The U.S. Presidential Race

Comparing Coverage Across Three Major Sources

Thomas Brawner

Galvanize, Inc.

September 2015

Objective

- Classification: *Is there a difference in coverage?*
- Topic Modeling: *What is the difference?*

Data collection

- Scrape *Guardian*, *New York Times*, *Wall Street Journal*
- January to August 2015
- Raw data → MongoDB
- Filter on candidates
- TF-IDF, 2000 terms

Multiclass classification accuracy

	Global	Guardian	New York Times	Wall Street Journal
Naive Bayes	0.703	0.414	0.965	0.340
Random Forest	0.905	0.783	0.998	0.809
Gradient Boosting	0.921	0.856	0.983	0.833
N	3757	1018	2087	652

Topic modeling strategy

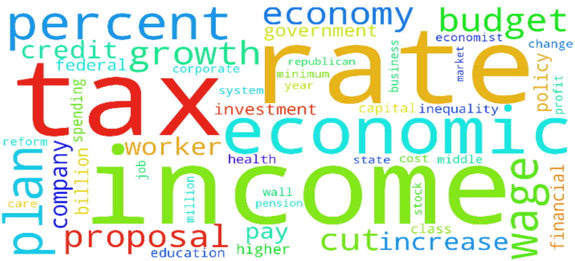
$$\begin{bmatrix} v_{11} & \cdots & v_{1t} & \cdots & v_{1T} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ v_{i1} & \cdots & v_{it} & \cdots & v_{iT} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ v_{N1} & \cdots & v_{Nt} & \cdots & v_{NT} \end{bmatrix} =$$

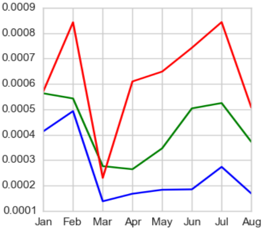
$$\begin{bmatrix} w_{11} & \cdots & w_{1K} \\ \vdots & \ddots & \vdots \\ w_{i1} & \cdots & w_{iK} \\ \vdots & \ddots & \vdots \\ w_{N1} & \cdots & w_{NK} \end{bmatrix} \times \begin{bmatrix} h_{11} & \cdots & h_{1t} & \cdots & h_{1T} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ h_{K1} & \cdots & h_{Kt} & \cdots & h_{KT} \end{bmatrix}$$

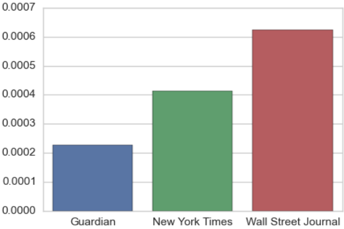
Topic modeling strategy (cont.)

- With $\mathbf{H}_{K \times T}$
  - sort descending for each k
  - top 50 terms → Word Cloud
- With $\mathbf{W}_{N \times K}$
  - group by source, month → time series
  - group by source → bar plots

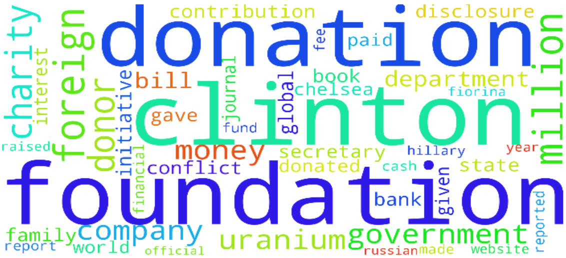
Economic Policy

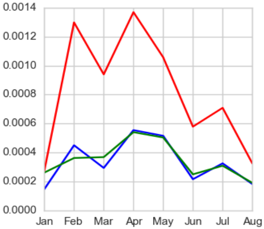


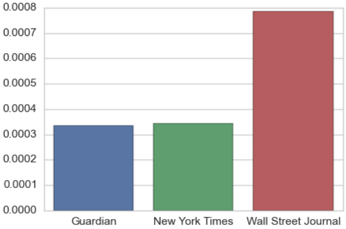




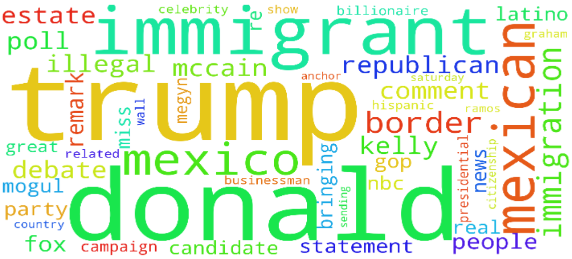
Hillary

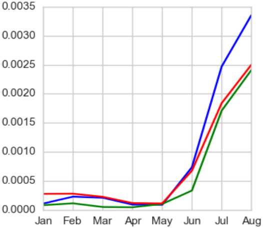

charity
interest
foreign
donor
initiative
financial
contribution
donation
fee
paid
disclosure
million
bill
journal
global
book
department
chelsea
fiorina
gave
fund
clinton
raised
money
secretary
hillary
year
conflict
donated
cash
state
foundation
bank
given
reported
family
report
company
world
official
uranium
government
russian
made
website

