Análisis de texto de las Cuentas Públicas Participativas de la gestión de Gendarmería de Chile entre los años 2017 y 2020

Autores: Fabián Álvarez / Roberto Rodríguez - Actualización: Fabián Álvarez

Primera versión: 09-11-2020 - Actualización: 23/05/2022

Introducción

Las Cuentas Públicas Participativas son mecanismos de diálogo abierto que vinculan a las autoridades de los órganos de la Administración del Estado con la ciudadanía y tienen como objetivo informar sobre la gestión de políticas públicas realizadas, generar un proceso de retroalimentación que permita recoger las inquietudes y aportes de quienes participen de éstas y dar respuesta organizada en plazos oportunos a las inquietudes surgidas en el proceso.

Para este análisis se aplican técnicas de análisis de texto en R a las cuentas públicas de Gendarmería de Chile correspondientes a los años 2018, 2019, 2020 y 2021 (gestión 2017, 2018, 2019 y 2020, respectivamente).

Lectura y limpieza de archivos

Primero, leemos la fuente de los datos que, en nuestro caso, se trata de archivos en formato pdf.

```
speech_2021 <- pdf_text("../data/discurso_2021_2020.pdf")
speech_2020 <- pdf_text("../data/discurso_2020_2019.pdf")
speech_2019 <- pdf_text("../data/discurso_2019_2018.pdf")
speech_2018 <- pdf_text("../data/discurso_2018_2017.pdf")
```

Luego, eliminamos las primeras páginas de los discursos ya que contienen elementos que no forman parte de la cuenta pública. Lo mismo hacemos para las últimas páginas; en ambos casos, solo cuando corresponda.

```
speech_2021 <- speech_2021 %>%
  .[-1] %>%
  .[-13]
speech_2020 <- speech_2020 %>%
  .[-1:-2] %>%
  .[-53:-56]
speech_2019 <- speech_2019 %>%
  .[-1:-2]
speech_2018 <- speech_2018 %>%
  .[-1]
```

Los objetos en cada discurso quedaron separados, por lo que los unimos en un solo objeto por cada discurso.

```
speech_2021 <- paste(speech_2021, collapse = " ")
speech_2020 <- paste(speech_2020, collapse = " ")
speech_2019 <- paste(speech_2019, collapse = " ")
speech_2018 <- paste(speech_2018, collapse = " ")
```

Las *stopwords* o *palabras vacías* son aquellas palabras que no tienen significado por sí mismas. Solo modifican o acompañan a otras, por lo cual debemos quitarlas para el análisis. Para ello, tomamos las palabras desde el repositorio AnaText que ya hemos descargado previamente. Agregamos palabras adicionales que vayamos encontrando irrelevantes.

```
stopwords_es <- read_csv("../data/vacias.txt", col_names = TRUE, show_col_types = FALSE)
my_stopwords <- tibble(palabra = c("mil", "millones", "año", "años", "chile", "dado",
                                   "dar", "debido", "decir", "acerca", "pesos",
                                   "fin", "ser", "respecto", "debe", "gran", "tiene",
                                   "tienen", "puede", "ir", "hace"))
more_stopwords <- tibble(palabra = c("cdp", "cerrado", "gendarmería", "fecha", "período",
                                     "cuenta", "informe", "viii", "monto", "diariamente",
                                     "diferentes", "impacta", "enfocar", "deberá"))
```

Expresiones Regulares y Palabras Vacías (*stopwords*)

Revisamos expresiones regulares y palabras vacías para descartar del documento.

Discurso 2021 (Gestión 2020)

Comenzamos eliminando las expresiones regulares para el Discurso 2020 y haciendo algunas correcciones.

```
speech_2021 <- speech_2021 %>%
  str_replace_all("\n", " ") %>% # Replace "\n" by space
  str_remove_all("""") %>% str_remove_all("""") %>% # Remove ""
  str_remove_all("INTRODUCCIÓN") %>%
  str_replace_all("\\d+\\s+\\S+gob\\.cl", " ") %>% # Pages
  str_replace_all("INFORME FINAL", " ") %>%
  str_replace_all("\\s\\S+\\.-", " ") %>% str_replace_all("VI\\.", " ") %>% # Numbering
  str_replace_all("\\s\\d+-\\s", " ") %>%
  str_replace_all("\\s\\d\\.\\s", " ") %>%
  str_replace_all("5\\.1", " ") %>% str_replace_all("5\\.2-", " ") %>%
  str_replace_all("\\s[abcde]+\\)\\s", " ") %>%
  str_replace_all("http\\S+\\s", " ") %>% # urls
  str_replace_all("SOCIEDAD\\S+\\s", " ") %>%
  str_replace_all("Gendarmería de Chile", " ") %>% str_replace_all("Gendarmería", " ") %>%
  str_replace_all("Institución.+Manríquez", " ") %>%
  str_replace_all("con 64\\s+.+Nacional de Chile", " ") %>% str_replace_all("64 \\(1\\)", " ") %>%
  str_replace_all("públicas65", "públicas") %>% str_replace_all("decisiones\\.66", "decisiones") %>%
  str_replace_all("65\\s+Instructivo.+Transparencia\\.2015\\.", " ") %>%
  str_replace_all("Informe Ejecutivo", " ") %>%
  str_replace_all("informe Resumen Ejecutivo", " ") %>%
  str_replace_all("\\Snforme final", " ") %>%
  str_replace_all("\\Snforme.", " ")
```

```
speech_2021 <- speech_2021 %>%
  str_replace_all("Cuentas Públicas Participativas", "Cuenta Pública Participativa ") %>%
  str_replace_all("cuentas públicas", "cuenta pública") %>%
  str_replace_all("COVID 19", "COVID-19") %>%
  str_replace_all("\\sCOVID\\s", " COVID-19 ") %>%
  str_remove_all("\\(Consejo para la transparencia\\.2015\\)") %>%
  str_replace_all("Consejo de la Sociedad Civil \\(COSOC\\)", "Consejo de la Sociedad Civil") %>%
  str_replace_all("Consejo de la Sociedad Civil", "COSOC") %>%
  str_replace_all("CONSEJO DE LA SOCIEDAD CIVIL", "COSOC") %>%
  str_remove_all("Unidad de Atención y Participación Ciudadana") %>%
  str_remove_all("Unidad de Comunicaciones") %>%
  str_replace_all("Unidades Penales", "Unidad Penal") %>%
  str_remove_all("Departamento de Estadística y Estudios penitenciarios") %>%
  str_remove_all("disponer") %>%
  str_replace_all("pacientes", "paciente") %>%
  str_replace_all("PACIENTES", "paciente") %>%
  str_replace_all("paciente", "pacientes") %>%
  str_replace_all("clínico", "clínica") %>%
  stripWhitespace()
```

Convertimos el discurso en un dataframe, separamos sus palabras y calculamos sus frecuencias.

```
frequencies_2021 <- tibble(speech = speech_2021) %>%
  unnest_tokens(output = palabra, input = speech, strip_numeric = TRUE) %>%
  count(palabra, sort = TRUE)
frequencies_2021
```

```
## # A tibble: 921 x 2
##    palabra     n
##    <chr>   <int>
##  1 de        255
##  2 la        151
##  3 y          97
##  4 en         96
##  5 el         79
##  6 a          78
##  7 que        72
##  8 se         44
##  9 las        41
## 10 del        40
## # ... with 911 more rows
```

