Text as Data Final Paper

Jacqueline Gutman, Alex Pine

May 10, 2016

```
# LDA model for debate: debate_LDA_15
# topic names for debate: debate_LDA_15_names
# data frame for twitter: twitter.df
# dfm for twitter: twitter_dfm
# posterior topic distribution (LDA) = @gamma
# LDA model for twitter: use simple_lda_20, simple_lda_15, simple_lda_10
# LDA posterior for twitter using debate topics: twitter.topics$topics
all(nrow(twitter_dfm) == sum(twitter.df$debate_topic != 0),
    nrow(twitter_dfm)  == nrow(twitter.topics$topics))
```

```
[1] TRUE
```

```
table(twitter.df$debate_topic)
```

```
   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14
2940  525  937  430  533 1387  618  490  617  622  673  406 2340  480  415
  15
 456
```

```
debate_LDA_15_names
```

```
 [1] "common core"  "mods1"        "foreign pol"  "social sec"
 [5] "mods1"        "immigration"  "economics"    "border"
 [9] "budget"       "Paul Ryan"    "military"     "mods2"
[13] "gen election" "iran"         "marriage"
```

```
# pos.neg <- dplyr::select(twitter.df[tweet_indices,], -tweet_created)
# pos.neg <- filter(pos.neg, sentiment != "Neutral")
# pos.neg$candidate[pos.neg$candidate == "OTHER"] <- NA
# pos.neg$subject_matter[pos.neg$subject_matter == "None of the above"] <- NA
# pos.neg <- droplevels(pos.neg)
levels(pos.neg$sentiment)
```

```
[1] "Negative" "Positive"
```

```
levels(pos.neg$candidate)
```

```
 [1] "Ben Carson"     "Chris Christie" "Donald Trump"   "Jeb Bush"
 [5] "John Kasich"    "Marco Rubio"    "Mike Huckabee"  "Rand Paul"
 [9] "Scott Walker"   "Ted Cruz"
```

```
levels(pos.neg$subject_matter)
```

```
 [1] "Abortion"
 [2] "Foreign Policy"
 [3] "FOX News or Moderators"
 [4] "Gun Control"
 [5] "Healthcare (including Medicare)"
 [6] "Immigration"
 [7] "Jobs and Economy"
 [8] "LGBT issues"
 [9] "Racial issues"
[10] "Religion"
[11] "Women's Issues (not abortion though)"
```

```
dropped.rows <- which(twitter.df[tweet_indices, "sentiment"] == "Neutral")
nrow(pos.neg) + length(dropped.rows) == nrow(twitter.topics$topics)
```

```
[1] TRUE
```

```
all(dim(simple_lda_15@gamma) == dim(twitter.topics$topics),
    class(simple_lda_15@gamma) == class(twitter.topics$topics))
```

```
[1] TRUE
```

```
dim(simple_lda_25@gamma[-dropped.rows,])
```

```
[1] 8722   25
```

```
dim(twitter.topics$topics[-dropped.rows,])
```

```
[1] 8722   15
```

```
all(abs(rowSums(simple_lda_25@gamma) - 1) < 1e-10)
```

```
[1] TRUE
```

```
all(abs(rowSums(twitter.topics$topics) - 1) < 1e-10)
```

```
[1] TRUE
```

```
# build a logistic regression from lda model parameters, additional predictors as parameter
glm_lda_model <- function(lda_model_post, modified_data,
                          predictors = c("candidate", "subject_matter")) {
    x <- lda_model_post[,-2] # need to drop one of the topics, I drop #2
    colnames(x) <- paste("topic", 1:(ncol(x)+1), sep=".")[-2]
    data <- cbind(modified_data, x)
    formula <- paste("sentiment ~ ",
                     paste(c(colnames(x), predictors), collapse = " + "))
```

```
    fit <- glm(as.formula(formula) , data = data, family = "binomial")
    print(summary(fit))
    fit
}

# use forward-backward stepwise procedure with AIC criterion to choose best model from full model
stepwise_twitter <- function(lda_model_post, modified_data,
                        predictors = c("candidate", "subject_matter")) {
  x <- lda_model_post # don't drop any topics
  colnames(x) <- paste("topic", 1:(ncol(x)), sep=".")
  data <- cbind(modified_data, x)
  formula <- paste("sentiment ~ ",
                   paste(c(colnames(x), predictors), collapse = " + "))
  fit <- glm(as.formula(formula) , data = data, family = "binomial")
  stepAIC(fit, trace = FALSE) # stops verbose printing
}
```

```
pos.neg.sub <- pos.neg[c("sentiment", "candidate", "subject_matter")]
levels(pos.neg.sub$candidate) <- c(levels(pos.neg.sub$candidate), "other")
pos.neg.sub$candidate <- relevel(pos.neg.sub$candidate, ref = "other")
pos.neg.sub$candidate[is.na(pos.neg.sub$candidate)] <- "other"
levels(pos.neg.sub$subject_matter) <- c(levels(pos.neg.sub$subject_matter), "other")
pos.neg.sub$subject_matter[is.na(pos.neg.sub$subject_matter)] <- "other"
pos.neg.sub$subject_matter <- relevel(pos.neg.sub$subject_matter, ref = "other")

dummy_candidate <- dummy(pos.neg.sub$candidate,
                         levels(pos.neg.sub$candidate)[-1])
dummy_subject_matter <- dummy(pos.neg.sub$subject_matter,
                         levels(pos.neg.sub$subject_matter)[-1])
candidate_only <- cv.glmnet(x = dummy_candidate, y = pos.neg.sub$sentiment,
          family = "binomial", alpha = 1, nfolds = 10)
candidate_subject_only <- cv.glmnet(x = cbind(dummy_candidate, dummy_subject_matter),
          y = pos.neg.sub$sentiment, family = "binomial", alpha = 1, nfolds = 10)
min(candidate_only$cvm)
```

```
    [1] 0.8932766
```

```
min(candidate_subject_only$cvm)
```

```
    [1] 0.870264
```

```
coef(candidate_only, s="lambda.min")
```

```
    11 x 1 sparse Matrix of class "dgCMatrix"
                             1
    (Intercept)    -2.1069223
    Ben Carson      1.9499006
    Chris Christie  0.2470594
    Donald Trump    1.0670563
    Jeb Bush       -0.6819826
    John Kasich     2.4153283
```

```
Marco Rubio     2.1394047
Mike Huckabee   0.6731933
Rand Paul       1.1462626
Scott Walker    0.4891074
Ted Cruz        2.2990564
```

```
coef(candidate_subject_only, s="lambda.min")
```

```
22 x 1 sparse Matrix of class "dgCMatrix"
                                                1
(Intercept)                           -1.76434654
Ben Carson                             1.97046000
Chris Christie                         0.04705098
Donald Trump                           0.95949844
Jeb Bush                              -0.90589523
John Kasich                            2.25006755
Marco Rubio                            2.00729576
Mike Huckabee                          0.60514038
Rand Paul                              0.93645986
Scott Walker                           0.45021499
Ted Cruz                               2.11877680
Abortion                              -0.70004793
Foreign Policy                        -0.66439762
FOX News or Moderators                -0.49851235
Gun Control                           -3.94610435
Healthcare (including Medicare)       -0.18812993
Immigration                            0.22064672
Jobs and Economy                      -0.63385942
LGBT issues                           -0.60615088
Racial issues                         -1.56539634
Religion                              -1.40575844
Women's Issues (not abortion though) -2.01680222
```

```
sentiment_twitter_candidate_10 <- glm_lda_model(simple_lda_10@gamma[-dropped.rows,] ,
                              modified_data = pos.neg.sub, predictors = "candidate")
sentiment_twitter_candidate_15 <- glm_lda_model(simple_lda_15@gamma[-dropped.rows,] ,
                              modified_data = pos.neg.sub, predictors = "candidate")
sentiment_twitter_candidate_20 <- glm_lda_model(simple_lda_20@gamma[-dropped.rows,] ,
                              modified_data = pos.neg.sub, predictors = "candidate")
