Proceedings of 2012 ACM SIGCHI Conference on Human Factors in Computing Systems, 2012. Austin, Texas.

Acknowledgements

Many thanks to the following members for the development of this project

- Contributors from USDOT:
  - Cha-Chi Fan
  - Ryan Grube
  - Young-Jun Kweon
  - Joseph McGill
  - Mike Carter
- U.S. Census Bureau contributor:
  - Berin Linfors

Thank you

for listening ☺

Post-presentation survey

U.S. Department of Transportation

Thank you for listening,

- Please scan QR code to complete a quick survey,
- Enjoy the rest of conference ☺

Contact

WenWei Zeng
ORISE Data Research Fellow
U.S. Dept. of Transportation OST
wenwei.zeng.ctr@dot.gov