Quitamos las *stopwords* y recalculamos las frecuencias.

```
frequencies_2021 <- frequencies_2021 %>%
  anti_join(stopwords_es) %>%
  anti_join(my_stopwords) %>%
  anti_join(more_stopwords)
head(frequencies_2021)
```

```
## # A tibble: 6 x 2
##   palabra         n
```

```
##   <chr>       <int>
## 1 salud          17
## 2 pública        16
## 3 mental         14
## 4 información    13
## 5 personas       13
## 6 atención       12
```

Discurso 2020 (Gestión 2019)

Comenzamos eliminando las expresiones regulares para el Discurso 2020 y haciendo algunas correcciones.

```
speech_2020 <- speech_2020 %>%
  str_replace_all("\n", " ") %>% # Replace "\n" by space
  str_remove_all("""") %>% str_remove_all("""") %>% # Remove ""
  str_replace_all("MINJU", "MINJUDDHH") %>% # Ministerio de Justicia y DDHH
  str_replace_all("MINJUDDHH-DDHH", "MINJUDDHH") %>%
  str_replace_all("\\s+•\\s+", " ") %>% # Bullets
  str_replace_all("\\s\\S+\\.-\\s", " ") %>%
  str_replace_all("\\s\\d-\\s", " ") %>% # Numbering
  str_replace_all("\\s\\d\\.\\s", " ") %>%
  str_replace_all("\\s\\d\\d\\.\\s", " ") %>%
  str_replace_all("\\s\\d\\.\\d\\s", " ") %>%
  str_replace_all("\\s+[abcde]+\\)+\\s", " ") %>%
  str_replace_all("\\s[abcde]\\.\\d\\)", " ") %>%
  str_replace_all("19\\.-", " ") %>% #
  str_remove_all("http\\S*") %>% # urls
  str_remove_all("www.\\S*") %>% # Remove web pages
  str_remove_all("Twitter.+gendarmeriacl") %>%  # Remove social networks
  str_remove_all("N° de internos heridos.+\\(S\\.I\\.G\\)") %>% # Remove Tables
  str_remove_all("A continuación, se expone un desglose.+Fuente: Departamento de Infraestructura") %>% 
  str_remove_all("MATRICULADOS EN EDUCACIÓN SUPERIOR DICIEMBRE 2019.+Total\\s+163\\s+Fuente: Departamen
  str_remove_all("\\s+Tabla Privados de Libertad Inscritos para dar PSU.+\\s+2046\\s+Fuente: Departamen
  str_remove_all("\\s+Tabla Resultados PSU 2019 de Privados de Libertad, por región:.+\\s+13\\s+Fuente:
  str_remove_all("\\s+INTERNOS PARTICIPANDO.+\\s+Automotriz\\s+Fuente: Departamento Sistema Cerrado") %
  str_remove_all("\\s+Eliminación de antecedentes:.+Fuente: Departamento Post Penitenciario") %>% # Rem
  str_remove_all("\\s+Intervención:.+\\s+37\\s+Fuente: Departamento Subsistema Cerrado") %>% # Remove T
  str_remove_all("\\s+CANTIDAD DE CELULARES.+\\s+256\\s+Fuente: Subdirección Operativa") %>% # Remove T
  str_remove_all("\\sCOVID: Estadística de contagios por región.+\\s+1357\\s+Fuente: Subdirección Opera
  str_remove_all("\\s+Catastro.+\\s+167\\s+Fuente: Subdirección Operativa") %>% # Remove Tables
  str_remove_all("\\sfecha:.+Fuente: Subdirección Operativa") %>% # Remove Tables
  stripWhitespace() # Remove unnecessary spaces
```

```
speech_2020 <- speech_2020 %>%
  str_replace_all("COVID\\s19", "COVID-19") %>% # Standardize COVID-19
  str_replace_all("COVID-\\s19", "COVID-19") %>%
  str_replace_all("COVID:", "COVID-19") %>%
  str_replace_all("Covid-19", "COVID-19") %>%
  str_replace_all("Covid19", "COVID-19") %>%
  str_replace_all("COVID y", "COVID-19 y") %>%
  str_replace_all("cas2", "cas") %>% # Others
  str_replace_all("ransparencia\\.", "ransparencia ") %>%
  str_replace_all("para la transparencia", "para la Transparencia") %>%
```

```
  str_replace_all("s e g u r i d a d in t e r na", "seguridad interna") %>%
  str_replace_all("d e l o s re c i n to s", "de los recintos") %>%
  str_replace_all("p e n i te nc i a r io s", "penitenciarios") %>%
  str_remove_all("Departamento Sistema Cerrado") %>%
  str_remove_all("Departamento de Salud") %>%
  str_remove_all("Departamento de DDHH") %>%
  str_remove_all("Departamento DDHH") %>%
  str_remove_all("Departamento de Promoción y Protección de los DDHH") %>%
  str_remove_all("Departamento de Promoción y Protección de Derechos Humanos") %>%
  str_remove_all("Departamento de Promoción y Protección de los Derechos Humanos") %>%
  str_remove_all("Departamento de Infraestructura") %>%
  str_remove_all("Departamento de Informática") %>%
  str_remove_all("Departamento en el Sistema Cerrado") %>%
  str_remove_all("Departamento") %>%
  str_replace_all("autoridades sanitarias", "autoridad sanitaria") %>%
  str_replace_all("condiciones sanitarias", "condición sanitaria") %>%
  str_replace_all("residencias sanitarias", "residencia sanitaria") %>%
  str_replace_all("restricciones sanitarias", "restricción sanitaria") %>%
  str_replace_all("sanitarias", "sanitaria") %>%
  str_replace_all("Sanitarias", "sanitaria") %>%
  str_replace_all("contagios", "contagio") %>% str_replace_all("contagio", "contagios") %>%
  str_remove_all("Departamento de Estadística y Estudios penitenciarios") %>%
  str_replace_all("lesión", "lesiones") %>%
  str_replace_all("Egresos", "egreso") %>% str_replace_all("egresos", "egreso") %>%
  str_replace_all("egreso", "egresos") %>%
  stripWhitespace()
```

Convertimos el discurso en un dataframe, separamos sus palabras y calculamos sus frecuencias.

```
frequencies_2020 <- tibble(speech = speech_2020) %>%
  unnest_tokens(output = palabra, input = speech, strip_numeric = TRUE) %>%
  count(palabra, sort = TRUE)
frequencies_2020
```

```
## # A tibble: 2,242 x 2
##    palabra     n
##    <chr>   <int>
##  1 de       1032
##  2 la        429
##  3 el        315
##  4 y         306
##  5 en        288
##  6 a         285
##  7 que       209
##  8 se        191
##  9 los       183
## 10 las       179
## # ... with 2,232 more rows
```

