Content Analysis of Scaling Up Nutrition (SUN) Movement Progress Reports from 2011-2017

Ernest Guevarra

17/05/2018

Re-structure dataset into a one-token-per-row format

```
tidy_reports <- progress_reports() %>%
  unnest_tokens(word, text)

tidy_reports
```

```
## # A tibble: 354,209 x 5
##     page linenumber chapter  year word
##    <int>      <int>   <int> <int> <chr>
##  1     3          1       0  2011 preface
##  2     3          2       0  2011 one
##  3     3          2       0  2011 year
##  4     3          2       0  2011 ago
##  5     3          2       0  2011 i
##  6     3          2       0  2011 joined
##  7     3          2       0  2011 a
##  8     3          2       0  2011 group
##  9     3          2       0  2011 of
## 10     3          2       0  2011 leaders
## # ... with 354,199 more rows
```

Remove stop words - words not useful in analysis

```
data(stop_words)

tidy_reports <- tidy_reports %>%
  anti_join(stop_words)
```

```
## Joining, by = "word"
```

Find the most common words in all the reports as a whole

```
tidy_reports %>%
  count(word, sort = TRUE)
```

```
## # A tibble: 12,659 x 2
##    word          n
```

```
##    <chr>     <int>
##  1 nutrition  5932
##  2 sun        3050
##  3 countries  2302
##  4 national   1914
##  5 movement   1843
##  6 country    1552
##  7 2015       1377
##  8 progress   1308
##  9 2016       1199
## 10 2014       1174
## # ... with 12,649 more rows
```

Visualise the most common words

```
tidy_reports %>%
  count(word, sort = TRUE) %>%
  filter(n > 600) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  xlab(NULL) +
  coord_flip()
```