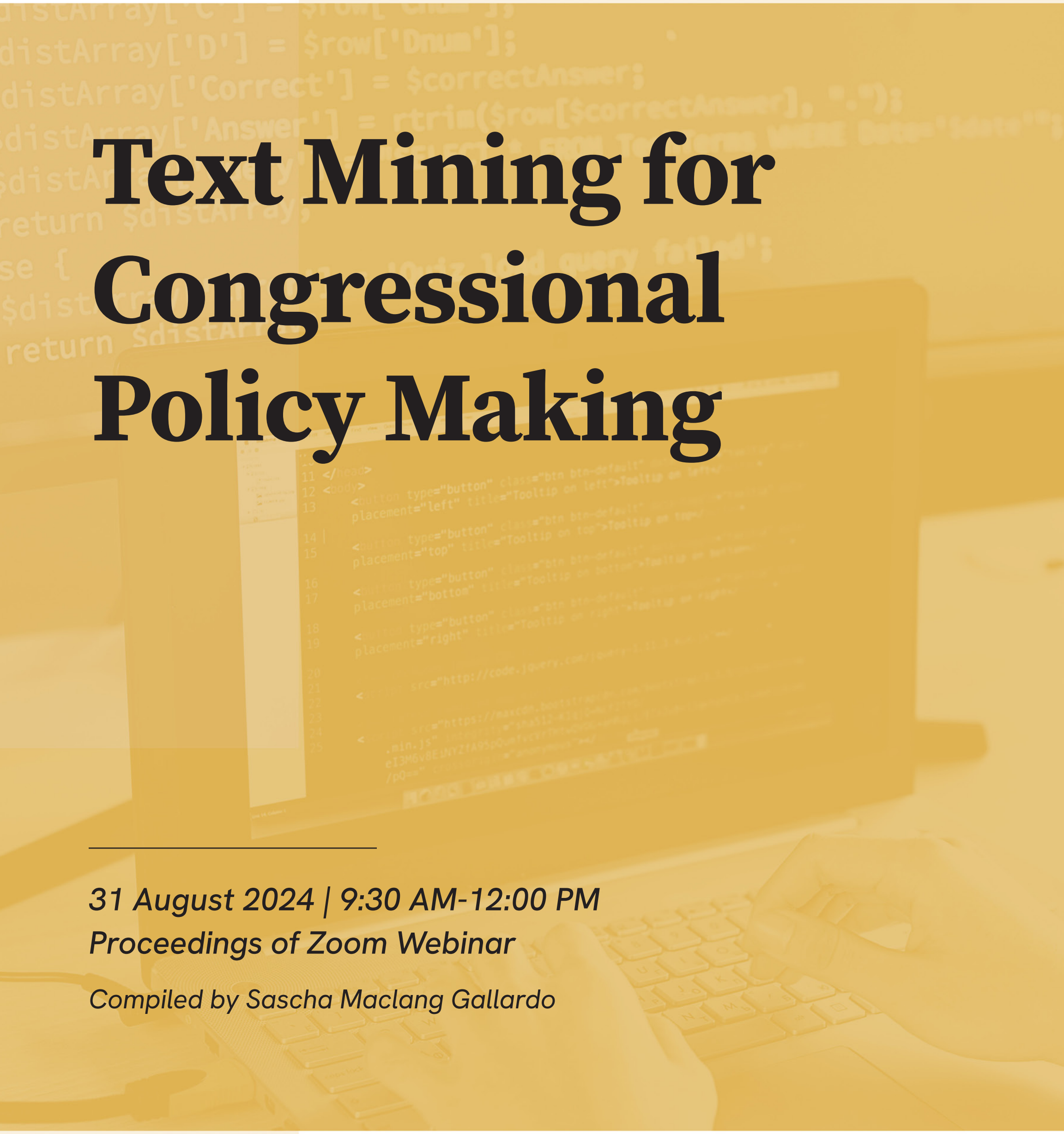
Proceedings 2025-06



# Text Mining for Congressional Policy Making

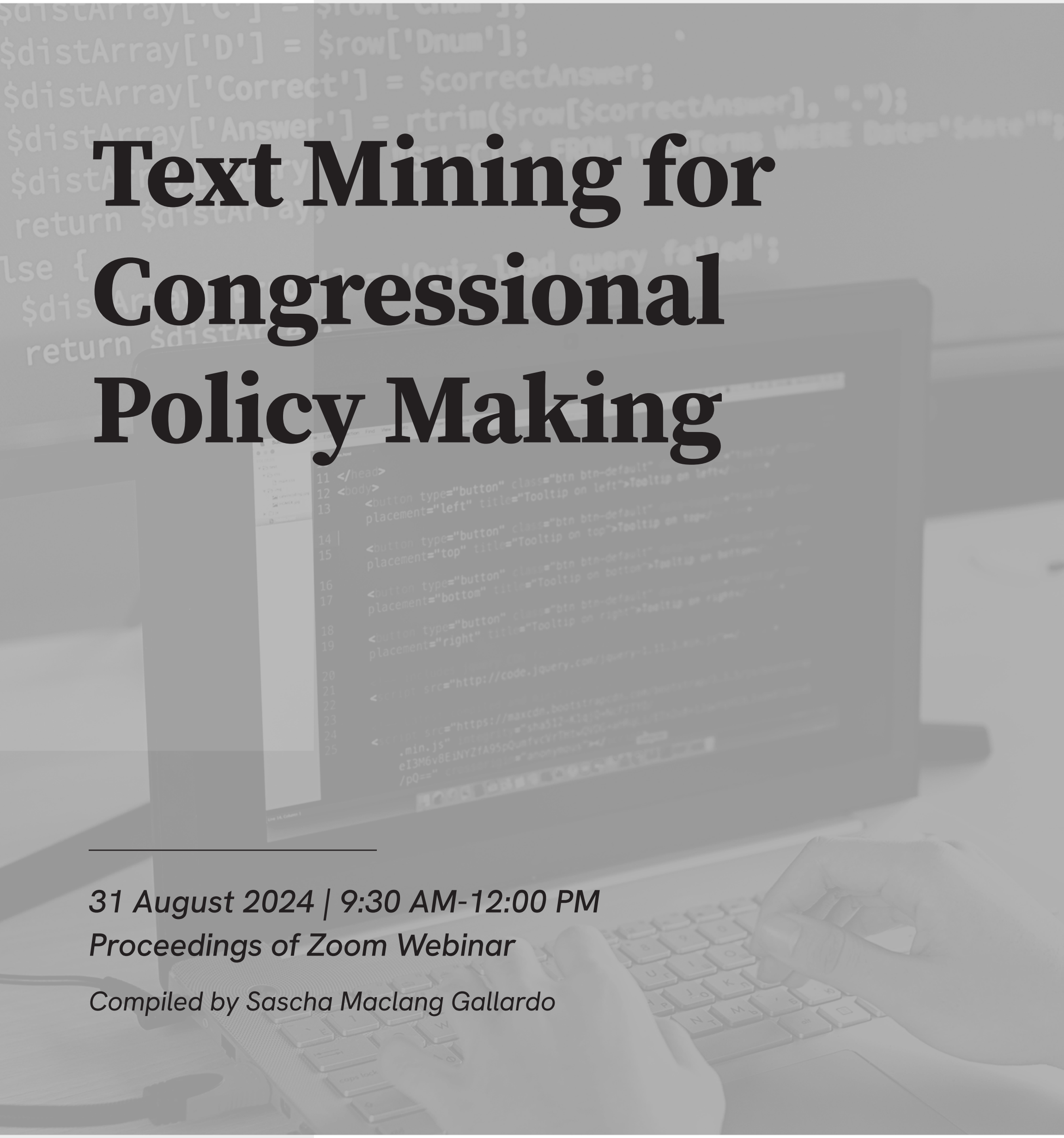
*31 August 2024 | 9:30 AM-12:00 PM*

*Proceedings of Zoom Webinar*

*Compiled by Sascha Maclang Gallardo*

upcids Social & Political Change

UNIVERSITY OF THE PHILIPPINES
CENTER FOR
INTEGRATIVE AND
DEVELOPMENT
STUDIES



# Text Mining for Congressional Policy Making

*31 August 2024 | 9:30 AM-12:00 PM*
*Proceedings of Zoom Webinar*

*Compiled by Sascha Maclang Gallardo*

UP CIDS | PROCEEDINGS | 2025-06

**UP CIDS Proceedings**
is published by the

**University of the Philippines**
**Center for Integrative and Development Studies**
Lower Ground Floor, Ang Bahay ng Alumni
Magsaysay Avenue, University of the Philippines
Diliman, Quezon City 1101

cids.up.edu.ph

**Telephone:** (02) 8981-8500 loc. 4266 to 4268 / (02) 8426-0955
**Email:** cidspublications@up.edu.ph
**Website:** cids.up.edu.ph

The views and opinions expressed in this discussion paper are those of the author/s and neither reflect nor represent those of the University of the Philippines or the UP Center for Integrative and Development Studies. No copies can be made in part or in whole without prior written permission from the authors/editors and the publisher.

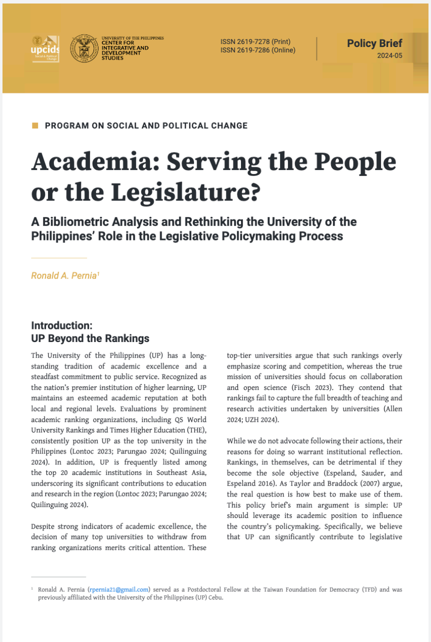
**ISSN 2718-9295 (Print)**
**ISSN 2718-9309 (Online)**

**Cover image credit**
"A Person Doing Computer Programming"

Mikhail Nilov, Pexels, https://www.pexels.com/photo/a-person-doing-computer-programming-7988086/

# Table of Contents

# Download related policy papers for *free*



UNIVERSITY OF THE PHILIPPINES
CENTER FOR INTEGRATIVE AND DEVELOPMENT STUDIES

ISSN 2619-7278 (Print)
ISSN 2619-7286 (Online)

Policy Brief
2024-05

PROGRAM ON SOCIAL AND POLITICAL CHANGE

## Academia: Serving the People or the Legislature?

### A Bibliometric Analysis and Rethinking the University of the Philippines' Role in the Legislative Policymaking Process

*Ronald A. Pernia*[1]

### Introduction: UP Beyond the Rankings

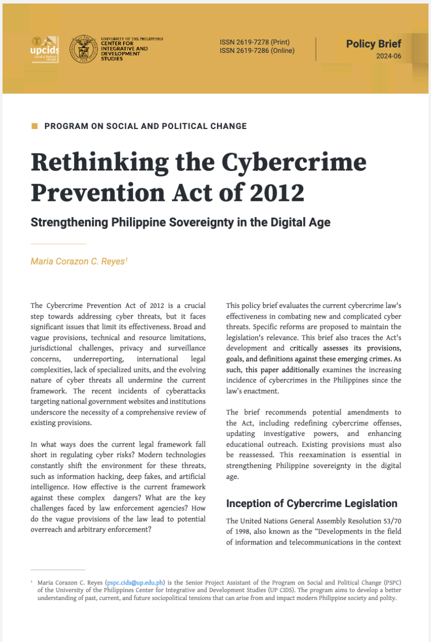
The University of the Philippines (UP) has a long-standing tradition of academic excellence and a steadfast commitment to public service. Recognized as the nation's premier institution of higher learning, UP maintains an esteemed academic reputation at both local and regional levels. Evaluations by prominent academic ranking organizations, including QS World University Rankings and Times Higher Education (THE), consistently position UP as the top university in the Philippines (Lontoc 2023; Parungao 2024; Quilinguing 2024). In addition, UP is frequently listed among the top 20 academic institutions in Southeast Asia, underscoring its significant contributions to education and research in the region (Lontoc 2023; Parungao 2024; Quilinguing 2024).