Quitamos las *stopwords* y recalculamos las frecuencias.

```
frequencies_2020 <- frequencies_2020 %>%
  anti_join(stopwords_es) %>%
  anti_join(my_stopwords) %>%
  anti_join(more_stopwords)
head(frequencies_2020)
```

```
## # A tibble: 6 x 2
##   palabra      n
##   <chr>    <int>
## 1 personas    62
## 2 libertad    47
## 3 internos    31
## 4 penal       31
## 5 visitas     28
## 6 derechos    25
```

Discurso 2019 (Gestión 2018)

Comenzamos eliminando las expresiones regulares para el Discurso 2019 y haciendo algunas correcciones.

```
speech_2019 <- speech_2019 %>%
  str_replace_all("\n", " ") %>% # Replace "\n" by space
  str_remove_all("Tabla rela.+\\s+Fuente: Departamento de Infraestructura de Gendarmería de Chile") %>%
  str_remove_all("Capacitaciones en cifras.+\\s+Fuente: Escuela Institucional") %>%
  str_remove_all("Presupuesto Inicial.+\\s+Contabilidad y Presupuesto, Gendarmería de Chile") %>%
  str_replace_all("\\s\\s\\s\\s\\d+\\s", " ") %>% # Page number
  str_replace_all("\\s\\d\\.\\s", " ") %>% # Numbering & bullets
  str_remove_all("1. PRESENTACIÓN ") %>%
  str_replace_all("\\s\\d\\.\\d\\.\\d\\.\\s", " ") %>%
  str_replace_all("\\s\\d\\.\\d\\.\\s", " ") %>%
  str_replace_all("\\s•\\s", " ") %>%
  str_replace_all("\\s-\\s", " ") %>%
  str_replace_all("\\s\\S\\.\\s", " ") %>%
  str_replace_all("Ministerio de Justicia y Derechos Humanos", "MINJUDDHH") %>%
  str_replace_all("Ministerio de Justicia y DD.HH.", "MINJUDDHH") %>%
  str_remove_all("N°")
```

```
speech_2019 <- speech_2019 %>%
  str_replace_all("ETIntervención", "Intervención") %>%
  str_replace_all("ETCapacitación", "Capacitación") %>%
  str_replace_all("ETColocación", "Colocación") %>%
  str_replace_all("APpsicosocial", "psicosocial") %>%
  str_replace_all("", " ") %>%
  str_replace_all("APestablecimiento", "establecimiento") %>%
  str_replace_all("Aapoyo", "apoyo") %>%
  str_replace_all("Aopermisos", "permisos") %>%
  str_replace_all("1terceros", "terceros") %>%
  str_remove_all("""") %>% str_remove_all("""") %>%
  str_replace_all("de \\+R", "de Proyecto +R") %>%
  str_replace_all("Departamento de Contabilidad y Presupuesto", " ") %>%
  str_replace_all("Departamento de Gestión de Personas", " ") %>%
  str_replace_all("Departamento de Gestión y Desarrollo de Personas", " ") %>%
```

```
  str_replace_all("Departamento de Salud", " ") %>%
  str_replace_all("Departamentos de Inteligencia Penitenciaria\\s+y de\\s+Investigación Criminal", " ")
  str_replace_all("Departamento de Investigación y Análisis Penitenciario \\(DIAP\\)", " ") %>%
  str_replace_all("departamentos de Salud e Informática", " ") %>%
  str_replace_all("Departamento de Promoción y Protección de los Derechos Humanos", " ") %>%
  str_replace_all("Departamento de Infraestructura", " ") %>%
  str_replace_all("Subdepartamento de Servicios\\s+Especializados", " ") %>%
  str_replace_all("Departamento de Control Penitenciario", " ") %>%
  str_replace_all("departamentos y/o unidades", " ") %>%
  str_remove_all("a Departamento") %>%
  str_remove_all("términos") %>%
  str_remove_all("mejorando") %>%
  str_remove_all("desarrollar") %>%
  str_remove_all("porcentaje") %>%
  str_remove_all("Departamento de Estadística y Estudios penitenciarios") %>%
  str_replace_all("\\Sstablecimientos", "establecimientos") %>%
  str_replace_all("\\Saborales", "laboral") %>%
  str_replace_all("\\Snternacionales", "internacional") %>%
  str_replace_all("\\s\\Sompras\\s", " compra ") %>% str_replace_all("\\s\\Sompra\\s", " compras ") %>%
  stripWhitespace()
```

Convertimos el discurso en un dataframe, separamos sus palabras y calculamos sus frecuencias.

```
frequencies_2019 <- tibble(speech = speech_2019) %>%
  unnest_tokens(output = palabra, input = speech, strip_numeric = TRUE) %>%
  count(palabra, sort = TRUE)
frequencies_2019
```

```
## # A tibble: 2,966 x 2
##    palabra     n
##    <chr>   <int>
##  1 de       1619
##  2 y         613
##  3 la        606
##  4 en        474
##  5 el        417
##  6 a         382
##  7 se        290
##  8 que       267
##  9 los       251
## 10 del       211
## # ... with 2,956 more rows
```

Quitamos las *stopwords* y recalculamos las frecuencias.

```
frequencies_2019 <- frequencies_2019 %>%
  anti_join(stopwords_es) %>%
  anti_join(my_stopwords) %>%
  anti_join(more_stopwords)
head(frequencies_2019)
```

```
## # A tibble: 6 x 2
```

```
##   palabra          n
##   <chr>        <int>
## 1 personas        52
## 2 población       48
## 3 trabajo         47
## 4 funcionarios    45
## 5 laboral         45
## 6 seguridad       45
```

Discurso 2018 (Gestión 2017)

Comenzamos eliminando las expresiones regulares para el Discurso 2018 y haciendo algunas correcciones.

```
speech_2018 <- speech_2018 %>%
  str_replace_all("\n", " ") %>% # Replace "\n" by space
  str_remove_all("\\s+Allanamientos\\s+.+854") %>% # Remove Tables
  str_remove_all("\\s+PROGRAMAS\\s+REINSERCIÓN\\s.+\\sLACTANTES") %>%
  str_remove_all("Distribución Regional.+NACIONAL\\s+43") %>%
  str_remove_all("siguiente detalle.+Total\\s+305") %>%
  str_remove_all("Aspirante Oficiales P.+327 Hombres\\)") %>%
  str_remove_all("\\s+N° de horas.+ARAUCANÍA\\)") %>%
  str_remove_all("AÑO\\s+NOMBRE.+PUBLICA\\s+30") %>%
  str_remove_all("II\\.\\s") %>% str_remove_all("I\\.\\s") %>%
  str_replace_all("\\s\\d\\.\\s+", " ") %>%
  str_replace_all("\\s•\\s", " ") %>%
  str_remove_all("1 Fuente de datos: Estadística General Penitenciaria\\. Unidad de Estadística\\. Genda
```

```
speech_2018 <- speech_2018 %>%
  str_remove_all("N°") %>%
  str_replace_all("Departamento del Sistema Cerrado\\.", " ") %>%
  str_replace_all("Departamento de Recursos Humanos", " ") %>%
  str_replace_all("multidisclinarios", "multidisciplinarios") %>%
  str_replace_all("asociadas:Implementación", "asociadas Implementación") %>%
  str_replace_all("pesos\\.Reposición", "pesos Reposición") %>%
  str_replace_all("incendios:Programa", "incendios Programa") %>%
  str_replace_all("Coyhaique\\.Mantención", "Coyhaique Mantención") %>%
  str_replace_all("intervenidos", "intervenido") %>% str_replace_all("intervenido", "intervenidos") %>%
  str_replace_all("\\Sstablecimientos", "establecimientos") %>%
  str_replace_all("Laborales", "laboral") %>%
  stripWhitespace()
```

