A Topic Modeling Approach To The Hillary Clinton Email Controversy

December 10, 2015

Team 9: Dongheon Lee, Michael Lenart, Yifeng Song, Nicholas Venuti

UNIVERSITY of VIRGINIA

Problem Background

- Large amount of political debate around the use of Clinton’s private email server
- Timeline
    - Jan 13, 2009- Clintonemail.com registered/ Clinton beings confirmation hearing for Secretary of State
    - Jan 21, 2009- Clinton confirmed as Secretary of State
    - Aug 2012- U.S. Ambassador to Kenya fired for handling “sensitive but unclassified” on private email account
    - Feb 1, 2013- Clinton leaves role as secretary of state
    - Mar 15, 2013- hacker “Guccifer” exposes hack of Sidney Blumenthal’s email, showing communication with hdr22@clintonemail.com- shows discussions of a number of sensitive foreign policy issues
- Clinton has claimed government has records of conversation as they were forwarded to government emails

Source: https://www.washingtonpost.com/news/fact-checker/wp/2015/03/10/hillary-clintons-emails-a-timeline-of-actions-and-regulations/ Photo/Kevin Lamarque/pool)

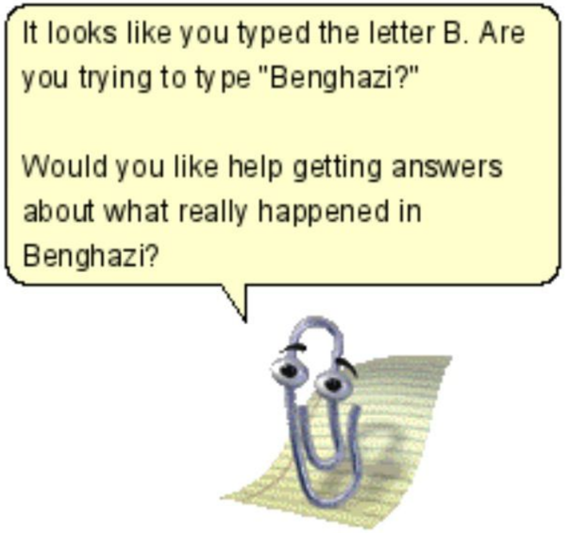
Problem Background- continued

- Kaggle has published a competition to "Uncover the political landscape in Hillary Clinton's emails"
- State Department has released about 7,000 pages of emails
- Kaggle has scraped PDFs, normalized and saved data into a csv file
- **Our goal** - *To substantiate or refute her claims through Latent Dirichlet Analysis (LDA) and K-means clustering*

Source:
AP Photo/Kevin Lamarque/pool

Previous Approaches

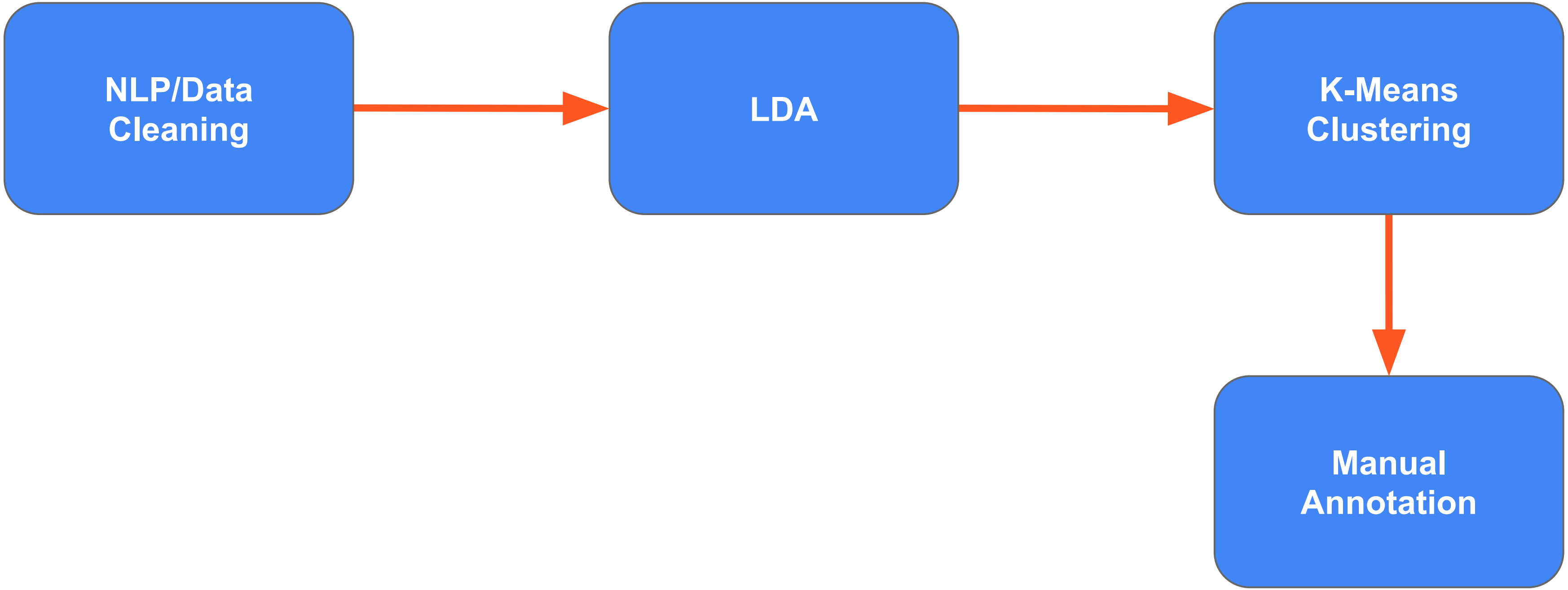
- Kaggle
  - Mostly data visualizations- word clouds, email frequency analysis
- Commercial/Government
  - Manually vetted, but only publications have been political sound bytes