Despite strong indicators of academic excellence, the decision of many top universities to withdraw from ranking organizations merits critical attention. These top-tier universities argue that such rankings overly emphasize scoring and competition, whereas the true mission of universities should focus on collaboration and open science (Fisch 2023). They contend that rankings fail to capture the full breadth of teaching and research activities undertaken by universities (Allen 2024; UZH 2024).

While we do not advocate following their actions, their reasons for doing so warrant institutional reflection. Rankings, in themselves, can be detrimental if they become the sole objective (Espeland, Sauder, and Espeland 2016). As Taylor and Braddock (2007) argue, the real question is how best to make use of them. This policy brief's main argument is simple: UP should leverage its academic position to influence the country's policymaking. Specifically, we believe that UP can significantly contribute to legislative

[1] Ronald A. Pernia (rpernia21@gmail.com) served as a Postdoctoral Fellow at the Taiwan Foundation for Democracy (TFD) and was previously affiliated with the University of the Philippines (UP) Cebu.

*Academia: Serving the People or Legislature?: A Bibliometric Analysis and Rethinking the University of the Philippines' Role in the Legislative Policymaking Process*



UNIVERSITY OF THE PHILIPPINES
CENTER FOR INTEGRATIVE AND DEVELOPMENT STUDIES

ISSN 2619-7278 (Print)
ISSN 2619-7286 (Online)

Policy Brief
2024-06

PROGRAM ON SOCIAL AND POLITICAL CHANGE

## Rethinking the Cybercrime Prevention Act of 2012

### Strengthening Philippine Sovereignty in the Digital Age

*Maria Corazon C. Reyes*[1]

The Cybercrime Prevention Act of 2012 is a crucial step towards addressing cyber threats, but it faces significant issues that limit its effectiveness. Broad and vague provisions, technical and resource limitations, jurisdictional challenges, privacy and surveillance concerns, underreporting, international legal complexities, lack of specialized units, and the evolving nature of cyber threats all undermine the current framework. The recent incidents of cyberattacks targeting national government websites and institutions underscore the necessity of a comprehensive review of existing provisions.

In what ways does the current legal framework fall short in regulating cyber risks? Modern technologies constantly shift the environment for these threats, such as information hacking, deep fakes, and artificial intelligence. How effective is the current framework against these complex  dangers? What are the key challenges faced by law enforcement agencies? How do the vague provisions of the law lead to potential overreach and arbitrary enforcement?

This policy brief evaluates the current cybercrime law's effectiveness in combating new and complicated cyber threats. Specific reforms are proposed to maintain the legislation's relevance. This brief also traces the Act's development and critically assesses its provisions, goals, and definitions against these emerging crimes. As such, this paper additionally examines the increasing incidence of cybercrimes in the Philippines since the law's enactment.

The brief recommends potential amendments to the Act, including redefining cybercrime offenses, updating investigative powers, and enhancing educational outreach. Existing provisions must also be reassessed. This reexamination is essential in strengthening Philippine sovereignty in the digital age.

### Inception of Cybercrime Legislation

The United Nations General Assembly Resolution 53/70 of 1998, also known as the "Developments in the field of information and telecommunications in the context

[1] Maria Corazon C. Reyes (pspc.cids@up.edu.ph) is the Senior Project Assistant of the Program on Social and Political Change (PSPC) of the University of the Philippines Center for Integrative and Development Studies (UP CIDS). The program aims to develop a better understanding of past, current, and future sociopolitical tensions that can arise from and impact modern Philippine society and polity.

*Rethinking the Cybercrime Prevention Act of 2012: Strengthening Philippine Sovereignty in the Digital Age*

**cids.up.edu.ph/collection-database**

## Get your policy papers published.

## Download open-access articles.

The Philippine Journal of Public Policy: Interdisciplinary Development Perspectives (PJPP), the annual peer-reviewed journal of the UP Center for Integrative and Development Studies (UP CIDS), welcomes submissions in the form of full-length policy-oriented manuscripts, book reviews, essays, and commentaries. The PJPP provides a multidisciplinary forum for examining contemporary social, cultural, economic, and political issues in the Philippines and elsewh ere. Submissions are welcome year-around.

***For more information, visit cids.up.edu.ph.***
***All issues/articles of the PJPP can be downloaded for free.***

## Get news and the latest publications.

Join our mailing list to get our publications delivered straight to your inbox! Also, you'll receive news of upcoming webinars and other updates.

***bit.ly/signup_cids***

## We need your feedback.

Have our publications been useful? Tell us what you think.

***bit.ly/dearcids***

# About the Proceedings

The “Text Mining for Congressional Policymaking” webinar was held on August 31, 2024 via Zoom. Organized by the Program on Social and Political Change (PSPC) of the University of the Philippines Center for Integrative and Development Studies (UP CIDS), the webinar aimed to provide foundational knowledge on text mining and highlight its potential for evidence-based policymaking.

The main speaker, Dr. Ronald Pernia, provided an overview of advanced text mining techniques for processing unstructured data in the social sciences. He focused on congressional documents from the Congressional Policy and Budget Research (CPBRD) website of the Philippine House of Representatives to demonstrate insights into policymaking. He demonstrated how tools like R can extract valuable information from large datasets.

The presentation was followed by a reaction from Prof. Dominador M. Gamboa, the Service Director of CPBRD. Director Gamboa emphasized the importance of predictive approaches in policymaking and underscored the value of data as a resource in the present time.

This webinar was moderated by Maria Corazon Reyes, Senior Project Assistant of UP CIDS Program on Social and Political Change and this event was documented by Sascha Maclang Gallardo.

# Welcome Remarks

Rogelio Alicor Panao, PhD[1]

In his welcome remarks, Dr. Rogelio Alicor Panao gave a brief overview of the history and mission of the UP Center for Integrative and Development Studies. He also highlighted the aim of the Program on Social and Political Change's (PSPC) "to allow experts from a variety of disciplines in the university to develop a better understanding of past, current, and future social and political tensions that can arise and impact on modern Philippine society and polity." In line with the webinar's focus on "Text Mining for Congressional Policymaking," Dr. Panao defined text mining by comparing it to "digging through a huge pile of written information to find useful insights." He further adds that text mining functions with "computer programs and algorithms to scan all those texts, pick out keywords, for example, and then sort or come up with patterns or themes and summarize the main ideas." On the benefits of text mining, Dr. Panao emphasized its potentials for both UP and the Philippines as a whole. According to him, text mining enables the University to "assess its influence on public policy." Meanwhile, for the country, text mining allows the government "to navigate the complexities of data scarcity, gain deeper insights into public sentiment, and enhance democratic engagement." Dr. Panao concluded by stating how embracing text mining and harnessing its potentials may lead "to more informed and effective decision making."

[1] Dr. Rogelio Alicor Panao is an associate professor at the Department of Political Science, UP Diliman. He is also the convenor of the Program on Social and Political Change at UP CIDS at the time of the webinar.

# “Text Mining in R”: Representing Text as Data

Dr. Ronald Pernia[2]

Dr. Pernia presented an overview of advanced text mining techniques and their significance in processing unstructured data within the realm of social sciences. He particularly focused on documents from the Congressional Policy and Budget Research Department (CPBRD) of the Philippine House of Representatives. His analysis aimed to uncover insights regarding congressional policymaking through the application of various text mining strategies. Utilizing software such as R, he demonstrated how these tools can be employed to extract valuable information from large datasets.

As a preamble to his presentation, Dr. Pernia first introduced R, which is an open-source software that can execute commands necessary for text mining. He also explained that his presentation title, “Representing Text As Data,” was inspired by the book of a similar name from which he drew a lot of information from.[3]

Dr. Pernia then proceeded with presentation and emphasized the importance of words in the political landscape, particularly in shaping public opinion on contentious issues such as divorce and sexual rights. He highlighted the

[2] Dr. Ronald Pernia is a senior lecturer at the National College of Public Administration and Governance, UP Diliman and a postdoctoral fellow at the Taiwan Foundation for Democracy.

[3] Justin Grimmer et al, *Text as Data: A New Framework for Machine Learning and the Social Sciences* (Princeton: Princeton University Press, 2022.).

challenges faced by academics in extracting information from unstructured data sources, which often include social media posts and written statements from politicians.

Traditionally, researchers resorted to labor-intensive methods, such as printing and manually analyzing texts, to comprehend public sentiments. He noted that recent advancements in computer science and programming have transformed the way researchers engage with data in the social sciences.

Dr. Pernia referenced a compelling article by Professor Gill[4] published in the *Journal of Politics* which underscored the staggering volume of text data generated daily—23 billion text messages and 40,000 searches per second, as of 2021. He also stressed that this was only four years ago and that this volume is increasing exponentially. This rapid surge in data necessitates advanced analytical methods, which computer scientists have now developed.

Dr. Pernia then defined text data, explaining that while it originates from various sources, it typically appears in a natural language, which humans use and are context-dependent. He noted challenges in working with certain languages like Japanese or Chinese which may need specialized R packages. Documents like congressional policies stored in PDFs demand additional methods to effectively extract relevant information. Other examples of text data include emails, online chats, phone transcripts, online news, forums or social media accounts, and press releases. These data need to be processed as we risk missing information that could help legislators and academics.

To acquaint the audience with the book that inspired his presentation title, Dr. Pernia introduced the general theme of "Text as Data,"[5] encouraging attendees to consult it as a valuable resource. He explained that while scarcity historically defined empirical work in the social sciences, recent years have seen a shift toward data abundance. However, he noted that the core of the big data revolution extends beyond sheer volume, focusing instead on adapting research methods and design to take advantage of the opportunities presented by this extensive data environment.

