Handout 11: NLP with cleanNLP

Taylor Arnold

Loading and parsing the data

The full text of all the State of the Union addresses through 2016 are available in the R package **sotu**, available on CRAN. The package also contains meta-data concerning each speech that we will add to the document table while annotating the corpus. The code to run this annotation is given by:

```
library(sotu)
library(cleanNLP)

data(sotu_text)
data(sotu_meta)
init_spaCy()
sotu <- cleanNLP::run_annotators(sotu_text, as_strings = TRUE,
                                 meta = sotu_meta)
```

The annotation object, which we will use in the example in the following analysis, is stored in the object `sotu`. We will create a single data frame with all of the tokens here and save the metadata as a seperate data set:

```
tokens <- cleanNLP::get_token(sotu, combine = TRUE)
tokens
```

```
## # A tibble: 2,181,493 x 14
##       id   sid   tid            word
##    <int> <int> <int>           <chr>
##  1     1     1     1          Fellow
##  2     1     1     2               -
##  3     1     1     3        Citizens
##  4     1     1     4              of
##  5     1     1     5             the
##  6     1     1     6          Senate
##  7     1     1     7             and
##  8     1     1     8           House
##  9     1     1     9              of
## 10     1     1    10 Representatives
## # ... with 2,181,483 more rows, and 10 more
## #   variables: lemma <chr>, upos <chr>,
## #   pos <chr>, cid <int>, source <int>,
## #   relation <chr>, word_source <chr>,
## #   lemma_source <chr>, entity_type <chr>,
## #   entity <chr>
```

```
doc <- cleanNLP::get_document(sotu)
```

Exploratory analysis

Simple summary statistics are easily computed off of the token table. To see the distribution of sentence length, the token table is grouped by the document and sentence id and the number of rows within each group are computed. The percentiles of these counts give a quick summary of the distribution.

```
library(ggplot2)
library(dplyr)
temp <- count(tokens, id, sid)
quantile(temp$n, seq(0,1,0.1))
```

```
##   0%  10%  20%  30%  40%  50%  60%  70%  80%
##    1   11   16   19   23   27   31   37   44
##  90% 100%
##   58  681
```

The median sentence has 28 tokens, whereas at least one has over 600 (this is due to a bulleted list in one of the written addresses being treated as a single sentence) To see the most frequently used nouns in the dataset, the token table is filtered on the universal part of speech field, grouped by lemma, and the number of rows in each group are once again calculated. Sorting the output and selecting the top 42 nouns, yields a high level summary of the topics of interest within this corpus.

```
temp <- filter(tokens, upos == "NOUN")
temp <- count(temp, lemma)
temp <- top_n(temp, n = 42, n)
arrange(temp, desc(n))$lemma
```

```
##  [1] "year"          "country"
##  [3] "people"        "government"
##  [5] "law"           "time"
##  [7] "nation"        "who"
##  [9] "power"         "interest"
## [11] "world"         "war"
## [13] "citizen"       "service"
## [15] "duty"          "part"
## [17] "system"        "peace"
## [19] "right"         "man"
## [21] "program"       "policy"
## [23] "work"          "act"
## [25] "state"         "condition"
## [27] "subject"       "legislation"
## [29] "force"         "effort"
## [31] "treaty"        "purpose"
## [33] "what"          "land"
## [35] "business"      "action"
## [37] "measure"       "tax"
## [39] "way"           "question"
## [41] "relation"      "consideration"
```

The result is generally as would be expected from a corpus of government speeches, with references to proper nouns representing various organizations within the government and non-proper nouns indicating general topics of interest such as `tax'',law`'', and “peace”.

The length in tokens of each address is calculated similarly by grouping and summarizing at the document id level. The results can be joined with the document table to get the year of the speech and then piped in a **ggplot2** command to illustrate how the length of the State of the Union has changed over time.

```
doc <- cleanNLP::get_document(sotu)
temp <- left_join(count(tokens, id), doc)

qplot(year, n, data = temp, color = sotu_type) +
  geom_line() +
  geom_smooth()
```

Here, color is used to represent whether the address was given as an oral address or a written document. The output shows that their are certainly time trends to the address length, with the form of the address (written versus spoken) also having a large effect on document length.

