Handout 12: Textual models

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 × 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)
```

Models

The `get_tfidf` function provided by **cleanNLP** converts a token table into a sparse matrix representing the term-frequency inverse document frequency matrix (or any intermediate part of that calculation). This is particularly useful when building models from a textual corpus. The `tidy_pca`, also included with the package, takes a matrix and returns a data frame containing the desired number of principal components. Dimension reduction involves piping the token table for a corpus into the `get_tfidf` function and passing the results to `tidy_pca`.

```
temp <- filter(tokens, pos %in% c("NN", "NNS"))
tfidf <- get_tfidf(temp, min_df = 0.05, max_df = 0.95,
                       type = "tfidf", tf_weight = "dnorm")$tfidf
pca <- tidy_pca(tfidf, get_document(sotu))
select(pca, year, PC1, PC2)
```

In this example only non-proper nouns have been included in order to minimize the stylistic attributes of the speeches in order to focus more on their content. A scatter plot of the speeches using these components is shown. There is a definitive temporal pattern to the documents, with the 20th century addresses forming a distinct cluster on the right side of the plot.

The output of the `get_tfidf` function may be given directly to the `LDA` function in the package **topicmodels**. The topic model function requires raw counts, so the type variable in `get_tfidf` is set to 'tf''; the results may then be directly piped toLDA'.

```
library(topicmodels)
tm <- cleanNLP::get_token(sotu) %>%

temp <- filter(tokens, pos %in% c("NN", "NNS"))
temp <- get_tfidf(temp, min_df = 0.05, max_df = 0.95,
                       type = "tf", tf_weight = "raw")
LDA(temp$tf, k = 16, control = list(verbose = 1))
```

The topics, ordered by approximate time period, are visualized in the Figure. Most topics persist for a few decades and then largely disappear, though some persist over non-contiguous periods of the presidency. The Energy topic, for example, appears during the 1950s

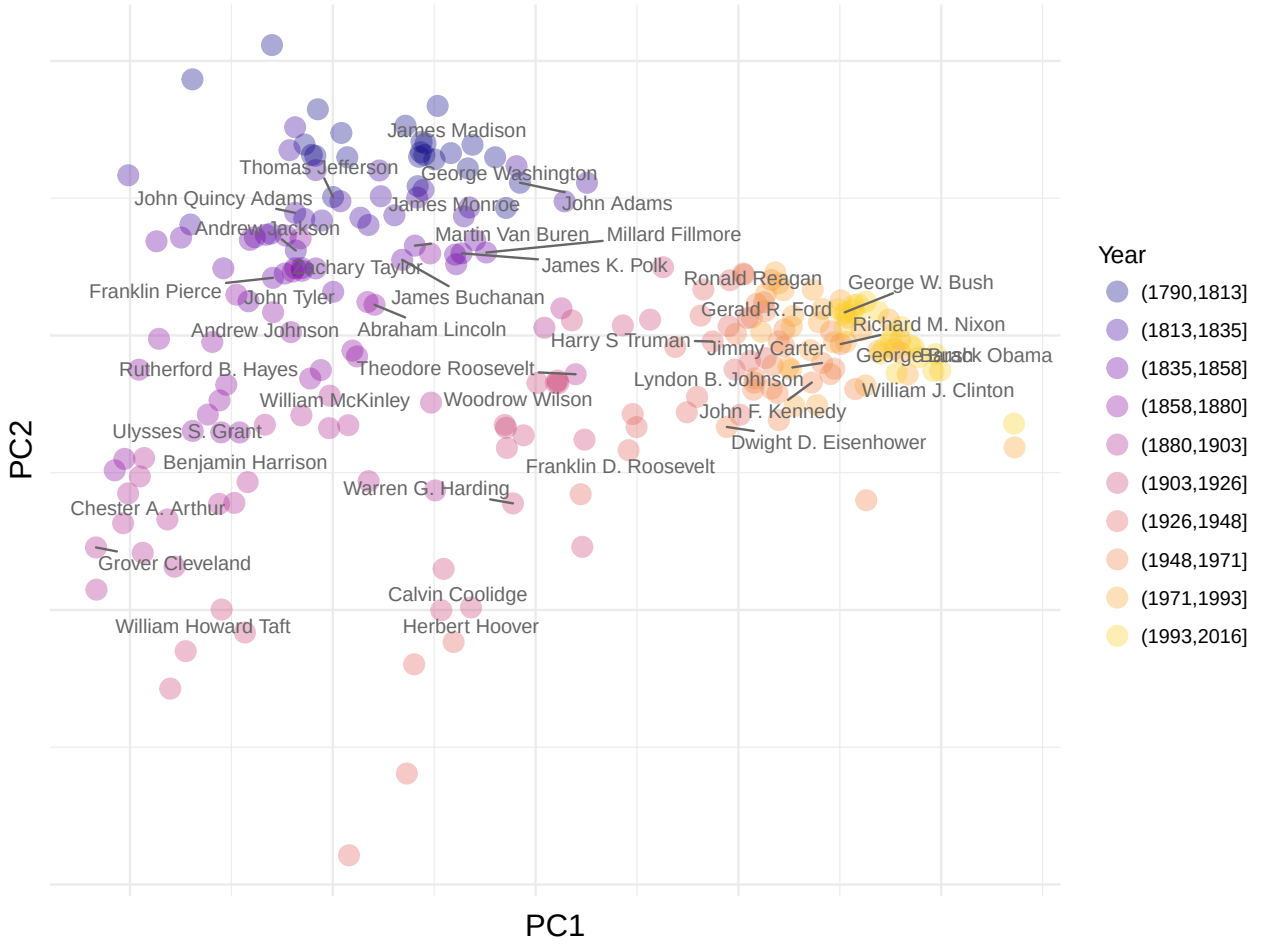

Figure 1: State of the Union Speeches, highlighting each President's first address, plotted using the first two principal components of the term frequency matrix of non-proper nouns.

and crops up again during the energy crisis of the 1970s. The "world, man, freedom, force, defense" topic peaks during both World Wars, but is absent during the 1920s and early 1930s.

Finally, the **cleanNLP** data model is also convenient for building predictive models. The State of the Union corpus does not lend itself to an obviously applicable prediction problem. A classifier that distinguishes speeches made by George W. Bush and Barrack Obama will be constructed here for the purpose of illustration. As a first step, a term-frequency matrix is extracted using the same technique as was used with the topic modeling function. However, here the frequency is computed for each sentence in the corpus rather than the document as a whole. The ability to do this seamlessly with a single additional `mutate` function defining a new id illustrates the flexibility of the `get_tfidf` function.