[4] Jeff Gill, "Political Science is a Data Science," *The Journal of Politics* 83:1 (January, 2021).

[5] Grimmer et al. (2022), Text as Data.

In a separate article, Justin Grimmer posits that essentially, “We are all scientists now.”[6] Dr. Pernia explained that this, coupled with data abundance, makes us forget that we still need a theory to make sense of the data that we have. Aside from this, he noted that the field is fundamentally interdisciplinary. This nature often leads researchers to feel overwhelmed by the statistical analyses required. However, understanding both computer programming and statistical analysis have become essential in the social sciences due to the evolving nature of data. Rather than dismissing one field for another, this necessity reflects the rich complexity of the information available to us today.

Furthermore, Dr. Pernia referenced Grimmer and Stewart's text[7] which contends that data science, while inherently complex, is fundamentally qualitative. He emphasized that algorithms and machines should not be relied upon to interpret data for us. Instead, it is the responsibility of experts, who possess contextual knowledge, to derive meaning from the data. This highlights the crucial importance of validation in the research process.

In discussing the third theme in “Text Mining and Big Data,” Dr. Pernia highlighted the importance of exploration within the research process. He noted that while traditional empirical social science often begins with assumptions or theories followed by data collection to test results, the current approach necessitates innovation through direct exploration of the data. This iterative process contrasts with the deductive approach, prompting researchers to rethink their methodologies. Rather than starting from a predefined theory, researchers should consider the data itself as a foundational element which allows for a more flexible research iteration.

After collecting and cleaning data, researchers may need to revisit their theoretical frameworks to find suitable explanations for observed patterns. This model fosters innovation as it allows for the development of hypotheses based on actual data rather than preconceived notions.

[6] Justin Grimmer, “We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together,” *PS: Political Science & Politics* 48:1 (January, 2015): 80-83. DOI: https://doi.org/10.1017/S1049096514001784.

[7] Justin Grimmer and Brandon M. Stewart, “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts,” *Political Analysis* 21:3 (2013): 267–297.

Dr. Pernia illustrated this concept using Hopkins and King's "Haystack metaphor,"[8] highlighting that while text mining enhances our ability to read and interpret trends, it cannot replace the human element of reading and contextualizing documents. But then, humans have tiny active memories. That is why ultimately, Dr. Pernia reiterated that research must remain a computer-assisted endeavor rather than a fully automated one.

Before discussing in detail the different steps researchers usually go through in text mining, Dr. Pernia first gave a brief overview of this process. As he has outlined, text mining begins with the collection of unstructured documents, which can include various formats such as text files, PDFs, and web pages. These documents are then transformed into tokens using the "bag-of-words" approach. Following this, a term-document matrix is created, which organizes these tokens into matrices that reflect the frequency of each term's occurrence within the documents. Researchers can then extract models tailored to their specific research questions. Dr. Pernia explained that essentially, all text mining approaches follow these systematic flowcharts.

Although there are several approaches, Dr. Pernia focused on the bag of words approach to illustrate the preprocessing methods essential for extracting useful information from unstructured data. This method begins with removing capitalization and punctuation from the collected data while discarding the order of words. The bag of words model operates under the assumption that the sequence of terms is less important than their individual locations and associations within the text. Additionally, common placeholder terms, known as stop words (e.g., "the," "it"), are typically removed, as they do not contribute meaningful content to the analysis.

In this approach, creating equivalence among terms is crucial. For example, variations of the word "run" such as "running" should be reduced to their base form, "run." Another process that needs to be performed, known as *lemmatization*, involves truncating words into their dictionary form such as changing "better" to "good." Special characters also need to be removed for the text to be read by computers.

[8] Daniel J. Hopkins and Gary King, "A Method of Automated Nonparametric Content Analysis for Social Science," *American Journal of Political Science* 54:1 (January, 2010): 229-247.

After these preprocessing steps, the output is a count vector or matrix, which summarizes the frequency of each token within the document.

Once the text has been preprocessed, the documents are *tokenized*. This means that they are split into individual words or terms. This process is generally straightforward in languages like English but may require specialized tools for languages such as Chinese or Japanese. It is also important to identify and combine terms that should remain together, like "White House."

To illustrate the text mining process presented so far, Dr. Pernia used the United States' State of the Union as an example. For instance, the documents being studied include the following paragraphs:

> Two example sentences include the word "manufacturing" from the SOTU corpus.
>
> Doc 1: It is undoubtedly in the power of Congress seriously to affect the agricultural and manufacturing interests of France by the passage of laws related to her trade with the United States. (President Jackson, 1831)
>
> Doc 2: And this Congress should make sure that no foreign company has an advantage over American manufacturing when it comes to accessing financing or new markets like Russia. (President Obama, 2012)

First, the term United States might need an underline for them to be combined into one token. Words should also be transformed into lower case and punctuation and stop words should be removed. After these steps, the text now becomes:

> undoubtedly power congress seriously affect agricultural manufacturing interests france passage laws relating trade united_states

And then the document term matrix produced is as follows:

**Table 1. Document-feature matrix W from two sentences containing the word manufacturing from the SOTU corpus**

| | UNDOUBT | POWER | CONGRESS | SERIOUS | AFFECT | AGRICULTUR | MANUFACTUR |
|---|---|---|---|---|---|---|---|
| Doc 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Doc 2 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |

This table shows the set of documents on the rows. If the documents being studied include speeches, the rows will include the speeches. On the columns then are the terms. Therefore, in the first document, the terms “undoubt” and “power” exist. But in document two, the terms “congress” and “manufacture” exist. Now, the document term matrix makes it easy to conduct modelling techniques.

Dr. Pernia emphasized that word frequency is an important heuristic in text analysis, as it reveals a central tendency within the text. Recurring terms such as “war,” “civil,” “died,” and “lives” can signal underlying themes, particularly those related to conflict. However, he stressed the necessity of validation through careful document review to avoid misleading interpretations. Additionally, human supervision is essential to enhance analysis, allowing the identification of subtle relationships between words and outcomes. While studies show that machines can excel at predicting patterns and tones, the integration of human judgment is crucial for a more nuanced understanding of the text.

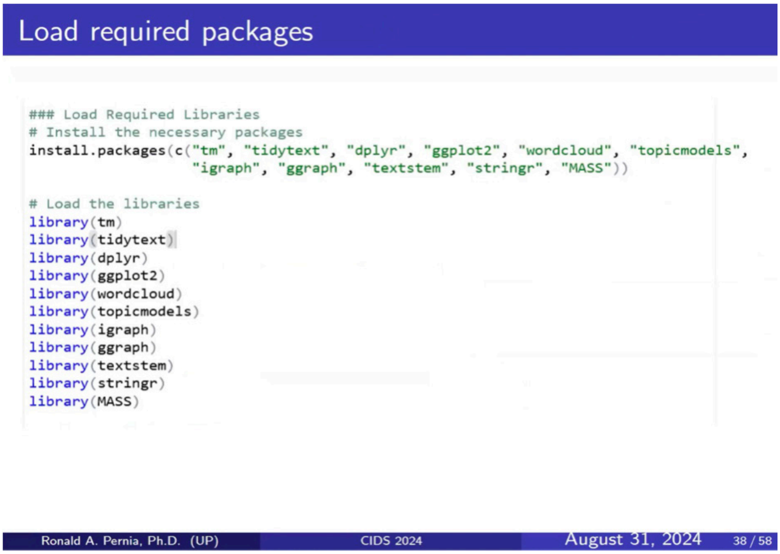
Moreover, he highlighted that while a word's repeated use can indicate significance, it must also possess meaningfulness. For example, a term may occur frequently in a document, but its prevalence alone does not yield valuable information such as when you have a document about Duterte or Marcos. Naturally, the words Duterte or Marcos will appear frequently because it is about them. To tackle this challenge, the Term Frequency-Inverse Document Frequency (TF-IDF) weighting method is utilized. This algorithm proportionately assesses a term's occurrence against the overall corpus size, ensuring that only truly relevant terms contribute to the analysis while filtering out less meaningful words. In some cases, these terms should just be listed as stop words because they do not provide meaningful information at all.

Dr. Pernia then proceeded to the main focus of his presentation, which is the Congressional Policy and Budget Research Department (CPBRD). He explained that after extracting information from the CPBRD website, he created a factor variable to categorize documents into policy notes, discussion papers, and other types, allowing him to analyze the differences in terminology.

He explained that it is essential to understand that R is an open-source software requiring coding, in contrast to a user-friendly interface like SPSS with drop-down menus. He encouraged participants to download both the R program and R Studio, the integrated development environment (IDE) necessary for running R scripts. He also mentioned the cloud version, Posit Cloud, where replication codes are stored.

During his presentation, Dr. Pernia demonstrated how to run the actual text, which involves loading required packages as shown below. It is important to note, however, that he has already extracted the data from CPBRD, transformed it into an excel file, and then loaded in R cloud prior to the presentation.

Dr. Pernia explained that the code below was from the text mining (TM) package (Fig. 1), which has a content transformer function. The researcher needs to "call in" this script.



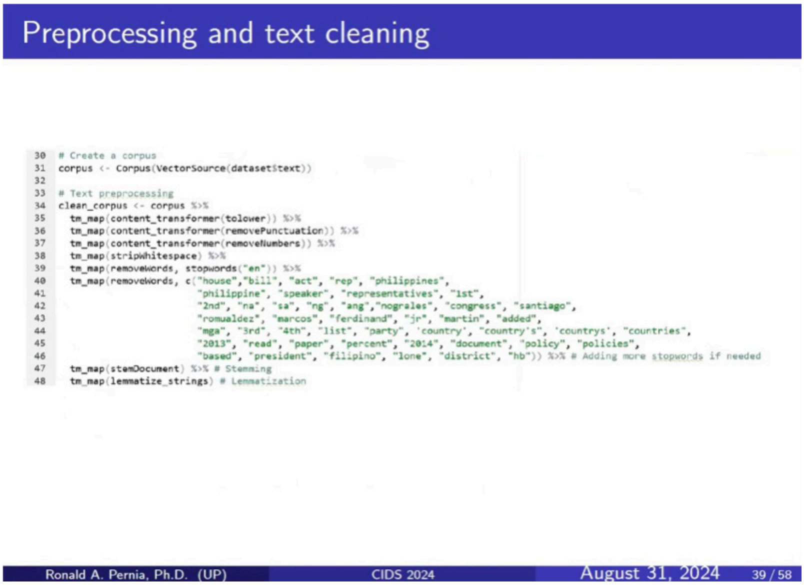
Load required packages

```
### Load Required Libraries
# Install the necessary packages
install.packages(c("tm", "tidytext", "dplyr", "ggplot2", "wordcloud", "topicmodels",
                   "igraph", "ggraph", "textstem", "stringr", "MASS"))

# Load the libraries
library(tm)
library(tidytext)
library(dplyr)
library(ggplot2)
library(wordcloud)
library(topicmodels)
library(igraph)
library(ggraph)
library(textstem)
library(stringr)
library(MASS)
```

