# Statistical Performance Indicators: Data Use Text Mining

SPI Team

2020-07-13

WORLD BANK GROUP

# Introduction

- Goal is to examine the use of text mining of National Development Plans (NDPs) & National Poverty Reduction Strategies for the SPI Data Use dimension
- In this presentation, will discuss:
  - Source for National Development Plan (FAOLEX)
  - Dictionary based approach for text mining
  - Initial results from text mining NDPs
- Methods and data rely heavily on Paris21 approach to text mining

# Background on Data Use Dimension

- Indicator intended to be Proxy for the intensity of use of official statistics by each respective user segment in country.

Data Use dimension contains 5 indicators:

- Data use by national legislature
- Data use by national executive branch
- Data use by civil society
- Data use by academia
- Data use by international organizations

# Data Use Indicator Methodology

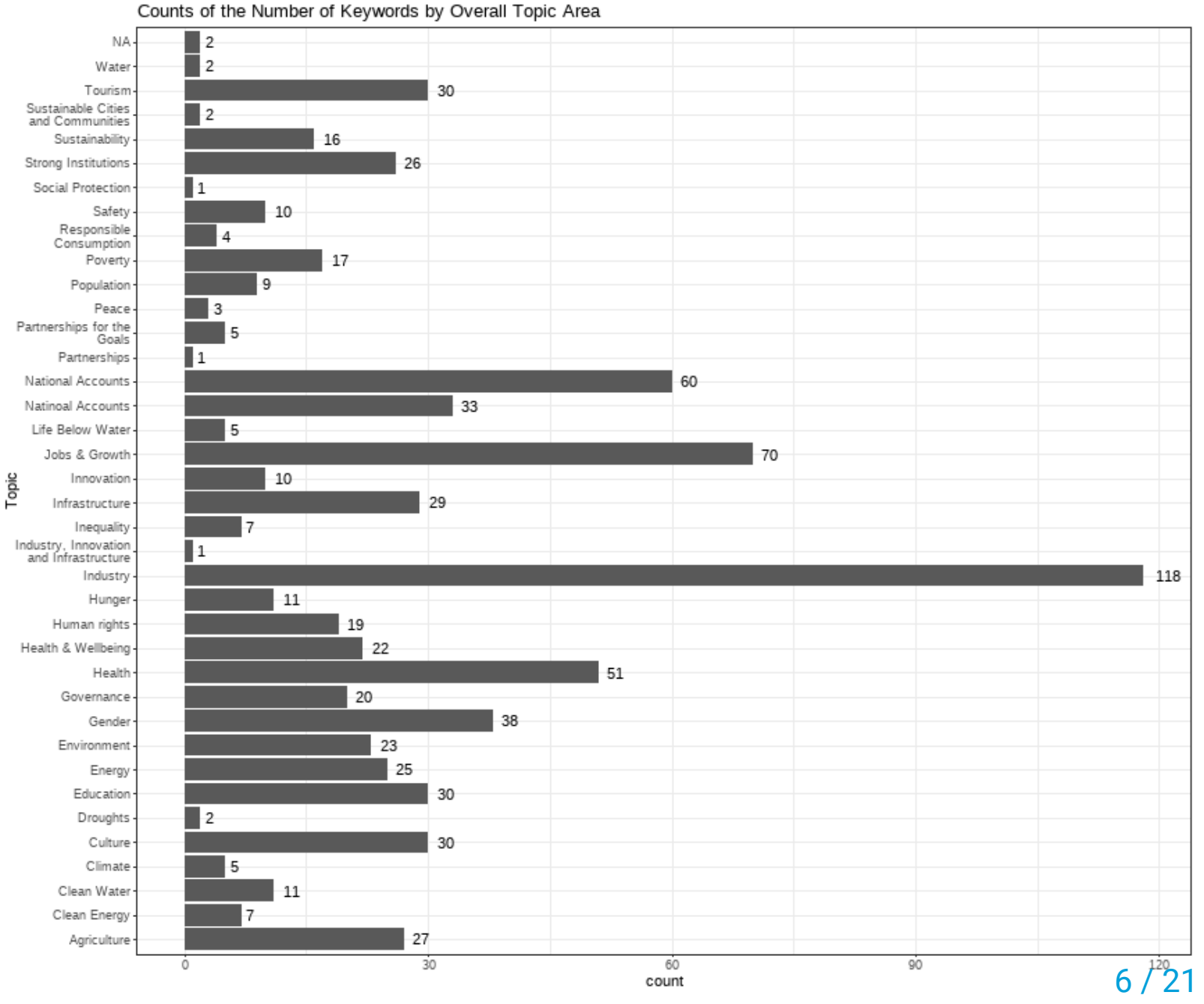
- Indicator for each of 5 areas is formed by:

