# Montana State Fund

CASE STUDY

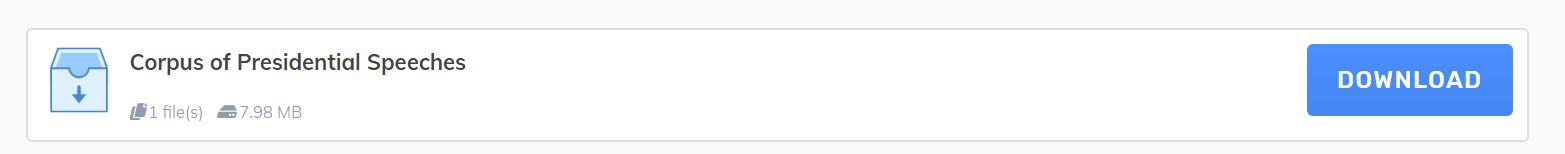
*Analysis of POTUS Speeches Using Natural Language Processing (NLP)*

***Objective & High-Level Goals:***

**Topic modeling** is an application of Natural Language Processing (NLP) and is an unsupervised machine learning technique that’s capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents.

The objective of this case study is to perform **Topic Modelling** using **R or Python** on a set of POTUS speeches of 44 Presidents (1789 – 2016) that can be downloaded from the website below

<http://www.thegrammarlab.com/?nor-portfolio=corpus-of-presidential-speeches-cops-and-a-clintontrump-corpus>



You are required to achieve two high level goals:

* Identify the dominant topic in speeches by every president
* Identify the dominant topic in speeches by the following time periods

1. 1789 – 1849
2. 1849 – 1893
3. 1897 – 1969
4. 1969 – 2016

List of Presidents and their terms -

|  |  |  |
| --- | --- | --- |
| FIRST NAME | LAST NAME | IN OFFICE |
| George | Washington | 1789-97 |
| John | Adams | 1797-1801 |
| Thomas | Jefferson | 1801-1809 |
| James | Madison | 1809-1817 |
| James | Monroe | 1817-1825 |
| John Quincy | Adams | 1825-1829 |
| Andrew | Jackson | 1829-1837 |
| Martin | Van Buren | 1837-1841 |
| William H. | Harrison | 1841 |
| John | Tyler | 1841-1845 |
| James K. | Polk | 1845-1849 |
| Zachary | Taylor | 1849-1850 |
| Millard | Fillmore | 1850-1853 |
| Franklin | Pierce | 1853-1857 |
| James | Buchanan | 1857-1861 |
| Abraham | Lincoln | 1861-1865 |
| Andrew | Johnson | 1865-1869 |
| Ulysses S. | Grant | 1869-1877 |
| Rutherford B. | Hayes | 1877-1881 |
| James A. | Garfield | 1881 |
| Chester A. | Arthur | 1881-1885 |
| Grover | Cleveland | 1885-1889 |
| Benjamin | Harrison | 1889-1893 |
| Grover | Cleveland | 1893-1897 |
| William | McKinley | 1897-1901 |
| Theodore | Roosevelt | 1901-1909 |
| William H. | Taft | 1909-1913 |
| Woodrow | Wilson | 1913-1921 |
| Warren G. | Harding | 1921-1923 |
| Calvin | Coolidge | 1923-1929 |
| Herbert | Hoover | 1929-1933 |
| Franklin D. | Roosevelt | 1933-1945 |
| Harry | Truman | 1945-1953 |
| Dwight D. | Eisenhower | 1953-1961 |
| John F. | Kennedy | 1961-1963 |
| Lyndon B. | Johnson | 1963-1969 |
| Richard M. | Nixon | 1969-1974 |
| Gerald R. | Ford | 1974-1977 |
| James E. | Carter | 1977-1981 |
| Ronald | Reagan | 1981-1989 |
| George | Bush | 1989-1993 |
| William Jefferson | Clinton | 1993-2001 |
| George | Bush | 2001-2009 |
| Barack | Obama | 2009-2016 |

***Detailed Requirements:***

Data Preparation:

1. Import the Dataset of documents ( each document contains title of the speech, year and text of speech)
2. Tokenize Sentences and Clean

Model Building

1. Build Bigram, Trigram Models and Lemmatize
2. Use any topic modelling technique ( For example LDA – Latent Direchlet Association) to determine the dominant topic and its percentage contribution
3. Across Speeches of a President
4. Across Speeches of a given time period

Visualize Model Outcomes

1. Obtain the following statistics
2. Most Representative sentence for each dominant topic in (A) and (B)
3. Word Cloud of Top 10 keywords for each dominant topic in (A) and (B)

Or

any other kind of visualization (example LDAVis /PyLDAVis) that depicts the Top 10 keywords for each dominant topic in (A) and (B)

**Deliverables**

1. Create an R-code or Python Code Executable depicting the steps above
2. A maximum of 2-page management summary of the outcomes.

( Both R-STUDIO and Jupiter / Anaconda editors used to develop R and Python programs respectively can be downloaded for free)

<https://rstudio.com/products/rstudio/download/>

<https://jupyter.org/install>

<https://www.anaconda.com/products/individual>