Convertimos el discurso en un dataframe, separamos sus palabras y calculamos sus frecuencias.

```
frequencies_2018 <- tibble(speech = speech_2018) %>%
  unnest_tokens(output = palabra, input = speech, strip_numeric = TRUE) %>%
  count(palabra, sort = TRUE)
frequencies_2018
```

```
## # A tibble: 1,597 x 2
##    palabra     n
##    <chr>   <int>
##  1 de        672
```

```
##  2 la        242
##  3 y         211
##  4 en        209
##  5 el        183
##  6 a         178
##  7 se        140
##  8 los       113
##  9 del        99
## 10 que        95
## # ... with 1,587 more rows
```

Quitamos las *stopwords* y recalculamos las frecuencias.

```
frequencies_2018 <- frequencies_2018 %>%
  anti_join(stopwords_es) %>%
  anti_join(my_stopwords) %>%
  anti_join(more_stopwords)
head(frequencies_2018)
```

```
## # A tibble: 6 x 2
##   palabra              n
##   <chr>            <int>
## 1 personas            38
## 2 programa            28
## 3 establecimientos    21
## 4 laboral             20
## 5 trabajo             20
## 6 capacitación        19
```

Análisis por palabra

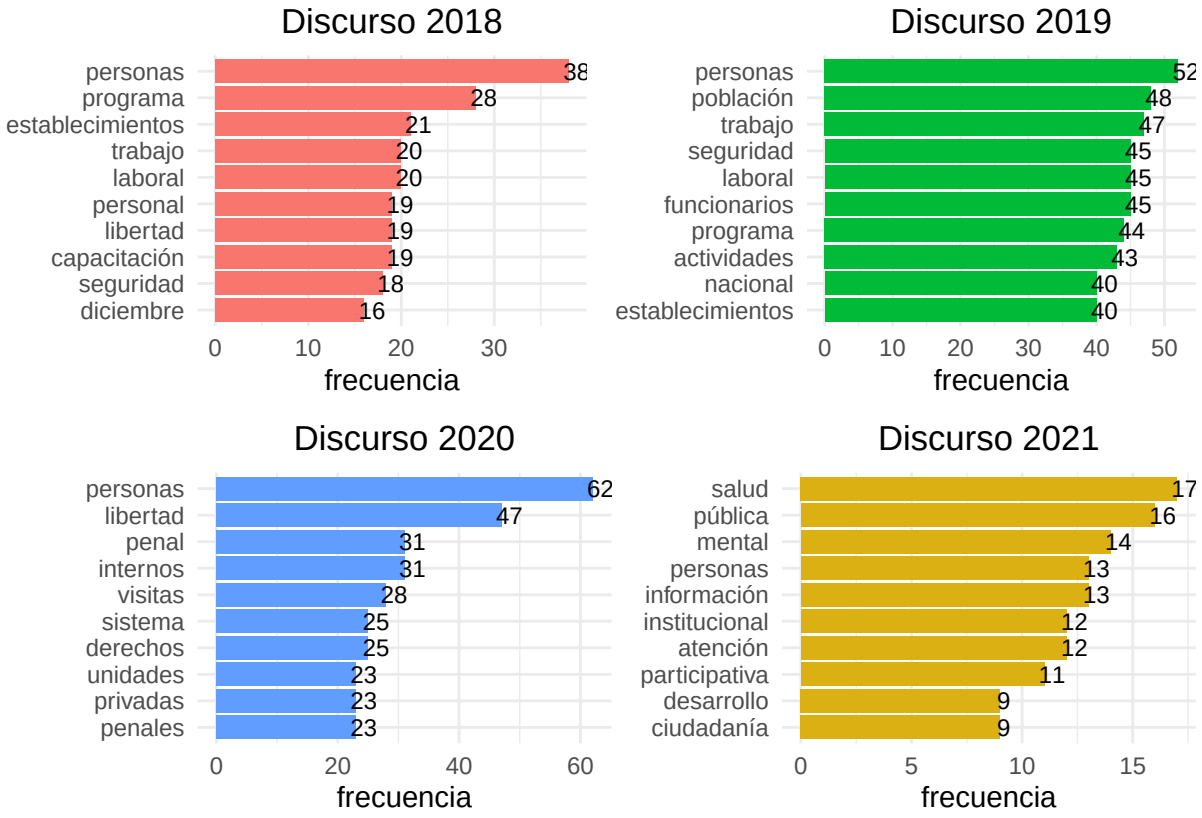
Primero calculamos y graficamos las palabras más frecuentes en cada discurso para, posteriormente, realizar un análisis TF-IDF. Este análisis permite determinar la relevancia de una palabra para un documento respecto de un conjunto de documentos.

Frecuencias

Graficamos las 10 palabras más frecuentes de cada discurso.

```
pic_2021 <- frequencies_2021 %>%
  slice_head(n = 10) %>%
  ggplot(aes(y = reorder(palabra, n), n)) +
  geom_col(fill = "#dbb012") +
  geom_text(aes(label = n), size = 3, hjust = 0.2) +
  theme_minimal() +
  labs(y = NULL, x = "frecuencia") +
  ggtitle("Discurso 2021") +
  theme(plot.title = element_text(hjust = 0.5))

pic_2020 <- frequencies_2020 %>%
  slice_head(n = 10) %>%
```

```
  ggplot(aes(y = reorder(palabra, n), n)) +
  geom_col(fill = "#619cff") +
  geom_text(aes(label = n), size = 3, hjust = 0.2) +
  theme_minimal() +
  labs(y = NULL, x = "frecuencia") +
  ggtitle("Discurso 2020") +
  theme(plot.title = element_text(hjust = 0.5))

pic_2019 <- frequencies_2019 %>%
  slice_head(n = 10) %>%
  ggplot(aes(y = reorder(palabra, n), n)) +
  geom_col(fill = "#00ba38") +
  geom_text(aes(label = n), size = 3, hjust = 0.2) +
  theme_minimal() +
  labs(y = NULL, x = "frecuencia") +
  ggtitle("Discurso 2019") +
  theme(plot.title = element_text(hjust = 0.5))

pic_2018 <- frequencies_2018 %>%
  slice_head(n = 10) %>%
  ggplot(aes(y = reorder(palabra, n), n)) +
  geom_col(fill = "#f8766d") +
  geom_text(aes(label = n), size = 3, hjust = 0.2) +
  theme_minimal() +
  labs(y = NULL, x = "frecuencia") +
  ggtitle("Discurso 2018") +
  theme(plot.title = element_text(hjust = 0.5))

(pic_2018 + pic_2019) / (pic_2020 + pic_2021)
```