sentiment_twitter_candidate_25 <- glm_lda_model(simple_lda_25@gamma[-dropped.rows,] ,
                              modified_data = pos.neg.sub, predictors = "candidate")
sentiment_twitter_candidate_30 <- glm_lda_model(simple_lda_30@gamma[-dropped.rows,] ,
                              modified_data = pos.neg.sub, predictors = "candidate")
sentiment_twitter_candidate_50 <- glm_lda_model(simple_lda_50@gamma[-dropped.rows,] ,
                              modified_data = pos.neg.sub, predictors = "candidate")
```

```
which.max(c(k10 = simple_lda_10@loglikelihood, k15 = simple_lda_15@loglikelihood,
            k20 = simple_lda_20@loglikelihood, k25 = simple_lda_25@loglikelihood,
            k30 = simple_lda_30@loglikelihood, k50 = simple_lda_50@loglikelihood))
```

```
k25
  4
```

```
sort(c(k10 = AIC(sentiment_twitter_candidate_10), k15 = AIC(sentiment_twitter_candidate_15),
          k20 = AIC(sentiment_twitter_candidate_20), k25 = AIC(sentiment_twitter_candidate_25),
          k30 = AIC(sentiment_twitter_candidate_30), k50 = AIC(sentiment_twitter_candidate_50)))
```

```
     k50      k25      k20      k30      k15      k10
7629.618 7634.541 7676.793 7688.978 7744.306 7780.977
```

```
sort(c(k10 = BIC(sentiment_twitter_candidate_10), k15 = BIC(sentiment_twitter_candidate_15),
           k20 = BIC(sentiment_twitter_candidate_20), k25 = BIC(sentiment_twitter_candidate_25),
           k30 = BIC(sentiment_twitter_candidate_30), k50 = BIC(sentiment_twitter_candidate_50)))
```

```
     k25      k20      k15      k10      k30      k50
7882.117 7889.001 7921.147 7922.449 7971.922 8054.034
```

```
anova(sentiment_twitter_candidate_20, sentiment_twitter_candidate_25, test="Chisq")
```

```
Analysis of Deviance Table

Model 1: sentiment ~ topic.1 + topic.3 + topic.4 + topic.5 + topic.6 +
    topic.7 + topic.8 + topic.9 + topic.10 + topic.11 + topic.12 +
    topic.13 + topic.14 + topic.15 + topic.16 + topic.17 + topic.18 +
    topic.19 + topic.20 + candidate
Model 2: sentiment ~ topic.1 + topic.3 + topic.4 + topic.5 + topic.6 +
    topic.7 + topic.8 + topic.9 + topic.10 + topic.11 + topic.12 +
    topic.13 + topic.14 + topic.15 + topic.16 + topic.17 + topic.18 +
    topic.19 + topic.20 + topic.21 + topic.22 + topic.23 + topic.24 +
    topic.25 + candidate
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      8692     7616.8
2      8687     7564.5  5   52.253 4.788e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(sentiment_twitter_candidate_25, sentiment_twitter_candidate_50, test="Chisq")
```

```
Analysis of Deviance Table

Model 1: sentiment ~ topic.1 + topic.3 + topic.4 + topic.5 + topic.6 +
    topic.7 + topic.8 + topic.9 + topic.10 + topic.11 + topic.12 +
    topic.13 + topic.14 + topic.15 + topic.16 + topic.17 + topic.18 +
    topic.19 + topic.20 + topic.21 + topic.22 + topic.23 + topic.24 +
    topic.25 + candidate
Model 2: sentiment ~ topic.1 + topic.3 + topic.4 + topic.5 + topic.6 +
    topic.7 + topic.8 + topic.9 + topic.10 + topic.11 + topic.12 +
    topic.13 + topic.14 + topic.15 + topic.16 + topic.17 + topic.18 +
    topic.19 + topic.20 + topic.21 + topic.22 + topic.23 + topic.24 +
    topic.25 + topic.26 + topic.27 + topic.28 + topic.29 + topic.30 +
    topic.31 + topic.32 + topic.33 + topic.34 + topic.35 + topic.36 +
    topic.37 + topic.38 + topic.39 + topic.40 + topic.41 + topic.42 +
    topic.43 + topic.44 + topic.45 + topic.46 + topic.47 + topic.48 +
    topic.49 + topic.50 + candidate
```

```
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      8687     7564.5
2      8662     7509.6 25   54.923 0.0005038 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
sentiment_debate_candidate <- glm_lda_model(twitter.topics$topics[-dropped.rows,] ,
                              modified_data = pos.neg.sub, predictors = "candidate")
```

```

Call:
glm(formula = as.formula(formula), family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4777  -0.6838  -0.4776  -0.3470   2.5639

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.17259    0.25615  -8.482  < 2e-16 ***
topic.1                  0.16577    0.43132   0.384  0.70073
topic.3                 -0.80692    0.49785  -1.621  0.10506
topic.4                 -1.33005    0.45888  -2.898  0.00375 **
topic.5                 -0.35243    0.32379  -1.088  0.27640
topic.6                  0.39584    0.37422   1.058  0.29016
topic.7                  0.49704    0.40951   1.214  0.22484
topic.8                  0.43118    0.38629   1.116  0.26433
topic.9                  0.71632    0.36074   1.986  0.04707 *
topic.10                 0.10242    0.35992   0.285  0.77597
topic.11                -0.22551    0.40652  -0.555  0.57908
topic.12                 0.10558    0.29100   0.363  0.71674
topic.13                 0.05758    0.37514   0.153  0.87802
topic.14                 0.36952    0.38745   0.954  0.34022
topic.15                -0.17283    0.38117  -0.453  0.65026
candidateBen Carson      2.02421    0.13584  14.901  < 2e-16 ***
candidateChris Christie  0.37805    0.20229   1.869  0.06165 .
candidateDonald Trump    1.14454    0.07596  15.067  < 2e-16 ***
candidateJeb Bush       -0.77960    0.19278  -4.044 5.25e-05 ***
candidateJohn Kasich     2.52487    0.16882  14.956  < 2e-16 ***
candidateMarco Rubio     2.17908    0.15288  14.254  < 2e-16 ***
candidateMike Huckabee   0.86048    0.16767   5.132 2.87e-07 ***
candidateRand Paul       1.19318    0.18081   6.599 4.14e-11 ***
candidateScott Walker    0.52126    0.20184   2.583  0.00981 **
candidateTed Cruz        2.33548    0.11218  20.818  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8774.8  on 8721  degrees of freedom
Residual deviance: 7726.9  on 8697  degrees of freedom
AIC: 7776.9

Number of Fisher Scoring iterations: 5
```

```
AIC(sentiment_debate_candidate); BIC(sentiment_debate_candidate)
```

```
[1] 7776.886
```

```
[1] 7953.726
```

```
AIC(sentiment_twitter_candidate_25); BIC(sentiment_twitter_candidate_25)
```

```
[1] 7634.541
```

```
[1] 7882.117
```

```
anova(sentiment_debate_candidate, sentiment_twitter_candidate_25, test="Chisq")
```

```
Analysis of Deviance Table

Model 1: sentiment ~ topic.1 + topic.3 + topic.4 + topic.5 + topic.6 +
    topic.7 + topic.8 + topic.9 + topic.10 + topic.11 + topic.12 +
    topic.13 + topic.14 + topic.15 + candidate
Model 2: sentiment ~ topic.1 + topic.3 + topic.4 + topic.5 + topic.6 +
    topic.7 + topic.8 + topic.9 + topic.10 + topic.11 + topic.12 +
    topic.13 + topic.14 + topic.15 + topic.16 + topic.17 + topic.18 +
    topic.19 + topic.20 + topic.21 + topic.22 + topic.23 + topic.24 +
    topic.25 + candidate
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      8697     7726.9
2      8687     7564.5 10   162.34 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
require(MASS)
```

```
Loading required package: MASS


Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select
```

```
step_25_candidate_subject <- stepwise_twitter(simple_lda_25@gamma[-dropped.rows,],
                            pos.neg.sub, predictors = c("candidate", "subject_matter"))
step_25_candidate_subject$anova
```