Finding the most used entities from the entity table over the time period of the corpus yields an alternative way to see the underlying topics. A slightly modified version of the code snippet used to find the top nouns in the dataset can be used to find the top entities. The `get_token` function is replaced by `get_entity` and the table is filtered on `entity_type` rather than the universal part of speech code.

```
temp <- filter(tokens, entity_type == "GPE")
temp <- count(temp, entity)
temp <- top_n(temp, n = 26, n)
arrange(temp, desc(n))$entity
```

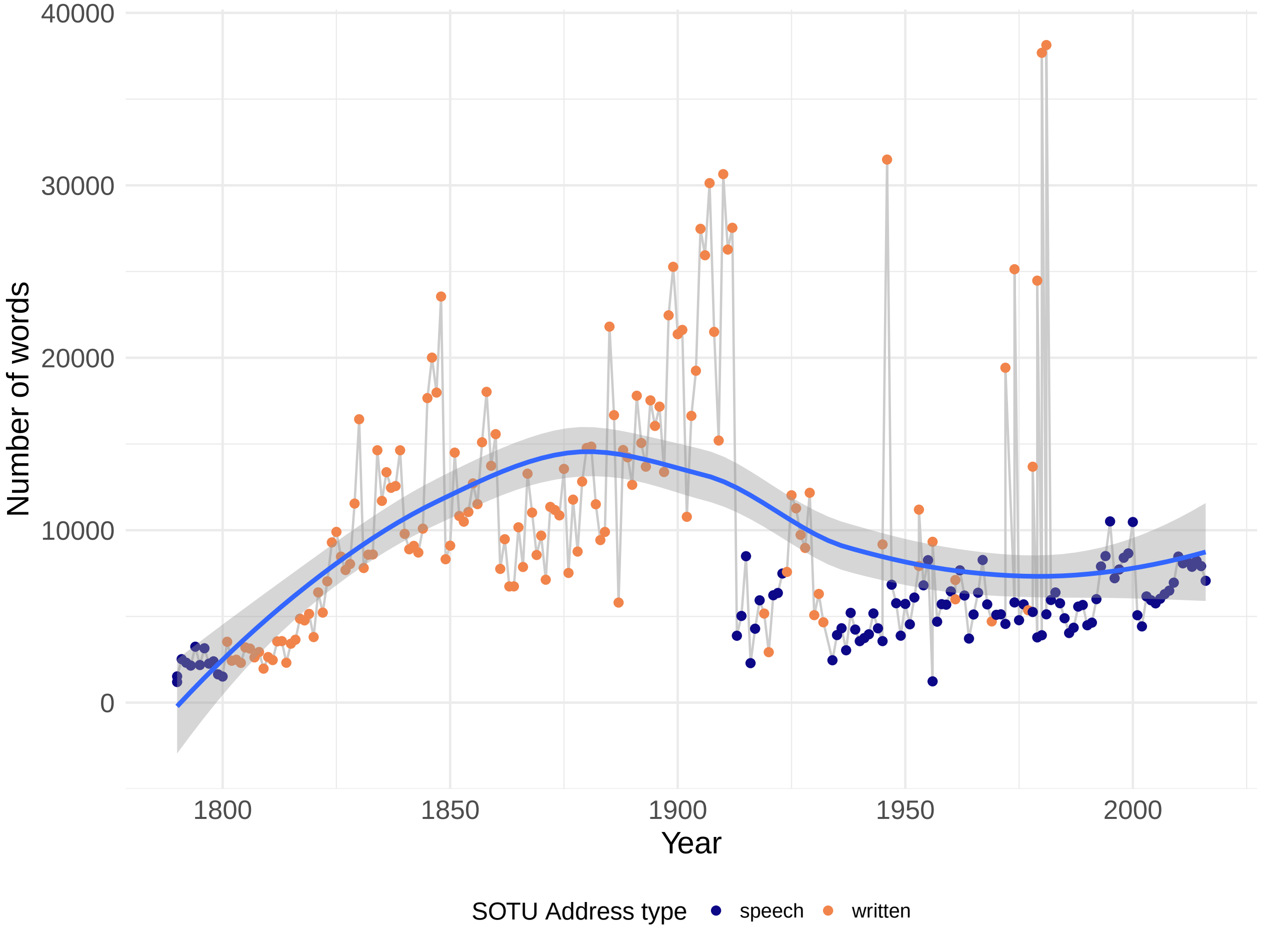

Figure 1: Length of each State of the Union address, in total number of tokens. Color shows whether the address was given as a speech or delivered as a written document.

```
##  [1] "the United States"
##  [2] "America"
##  [3] "States"
##  [4] "Mexico"
##  [5] "Great Britain"
##  [6] "Spain"
##  [7] "Washington"
##  [8] "China"
##  [9] "Executive"
## [10] "France"
## [11] "Cuba"
## [12] "Japan"
## [13] "Texas"
## [14] "Russia"
## [15] "The United States"
## [16] "Germany"
## [17] "United States"
## [18] "California"
## [19] "Nicaragua"
## [20] "the Soviet Union"
## [21] "Mississippi"
## [22] "Iraq"
## [23] "Alaska"
## [24] "U.S."
## [25] "Philippines"
## [26] "Panama"
## [27] "the District of Columbia"
```

The ability to redo analyses from a slightly different perspective is a direct consequence of the tidy data model supplied by **cleanNLP**. The top locations include some obvious and some less obvious instances. Those sovereign nations included such as Great Britain, Mexico, Germany, and Japan seem as expected given either the United State's close ties or periods of war with them. The top states include the most populous regions (New York, California, and Texas) but also smaller states (Kansas, Oregon, Mississippi), the latter being more surprising.

One of the most straightforward way of extracting a high-level summary of the content of a speech is to extract all direct object object dependencies where the target noun is not a very common word. In order to do this for a particular speech, the dependency table is joined to the document table, a particular document is selected, and relationships of type '`dobj'' (direct object) are filtered out. The result is then joined to the data set`word_frequency', which is included with **cleanNLP**, and pairs with a target occurring less than

0.5% of the time are selected to give the final result. Here is an example of this using the first address made by George W. Bush in 2001:

```
temp <- left_join(tokens, doc)
temp <- filter(temp, year == 2001, relation == "dobj")
temp <- select(temp, id = id, start = word, word = lemma_source)
temp <- left_join(temp, word_frequency)
temp <- filter(temp, frequency < 0.001)
temp <- select(temp, id, start, word)
sprintf("%s => %s", temp$start, temp$word)
```

```
##  [1] "country => govern"
##  [2] "world => revolutionize"
##  [3] "other => repaint"
##  [4] "teachers => recruit"
##  [5] "one => marry"
##  [6] "Americans => recruit"
##  [7] "dozens => streamline"
##  [8] "billion => dedicate"
##  [9] "support => owe"
## [10] "cleanup => accelerate"
## [11] "contributions => deduct"
## [12] "work => hinder"
## [13] "it => owe"
## [14] "Code => simplify"
## [15] "marriage => discourage"
## [16] "tax => repeal"
## [17] "work => punish"
## [18] "it => punish"
## [19] "recovery => stimulate"
## [20] "defenses => restructure"
## [21] "challenges => confront"
## [22] "threats => confront"
## [23] "intent => threaten"
## [24] "defenses => deploy"
## [25] "relics => discard"
## [26] "agreements => negotiate"
## [27] "shortage => confront"
## [28] "challenge => confront"
## [29] "election => enact"
```

Most of these phrases correspond with the “compassionate conservatism" that George W. Bush ran under in the preceding 2000 election. Applying the same analysis to the 2002 State of the Union, which came under the shadow of the September 11th terrorist attacks, shows a drastic shift in focus.

```
temp <- left_join(tokens, doc)
temp <- filter(temp, year == 2002, relation == "dobj")
temp <- select(temp, id = id, start = word, word = lemma_source)
temp <- left_join(temp, word_frequency)
temp <- filter(temp, frequency < 0.0005)
temp <- select(temp, id, start, word)
sprintf("%s => %s", temp$start, temp$word)
```

```
##  [1] "Afghanistan => occupy"
##  [2] "cells => occupy"
##  [3] "planes => hijack"
##  [4] "plans => disrupt"
##  [5] "call => heed"
##  [6] "leadership => admire"
##  [7] "citizens => starve"
##  [8] "hope => repress"
##  [9] "hostility => flaunt"
## [10] "States => blackmail"
## [11] "economy => revive"
## [12] "unity => applaud"
## [13] "man => subdue"
## [14] "budget => revitalize"
## [15] "spending => restrain"
## [16] "reforms => reauthorize"
## [17] "bill => enact"
## [18] "safeguards => enact"
## [19] "volunteers => mobilize"
## [20] "knock => await"
## [21] "fall => greet"
## [22] "it => affirm"
```

Here the topics have almost entirely shifted to counter-terrorism and national security efforts.