Análisis TF-IDF

Unimos las frecuencias en un solo data frame, identificando palabras y frecuencias con sus respectivos discursos.

```
frequencies_2021 <- frequencies_2021 %>%
  mutate(discurso = "C.P.P. 2021", .before = palabra)
frequencies_2020 <- frequencies_2020 %>%
  mutate(discurso = "C.P.P. 2020", .before = palabra)
frequencies_2019 <- frequencies_2019 %>%
  mutate(discurso = "C.P.P. 2019", .before = palabra)
frequencies_2018 <- frequencies_2018 %>%
  mutate(discurso = "C.P.P. 2018", .before = palabra)

messages <- bind_rows(frequencies_2018, frequencies_2019, frequencies_2020, frequencies_2021)
head(messages)
```

```
## # A tibble: 6 x 3
##   discurso    palabra             n
##   <chr>       <chr>           <int>
## 1 C.P.P. 2018 personas           38
## 2 C.P.P. 2018 programa           28
## 3 C.P.P. 2018 establecimientos   21
## 4 C.P.P. 2018 laboral            20
```

```
## 5 C.P.P. 2018 trabajo             20
## 6 C.P.P. 2018 capacitación        19
```

Luego, calculamos las frecuencias inversas de las palabras.

```
messages_tfidf <- bind_tf_idf(messages,
                              term = palabra,
                              document = discurso,
                              n = n)
head(messages_tfidf)
```

```
## # A tibble: 6 x 6
##   discurso    palabra              n      tf   idf  tf_idf
##   <chr>       <chr>            <int>   <dbl> <dbl>   <dbl>
## 1 C.P.P. 2018 personas            38 0.0124  0     0
## 2 C.P.P. 2018 programa            28 0.00915 0     0
## 3 C.P.P. 2018 establecimientos    21 0.00686 0.288 0.00197
## 4 C.P.P. 2018 laboral             20 0.00654 0.693 0.00453
## 5 C.P.P. 2018 trabajo             20 0.00654 0     0
## 6 C.P.P. 2018 capacitación        19 0.00621 0.288 0.00179
```

Y comparamos gráficamente.

```
one_word_plot <-
  messages_tfidf %>%
    group_by(discurso) %>%
    top_n(5) %>%
    ungroup %>%
    mutate(discurso = as.factor(discurso),
           palabra = reorder_within(palabra, tf_idf, discurso)) %>%
    ggplot(aes(palabra, tf_idf, fill = discurso)) +
    geom_col(show.legend = FALSE) +
    facet_wrap(~discurso, scales = "free_y") +
    coord_flip() +
    scale_x_reordered() +
    scale_y_continuous(expand = c(0,0)) +
    labs(y = "tf-idf", x = NULL)
one_word_plot
```

En el discurso de la Cuenta Pública Participativa (C.P.P.) 2018 se dio énfasis a la actividad, capacitación y colocación *laboral*, y a aquellos *penados* que han sido intervenidos. Asimismo, se rindió cuenta de *recursos* ejecutados durante el 2017, financiados por Ley de Presupuesto. También se menciona la política de protección de *armamento* en las unidades especiales.

Durante el 2019, el discurso de la C.P.P. se enfocó en destacar los programas de intervención, capacitación y colocación *laboral*, así como también las actividades *laborales* realizadas al interior de los *establecimientos* penitenciarios durante el 2018. De la misma forma, se dio énfasis a la *vigilancia* de las personas privadas de libertad, a la adopción de normas y buenas prácticas ajustadas a derecho *internacional* y cumplimiento de compromisos con organismos *internacionales*.

Vemos que el término *contagios* y *covid* se tomó la atención del discurso 2020, lo cual tiene mucho sentido en atención a la emergencia *sanitaria* de pandemia que comenzó en dicho año. Por otro lado, se informa sobre la publicación de un estudio de reincidencia que consideró *egresos* de privados de libertad en el año 2016. Además, se puso atención a la gran cantidad de detenidos durante el 2019 que fueron recibidos con *lesiones* evidentes desde las instituciones policiales, para lo cual se aprobaron protocolos de constatación de estado de salud.

Por último, el discurso del 2021 de la C.P.P. enfatizó dos grandes temáticas. La primera, relacionada con la salud *mental* y la derivación de posibles *pacientes* privados de libertad con necesidades de atención *clínica* psicológica en el ámbito de la *psicoterapia* individual. Esto, en atención especialmente a que durante el 2020 se mantuvo el riesgo sanitario de pandemia. El segundo tópico abordó el desarrollo de una Cuenta Pública *Participativa* con mayor inclusión social y con transmisión vía streaming por segundo año consecutivo.