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
```

```
sentiment ~ topic.1 + topic.2 + topic.3 + topic.4 + topic.5 +
    topic.6 + topic.7 + topic.8 + topic.9 + topic.10 + topic.11 +
    topic.12 + topic.13 + topic.14 + topic.15 + topic.16 + topic.17 +
    topic.18 + topic.19 + topic.20 + topic.21 + topic.22 + topic.23 +
    topic.24 + topic.25 + candidate + subject_matter

Final Model:
sentiment ~ topic.1 + topic.2 + topic.3 + topic.4 + topic.6 +
    topic.8 + topic.11 + topic.13 + topic.14 + topic.15 + topic.16 +
    topic.17 + topic.19 + topic.22 + topic.23 + topic.24 + candidate +
    subject_matter


         Step Df   Deviance Resid. Df Resid. Dev      AIC
1                                 8676   7371.464 7463.464
2  - topic.25  0 0.00000000      8676   7371.464 7463.464
3   - topic.5  1 0.04248644      8677   7371.506 7461.506
4  - topic.10  1 0.06326254      8678   7371.570 7459.570
5   - topic.9  1 0.06924882      8679   7371.639 7457.639
6  - topic.21  1 0.16682058      8680   7371.806 7455.806
7   - topic.7  1 0.49511827      8681   7372.301 7454.301
8  - topic.18  1 0.72221103      8682   7373.023 7453.023
9  - topic.20  1 1.08137289      8683   7374.104 7452.104
10 - topic.12  1 1.47842730      8684   7375.583 7451.583
```