    1. Searches the site for references to keywords related to each SDG goal
    2. Searches the site for the subset of references that also include the word statistics
    3. Calculates the ratio between the two

- Algorithm captures density of discussion of each topic in each organization assessed

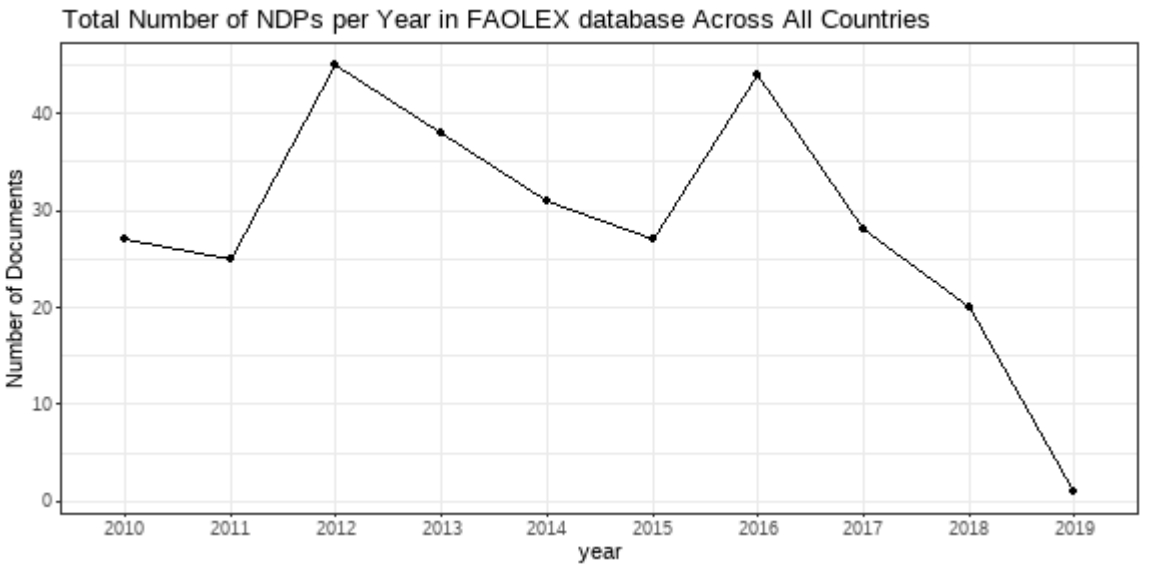
# Dictionary Approach to Text Mining

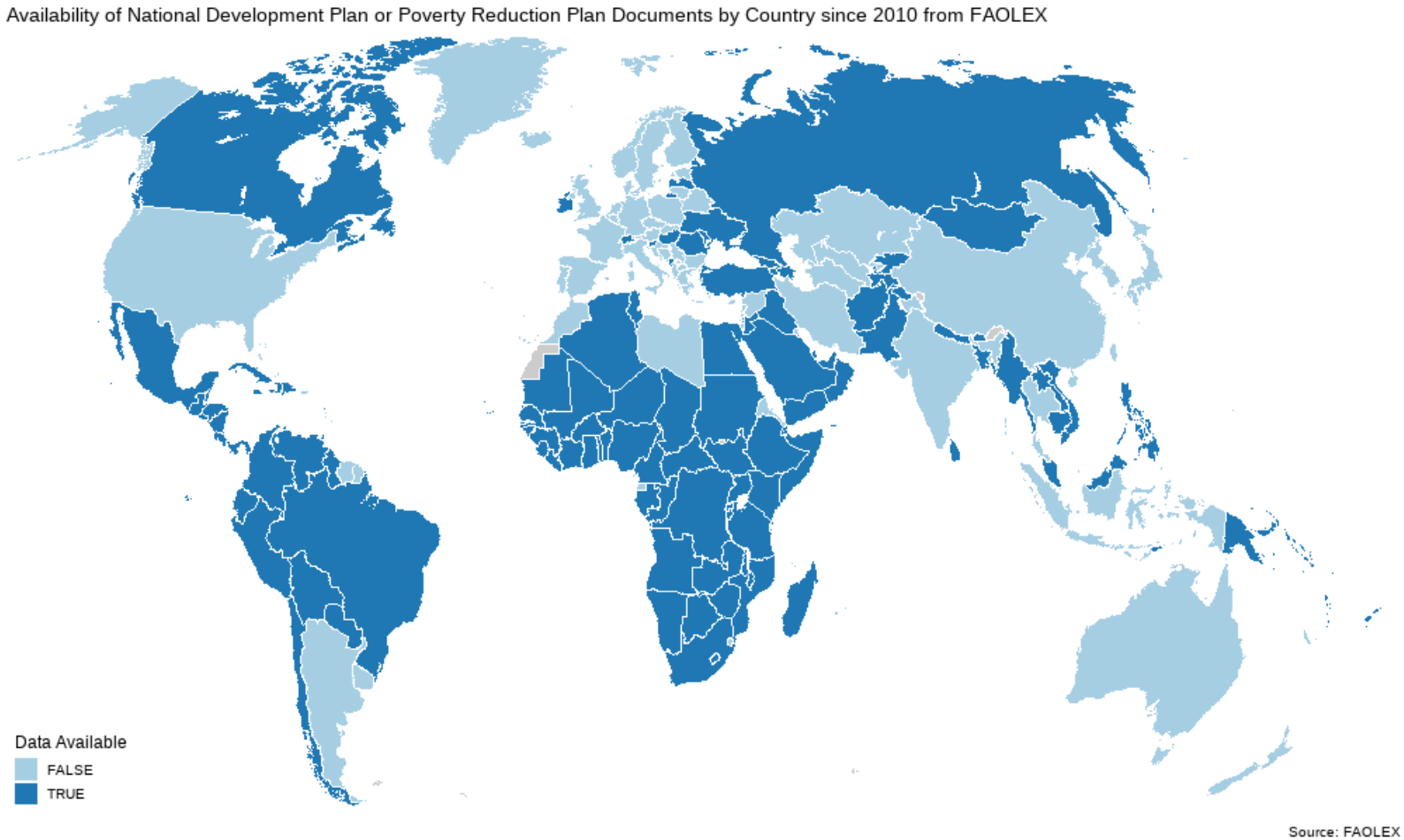
- Paris21 has developed set of text mining tools for analyzing NDPs
- Created dictionary of 782 search terms related to different topic areas
- I categorized keywords into 38 topic areas


Counts of the Number of Keywords by Overall Topic Area
Topic
NA 2
Water 2
Tourism 30
Sustainable Cities and Communities 2
Sustainability 16
Strong Institutions 26
Social Protection 1
Safety 10
Responsible Consumption 4
Poverty 17
Population 9
Peace 3
Partnerships for the Goals 5
Partnerships 1
National Accounts 60
Natinoal Accounts 33
Life Below Water 5
Jobs & Growth 70
Innovation 10
Infrastructure 29
Inequality 7
Industry, Innovation and Infrastructure 1
Industry 118
Hunger 11
Human rights 19
Health & Wellbeing 22
Health 51
Governance 20
Gender 38
Environment 23
Energy 25
Education 30
Droughts 2
Culture 30
Climate 5
Clean Water 11
Clean Energy 7
Agriculture 27
0
30
60
90
120
count

# Data Sources for NDPs

- Paris21 used the FAOLEX database:
  - Database of national legislation, policies and other
  - Updated with an average of 8,000 new entries per year
  - Includes 286 NDPs since 2010 from 123 Countries
  - Limited number of documents per year




Availability of National Development Plan or Poverty Reduction Plan Documents by Country since 2010 from FAOLEX
Data Available
FALSE
TRUE
Source: FAOLEX

# Text Mining

- Using set of dictionary terms, we can look for sentences in NDPs containing these terms
- Then we can check to see if a number is cited in that sentence, as a measure of data use
- As an example going forward, we will consider the "National Strategic Development Plan 2014-2018" from Cambodia

# Example sentence 1

- The sentence in the document: "in contrast, malaysia and korea have this proportion in the range of 5% of gdp."
- Contains the keyword: gdp
- Did the sentence use data according to model?: TRUE