Análisis por bigramas

Definimos algunas *stopwords* para descartar adicionalmente a las ya definidas, realizamos algunas correcciones adicionales y calculamos la frecuencia de los bigramas para cada documento.

```
bi_stopwords <- tibble(palabra = c("cuenta", "cuentas", "deberán", "realizar", "diferentes"))
```

```
speech_2021 <- speech_2021 %>%
  str_remove_all("políticas,") %>% str_remove_all("planes") %>% str_remove_all("programas") %>%
  str_remove_all("Intrapenitenciaria") %>% str_remove_all("INTRAPENITENCIARIA") %>%
  str_replace_all("\\Snidad \\Senal", "establecimientos penitenciarios") %>%
  str_replace_all("del establecimiento", "de los establecimientos penitenciarios") %>%
  str_remove_all("\\Suenta \\Sública \\Sarticipativa") %>%
  stripWhitespace()
```

```
speech_2020 <- speech_2020 %>%
  str_replace_all("\\Snidad \\Senal", "unidades penales") %>%
  str_replace_all("\\Snidades \\Senales", "establecimientos penitenciarios") %>%
  str_replace_all("visita virtuales", "visitas virtuales") %>%
  str_replace_all("visita virtual", "visitas virtuales") %>%
  str_replace_all("espacio físico", "espacios físicos") %>%
  str_replace_all("\\Sstablecimiento \\Senitenciario", "establecimientos penitenciarios") %>%
  str_replace_all("establecimiento penal", "establecimientos penitenciarios") %>%
  str_replace_all("teléfono celular", "teléfonos móviles") %>%
  str_replace_all("teléfono móvil", "teléfonos móviles") %>%
  str_remove_all("Centro de Estudios Justicia & Sociedad") %>%
  stripWhitespace()
```

```
speech_2019 <- speech_2019 %>%
  str_remove_all("igual forma") %>%
  str_remove_all("Centro de Cumplimiento Penitenciario") %>%
  str_remove_all("Centros de Cumplimiento Penitenciario") %>%
  str_replace_all("unidades penales", "establecimientos penitenciarios") %>%
  str_replace_all("unidad penal", "establecimientos penitenciarios") %>%
  str_replace_all("\\Sstablecimiento \\Senal", "establecimientos penitenciarios") %>%
  str_replace_all("\\Sstablecimientos \\Senales", "establecimientos penitenciarios") %>%
  str_replace_all("diferentes penales", "diferentes establecimientos penitenciarios") %>%
  str_replace_all("\\Sirección \\Segional", "direcciones regionales") %>%
  stripWhitespace()
```

```
speech_2018 <- speech_2018 %>%
  str_replace_all("establecimiento penitenciario", "establecimientos penitenciarios") %>%
  str_replace_all("establecimientos penales", "establecimientos penitenciarios") %>%
  str_replace_all("algunos establecimientos", "algunos establecimientos penitenciarios") %>%
  str_replace_all("73 establecimientos", "73 establecimientos penitenciarios") %>%
  str_replace_all("prohibidas en establecimientos", "prohibidas en establecimientos penitenciarios") %>%
  str_replace_all("establecimientos institucionales", "establecimientos penitenciarios") %>%
  str_replace_all("condena en establecimientos", "condena en establecimientos penitenciarios") %>%
  str_replace_all("otros establecimientos", "otros establecimientos penitenciarios") %>%
  str_replace_all("últimos establecimientos", "últimos establecimientos penitenciarios") %>%
  str_replace_all("equipamiento del establecimiento", "equipamiento de establecimientos penitenciarios"
  str_replace_all("nuevo establecimiento", "nuevos establecimientos penitenciarios") %>%
```

```
  str_replace_all("tener un establecimiento", "tener establecimientos penitenciarios") %>%
  str_replace_all("10 establecimientos", "10 establecimientos penitenciarios") %>%
  str_remove_all("Departamento del \\Sistema \\Serrado") %>%
  str_replace_all("\\s\\Sistema \\Serrado", " subsistema cerrado") %>%
  str_replace_all("unidades penales", "establecimientos penitenciarios") %>%
  str_remove_all("siguientes iniciativas") %>%
  str_replace_all("direcciones técnicas regionales", "unidades técnicas") %>%
  str_replace_all("unidades técnicas regionales", "unidades técnicas") %>%
  str_replace_all("\\s\\Sonvenio\\s", " convenios suscritos ") %>%
  str_replace_all("convenios con", "convenios suscritos con") %>%
  str_replace_all("convenios en", "convenios suscritos en") %>%
  str_replace_all("de allanamiento en", "de allanamientos simultáneos en") %>%
  str_replace_all("estos allanamientos", "estos allanamientos simultáneos") %>%
  stripWhitespace()
```

Frecuencias

```
bigram_2021 <- speech_2021 %>%
  tibble(speech = speech_2021) %>%
  unnest_tokens(input = speech,
                output = palabra,
                token = "ngrams",
                n = 2) %>%
  filter(!is.na(palabra)) %>%
  count(palabra, sort = TRUE) %>%
  separate(palabra,
           into = c("palabra_1", "palabra_2"),
           sep = " ") %>%
  filter(!palabra_1 %in% stopwords_es$palabra) %>%
  filter(!palabra_2 %in% stopwords_es$palabra) %>%
  filter(!palabra_1 %in% my_stopwords$palabra) %>%
  filter(!palabra_2 %in% my_stopwords$palabra) %>%
  filter(!palabra_1 %in% bi_stopwords$palabra) %>%
  filter(!palabra_2 %in% bi_stopwords$palabra) %>%
  mutate(palabra = paste(palabra_1, palabra_2, sep = " "), .before = n) %>%
  dplyr::select(-c(palabra_1, palabra_2))
head(bigram_2021)
```

```
## # A tibble: 6 x 2
##   palabra                            n
##   <chr>                          <int>
## 1 salud mental                      12
## 2 web institucional                  5
## 3 ejecución presupuestaria           4
## 4 establecimientos penitenciarios    4
## 5 participación ciudadana            4
## 6 personas privadas                  4
```

```
bigram_2020 <- speech_2020 %>%
  tibble(speech = speech_2020) %>%
  unnest_tokens(input = speech,
```

```
                output = palabra,
                token = "ngrams",
                n = 2) %>%
  filter(!is.na(palabra)) %>%
  count(palabra, sort = TRUE) %>%
  separate(palabra,
           into = c("palabra_1", "palabra_2"),
           sep = " ") %>%
  filter(!palabra_1 %in% stopwords_es$palabra) %>%
  filter(!palabra_2 %in% stopwords_es$palabra) %>%
  filter(!palabra_1 %in% my_stopwords$palabra) %>%
  filter(!palabra_2 %in% my_stopwords$palabra) %>%
  filter(!palabra_1 %in% bi_stopwords$palabra) %>%
  filter(!palabra_2 %in% bi_stopwords$palabra) %>%
  mutate(palabra = paste(palabra_1, palabra_2, sep = " "), .before = n) %>%
  dplyr::select(-c(palabra_1, palabra_2))
head(bigram_2020)
```

```
## # A tibble: 6 x 2
##   palabra                            n
##   <chr>                          <int>
## 1 establecimientos penitenciarios    33
## 2 covid 19                           22
## 3 derechos humanos                   22
## 4 personas privadas                  17
## 5 visitas virtuales                  13
## 6 reinserción social                 12
```

```
bigram_2019 <- speech_2019 %>%
  tibble(speech = speech_2019) %>%
  unnest_tokens(input = speech,
                output = palabra,
                token = "ngrams",
                n = 2) %>%
  filter(!is.na(palabra)) %>%
  count(palabra, sort = TRUE) %>%
  separate(palabra,
           into = c("palabra_1", "palabra_2"),
           sep = " ") %>%
  filter(!palabra_1 %in% stopwords_es$palabra) %>%
  filter(!palabra_2 %in% stopwords_es$palabra) %>%
  filter(!palabra_1 %in% my_stopwords$palabra) %>%
  filter(!palabra_2 %in% my_stopwords$palabra) %>%
  filter(!palabra_1 %in% bi_stopwords$palabra) %>%
  filter(!palabra_2 %in% bi_stopwords$palabra) %>%
  mutate(palabra = paste(palabra_1, palabra_2, sep = " "), .before = n) %>%
  dplyr::select(-c(palabra_1, palabra_2))
head(bigram_2019)
```

```
## # A tibble: 6 x 2
##   palabra                          n
##   <chr>                        <int>
```

```
## 1 establecimientos penitenciarios    46
## 2 derechos humanos                   34
## 3 población penal                    25
## 4 reinserción social                 25
## 5 nivel nacional                     15
## 6 personas privadas                  14
```