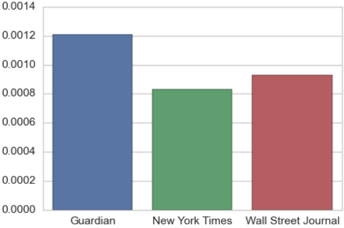

0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug


0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
Guardian
New York Times
Wall Street Journal

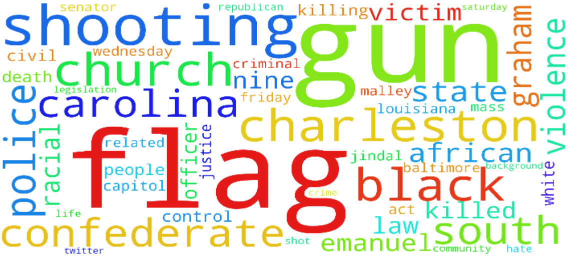
The Rise of Trump

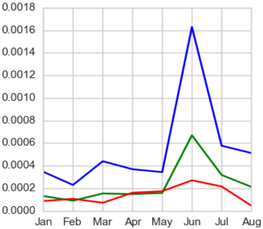


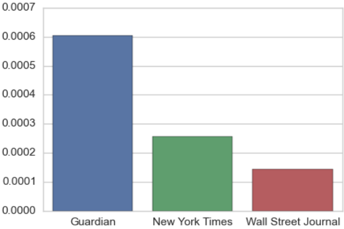




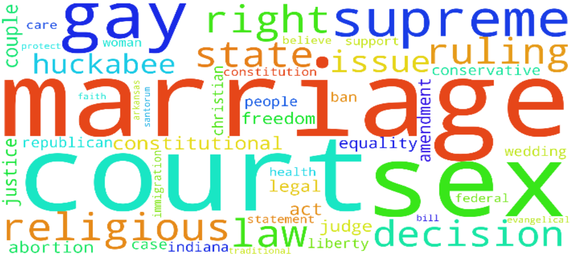
Gun Violence

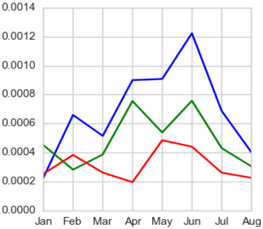


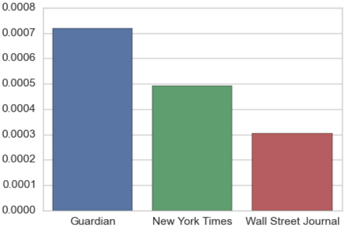




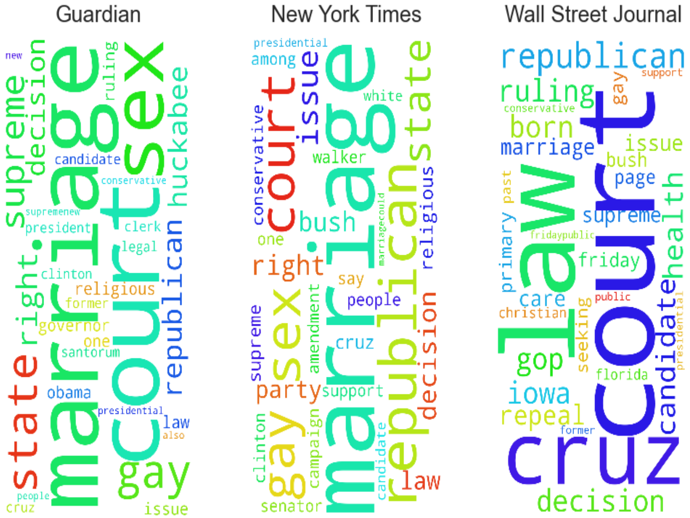
Gay Marriage







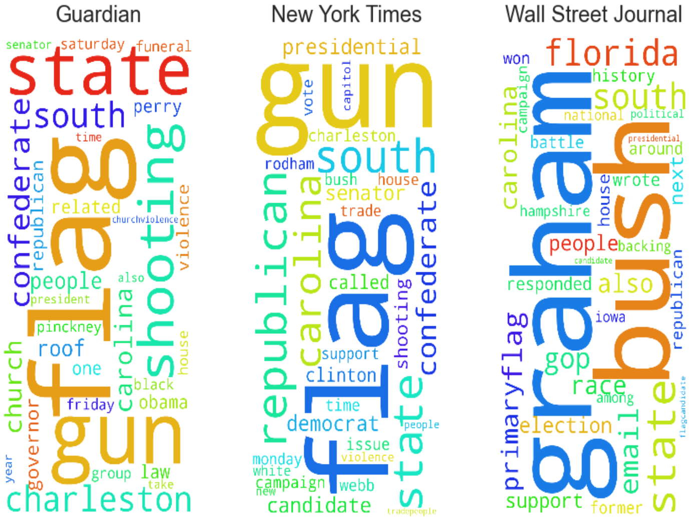
Gay Marriage by Outlet


Guardian

New York Times

Wall Street Journal

Gun Violence by Outlet



Next steps

- ♯ More data: collect through duration of election
- ♯ More data: *Washington Post*, *Financial Times*, etc.
- ♯ Finer classification model tuning
- ♯ What are we misclassifying?