# Example sentence 2

- The sentence in the document: "¾ established 14 border liaison offices (blo's) along the border of the country and the mekong river to control and curb drug trafficking and precursor chemicals across the border."
- Contains the keyword: traffic
- Did the sentence use data according to model?: TRUE

# Example sentence 3

- The sentence in the document: "some citizens did not take advantage of employment services, and there was little cooperation between the concerned parties such as trainers, public citizens, and authorities at all levels. x limited cooperation between service providers, ministries, agencies, local authorities, and employers, and the national job and employment agency."
- Contains the keyword: services
- Did the sentence use data according to model?: FALSE

# Example sentence 4

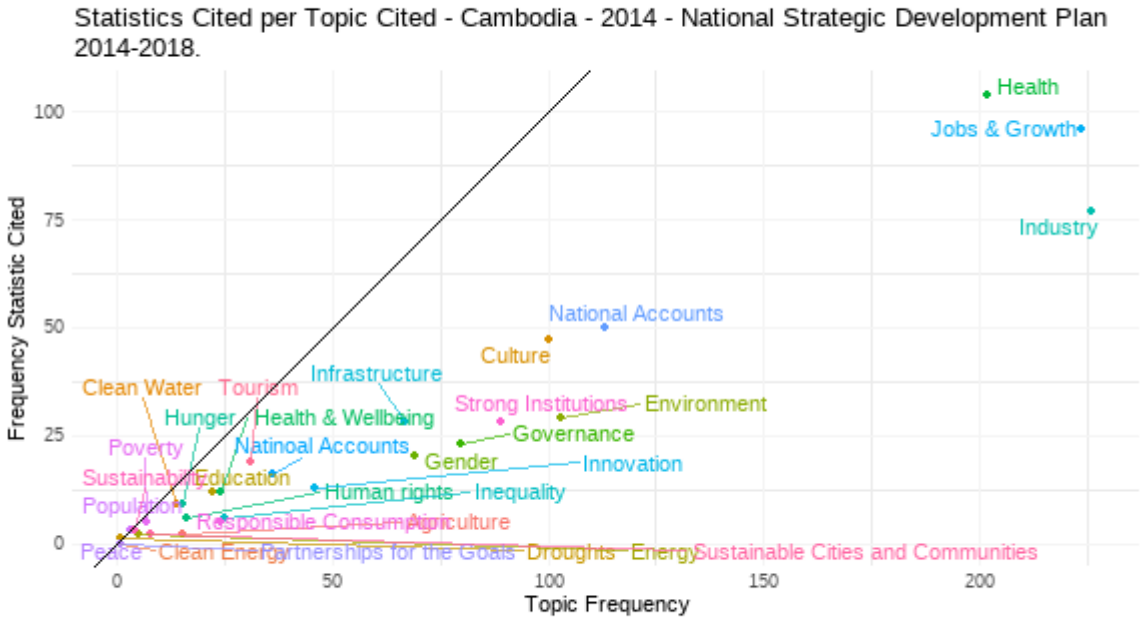
- The sentence in the document: "¾ study the root causes of the problems hindering the waterworks' revenue from covering their expenditures and try all possible solutions to resolve the problems."
- Contains the keyword: revenue
- Did the sentence use data according to model?: FALSE