```
summary(step_25_candidate_subject)
```

```
Call:
glm(formula = sentiment ~ topic.1 + topic.2 + topic.3 + topic.4 +
    topic.6 + topic.8 + topic.11 + topic.13 + topic.14 + topic.15 +
    topic.16 + topic.17 + topic.19 + topic.22 + topic.23 + topic.24 +
    candidate + subject_matter, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9741  -0.6583  -0.4530  -0.2559   2.8157

Coefficients:
                                                 Estimate Std. Error
(Intercept)                                      -6.65511    1.34959
topic.1                                          13.71110    4.04975
topic.2                                         -11.62051    5.94830
topic.3                                          10.87861    4.87992
topic.4                                          10.57973    3.77652
topic.6                                          11.55361    4.85279
topic.8                                           8.86800    5.84793
topic.11                                         -8.92579    5.47661
topic.13                                          9.72193    5.51191
topic.14                                         27.27840    4.88334
topic.15                                         30.94867    4.90200
topic.16                                        -13.84200    7.14444
topic.17                                        -11.05920    5.43271
topic.19                                         22.45160    4.80179
```

```
topic.22                                          15.04853    3.70397
topic.23                                         -19.98853    6.33607
topic.24                                          24.27808    5.17896
candidateBen Carson                                2.11072    0.13688
candidateChris Christie                            0.55694    0.22785
candidateDonald Trump                              0.93028    0.07821
candidateJeb Bush                                 -0.46331    0.23543
candidateJohn Kasich                               2.42913    0.17562
candidateMarco Rubio                               1.90885    0.16544
candidateMike Huckabee                             1.11084    0.19592
candidateRand Paul                                 1.05870    0.18290
candidateScott Walker                              0.48071    0.20487
candidateTed Cruz                                  1.98748    0.12833
subject_matterAbortion                            -0.73770    0.23606
subject_matterForeign Policy                      -0.59055    0.19643
subject_matterFOX News or Moderators              -0.44444    0.08223
subject_matterGun Control                        -13.93138  215.21526
subject_matterHealthcare (including Medicare)     -0.25037    0.37166
subject_matterImmigration                          0.17490    0.22028
subject_matterJobs and Economy                    -0.73950    0.22430
subject_matterLGBT issues                         -0.58931    0.28809
subject_matterRacial issues                       -1.56566    0.25429
subject_matterReligion                            -1.44906    0.25492
subject_matterWomen's Issues (not abortion though) -2.04455    0.33195
                                                  z value Pr(>|z|)
(Intercept)                                        -4.931 8.17e-07 ***
topic.1                                             3.386 0.000710 ***
topic.2                                            -1.954 0.050750 .
topic.3                                             2.229 0.025797 *
topic.4                                             2.801 0.005087 **
topic.6                                             2.381 0.017274 *
topic.8                                             1.516 0.129410
topic.11                                           -1.630 0.103143
topic.13                                            1.764 0.077765 .
topic.14                                            5.586 2.32e-08 ***
topic.15                                            6.313 2.73e-10 ***
topic.16                                           -1.937 0.052690 .
topic.17                                           -2.036 0.041783 *
topic.19                                            4.676 2.93e-06 ***
topic.22                                            4.063 4.85e-05 ***
topic.23                                           -3.155 0.001607 **
topic.24                                            4.688 2.76e-06 ***
candidateBen Carson                                15.420  < 2e-16 ***
candidateChris Christie                             2.444 0.014512 *
candidateDonald Trump                              11.895  < 2e-16 ***
candidateJeb Bush                                  -1.968 0.049070 *
candidateJohn Kasich                               13.832  < 2e-16 ***
candidateMarco Rubio                               11.538  < 2e-16 ***
candidateMike Huckabee                              5.670 1.43e-08 ***
candidateRand Paul                                  5.788 7.10e-09 ***
candidateScott Walker                               2.346 0.018957 *
candidateTed Cruz                                  15.488  < 2e-16 ***
subject_matterAbortion                             -3.125 0.001778 **
subject_matterForeign Policy                       -3.006 0.002644 **
```

```
    subject_matterFOX News or Moderators                -5.405 6.48e-08 ***
    subject_matterGun Control                           -0.065 0.948387
    subject_matterHealthcare (including Medicare)       -0.674 0.500531
    subject_matterImmigration                            0.794 0.427200
    subject_matterJobs and Economy                      -3.297 0.000977 ***
    subject_matterLGBT issues                           -2.046 0.040796 *
    subject_matterRacial issues                         -6.157 7.42e-10 ***
    subject_matterReligion                              -5.684 1.31e-08 ***
    subject_matterWomen's Issues (not abortion though)  -6.159 7.31e-10 ***
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    (Dispersion parameter for binomial family taken to be 1)

        Null deviance: 8774.8  on 8721  degrees of freedom
    Residual deviance: 7375.6  on 8684  degrees of freedom
    AIC: 7451.6

    Number of Fisher Scoring iterations: 14
```

```
step_debate_topics <- stepwise_twitter(twitter.topics$topics[-dropped.rows,],
                       pos.neg.sub, predictors = c("candidate", "subject_matter"))
step_debate_topics$anova
```

```
    Stepwise Model Path
    Analysis of Deviance Table

    Initial Model:
    sentiment ~ topic.1 + topic.2 + topic.3 + topic.4 + topic.5 +
        topic.6 + topic.7 + topic.8 + topic.9 + topic.10 + topic.11 +
        topic.12 + topic.13 + topic.14 + topic.15 + candidate + subject_matter

    Final Model:
    sentiment ~ topic.3 + topic.4 + topic.5 + topic.9 + candidate +
        subject_matter