```
library(Matrix)
temp <- left_join(tokens, doc)
temp <- filter(temp, year > 2000)
temp <- filter(temp, pos %in% c("NN", "NNS"))
temp <- mutate(temp, new_id = paste(id, sid, sep = "-"))
mat <- get_tfidf(temp, min_df = 0, max_df = 1, type = "tf",
                 tf_weight = "raw", doc_var = "new_id")
```

It will be nessisary to define a response variable `y` indicating whether this is a speech made by President Obama as well as a training flag indicating which speeches were made in odd numbered years. This is done via a separate table join and a pair of mutations.

```
meta <- left_join(data_frame(new_id = mat$id),
                  temp[!duplicated(temp$new_id),])
meta <- mutate(meta, y = as.numeric(president == "Barack Obama"))
meta <- mutate(meta, train = year %in% seq(2001,2016, by = 2))
```

The output may now be used as input to the elastic net function provided by the **glmnet** package. The response is set to the binomial family given the binary nature of the response and training is done on only those speeches occurring in odd-numbered years. Cross-validation is used in order to select the best value of the model's tuning parameter.

```
library(glmnet)
model <- cv.glmnet(mat$tf[meta$train,], meta$y[meta$train], family = "binomial")
```

The algorithm does a very good job of separating the speeches. Looking at the odd years versus even years (the training and testing sets, respectively) indicates that the model has not been over-fit.

One benefit of the penalized linear regression model is that it is possible to interpret the coefficients in a meaningful way. Here are the

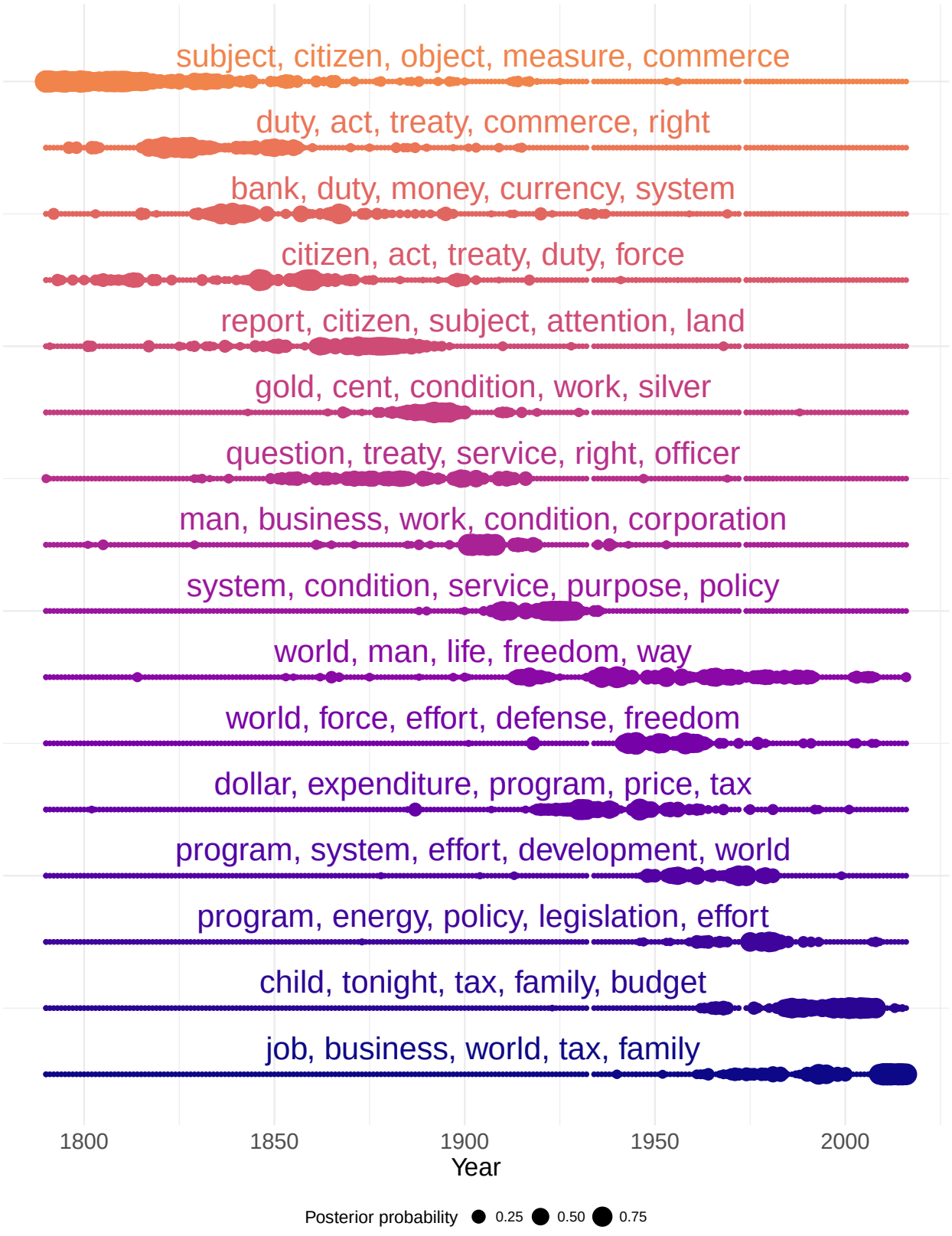

Figure 2: Distribution of topic model posterior probabilities over time on the State of the Union corpus. Top five words associated with each topic are displayed, with topics sorted by the median year of all documents placed into the respective topic using the maximum posterior probabilities.

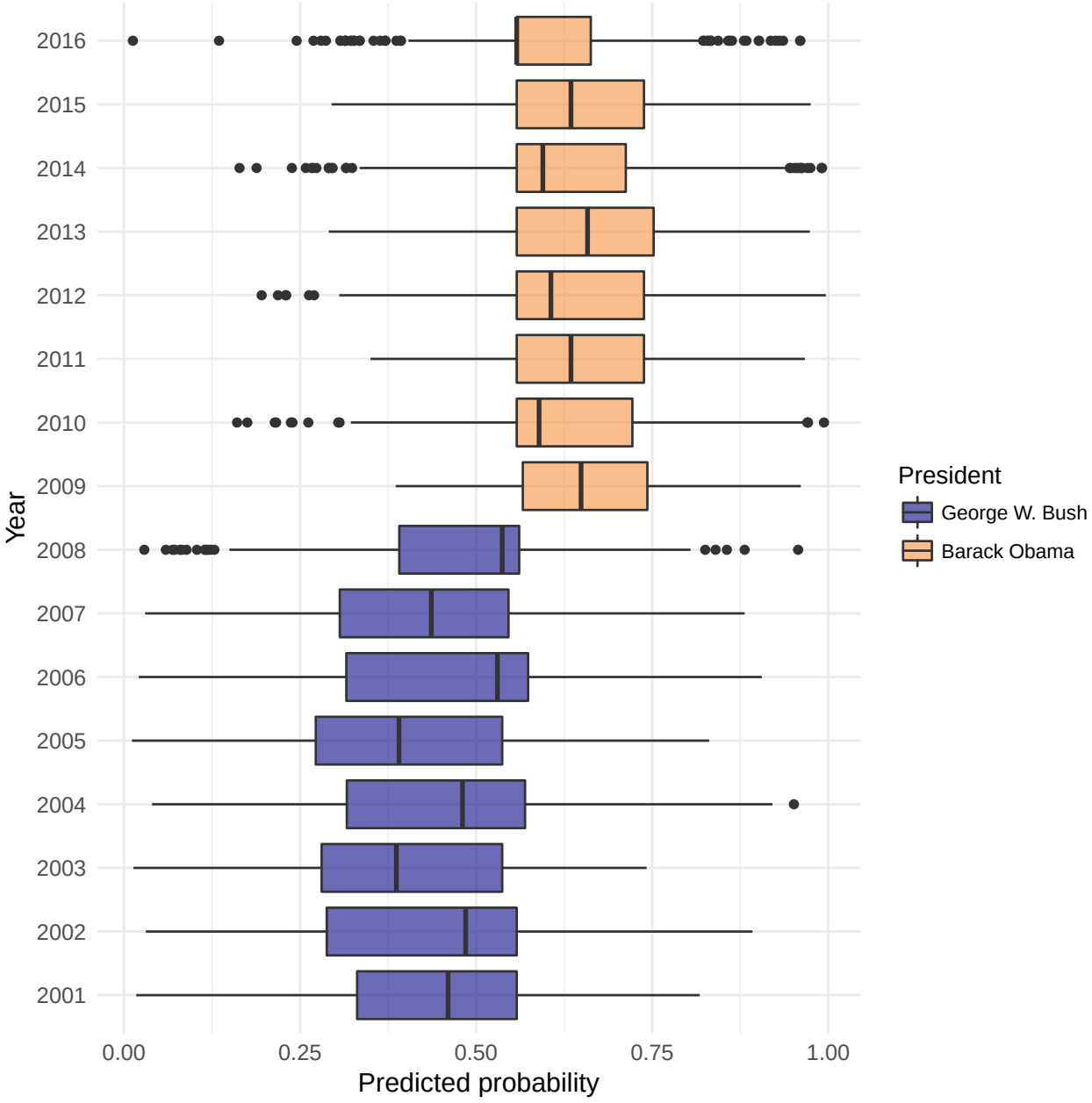

Figure 3: Boxplot of predicted probabilities, at the sentence level, for all 16 State of the Union addresses by Presidents George W. Bush and Barack Obama. The probability represents the extent to which the model believe the sentence was spoken by President Obama. Odd years were used for training and even years for testing. Cross-validation on the training set was used, with the one standard error rule, to set the lambda tuning parameter.