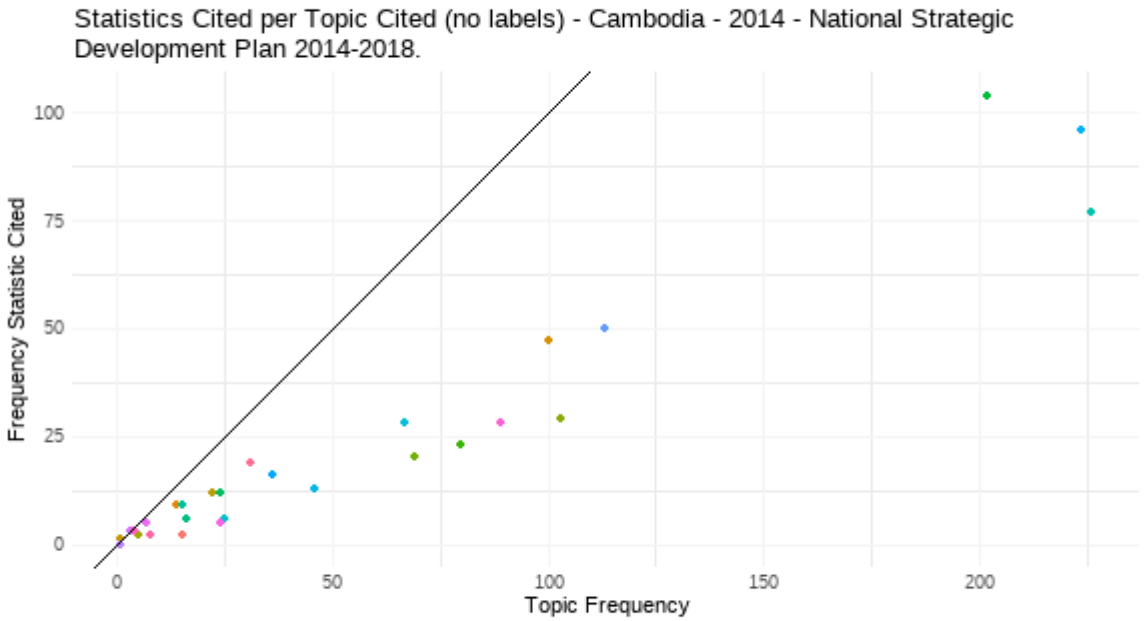
# Example sentence 5

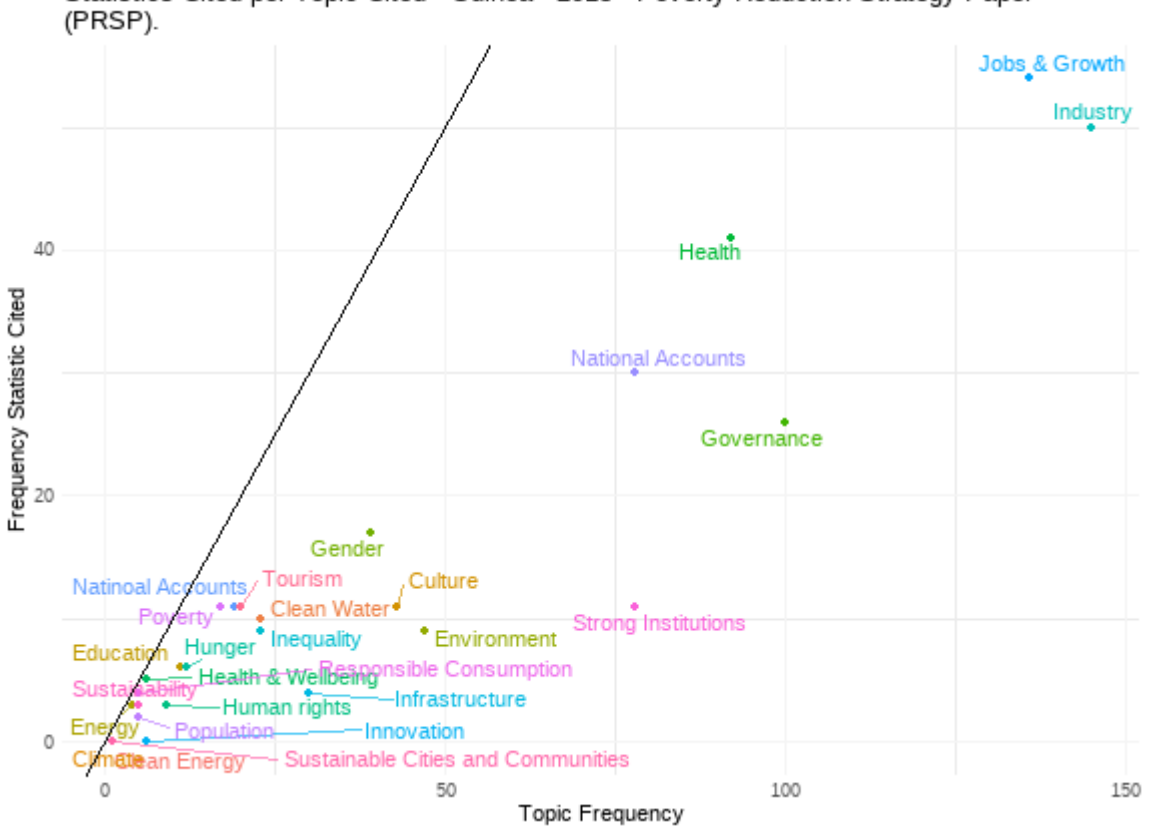
- The sentence in the document: "cambodia’s gdp per capita is approximated at usd 1,036 in 2013 (estimated)."
- Contains the keyword: gdp
- Did the sentence use data according to model?: TRUE