             Step Df  Deviance Resid. Df Resid. Dev      AIC
    1                               8686   7506.915 7578.915
    2  - topic.15  0 0.0000000      8686   7506.915 7578.915
    3  - topic.11  1 0.0361263      8687   7506.951 7576.951
    4  - topic.13  1 0.1544666      8688   7507.105 7575.105
    5   - topic.2  1 0.1144348      8689   7507.220 7573.220
    6   - topic.1  1 0.1522128      8690   7507.372 7571.372
    7  - topic.10  1 1.0886845      8691   7508.461 7570.461
    8  - topic.12  1 1.2258320      8692   7509.687 7569.687
    9   - topic.6  1 0.8967099      8693   7510.583 7568.583
    10 - topic.14  1 0.9235075      8694   7511.507 7567.507
    11  - topic.8  1 0.9577672      8695   7512.465 7566.465
    12  - topic.7  1 1.3343270      8696   7513.799 7565.799
```

```
summary(step_debate_topics)
```

```
Call:
glm(formula = sentiment ~ topic.3 + topic.4 + topic.5 + topic.9 +
    candidate + subject_matter, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4871  -0.6432  -0.4516  -0.2665   3.0724

Coefficients:
                                                  Estimate Std. Error
(Intercept)                                       -1.65379    0.07114
topic.3                                           -0.84461    0.44540
topic.4                                           -1.38034    0.39837
topic.5                                           -0.54730    0.20716
topic.9                                            0.44076    0.26667
candidateBen Carson                                1.97076    0.13278
candidateChris Christie                            0.13934    0.20448
candidateDonald Trump                              1.01733    0.07611
candidateJeb Bush                                 -0.98970    0.18763
candidateJohn Kasich                               2.28566    0.17181
candidateMarco Rubio                               2.00820    0.15789
candidateMike Huckabee                             0.69399    0.16685
candidateRand Paul                                 0.95132    0.18006
candidateScott Walker                              0.42488    0.20248
candidateTed Cruz                                  2.11554    0.11252
subject_matterAbortion                            -0.69955    0.23190
subject_matterForeign Policy                      -0.66643    0.19212
subject_matterFOX News or Moderators              -0.51114    0.07579
subject_matterGun Control                        -13.10766  131.77497
subject_matterHealthcare (including Medicare)     -0.19179    0.36892
subject_matterImmigration                          0.23802    0.19145
subject_matterJobs and Economy                    -0.63648    0.22177
subject_matterLGBT issues                         -0.61507    0.28052
subject_matterRacial issues                       -1.59946    0.25355
subject_matterReligion                            -1.46172    0.25159
subject_matterWomen's Issues (not abortion though) -2.02505    0.32657
                                                  z value Pr(>|z|)
(Intercept)                                       -23.246  < 2e-16 ***
topic.3                                            -1.896 0.057922 .
topic.4                                            -3.465 0.000530 ***
topic.5                                            -2.642 0.008245 **
topic.9                                             1.653 0.098367 .
candidateBen Carson                                14.843  < 2e-16 ***
candidateChris Christie                             0.681 0.495610
candidateDonald Trump                              13.366  < 2e-16 ***
candidateJeb Bush                                  -5.275 1.33e-07 ***
candidateJohn Kasich                               13.303  < 2e-16 ***
candidateMarco Rubio                               12.719  < 2e-16 ***
candidateMike Huckabee                              4.159 3.19e-05 ***
candidateRand Paul                                  5.283 1.27e-07 ***
candidateScott Walker                               2.098 0.035871 *
candidateTed Cruz                                  18.802  < 2e-16 ***
subject_matterAbortion                             -3.017 0.002557 **
```

```
subject_matterForeign Policy                          -3.469 0.000523 ***
subject_matterFOX News or Moderators                  -6.744 1.54e-11 ***
subject_matterGun Control                             -0.099 0.920765
subject_matterHealthcare (including Medicare)         -0.520 0.603152
subject_matterImmigration                              1.243 0.213774
subject_matterJobs and Economy                        -2.870 0.004105 **
subject_matterLGBT issues                             -2.193 0.028339 *
subject_matterRacial issues                           -6.308 2.82e-10 ***
subject_matterReligion                                -5.810 6.25e-09 ***
subject_matterWomen's Issues (not abortion though)    -6.201 5.61e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8774.8  on 8721  degrees of freedom
Residual deviance: 7513.8  on 8696  degrees of freedom
AIC: 7565.8