Ronald A. Pernia, Ph.D. (UP) CIDS 2024 August 31, 2024 38 / 58

■ **Figure 1.** Load necessary libraries

The next step is to clean the data, such as removing the punctuation, the numbers, the white space, and the stop words. For example, the words that need to be removed are listed below in Fig. 2:



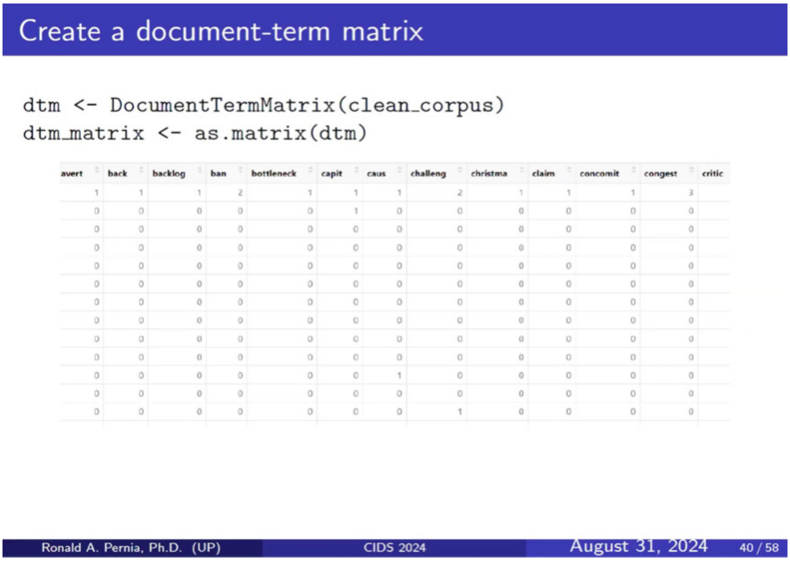
## Preprocessing and text cleaning

```
# Create a corpus
corpus <- Corpus(VectorSource(dataset$text))

# Text preprocessing
clean_corpus <- corpus %>%
  tm_map(content_transformer(tolower)) %>%
  tm_map(content_transformer(removePunctuation)) %>%
  tm_map(content_transformer(removeNumbers)) %>%
  tm_map(stripWhitespace) %>%
  tm_map(removeWords, stopwords("en")) %>%
  tm_map(removeWords, c("house","bill", "act", "rep", "philippines",
                        "philippine", "speaker", "representatives", "1st",
                        "2nd", "na", "sa", "ng", "ang","nograles", "congress", "santiago",
                        "romualdez", "marcos", "ferdinand", "jr", "martin", "added",
                        "mga", "3rd", "4th", "list", "party", 'country', "country's", 'countrys', "countries",
                        "2013", "read", "paper", "percent", "2014", "document", "policy", "policies",
                        "based", "president", "filipino", "lone", "district", "hb")) %>% # Adding more stopwords if needed
  tm_map(stemDocument) %>% # Stemming
  tm_map(lemmatize_strings) # Lemmatization
```

Ronald A. Pernia, Ph.D. (UP) CIDS 2024 August 31, 2024 39 / 58

■ **Figure 2.** Cleaning the unstructured data

After cleaning, the document-term matrix code needs to be executed. This will show the set of documents on the rows and the terms on the columns. For instance, the terms 'ban' and 'challenge' occurred twice as shown in the matrix below:



## Create a document-term matrix

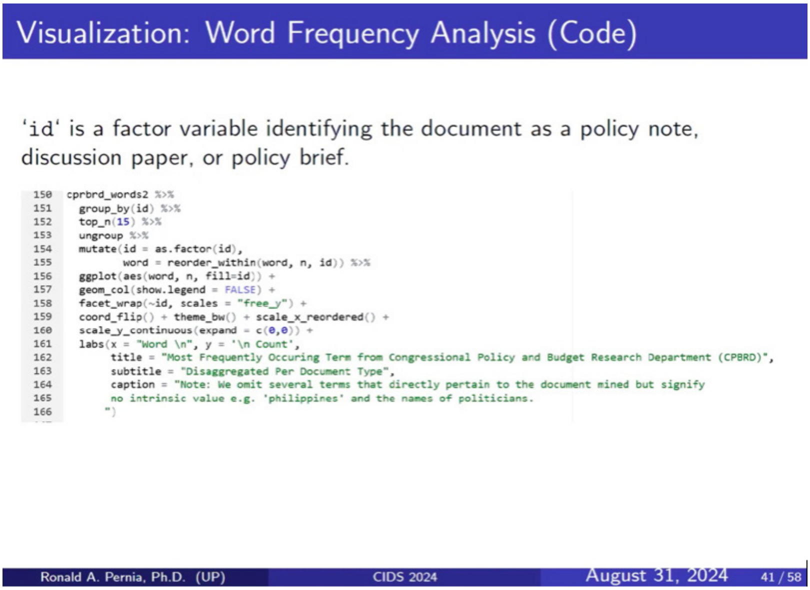
```
dtm <- DocumentTermMatrix(clean_corpus)
dtm_matrix <- as.matrix(dtm)
```

| avert | back | backlog | ban | bottleneck | capit | caus | challeng | christma | claim | concomit | congest | critic |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 3 | |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | |

Ronald A. Pernia, Ph.D. (UP) CIDS 2024 August 31, 2024 40 / 58

■ **Figure 3.** A matrix of words by documents

This particular sample code below will then map out the most occurring terms in the document, disaggregated by the kind of document that you have.



Visualization: Word Frequency Analysis (Code)

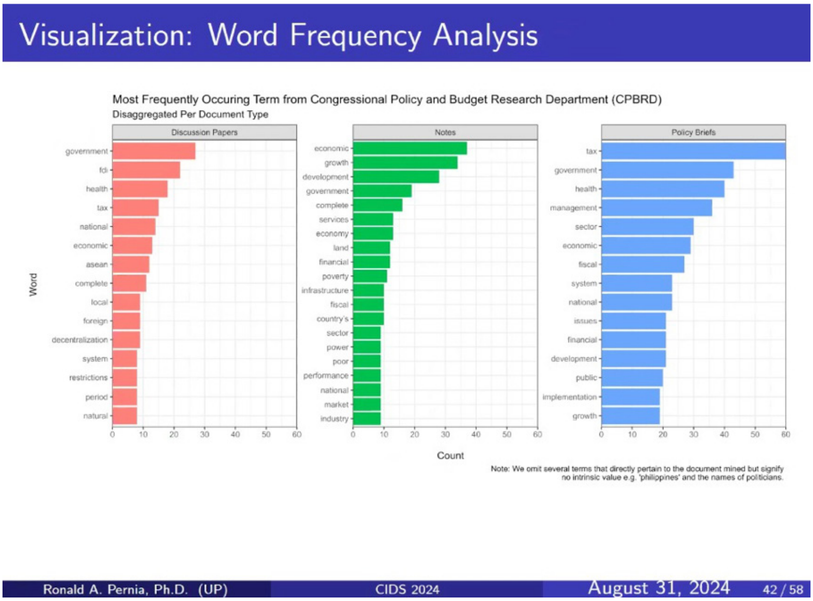
'id' is a factor variable identifying the document as a policy note, discussion paper, or policy brief.

```
cprbrd_words2 %>%
  group_by(id) %>%
  top_n(15) %>%
  ungroup %>%
  mutate(id = as.factor(id),
         word = reorder_within(word, n, id)) %>%
  ggplot(aes(word, n, fill=id)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~id, scales = "free_y") +
  coord_flip() + theme_bw() + scale_x_reordered() +
  scale_y_continuous(expand = c(0,0)) +
  labs(x = "Word \n", y = '\n Count',
       title = "Most Frequently Occuring Term from Congressional Policy and Budget Research Department (CPBRD)",
       subtitle = "Disaggregated Per Document Type",
       caption = "Note: We omit several terms that directly pertain to the document mined but signify
       no intrinsic value e.g. 'philippines' and the names of politicians.
      ")
```

Ronald A. Pernia, Ph.D. (UP) CIDS 2024 August 31, 2024 41 / 58

■ **Figure 4.** R Codes

This code will then produce the following. In this example, the terms “government,” “FDI,” “health,” “tax,” “national,” “economic,” were frequently mentioned in the discussion papers. In the policy notes, the terms “economic,” “growth,” and “development” have the highest frequency. Meanwhile, the terms “tax,” “government,” “health,” “management,” and “sector” received the highest occurrence in policy briefs. From this, we can understand that the main focus remains on the economy and the role of the government.

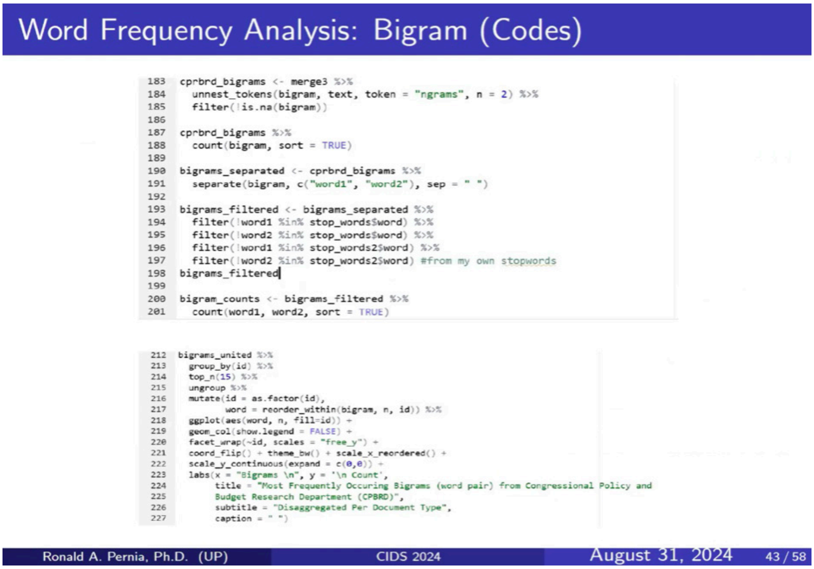


Visualization: Word Frequency Analysis

Ronald A. Pernia, Ph.D. (UP) CIDS 2024 August 31, 2024 42 / 58

■ **Figure 5.** Word frequency analysis visualization

Sometimes, there is a need to combine or pull together two words, such as "economic" and "growth." Therefore, instead of one token, they need to be transformed into two tokens. In such cases, researchers need to do *bigrams*.