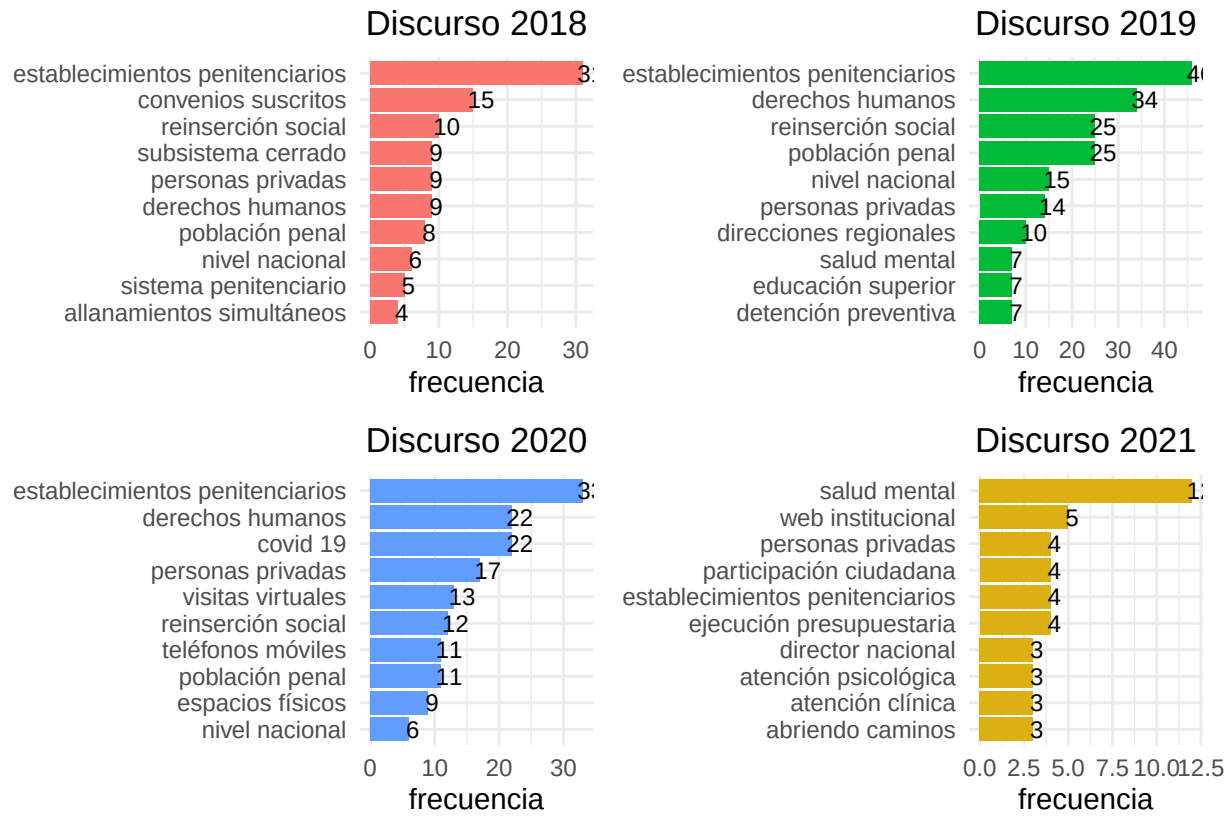
```
bigram_2018 <- speech_2018 %>%
  tibble(speech = speech_2018) %>%
  unnest_tokens(input = speech,
                output = palabra,
                token = "ngrams",
                n = 2) %>%
  filter(!is.na(palabra)) %>%
  count(palabra, sort = TRUE) %>%
  separate(palabra,
           into = c("palabra_1", "palabra_2"),
           sep = " ") %>%
  filter(!palabra_1 %in% stopwords_es$palabra) %>%
  filter(!palabra_2 %in% stopwords_es$palabra) %>%
  filter(!palabra_1 %in% my_stopwords$palabra) %>%
  filter(!palabra_2 %in% my_stopwords$palabra) %>%
  filter(!palabra_1 %in% bi_stopwords$palabra) %>%
  filter(!palabra_2 %in% bi_stopwords$palabra) %>%
  mutate(palabra = paste(palabra_1, palabra_2, sep = " "), .before = n) %>%
  dplyr::select(-c(palabra_1, palabra_2))
head(bigram_2018)
```

```
## # A tibble: 6 x 2
##   palabra                            n
##   <chr>                          <int>
## 1 establecimientos penitenciarios    31
## 2 convenios suscritos                15
## 3 reinserción social                 10
## 4 derechos humanos                    9
## 5 personas privadas                   9
## 6 subsistema cerrado                  9
```

Y graficamos los bigramas más frecuentes.

```
fig_2021 <- bigram_2021 %>%
  slice_head(n = 10) %>%
  ggplot(aes(y = reorder(palabra, n), n)) +
  geom_col(fill = "#dbb012") +
  geom_text(aes(label = n), size = 3, hjust = 0.2) +
  theme_minimal() +
  labs(y = NULL, x = "frecuencia") +
  ggtitle("Discurso 2021") +
  theme(plot.title = element_text(hjust = 0.5))

fig_2020 <- bigram_2020 %>%
  slice_head(n = 10) %>%
  ggplot(aes(y = reorder(palabra, n), n)) +
```

```
  geom_col(fill = "#619cff") +
  geom_text(aes(label = n), size = 3, hjust = 0.2) +
  theme_minimal() +
  labs(y = NULL, x = "frecuencia") +
  ggtitle("Discurso 2020") +
  theme(plot.title = element_text(hjust = 0.5))

fig_2019 <- bigram_2019 %>%
  slice_head(n = 10) %>%
  ggplot(aes(y = reorder(palabra, n), n)) +
  geom_col(fill = "#00ba38") +
  geom_text(aes(label = n), size = 3, hjust = 0.2) +
  theme_minimal() +
  labs(y = NULL, x = "frecuencia") +
  ggtitle("Discurso 2019") +
  theme(plot.title = element_text(hjust = 0.5))

fig_2018 <- bigram_2018 %>%
  slice_head(n = 10) %>%
  ggplot(aes(y = reorder(palabra, n), n)) +
  geom_col(fill = "#f8766d") +
  geom_text(aes(label = n), size = 3, hjust = 0.2) +
  theme_minimal() +
  labs(y = NULL, x = "frecuencia") +
  ggtitle("Discurso 2018") +
  theme(plot.title = element_text(hjust = 0.5))

(fig_2018 + fig_2019) / (fig_2020 + fig_2021)
```



Análisis TF-IDF

Juntamos las frecuencias de los bigramas en un solo data frame, identificando a qué discurso pertenece.

```
frequencies_2021 <- bigram_2021 %>%
  mutate(discurso = "C.P.P. 2021", .before = palabra)
frequencies_2020 <- bigram_2020 %>%
  mutate(discurso = "C.P.P. 2020", .before = palabra)
frequencies_2019 <- bigram_2019 %>%
  mutate(discurso = "C.P.P. 2019", .before = palabra)
frequencies_2018 <- bigram_2018 %>%
  mutate(discurso = "C.P.P. 2018", .before = palabra)

messages <- bind_rows(frequencies_2018, frequencies_2019, frequencies_2020, frequencies_2021)
head(messages)
```

```
## # A tibble: 6 x 3
##   discurso    palabra                          n
##   <chr>       <chr>                        <int>
## 1 C.P.P. 2018 establecimientos penitenciarios    31
## 2 C.P.P. 2018 convenios suscritos                15
## 3 C.P.P. 2018 reinserción social                 10
## 4 C.P.P. 2018 derechos humanos                    9
## 5 C.P.P. 2018 personas privadas                   9
## 6 C.P.P. 2018 subsistema cerrado                  9
```