Number of Fisher Scoring iterations: 13
```

```
anova(step_debate_topics, step_25_candidate_subject, test="Chisq")
```

```
Analysis of Deviance Table

Model 1: sentiment ~ topic.3 + topic.4 + topic.5 + topic.9 + candidate +
    subject_matter
Model 2: sentiment ~ topic.1 + topic.2 + topic.3 + topic.4 + topic.6 +
    topic.8 + topic.11 + topic.13 + topic.14 + topic.15 + topic.16 +
    topic.17 + topic.19 + topic.22 + topic.23 + topic.24 + candidate +
    subject_matter
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      8696     7513.8
2      8684     7375.6 12   138.22 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
step_25_candidate <- stepwise_twitter(simple_lda_25@gamma[-dropped.rows,],
                          pos.neg.sub, predictors = c("candidate"))
step_25_candidate$anova
```

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
sentiment ~ topic.1 + topic.2 + topic.3 + topic.4 + topic.5 +
    topic.6 + topic.7 + topic.8 + topic.9 + topic.10 + topic.11 +
    topic.12 + topic.13 + topic.14 + topic.15 + topic.16 + topic.17 +
    topic.18 + topic.19 + topic.20 + topic.21 + topic.22 + topic.23 +
    topic.24 + topic.25 + candidate

Final Model:
sentiment ~ topic.1 + topic.2 + topic.3 + topic.4 + topic.6 +
```

```
    topic.10 + topic.13 + topic.14 + topic.15 + topic.17 + topic.19 +
    topic.20 + topic.22 + topic.23 + topic.24 + candidate


         Step Df    Deviance Resid. Df Resid. Dev      AIC
1                                 8687   7564.541 7634.541
2  - topic.25  0 0.000000000      8687   7564.541 7634.541
3  - topic.21  1 0.007555184      8688   7564.548 7632.548
4  - topic.18  1 0.064860190      8689   7564.613 7630.613
5   - topic.7  1 0.057461514      8690   7564.670 7628.670
6   - topic.5  1 0.221742597      8691   7564.892 7626.892
7   - topic.8  1 0.196456490      8692   7565.089 7625.089
8   - topic.9  1 0.240554715      8693   7565.329 7623.329
9  - topic.12  1 0.788789524      8694   7566.118 7622.118
10 - topic.16  1 1.148051505      8695   7567.266 7621.266
11 - topic.11  1 1.175723099      8696   7568.442 7620.442
```

```
summary(step_25_candidate)
```

```
Call:
glm(formula = sentiment ~ topic.1 + topic.2 + topic.3 + topic.4 +
    topic.6 + topic.10 + topic.13 + topic.14 + topic.15 + topic.17 +
    topic.19 + topic.20 + topic.22 + topic.23 + topic.24 + candidate,
    family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9562  -0.6544  -0.4419  -0.3167   2.6099

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              -9.03446    1.19640  -7.551 4.31e-14 ***
topic.1                  13.80048    3.87768   3.559 0.000372 ***
topic.2                  -8.60373    5.68267  -1.514 0.130018
topic.3                  16.81774    4.77665   3.521 0.000430 ***
topic.4                  10.47045    3.65772   2.863 0.004202 **
topic.6                  16.45911    4.85769   3.388 0.000703 ***
topic.10                 -9.32921    5.79380  -1.610 0.107353
topic.13                 14.43949    4.77699   3.023 0.002505 **
topic.14                 29.69177    4.75588   6.243 4.29e-10 ***
topic.15                 34.93382    4.77433   7.317 2.54e-13 ***
topic.17                -11.08486    5.24679  -2.113 0.034627 *
topic.19                 23.95011    4.65096   5.149 2.61e-07 ***
topic.20                  9.29034    5.03864   1.844 0.065210 .
topic.22                 17.96145    3.61984   4.962 6.98e-07 ***
topic.23                -17.66470    6.21845  -2.841 0.004502 **
topic.24                 30.67275    5.09394   6.021 1.73e-09 ***
candidateBen Carson       2.10995    0.12793  16.493  < 2e-16 ***
candidateChris Christie   0.62827    0.21100   2.978 0.002905 **
candidateDonald Trump     0.99441    0.07994  12.439  < 2e-16 ***
candidateJeb Bush        -0.56764    0.19197  -2.957 0.003108 **
candidateJohn Kasich      2.60234    0.16782  15.507  < 2e-16 ***
candidateMarco Rubio      2.02145    0.16007  12.629  < 2e-16 ***
```

```
    candidateMike Huckabee    1.19364    0.19337   6.173 6.70e-10 ***
    candidateRand Paul        1.24053    0.17967   6.904 5.04e-12 ***
    candidateScott Walker     0.52249    0.20098   2.600 0.009330 **
    candidateTed Cruz         2.10374    0.12610  16.683  < 2e-16 ***
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    (Dispersion parameter for binomial family taken to be 1)

        Null deviance: 8774.8  on 8721  degrees of freedom
    Residual deviance: 7568.4  on 8696  degrees of freedom
    AIC: 7620.4

    Number of Fisher Scoring iterations: 5
```

```
step_debate_candidate_topics <- stepwise_twitter(twitter.topics$topics[-dropped.rows,],