Word Frequency Analysis: Bigram (Codes)

```r
cprbrd_bigrams <- merge3 %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2) %>%
  filter(!is.na(bigram))

cprbrd_bigrams %>%
  count(bigram, sort = TRUE)

bigrams_separated <- cprbrd_bigrams %>%
  separate(bigram, c("word1", "word2"), sep = " ")

bigrams_filtered <- bigrams_separated %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  filter(!word1 %in% stop_words2$word) %>%
  filter(!word2 %in% stop_words2$word) #from my own stopwords
bigrams_filtered

bigram_counts <- bigrams_filtered %>%
  count(word1, word2, sort = TRUE)
```

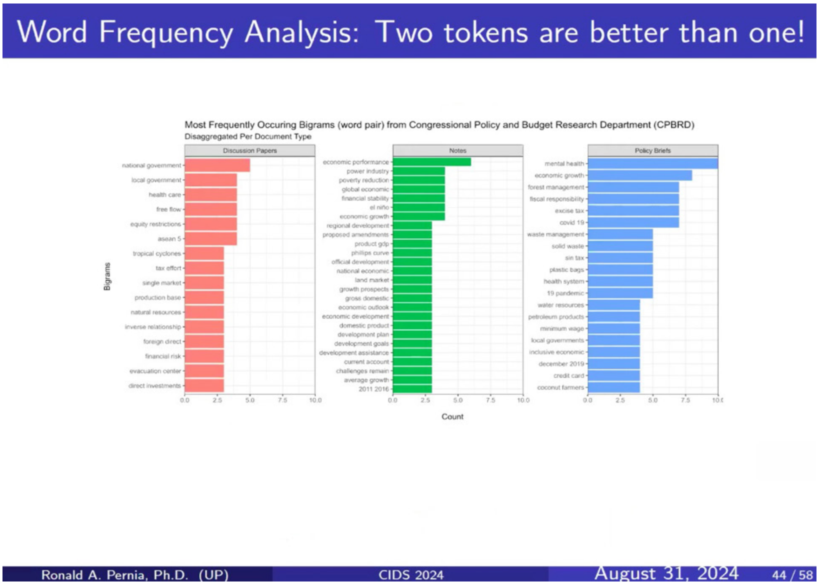
```r
bigrams_united %>%
  group_by(id) %>%
  top_n(15) %>%
  ungroup %>%
  mutate(id = as.factor(id),
         word = reorder_within(bigram, n, id)) %>%
  ggplot(aes(word, n, fill=id)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~id, scales = "free_y") +
  coord_flip() + theme_bw() + scale_x_reordered() +
  scale_y_continuous(expand = c(0,0)) +
  labs(x = "Bigrams \n", y = '\n Count',
       title = "Most Frequently Occuring Bigrams (word pair) from Congressional Policy and
       Budget Research Department (CPBRD)",
       subtitle = "Disaggregated Per Document Type",
       caption = " ")
```

Ronald A. Pernia, Ph.D. (UP) CIDS 2024 August 31, 2024 43 / 58

■ **Figure 6.** Bigrams

This particular code above shows the number of codes, which in this case is two. Essentially, the terms were combined, instead of being split.

Next, Dr. Pernia plotted them according to the different policy documents which produced the following:



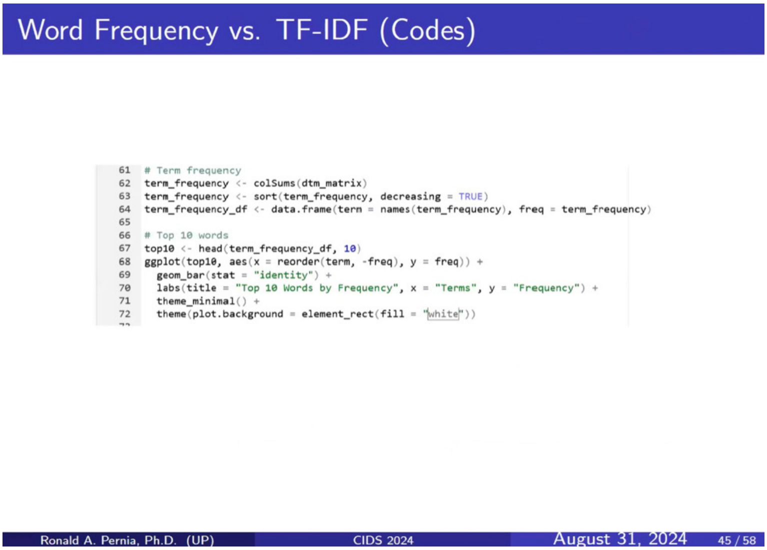
Word Frequency Analysis: Two tokens are better than one!

Ronald A. Pernia, Ph.D. (UP) CIDS 2024 August 31, 2024 44 / 58

■ **Figure 7.** Bigrams make much sense (not displayed in the paper)

What this example shows is that in the discussion papers, terms like "national government," "local government," "healthcare," and "free flow" appeared frequently. For policy notes, phrases such as "economic performance," "power industry," and "power reduction" were common. In policy briefs, terms like "mental health," "economic growth," "forest management," and "fiscal responsibility" were prominent. These demonstrate that combining certain terms can provide more nuanced information. This again underscores the need for researchers to actually read the material and not simply rely on the software. Aside from this, the example demonstrates that word frequency is not enough. The terms also need to be weighed against the whole corpus.

Dr. Pernia then showed how he extracted the top ten words by frequency, noting that he applied the codes (Fig. 8) to the whole document without splitting them by document types.



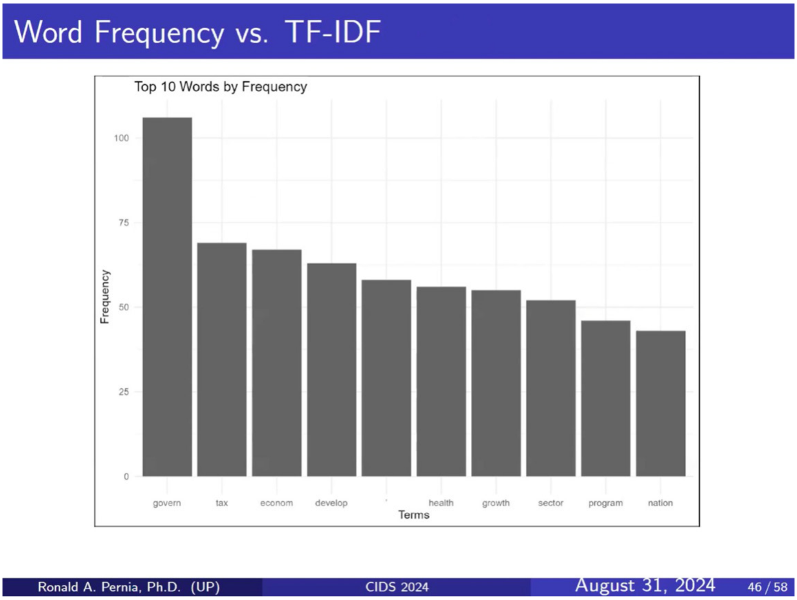
Word Frequency vs. TF-IDF (Codes)

```r
# Term frequency
term_frequency <- colSums(dtm_matrix)
term_frequency <- sort(term_frequency, decreasing = TRUE)
term_frequency_df <- data.frame(term = names(term_frequency), freq = term_frequency)

# Top 10 words
top10 <- head(term_frequency_df, 10)
ggplot(top10, aes(x = reorder(term, -freq), y = freq)) +
  geom_bar(stat = "identity") +
  labs(title = "Top 10 Words by Frequency", x = "Terms", y = "Frequency") +
  theme_minimal() +
  theme(plot.background = element_rect(fill = "white"))
```

Ronald A. Pernia, Ph.D. (UP) CIDS 2024 August 31, 2024 45 / 58

- **Figure 8.** R Codes

The result is shown in Fig. 9 below. Here, the term "govern" was the most frequently mentioned, followed by "tax" and "economy" albeit with a substantial gap.

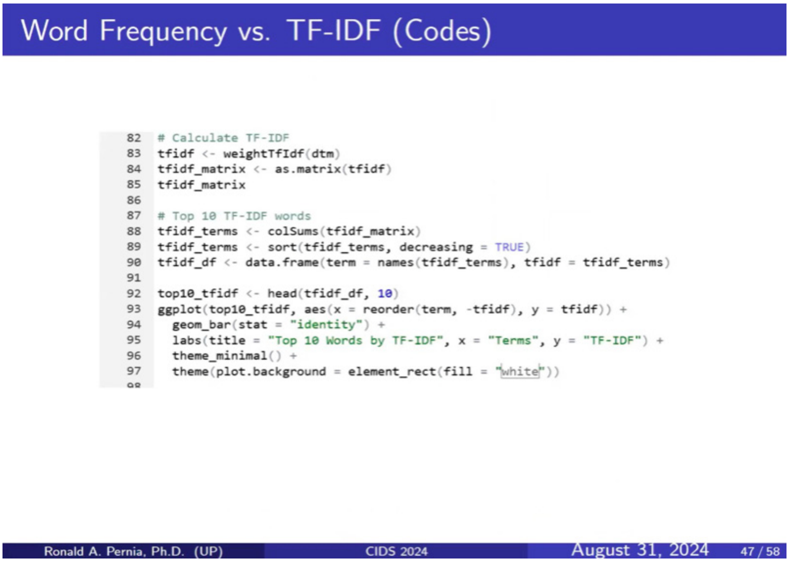


Word Frequency vs. TF-IDF

Ronald A. Pernia, Ph.D. (UP) CIDS 2024 August 31, 2024 46 / 58

- **Figure 9.** For the whole document (not disaggregated)

On the other hand, when weighing the terms vis-a-vis the entire document using the codes in Fig. 10, the terms “tax,” “growth,” and “forest” were the most frequently mentioned.