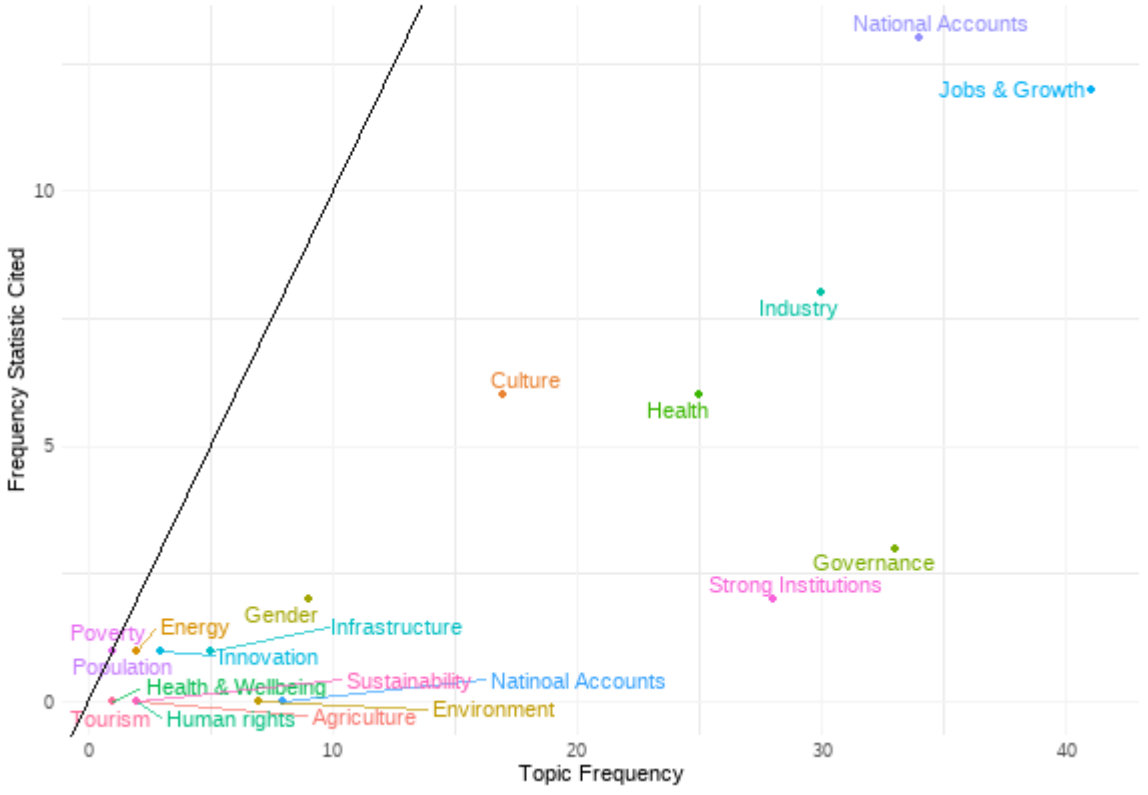
# Terms Usage & Data Mentions

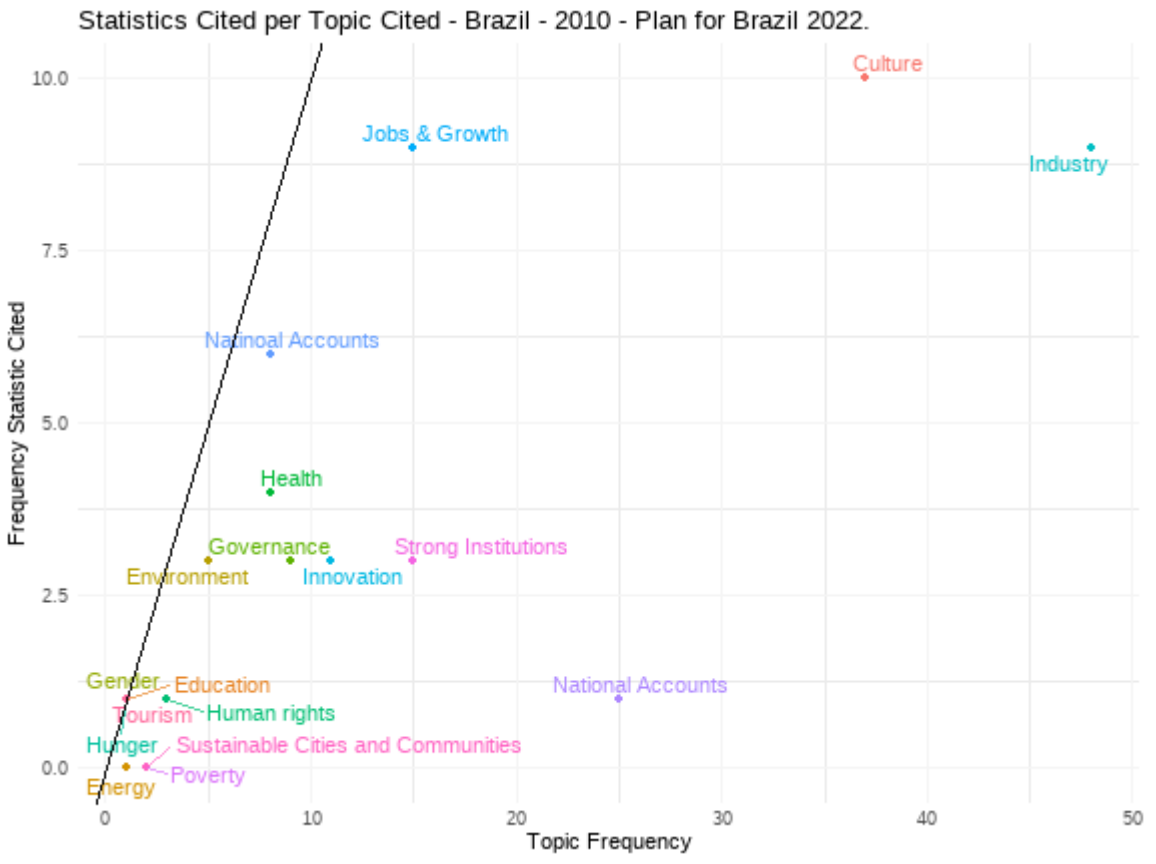
- Next, we will plot the number of times a topic was mentioned (on x-axis) against the of times statistics were cited according to our algorithm