                        pos.neg.sub, predictors = c("candidate"))
step_debate_candidate_topics$anova
```

```
    Stepwise Model Path
    Analysis of Deviance Table

    Initial Model:
    sentiment ~ topic.1 + topic.2 + topic.3 + topic.4 + topic.5 +
        topic.6 + topic.7 + topic.8 + topic.9 + topic.10 + topic.11 +
        topic.12 + topic.13 + topic.14 + topic.15 + candidate

    Final Model:
    sentiment ~ topic.3 + topic.4 + topic.9 + candidate


             Step Df  Deviance Resid. Df Resid. Dev      AIC
    1                               8697   7726.886 7776.886
    2  - topic.15  0 0.0000000      8697   7726.886 7776.886
    3  - topic.11  1 0.0144929      8698   7726.900 7774.900
    4   - topic.5  1 0.2962551      8699   7727.196 7773.196
    5   - topic.2  1 0.9019338      8700   7728.098 7772.098
    6  - topic.13  1 0.8188439      8701   7728.917 7770.917
    7   - topic.1  1 0.8072849      8702   7729.725 7769.725
    8  - topic.10  1 0.7991876      8703   7730.524 7768.524
    9  - topic.12  1 1.8884422      8704   7732.412 7768.412
    10 - topic.14  1 1.6095061      8705   7734.022 7768.022
    11  - topic.8  1 1.9694914      8706   7735.991 7767.991
    12  - topic.7  1 1.8863938      8707   7737.878 7767.878
    13  - topic.6  1 1.7458130      8708   7739.623 7767.623
```

```
summary(step_debate_candidate_topics)
```

```
    Call:
    glm(formula = sentiment ~ topic.3 + topic.4 + topic.9 + candidate,
        family = "binomial", data = data)
```

```
    Deviance Residuals:
        Min       1Q   Median       3Q      Max
    -1.4762  -0.6914  -0.4795  -0.3339   2.5343

    Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
    (Intercept)            -2.05743    0.06233 -33.011  < 2e-16 ***
    topic.3                -0.89249    0.43612  -2.046 0.040714 *
    topic.4                -1.40060    0.39048  -3.587 0.000335 ***
    topic.9                 0.61697    0.26169   2.358 0.018392 *
    candidateBen Carson     1.97682    0.12298  16.075  < 2e-16 ***
    candidateChris Christie 0.30507    0.20043   1.522 0.127994
    candidateDonald Trump   1.08617    0.07184  15.119  < 2e-16 ***
    candidateJeb Bush      -0.76441    0.18657  -4.097 4.18e-05 ***
    candidateJohn Kasich    2.44788    0.16325  14.995  < 2e-16 ***
    candidateMarco Rubio    2.17735    0.15183  14.341  < 2e-16 ***
    candidateMike Huckabee  0.79585    0.16158   4.925 8.42e-07 ***
    candidateRand Paul      1.18491    0.17640   6.717 1.85e-11 ***
    candidateScott Walker   0.50484    0.19810   2.548 0.010823 *
    candidateTed Cruz       2.31882    0.10956  21.165  < 2e-16 ***
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    (Dispersion parameter for binomial family taken to be 1)

        Null deviance: 8774.8  on 8721  degrees of freedom
    Residual deviance: 7739.6  on 8708  degrees of freedom
    AIC: 7767.6

    Number of Fisher Scoring iterations: 5
```

```
anova(step_debate_candidate_topics, step_25_candidate, test="Chisq")
```

```
    Analysis of Deviance Table

    Model 1: sentiment ~ topic.3 + topic.4 + topic.9 + candidate
    Model 2: sentiment ~ topic.1 + topic.2 + topic.3 + topic.4 + topic.6 +
        topic.10 + topic.13 + topic.14 + topic.15 + topic.17 + topic.19 +
        topic.20 + topic.22 + topic.23 + topic.24 + candidate
      Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
    1      8708     7739.6
    2      8696     7568.4 12   171.18 < 2.2e-16 ***
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
dummy_candidate <- dummy(pos.neg.sub$candidate,
                         levels(pos.neg.sub$candidate)[-1])
dummy_subject_matter <- dummy(pos.neg.sub$subject_matter,
                         levels(pos.neg.sub$subject_matter)[-1])
candidate_only <- cv.glmnet(x = dummy_candidate, y = pos.neg.sub$sentiment,
          family = "binomial", alpha = 1, nfolds = 10)
candidate_subject_only <- cv.glmnet(x = cbind(dummy_candidate, dummy_subject_matter),
          y = pos.neg.sub$sentiment, family = "binomial", alpha = 1, nfolds = 10)
min(candidate_only$cvm)
```

```
    [1] 0.893932
```

```
min(candidate_subject_only$cvm)
```

```
    [1] 0.8688909
```

```
coef(candidate_only, s="lambda.min")
```

```
    11 x 1 sparse Matrix of class "dgCMatrix"
                            1
    (Intercept)    -2.1069223
    Ben Carson      1.9499006
    Chris Christie  0.2470594
    Donald Trump    1.0670563
    Jeb Bush       -0.6819826
    John Kasich     2.4153283
    Marco Rubio     2.1394047
    Mike Huckabee   0.6731933
    Rand Paul       1.1462626
    Scott Walker    0.4891074
    Ted Cruz        2.2990564
```