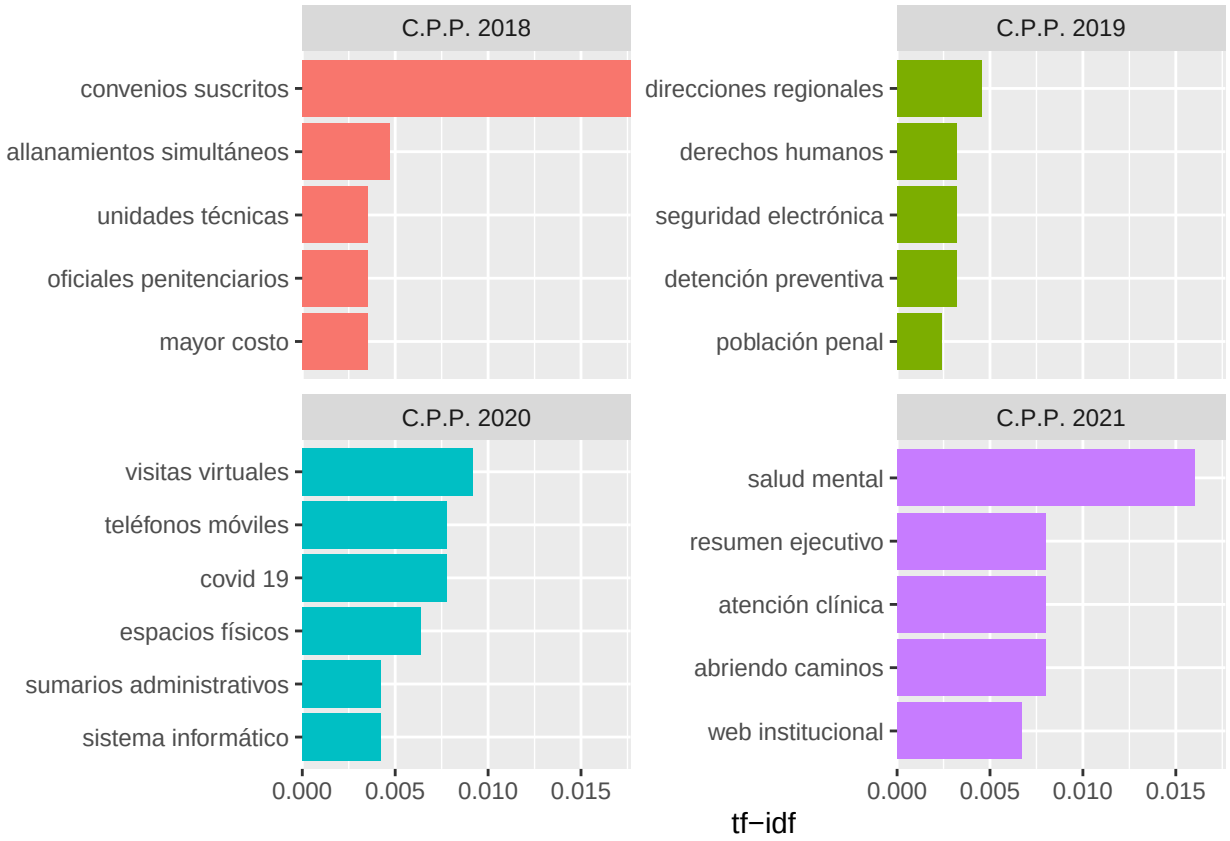
Calculamos las frecuencias inversas de los bigramas.

```
messages_tfidf <- bind_tf_idf(messages,
                              term = palabra,
                              document = discurso,
                              n = n)
head(messages_tfidf)
```

```
## # A tibble: 6 x 6
##   discurso    palabra                              n      tf   idf  tf_idf
##   <chr>       <chr>                            <int>   <dbl> <dbl>   <dbl>
## 1 C.P.P. 2018 establecimientos penitenciarios     31 0.0263  0     0
## 2 C.P.P. 2018 convenios suscritos                 15 0.0127  1.39  0.0177
## 3 C.P.P. 2018 reinserción social                  10 0.00850 0     0
## 4 C.P.P. 2018 derechos humanos                     9 0.00765 0.288 0.00220
## 5 C.P.P. 2018 personas privadas                    9 0.00765 0     0
## 6 C.P.P. 2018 subsistema cerrado                   9 0.00765 0.288 0.00220
```

Y hacemos la comparación gráfica.

```
two_words_plot <-
  messages_tfidf %>%
    group_by(discurso) %>%
    top_n(5) %>%
    ungroup %>%
    mutate(discurso = as.factor(discurso),
           palabra = reorder_within(palabra, tf_idf, discurso)) %>%
    ggplot(aes(palabra, tf_idf, fill = discurso)) +
    geom_col(show.legend = FALSE) +
    facet_wrap(~discurso, scales = "free_y") +
    coord_flip() +
    scale_x_reordered() +
    scale_y_continuous(expand = c(0,0)) +
    labs(y = "tf-idf", x = NULL)
two_words_plot
```



El discurso de la Cuenta Pública Participativa (C.P.P.) del año 2018 destacó la ejecución de actividades mediante *convenios suscritos* en 2017 con diversas entidades públicas, corporaciones, fundaciones y universidades, tanto a nivel central como regional, gestionadas e implementadas por las distintas *unidades técnicas* regionales y de los establecimientos penitenciarios. Durante este año se desarrollaron *allanamientos simultáneos* en diversas regiones, permitiendo incautar variadas especies ilícitas y prohibidas.

En el discurso del 2019, se enfatizó la coordinación de las *direcciones regionales* con las unidades de los subsistemas cerrado, abierto y postpenitenciario, con el fin de realizar actividades conjuntas para la vinculación público-privada. Otras de las actividades realizadas en 2018 estuvieron en el ámbito de los *Derechos Humanos* a través de capacitación a funcionarios y actualización de resoluciones y actos administrativos. También hubo ejecución de proyectos para implementar, ampliar, reponer y mantener equipos de *seguridad electrónica*, con el objetivo de complementar la vigilancia perimetral humana.

El análisis de bigramas destaca nuevamente que el discurso de la C.P.P. 2020 se centró en la situación pandémica por *covid 19* y en algunas las acciones tomadas a consecuencia de ella, tales como la implementación de salas para *visitas virtuales*, permitir el uso limitado de *teléfonos móviles* para la comunicación de los internos con sus familiares y la reubicación de internos para disminuir la probabilidad de contagios por falta de *espacios físicos* y su sanitización.

Finalmente, en el discurso del 2021 se abordaron las actividades realizadas en 2020 de *atención clínica* y derivación de pacientes por *salud mental* en contexto de pandemia. Por otro lado, se destacó la participación de la ciudadanía en la estructura y contenido del *resumen ejecutivo* de la C.P.P., y su disponibilidad para descargar desde la *web institucional*. También se respondió una consulta ciudadana sobre el programa *Abriendo Caminos* del Ministerio de Desarrollo Social.

– *Nota: a la fecha de publicación de este proyecto la C.P.P. 2022 respecto de la gestión 2021 aún se encontraba en desarrollo por parte de Gendarmería de Chile.*