non-zero elements of the regression vector, coded as whether the have a positive (more Obama) or negative (more Bush) sign:

```
beta <- coef(model, s = model[["lambda"]][11])[-1]
sprintf("%s (%d)", mat$vocab, sign(beta))[beta != 0]
```

```
 [1] "job (1)"          "business (1)"     "citizen (-1)"
 [4] "terrorist (-1)"   "government (-1)"  "freedom (-1)"
 [7] "home (1)"         "college (1)"      "weapon (-1)"
[10] "deficit (1)"      "company (1)"      "peace (-1)"
[13] "enemy (-1)"       "terror (-1)"      "income (-1)"
[16] "drug (-1)"        "kid (1)"          "regime (-1)"
[19] "class (1)"        "crisis (1)"       "industry (1)"
[22] "need (-1)"        "fact (1)"         "relief (-1)"
[25] "bank (1)"         "liberty (-1)"     "society (-1)"
[28] "duty (-1)"        "folk (1)"         "account (-1)"
[31] "compassion (-1)"  "environment (-1)" "inspector (-1)"
```

These generally seem as expected given the main policy topics of focus under each administration. During most of the Bush presidency, as mentioned before, the focus was on national security and foreign policy. Obama, on the other hand, inherited the recession of 2008 and was far more focused on the overall economic policy.