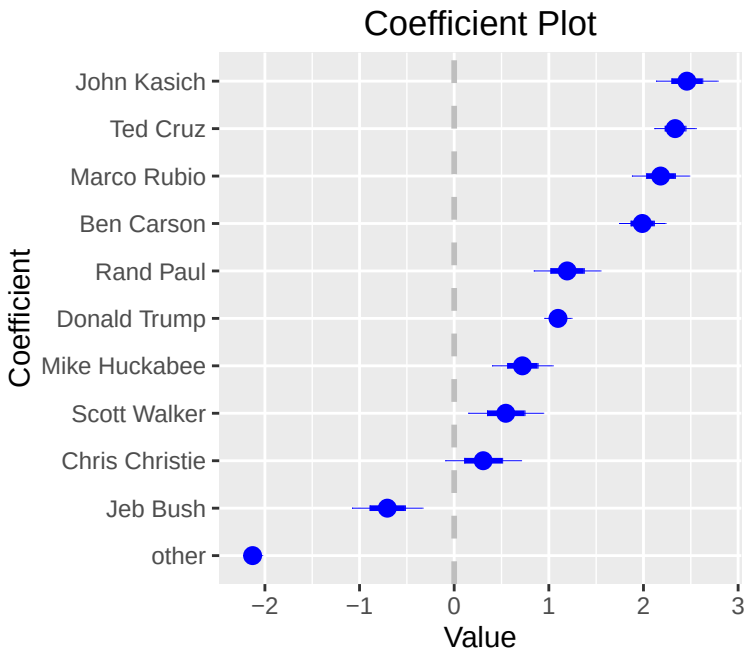
```
coef(candidate_subject_only, s="lambda.min")
```

```
    22 x 1 sparse Matrix of class "dgCMatrix"
                                                  1
    (Intercept)                          -1.76380655
    Ben Carson                            1.96841918
    Chris Christie                        0.04446841
    Donald Trump                          0.95836625
    Jeb Bush                             -0.90440262
    John Kasich                           2.24777913
    Marco Rubio                           2.00529504
    Mike Huckabee                         0.60256605
    Rand Paul                             0.93444710
    Scott Walker                          0.44740034
    Ted Cruz                              2.11728825
    Abortion                             -0.69735251
    Foreign Policy                       -0.66187811
    FOX News or Moderators               -0.49796323
    Gun Control                          -3.85199554
    Healthcare (including Medicare)      -0.18456509
    Immigration                           0.21963432
    Jobs and Economy                     -0.63133633
    LGBT issues                          -0.60277641
    Racial issues                        -1.56186113
    Religion                             -1.40245488
    Women's Issues (not abortion though) -2.01149524
```

```
require(coefplot)
candidate_only2 <- glm(sentiment ~ candidate, data = pos.neg.sub, family = "binomial")
candidate_subject_only2 <- glm(sentiment ~ candidate + subject_matter, data = pos.neg.sub,
                               family = "binomial")
summary(candidate_only2)
```

```
Call:
glm(formula = sentiment ~ candidate, family = "binomial", data = pos.neg.sub)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3197  -0.6612  -0.4739  -0.3375   2.4057

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.12999    0.05077 -41.951  < 2e-16 ***
candidateBen Carson      1.98733    0.12258  16.213  < 2e-16 ***
candidateChris Christie  0.30648    0.20011   1.532 0.125636
candidateDonald Trump    1.09666    0.07140  15.359  < 2e-16 ***
candidateJeb Bush       -0.70679    0.18616  -3.797 0.000147 ***
candidateJohn Kasich     2.45849    0.16269  15.112  < 2e-16 ***
candidateMarco Rubio     2.18076    0.15131  14.412  < 2e-16 ***
candidateMike Huckabee   0.72081    0.16007   4.503 6.69e-06 ***
candidateRand Paul       1.19349    0.17603   6.780 1.20e-11 ***
candidateScott Walker    0.54509    0.19772   2.757 0.005835 **
candidateTed Cruz        2.33440    0.10924  21.369  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8774.8  on 8721  degrees of freedom
Residual deviance: 7765.0  on 8711  degrees of freedom
AIC: 7787

Number of Fisher Scoring iterations: 5
```

```
rename_candidate <- c("other", gsub("candidate*", "", names(coef(candidate_only2))[2:11]))
names(rename_candidate) <- names(coef(candidate_only2))
coefplot::coefplot(candidate_only2, sort="magnitude", newNames = rename_candidate)
```




```
summary(candidate_subject_only2)
```

```
Call:
glm(formula = sentiment ~ candidate + subject_matter, family = "binomial",
    data = pos.neg.sub)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4383  -0.6698  -0.4426  -0.2631   3.0895

Coefficients:
                                                 Estimate Std. Error
(Intercept)                                     -1.76996    0.05836
candidateBen Carson                              1.99153    0.13236
candidateChris Christie                          0.07349    0.20268
candidateDonald Trump                            0.97118    0.07272
candidateJeb Bush                               -0.92140    0.18712
candidateJohn Kasich                             2.27370    0.17085
candidateMarco Rubio                             2.02796    0.15731
candidateMike Huckabee                           0.63158    0.16539
candidateRand Paul                               0.95716    0.17968
candidateScott Walker                            0.47909    0.20199
candidateTed Cruz                                2.13415    0.11210
subject_matterAbortion                          -0.72792    0.23120
subject_matterForeign Policy                    -0.69038    0.19146
subject_matterFOX News or Moderators            -0.50414    0.07554
subject_matterGun Control                      -13.11555  132.43277
subject_matterHealthcare (including Medicare)   -0.22495    0.36744
subject_matterImmigration                        0.23104    0.19125
subject_matterJobs and Economy                  -0.65994    0.21996
subject_matterLGBT issues                       -0.64097    0.27913
```

```
subject_matterRacial issues                          -1.60209    0.25292
subject_matterReligion                               -1.44012    0.25163
subject_matterWomen's Issues (not abortion though)   -2.07271    0.32617
                                                    z value Pr(>|z|)
(Intercept)                                         -30.327  < 2e-16 ***
candidateBen Carson                                  15.046  < 2e-16 ***
candidateChris Christie                               0.363 0.716921
candidateDonald Trump                                13.355  < 2e-16 ***
candidateJeb Bush                                    -4.924 8.47e-07 ***
candidateJohn Kasich                                 13.308  < 2e-16 ***
candidateMarco Rubio                                 12.891  < 2e-16 ***
candidateMike Huckabee                                3.819 0.000134 ***
candidateRand Paul                                    5.327 9.99e-08 ***
candidateScott Walker                                 2.372 0.017703 *
candidateTed Cruz                                    19.038  < 2e-16 ***
subject_matterAbortion                               -3.148 0.001642 **
subject_matterForeign Policy                         -3.606 0.000311 ***
subject_matterFOX News or Moderators                 -6.674 2.49e-11 ***
subject_matterGun Control                            -0.099 0.921110
subject_matterHealthcare (including Medicare)        -0.612 0.540398
subject_matterImmigration                             1.208 0.227023
subject_matterJobs and Economy                       -3.000 0.002698 **
subject_matterLGBT issues                            -2.296 0.021657 *
subject_matterRacial issues                          -6.334 2.38e-10 ***
subject_matterReligion                               -5.723 1.05e-08 ***
subject_matterWomen's Issues (not abortion though)   -6.355 2.09e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8774.8  on 8721  degrees of freedom
Residual deviance: 7540.6  on 8700  degrees of freedom
AIC: 7584.6

Number of Fisher Scoring iterations: 13
```

```r
rename_subject <- c("other", gsub("subject_matter*", "", names(coef(candidate_subject_only2))[12:22]))
names(rename_subject) <- names(coef(candidate_subject_only2))[c(1,12:22)]
#coefplot::coefplot(candidate_subject_only2, predictors = "subject_matter", sort="magnitude", newNames =
anova(candidate_only2, candidate_subject_only2, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: sentiment ~ candidate
Model 2: sentiment ~ candidate + subject_matter
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      8711     7765.0
2      8700     7540.6 11   224.32 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
require(dplyr)
require(reshape2)
```

```
    Loading required package: reshape2
```

```
require(ggplot2)
debate_LDA_15_names[2] <- "mods3"
topic_theta_by_speaker <- data.frame(debate_LDA_15@gamma, speaker = debate_corpus$documents$speaker)
# come up with descriptive names for topics
colnames(topic_theta_by_speaker) <- c(debate_LDA_15_names, "speaker")
grouped <- group_by(topic_theta_by_speaker, speaker)
topic_means_by_speaker <- as.data.frame(grouped %>% summarize_each(funs(mean)))
melted <- reshape2::melt(topic_means_by_speaker, id.vars = "speaker")
melted.candidate <- filter(melted, speaker != "OTHER" & speaker != "MODERATOR")
p <- ggplot(melted.candidate, aes(x = speaker, y = value, fill = variable))
p <- p + geom_bar(stat="identity")
p <- p + theme(axis.text.x=element_text(angle = 90, vjust = 0.5))
p <- p + labs(fill = "Topic", x = "Candidate", y = "Mean Theta by Topic")
p
```