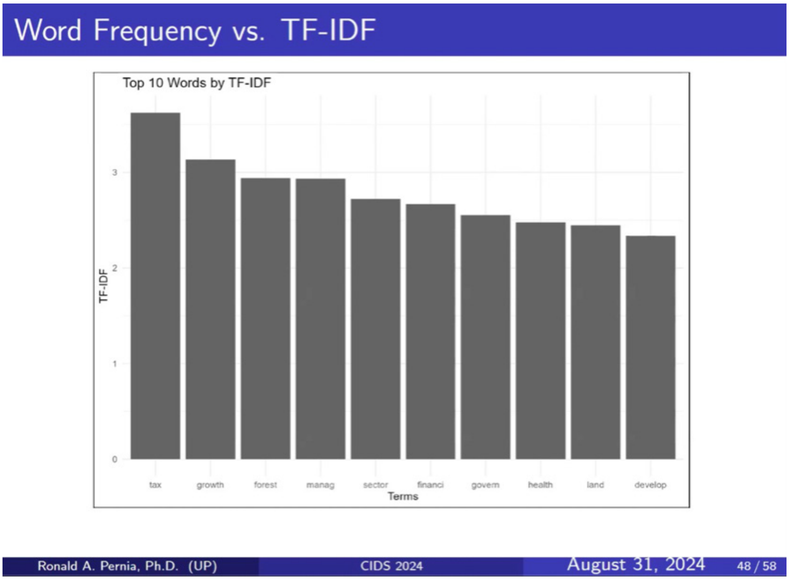


## Word Frequency vs. TF-IDF (Codes)

```r
# Calculate TF-IDF
tfidf <- weightTfIdf(dtm)
tfidf_matrix <- as.matrix(tfidf)
tfidf_matrix

# Top 10 TF-IDF words
tfidf_terms <- colSums(tfidf_matrix)
tfidf_terms <- sort(tfidf_terms, decreasing = TRUE)
tfidf_df <- data.frame(term = names(tfidf_terms), tfidf = tfidf_terms)

top10_tfidf <- head(tfidf_df, 10)
ggplot(top10_tfidf, aes(x = reorder(term, -tfidf), y = tfidf)) +
  geom_bar(stat = "identity") +
  labs(title = "Top 10 Words by TF-IDF", x = "Terms", y = "TF-IDF") +
  theme_minimal() +
  theme(plot.background = element_rect(fill = "white"))
```

Ronald A. Pernia, Ph.D. (UP) CIDS 2024 August 31, 2024 47 / 58

- **Figure 10.** R Codes



## Word Frequency vs. TF-IDF

Ronald A. Pernia, Ph.D. (UP) CIDS 2024 August 31, 2024 48 / 58

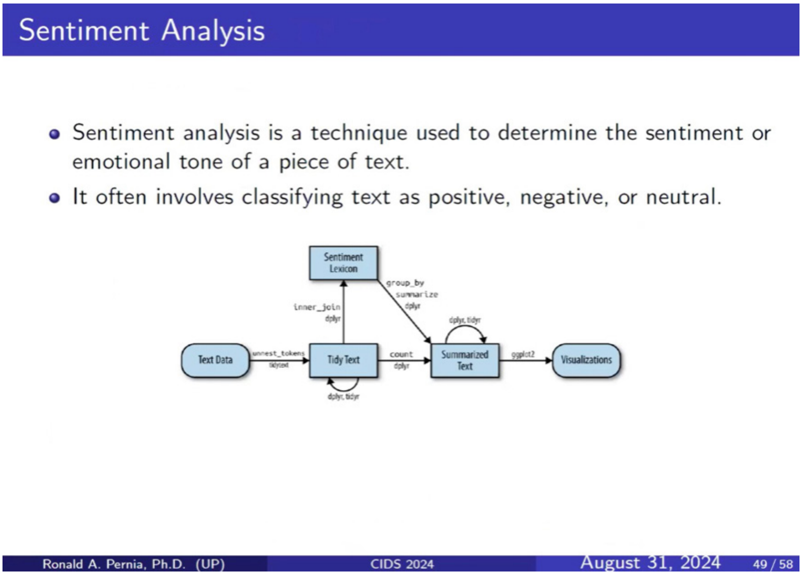
- **Figure 11.** For the whole document (not disaggregated)

These differences in results, according to Dr. Pernia, raise the need to ask certain questions such as “Why is this the case and what can we do?” and “What does this say as far as how we process documents?”

After this, Dr. Pernia discussed two more text mining techniques that are central to analyzing unstructured data: sentiment analysis and topic modelling.

Quoting Dr. Rogelio Panao, Dr. Pernia said that “it is very important to get the opinion of the public regarding certain policy interventions of the government. And politicians right now are employing a lot of data scientists to get to the bottom of the sentiment of the public.” Similarly, it is important for academicians to know the sentiment of the public regarding the documents that they produce.

Using the flow chart from Silge and Robinson,[9] Dr. Pernia explained that sentiment analysis classifies text according to whether it is positively worded, negatively worded, or neutral.



Sentiment Analysis

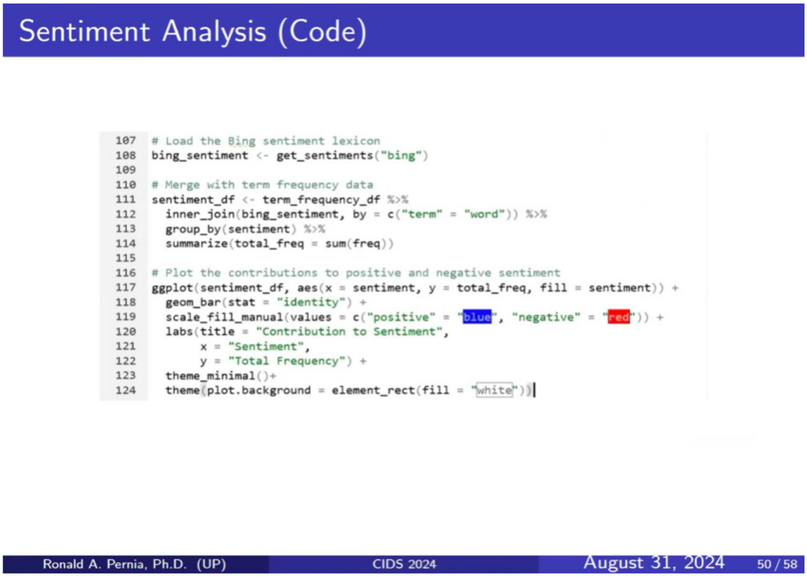
- Sentiment analysis is a technique used to determine the sentiment or emotional tone of a piece of text.
- It often involves classifying text as positive, negative, or neutral.

Ronald A. Pernia, Ph.D. (UP) CIDS 2024 August 31, 2024 49 / 58

■ **Figure 12.** Sentiment analysis flowchart from Silge and Robinson (2017)

[9] Julia Silge and David Robinson, *Text Mining with R: A Tidy Approach* (O'Reilly Media, 2017).

He used the following code, which is from the tidy text package that he ran earlier, to conduct sentiment analysis. He explained that this tidy text package includes a “get_sentiments” function, with Bing as the most common of three available sentiment analysis methods. This code provides scores for the terms, such as if it is a positive term, it will give it a positive score.



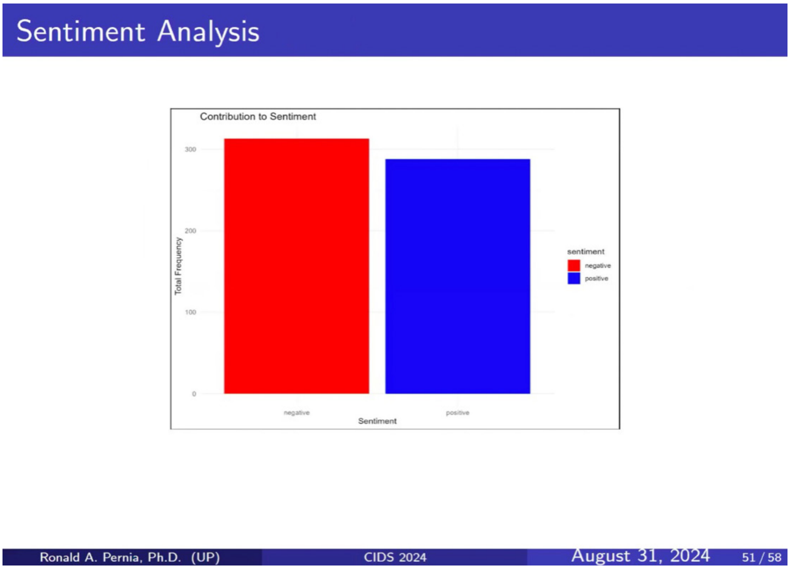
Sentiment Analysis (Code)

```r
# Load the Bing sentiment lexicon
bing_sentiment <- get_sentiments("bing")

# Merge with term frequency data
sentiment_df <- term_frequency_df %>%
  inner_join(bing_sentiment, by = c("term" = "word")) %>%
  group_by(sentiment) %>%
  summarize(total_freq = sum(freq))

# Plot the contributions to positive and negative sentiment
ggplot(sentiment_df, aes(x = sentiment, y = total_freq, fill = sentiment)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = c("positive" = "blue", "negative" = "red")) +
  labs(title = "Contribution to Sentiment",
       x = "Sentiment",
       y = "Total Frequency") +
  theme_minimal()+
  theme(plot.background = element_rect(fill = "white"))
```

Ronald A. Pernia, Ph.D. (UP) CIDS 2024 August 31, 2024 50 / 58

■ **Figure 13.** R Codes for the whole corpus

After running the code above, the result is as follows. This indicates that, the terms are more on the negative side.

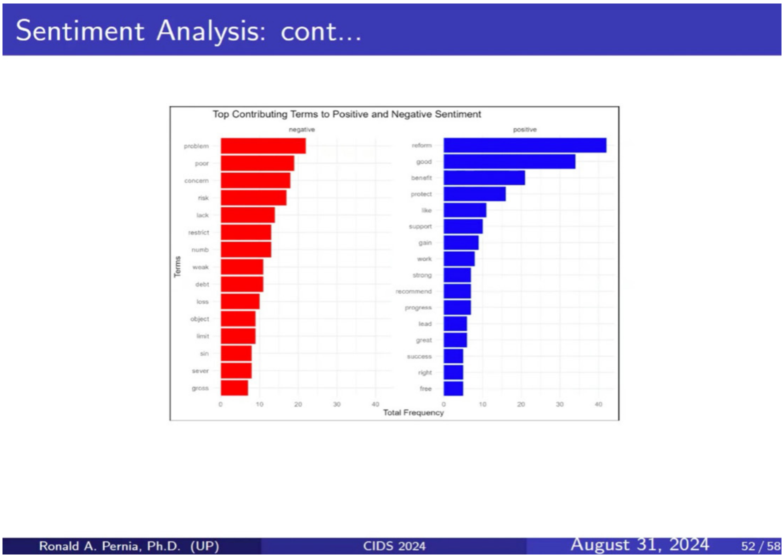


### Sentiment Analysis

Ronald A. Pernia, Ph.D. (UP) CIDS 2024 August 31, 2024 51 / 58

- **Figure 14.** Sentiment as a whole corpus

However, if the particular words were analyzed, it shows that the terms "reform," "good," "benefit," "protect," "like," and "support" contribute to the positive sentiment, while "problem," "poor," "concern," "risk," "lack," "restrict," "numb," and "weak" contribute to the negative sentiment.

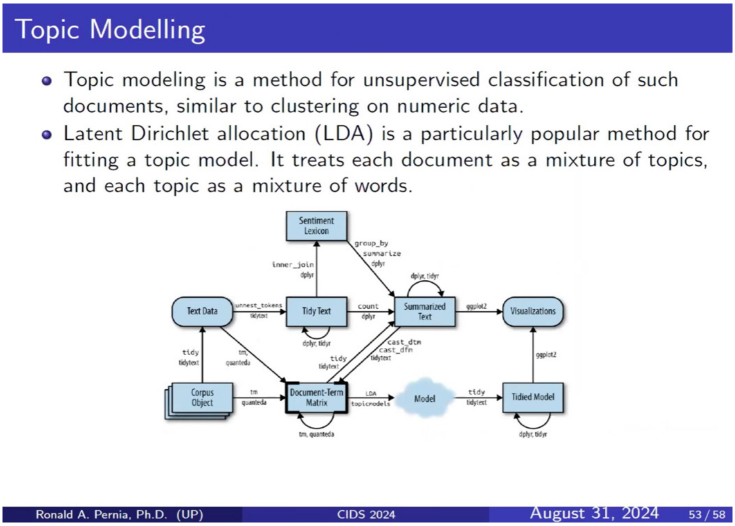


### Sentiment Analysis: cont...

Ronald A. Pernia, Ph.D. (UP) CIDS 2024 August 31, 2024 52 / 58

- **Figure 15.** Top 15 words that contribute highly to positive and negative terms

Lastly, Dr. Pernia explained topic modelling which groups documents into clusters, with the well-known method Latent Dirichlet Allocation (LDA) treating each document as a mix of topics, and each topic as a mix of words.