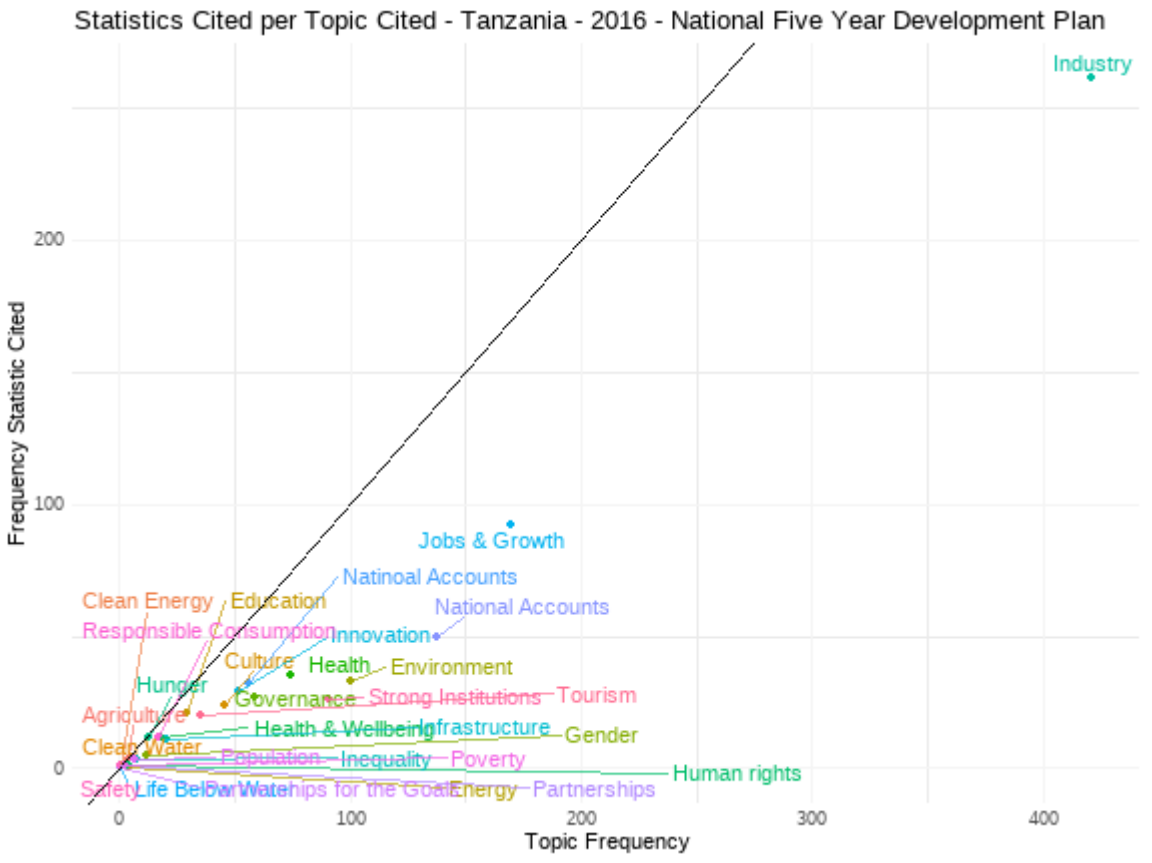

Statistics Cited per Topic Cited - Cambodia - 2014 - National Strategic Development Plan 2014-2018.
Frequency Statistic Cited
Topic Frequency
0
25
50
75
100
0
50
100
150
200
Health
Jobs & Growth
Industry
National Accounts
Culture
Infrastructure
Clean Water
Tourism
Strong Institutions
Environment
Hunger
Health & Wellbeing
Governance
Poverty
Natinoal Accounts
Gender
Innovation
Sustainability
Education
Human rights
Inequality
Population
Responsible Consumption
Agriculture
Peace
Clean Energy
Partnerships for the Goals
Droughts
Energy
Sustainable Cities and Communities


Statistics Cited per Topic Cited (no labels) - Cambodia - 2014 - National Strategic Development Plan 2014-2018.
Frequency Statistic Cited
100
75
50
25
0
0
50
100
150
200
Topic Frequency


Statistics Cited per Topic Cited - Guinea - 2013 - Poverty Reduction Strategy Paper (PRSP).
Jobs & Growth
Industry
Health
National Accounts
Governance
Gender
Natinoal Accounts
Tourism
Culture
Poverty
Clean Water
Strong Institutions
Inequality
Environment
Education
Hunger
Responsible Consumption
Health & Wellbeing
Sustainability
Infrastructure
Human rights
Energy
Population
Innovation
Climate
Clean Energy
Sustainable Cities and Communities
Frequency Statistic Cited
Topic Frequency
0
20
40
0
50
100
150


Statistics Cited per Topic Cited - Afghanistan - 2017 - Afghanistan National Peace and Development Framework (ANPDF) 2017 to 2021.
National Accounts
Jobs & Growth
Industry
Culture
Health
Governance
Strong Institutions
Gender
Energy
Poverty
Infrastructure
Innovation
Population
Health & Wellbeing
Sustainability
Natinoal Accounts
Environment
Agriculture
Human rights
Tourism
Frequency Statistic Cited
Topic Frequency
0
5
10
0
10
20
30
40


Statistics Cited per Topic Cited - Brazil - 2010 - Plan for Brazil 2022.
Culture
Jobs & Growth
Industry
Nainoal Accounts
Health
Governance
Strong Institutions
Environment
Innovation
Gender
Education
National Accounts
Tourism
Human rights
Hunger
Sustainable Cities and Communities
Poverty
Energy
Frequency Statistic Cited
Topic Frequency
10.0
7.5
5.0
2.5
0.0
0
10
20
30
40
50


Statistics Cited per Topic Cited - Tanzania - 2016 - National Five Year Development Plan
Frequency Statistic Cited
Topic Frequency
0
100
200
300
400
Industry
Jobs & Growth
Natinoal Accounts
National Accounts
Innovation
Clean Energy
Education
Responsible Consumption
Culture
Health
Environment
Hunger
Governance
Strong Institutions
Tourism
Agriculture
Health & Wellbeing
Infrastructure
Gender
Clean Water
Population
Inequality
Poverty
Human rights
Safety
Life Below Water
hips for the Goals
Energy
Partnerships