# shenghan_project4

*Shenghan Yu*

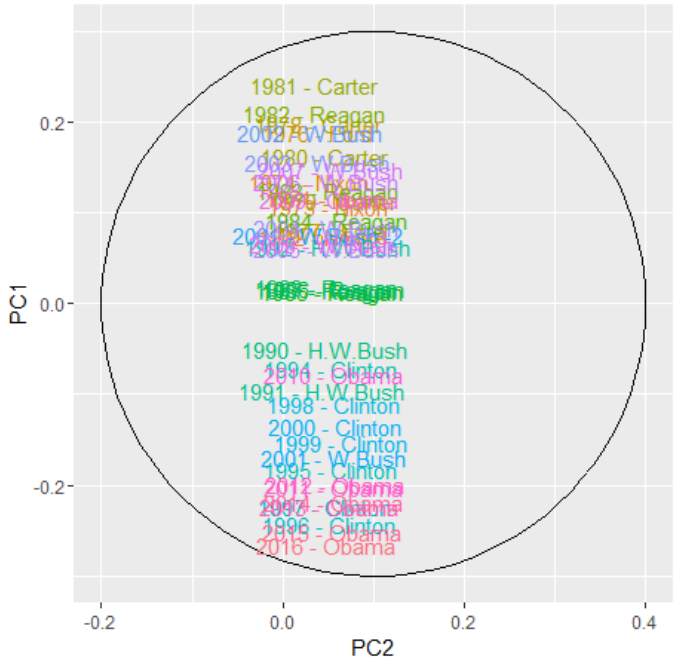
*May 12, 2016*

## PCA Analysis on Corpus

In this section, we conducted a principle component analysis on the corpus of state of union speech. The result shows that the the first principal component explains about 97% of the total variance in the data. We also plotteed the all the speeches in the 2D space formed by the first two principal components.

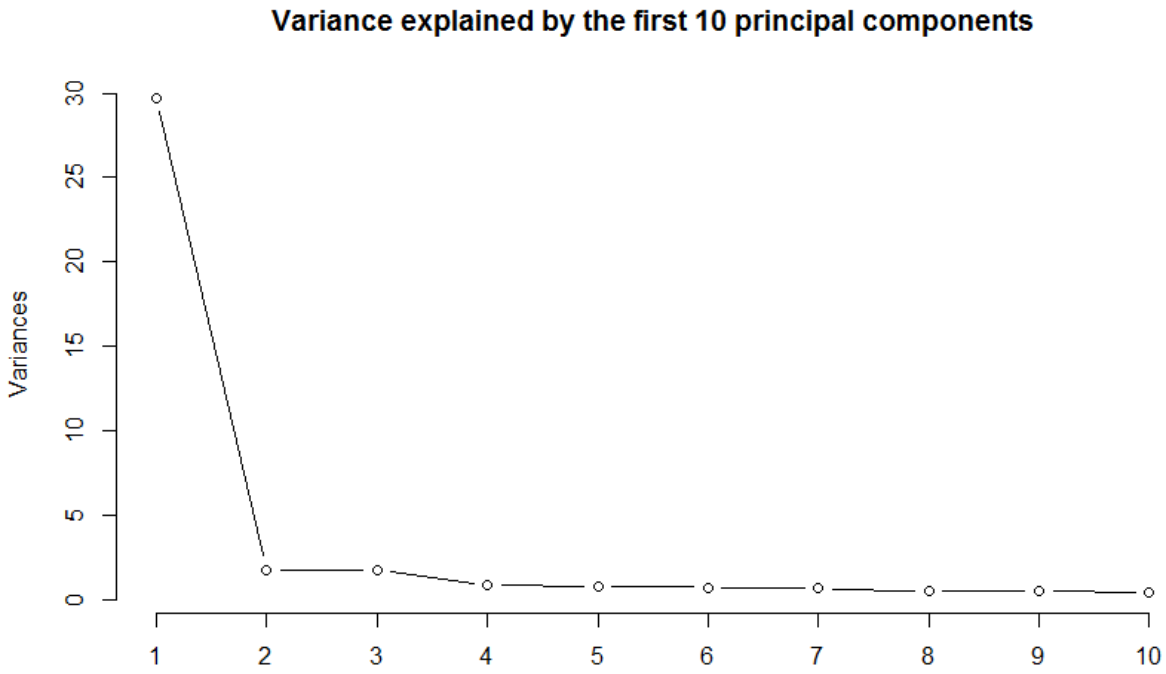
There are several patterns that is worth disucssing here. Overall, there are three clusters in this 2D space, one of which consists of President Ronald Reagan alone. Besides, 6 out of 7 Reagan's speeches are in the cluster. This indicates that the style of Reagon's state of union is signiificantly different from all other presidents in terms of vocubulary and words picking. Another intriguing finding is that the a majority of Bill Clinton's and Barack Obama's deliveries lies in the same cluster, along with those of George H.W. Bush in 1990 and 1991. This may be attributed to the the three presidents' common emphasis on domestic economy and development during there administration, whether in the golden times of 90s or in recovery from the financial crisis in 2008.

In addition, we can see that the first state of union that George W. Bush delivered (in Feb 2001) lies in the same cluster of Clinton and Obama's, while the latter speeches fall into the opposite cluster. A critical factor here is the 9-11 attach happened on Sep 11th, 2001. George W. Bush delivered an "extra" state-of-union to declare a war on terrorism. This event radically changes the major objective of Bush's admisitration from domestic economy to wars, foreign affairs and securities. That's a sound reason why, after 2001, all the Bush's deliveries lie in the other cluster.




.names

| | | | | | |
|---|---|---|---|---|---|
| a | 1971 - Nixon | a | 1986 - Reagan | a | 2002 - W.Bush |
| a | 1972 - Nixon | a | 1987 - Reagan | a | 2003 - W.Bush |
| a | 1973 - Nixon | a | 1988 - Reagan | a | 2004 - W.Bush |
| a | 1974 - Nixon | a | 1990 - H.W.Bush | a | 2005 - W.Bush |
| a | 1975 - Ford | a | 1991 - H.W.Bush | a | 2006 - W.Bush |
| a | 1976 - Ford | a | 1992 - H.W.Bush | a | 2007 - W.Bush |
| a | 1977 - Ford | a | 1994 - Clinton | a | 2008 - W.Bush |
| a | 1978 - Carter | a | 1995 - Clinton | a | 2009 - Obama |
| a | 1979 - Carter | a | 1996 - Clinton | a | 2010 - Obama |
| a | 1980 - Carter | a | 1997 - Clinton | a | 2011 - Obama |
| a | 1981 - Carter | a | 1998 - Clinton | a | 2012 - Obama |
| a | 1982 - Reagan | a | 1999 - Clinton | a | 2013 - Obama |
| a | 1983 - Reagan | a | 2000 - Clinton | a | 2014 - Obama |
| a | 1984 - Reagan | a | 2001 - W.Bush | a | 2015 - Obama |
| a | 1985 - Reagan | a | 2001 - W.Bush -2 | a | 2016 - Obama |


Variance explained by the first 10 principal components
Variances
0
5
10
15
20
25
30
1
2
3
4
5
6
7
8
9
10