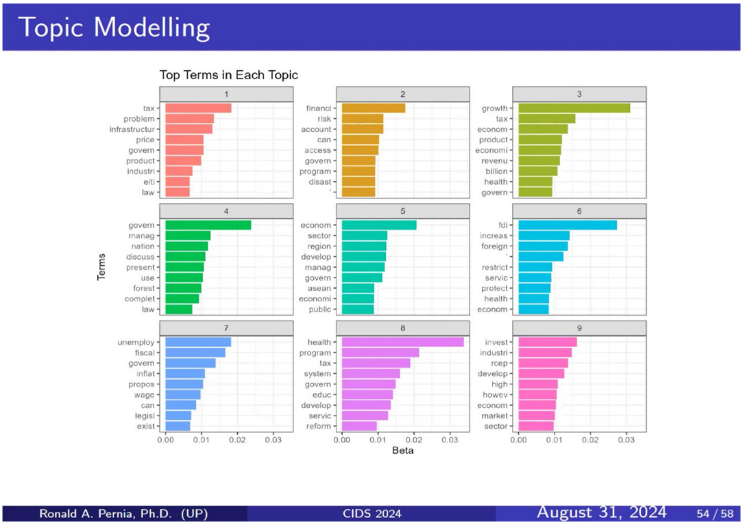
## Topic Modelling

- Topic modeling is a method for unsupervised classification of such documents, similar to clustering on numeric data.
- Latent Dirichlet allocation (LDA) is a particularly popular method for fitting a topic model. It treats each document as a mixture of topics, and each topic as a mixture of words.

Ronald A. Pernia, Ph.D. (UP) CIDS 2024 August 31, 2024 53 / 58

■ **Figure 16.** Top modelling flowchart from Silge and Robinson (2017)

Using the topic modelling technique with the CPBRD documents, Dr. Pernia found that the words "tax," "problem," "infrastructure," "price," "govern," and "product" are related to each other in Topic 1.



## Topic Modelling

Ronald A. Pernia, Ph.D. (UP) CIDS 2024 August 31, 2024 54 / 58

■ **Figure 17.** Top 9 Topics: This is where expertise matters!

What topic modelling does, therefore, is it allows researchers to identify and classify themes or topics within the text, though he noted the necessity of human interpretation to contextualize the findings. Again, Dr. Pernia stressed the importance of qualitative analysis in defining these topics to have a deeper understanding of the documents' content.

Dr. Pernia concluded his presentation by encouraging participants to embrace new methodologies in social science research, emphasizing the value of learning programming languages like R and Python to keep up with the advancements in data analysis.

# Reaction to Dr. Ronald Pernia's Presentation on Text Mining

Prof. Dominador Gamboa[10]

Prof. Dominador Gamboa began his presentation by posing the following question: “Is truth stranger than fiction?” to introduce an intriguing parallel to Philip K. Dick's[11] 1956 science fiction thriller, “The Minority Report.” He described how, in the story set in 2054, advanced technology enables the “pre-crime police” to foresee crimes and arrest suspects before any offense is committed. Drawing from this concept, Prof. Gamboa underscored the significance of predictive analysis in real-world contexts, especially within the data and text mining field.

Prof. Gamboa referenced *The Wisdom of Crowds*[12] by James Surowiecki to emphasize how text mining taps into collective feedback. He suggested that large groups often hold insights surpassing those of a few experts. He discussed how text mining research can synthesize varied perspectives to make meaningful sense of complex issues, which he highlighted as a critical skill for leadership especially in “a world scarred by the pandemic.”

[10] Prof. Dominador Gamboa is the Service Director of the Congressional Policy and Budget Research (CPBRD) Office at the House of Representatives, Republic of the Philippines.

[11] Philip K. Dick, “The Minority Report,” *Fantastic Universe* 8:4 (January, 1956): 4–29.

[12] James Surowiecki, *The Wisdom of Crowds* (Anchor Books, 2005).

Prof. Gamboa commended Dr. Pernia's innovative application of text mining and predictive analytics within public service, which he sees as a promising shift towards anticipatory governance.[13] By fostering a governance model that adapts to change and proactively addresses emerging issues, he noted that text mining enables a more responsive approach to stakeholder needs, transforming governance from being reactive to proactive. This emphasis on adaptation was further supported by identifying three core elements for future research: data, digital identity, and artificial intelligence.

In his commentary, Prof. Gamboa remarked on the rising value of data in contemporary society. He referenced an article from The Economist,[14] emphasizing that data has surpassed oil as the world's most valuable resource. This transformation highlights the importance of effectively processing unstructured data to extract meaningful information, which can subsequently inform knowledge and wisdom in policymaking.

Illustrating this transformation, he used the concept of the wisdom pyramid. The pyramid outlines a progression, beginning with raw data, which evolves into information, then knowledge, and finally wisdom. He highlighted the exponential growth of the global data sphere, projected to reach 181 zettabytes by 2025—a scale that reflects the need for careful data management.

He underscored that this management requires adhering to the principle of "Garbage In, Garbage Out" (GiGo), which emphasizes that high-quality data inputs are essential for credible outputs. To further reinforce this, he pointed to the importance of data provenance. He stressed that the source and transformation of data must be scrutinized to ensure reliability. He suggested that policy elites, stakeholders, academia, and the media could serve as valuable sources for such data, ultimately supporting the integrity of the analysis.

Prof. Gamboa further emphasized the importance of considering various factors like opinions, impressions, sentiments, interests, and perceptions,

[13] Jose Ramos, Ida Uusikyla, and Nguyen Tuan Luong, Anticipatory Governance — A Primer (UNDP, 2020), https://www.undp.org/vietnam/blog/anticipatory-governance-primer.

[14] The Economist, "The World's Most Valuable Resource: Data and the New Rules of Competition." *The Economist*, 6 May 2017.

collectively referred to as OISIP. He noted that these elements are crucial when mining texts through available software technologies. By leveraging these tools, it is possible to enhance current experiences and prepare for future actions that can inform effective decision-making.

Prof. Gamboa then expanded on the critical role of digital identity in today's data-driven world. He referenced historian Yuval Noah Harari's[15] assertion that "those who own the data, own the future." He posed thought-provoking questions about personal data ownership and highlighted how digital identities—comprising browsing habits, online purchases, and social media activity—form what he called "data-exhaust," a powerful economic asset. In the age of "surveillance capitalism,"[16] tech companies capitalize on this data, turning it into predictive products for targeted advertising.

Prof. Gamboa questioned whether individuals should receive compensation for the data they provide freely. He suggested that inclusive economic growth could arise if people were paid for their data contributions. Building on Dr. Pernia's work, he noted that advancements from descriptive to prescriptive analytics enable us to harness data for proactive governance.

Shifting focus to artificial intelligence, Prof. Gamboa discussed the rapid advancements in artificial intelligence (AI), which mimic human cognitive abilities. Referencing AI expert Ray Kurzweil,[17] he suggested that by 2300, humans might achieve "singularity," where intelligence merges with AI at unprecedented scales. This potential merging of human cognition with cloud technology, he argued, presents both opportunities and risks.

In conclusion, Prof. Gamboa reiterated the importance of Dr. Pernia's work, which provides the tools necessary for predictive and prescriptive analytics

[15] Yuval N. Harari, *21 Lessons for the 21st Century* (Random House, 2018).

[16] A term which Harvard Professor Shoshana Zuboff defines as "the unilateral claiming of private human experience as free raw material for translation into behavioral data." Shoshana Zuboff, "In new book, Business School professor emerita says surveillance capitalism undermines autonomy — and democracy," The Harvard Gazette, 4 March 2019. https://news.harvard.edu/gazette/story/2019/03/harvard-professor-says-surveillance-capitalism-is-undermining-democracy/

[17] Ray Kurzweil, *The Singularity is Nearer: When We Merge with AI* (Viking, 2024).

in policymaking. He encouraged participants to embrace the emerging methodologies and technologies facilitating informed decision-making, ultimately reinforcing the idea that the future of governance lies in our hands.

# Open Forum

Maria Corazon Reyes[18]
*Moderator*

## Dr. Pernia's Response to Director Gamboa's Presentation

The open forum began with Dr. Pernia's response to Prof. Gamboa's presentation. He highlighted two key points that resonated with him. First, he noted the emerging trend towards evidence-based social science, particularly in the context of digital identity and artificial intelligence. He emphasized the importance of utilizing existing data, such as public opinions gathered from social media platforms and comments, to substantiate claims rather than relying solely on statements from authority figures. Addressing the ethical implications of using government data, he asserted that, since such information is publicly available, its use for academic and educational purposes is justifiable.

Furthermore, Dr. Pernia shared his own experiences as a newcomer to the data science field, discussing the challenges of data extraction due to the complexity of various websites and their security features. He illustrated this with an example of using different Python packages to extract press releases from Congress. He acknowledged the learning curve associated with programming languages like Python and R, but also highlighted the supportive community that is ready to help those who are experiencing challenges with these programs.

He concluded by sharing how his engagement with data science has significantly enhanced his publication output and encouraged students to invest time in learning these methods.

[18] Maria Corazon Reyes is the Senior Project Assistant of the Program on Social and Political Change at UP CIDS.

## Using Text Mining to Examine the Concept of *Pakikipagkapwa* (relating with others) in Filipino Culture

During the open forum , multiple questions and comments were raised. First, Dr. Maria Margarita Lavides, who is also a research fellow of UP CIDS, shared information about an upcoming publication that applies text mining to the concept of “pakikipagkapwa” (relating with others) in Filipino culture. She shared that while previous studies on pakikipagkapwa in Filipino psychology primarily employed quantitative or qualitative approaches, her team's upcoming paper introduces a new model developed through text mining and big data analysis. This model uncovers emerging clusters related to pakikipagkapwa and highlights insights with potential policy implications, particularly concerning the recently proposed anti-discrimination law in the Senate. She refrained, however, from disclosing specifics to avoid preempting the findings. She mentioned that these details would be covered in an upcoming webinar the following month.

In response, Dr. Pernia expressed his appreciation for the shared perspectives, noting his eagerness to read Dr. Lavides's work. He pointed out that while text mining represents a progressive approach, one significant challenge remains: the inadequacy of data storage on government websites. Echoing Prof. Gamboa’s sentiment regarding coordination with various government offices, Dr. Pernia emphasized the need for these offices to make data publicly available. This accessibility would facilitate data scraping and enhance research efforts. Dr. Pernia highlighted the mutual responsibility of both educators and government entities to adapt and improve governance through shared information.

## Ethical Considerations in Text Mining

Dr. Lavides then shared that even when employing text mining techniques, researchers must not overlook ethical considerations. She lamented the scarcity of ethics committees in the Philippines, though she acknowledged the recent establishment of an ethics committee at UP Diliman. Dr. Lavides underscored the necessity of applying for ethics clearance before embarking on text mining research, regardless of the absence of direct human interaction.

In response, Dr. Pernia acknowledged Dr. Lavides's point, clarifying that while some publications may not require ethics clearance, top-tier journals often demand replication codes. He noted the difficulty of obtaining ethics clearance, particularly in cases involving data from Chinese websites, which further complicates the research process. Dr. Pernia highlighted the need for improvements in the ethics clearance process, especially considering its associated costs and the rapid evolution of research trends.

## Digitization of Filipiniana Corpus

An audience member inquired about how the challenges of digitizing Filipiniana materials still in analog format can be resolved.

Dr. Pernia confirmed this as a key obstacle, mentioning the high costs of software for digitization. He suggested that dialogue between academia and government is crucial. It is important for government entities to understand the benefits of digitization which will improve data management and processing.

## Broader Applications in Data Science

Dr. Lavides added to the discussion by noting that while text mining is a key focus, the broader field of data science encompasses various analytical techniques, including image analysis. She encouraged consideration of alternative methodologies in future investigations.

## Data Integrity and Free Speech

Another member of the audience raised a question considering the inevitable synergy between technology by 2030. The attendee asked how to ensure data integrity and authenticity while maintaining free speech within a democratic society.

Prof. Gamboa responded by emphasizing the complexity of ensuring data integrity, noting that there is no single entity responsible for verifying data authenticity. Trust plays a crucial role in assessing data reliability. Referencing his earlier point on data provenance, he highlighted that understanding both the origin and subsequent transformations of data is essential. Any initial doubts about data origins can be further amplified as the data undergoes modifications.

Prof. Gamboa contrasted policy approaches, stating that the European Union has taken more proactive measures on data regulation, transparency, and accountability compared to the United States, despite the latter's role as the home base for major tech companies. He cited Meta (Facebook) as an example and acknowledged recent policy interventions in the Philippines, while recognizing that further progress is needed.

He raised critical questions about data ownership and stressed the challenges individuals face when they effectively sign away rights upon entering websites. He also encouraged a thorough examination of data ownership and compensation, urging attendees to consider whether they wish to be part of a system that primarily benefits large tech enterprises.

## Establishment of the UP Data Commons

Mr. Dan Dorado discussed the UP Data Commons initiative which is intended to enhance data sharing and integrity within the UP system and encouraged those who attended the webinar to contribute to the repository.

In response, Dr. Panao affirmed that the Intelligent Systems Center (ISC) plans to upload data science materials to the Data Commons once established. He added that they are currently working on another project relating to the local government that hopefully will be available via the Data Commons the following year.

## Clarification on Stop Words and Preprocessing in Taglish/Filipino Text

Another member of the audience, Mr. Jim West Celeste, inquired about the methodology behind the removal of additional words in Dr. Pernia's presentation. He cited as an example the removal of “Ferdinand” and "Philippines.” In particular, he asked how additional words were identified, whether statistically or based on domain knowledge.

Dr. Pernia explained that identifying these words was an iterative process based on word frequency analysis. He noted that while the stop words list originated from a different dataset, it required human intervention to ensure relevant terms were either removed or retained based on contextual significance.

## Preprocessing in Taglish/Filipino Text

A question was raised regarding strategies for preprocessing data from social media comments, particularly when posts are written in multiple languages such as Taglish, English, or Filipino.

Dr. Pernia’s response noted that R offers packages capable of handling stop words for various languages, including Turkish and Chinese, with Tagalog as the default language setting for the Philippines. However, he explained that these tools currently lack built-in support for Bisaya, a limitation his team encountered while analyzing President Duterte’s speeches. Given that many of Duterte's statements contain Bisaya phrases and, at times, offensive language, processing this type of multilingual content presents unique challenges. Dr. Pernia highlighted the importance of researchers’ contextual knowledge in dealing with linguistic complexities.

## Emoticons and GIFs in Text Mining

In the context of the tourism industry, a question was posed on how to utilize GIFs and emoticons in developing policies.

Dr. Pernia acknowledged that he has not yet conducted studies specifically on emoticons but suggested that these visual elements could indeed serve as unique units of analysis beyond traditional text. He noted that, while most text mining packages typically filter out emoticons, tools or packages may exist to handle them separately. He then recommended exploring Stack Overflow, an online forum where programmers share open-source solutions, as a potential resource for identifying appropriate tools or packages that could process emoticons and GIFs.

## Handling Sarcasm in Sentiment Analysis

A question was raised regarding the handling of sarcasm within sentiment analysis.

Dr. Pernia explained that while sentiment analysis holds promise, it also has limitations—especially with sarcasm, which current tools cannot accurately interpret. To address this issue, Dr. Pernia suggested possible strategies such as manually reviewing texts to identify sarcasm, recoding or excluding sarcastic comments, or categorizing them separately from purely negative or positive sentiments. He noted that while he personally has not yet employed such techniques, specific software or packages may exist that are better equipped to detect and handle these sentiment polarities.

## Distinction Between Web Scraping and Text Mining

In response to a question raised about the distinction between text mining and web scraping, Dr. Pernia clarified the sequential and functional differences between the two.

He explained that web scraping essentially focuses on data collection, often when valuable information needs to be systematically gathered. This results in an output—typically in formats like CSV or Excel files—that forms the corpus for further analysis.

This is followed by text mining, where the collected unstructured data undergoes preprocessing to transform it into a readable format by segmenting it into tokens.

## GUI Versions for Text Mining Software

A member of the audience asked if GUI options are available for text mining to support those unfamiliar with programming, especially within the public and non-government organization (NGO) sectors. Dr. Pernia confirmed that GUI versions exist, but noted they are often labor-intensive and generally not recommended. He highlighted the advantages of cloud-based platforms like R Studio's cloud version, which can prompt users through installation issues and enhance ease of use compared to offline options.

Sharing his own experience, Dr. Pernia described how he initially learned by adapting existing scripts—a method he recommended for those new to coding. He emphasized the collaborative nature of data science, where researchers contribute unique skills that others might need.

Discussing the importance of data sharing, Dr. Pernia pointed to the role of open repositories, which allow scholars to store data and facilitate replication. This transparency enables others to verify research, catch errors, and even prompt journal retractions if necessary.

## Extraction of URL IDs for Network Visualization

An audience member inquired about techniques to extract specific URL IDs when using web scraping to study narratives. He noted that URLs could serve as reference points to visualize associations between conversations and user interactions in network visualization. Despite attempts with various algorithms, he was still unable to obtain this URL data for their analysis and sought guidance on achieving this.

In response, Dr. Pernia explained that to gather URLs from a government site, such as the

Philippine Congress website, one could use keywords to filter documents relevant to specific topics, like legislation on China. Each document accessed will generate a URL.

Dr. Pernia also mentioned that Python has a feature that can store URLs as a separate variable. Alternatively, open-source software is available for users less familiar with Python.

# REFERENCES

Dick, Philip K. "The Minority Report." *Fantastic Universe* 8, no. 4 (January 1956): 4–29.

Gill, Jeff. "Political Science Is a Data Science." *The Journal of Politics* 83, no. 1 (January 2021): 1–7. https://doi.org/10.1086/712132.

Grimmer, Justin. "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together." *PS: Political Science & Politics* 48, no. 1 (January 2015): 80–83. https://doi.org/10.1017/S1049096514001784.

Grimmer, Justin, and Brandon M. Stewart. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21, no. 3 (Summer 2013): 267–97. https://doi.org/10.1093/pan/mps028.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton, NJ: Princeton University Press, 2022.

Harari, Yuval N. *21 Lessons for the 21st Century*. New York: Spiegel & Grau, 2018.

Hopkins, Daniel J., and Gary King. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54, no. 1 (January 2010): 229–47. https://doi.org/10.1111/j.1540-5907.2009.00428.x.

Kurzweil, Ray. *The Singularity Is Nearer: When We Merge with AI*. New York: Viking, 2024.

Ramos, Jose, Ida Uusikyla, and Nguyen Tuan Luong. *Anticipatory Governance—A Primer*. United Nations Development Programme, 2020. https://www.undp.org/vietnam/blog/anticipatory-governance-primer.

Silge, Julia, and David Robinson. *Text Mining with R: A Tidy Approach*. Sebastopol, CA: O'Reilly Media, 2017.

Surowiecki, James. *The Wisdom of Crowds*. New York: Anchor Books, 2005.

The Economist. "The World's Most Valuable Resource Is No Longer Oil, but Data." *The Economist*, May 6, 2017. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data.

Walsh, Colleen. "'Surveillance Capitalism' Is Undermining Democracy." *The Harvard Gazette*, March 4, 2019. https://news.harvard.edu/gazette/story/2019/03/harvard-professor-says-surveillance-capitalism-is-undermining-democracy/.

# CENTER FOR INTEGRATIVE AND DEVELOPMENT STUDIES

Established in 1985 by University of the Philippines (UP) President Edgardo J. Angara, the UP Center for Integrative and Development Studies (UP CIDS) is the policy research unit of the University that connects disciplines and scholars across the several units of the UP System. It is mandated to encourage collaborative and rigorous research addressing issues of national significance by supporting scholars and securing funding, enabling them to produce outputs and recommendations for public policy.

The UP CIDS currently has twelve research programs that are clustered under the areas of education and capacity building, development, and social, political, and cultural studies. It publishes policy briefs, monographs, webinar/conference/forum proceedings, and the Philippine Journal for Public Policy, all of which can be downloaded free from the UP CIDS website.

# THE PROGRAM

The **Program on Social and Political Change (PSPC)** provides a platform for understanding the varied social and political challenges facing modern Philippine society and polity from a multidisciplinary perspective. In relation to this, the Program also designs empirical studies using a variety of methods and approaches which form the basis for policy inputs and discussions at the local, national, and international levels.

UNIVERSITY OF THE PHILIPPINES
CENTER FOR INTEGRATIVE AND DEVELOPMENT STUDIES
Lower Ground Floor, Ang Bahay ng Alumni, Magsaysay Avenue
University of the Philippines Diliman, Quezon City 1101

**Telephone** (02) 8981-8500 loc. 4266 to 4268
(02) 8426-0955

**Email** cids@up.edu.ph
cidspublications@up.edu.ph

**Website** cids.up.edu.ph