It looks like you typed the letter B. Are you trying to type "Benghazi?"

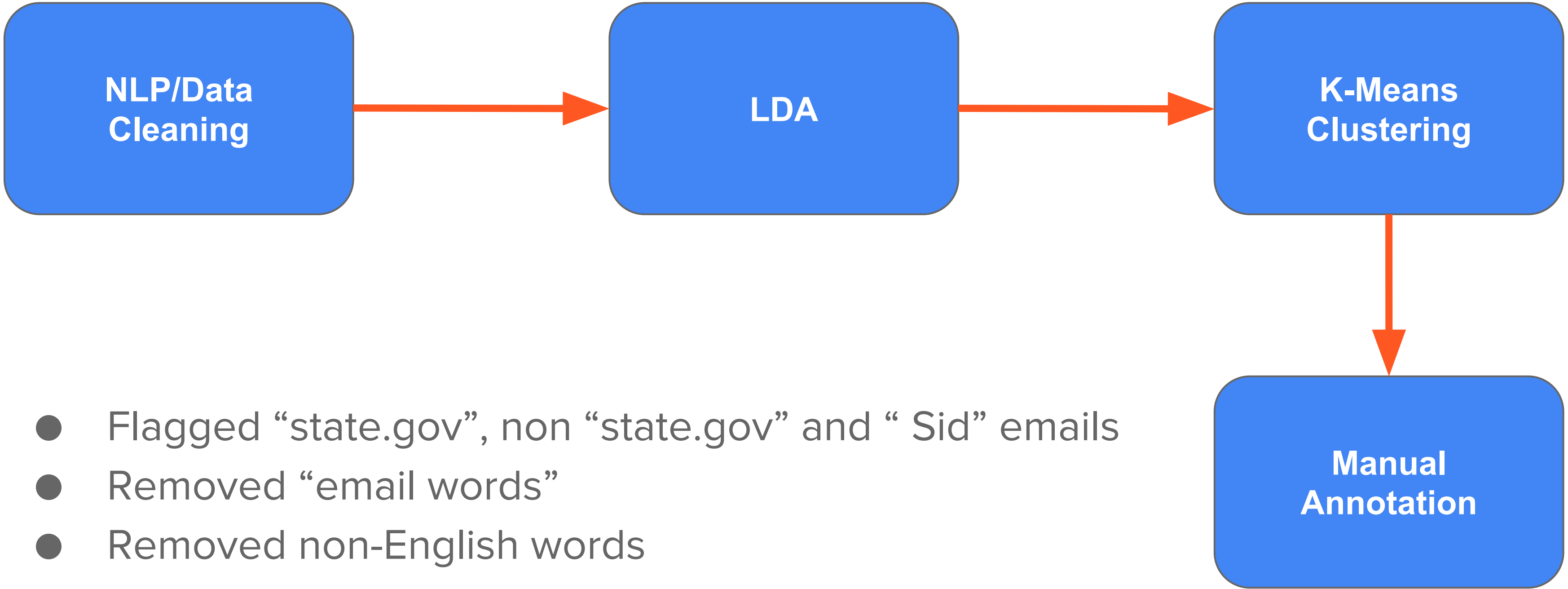
Would you like help getting answers about what really happened in Benghazi?

Source: *twitter.com/lawblob*

Approach- Overview

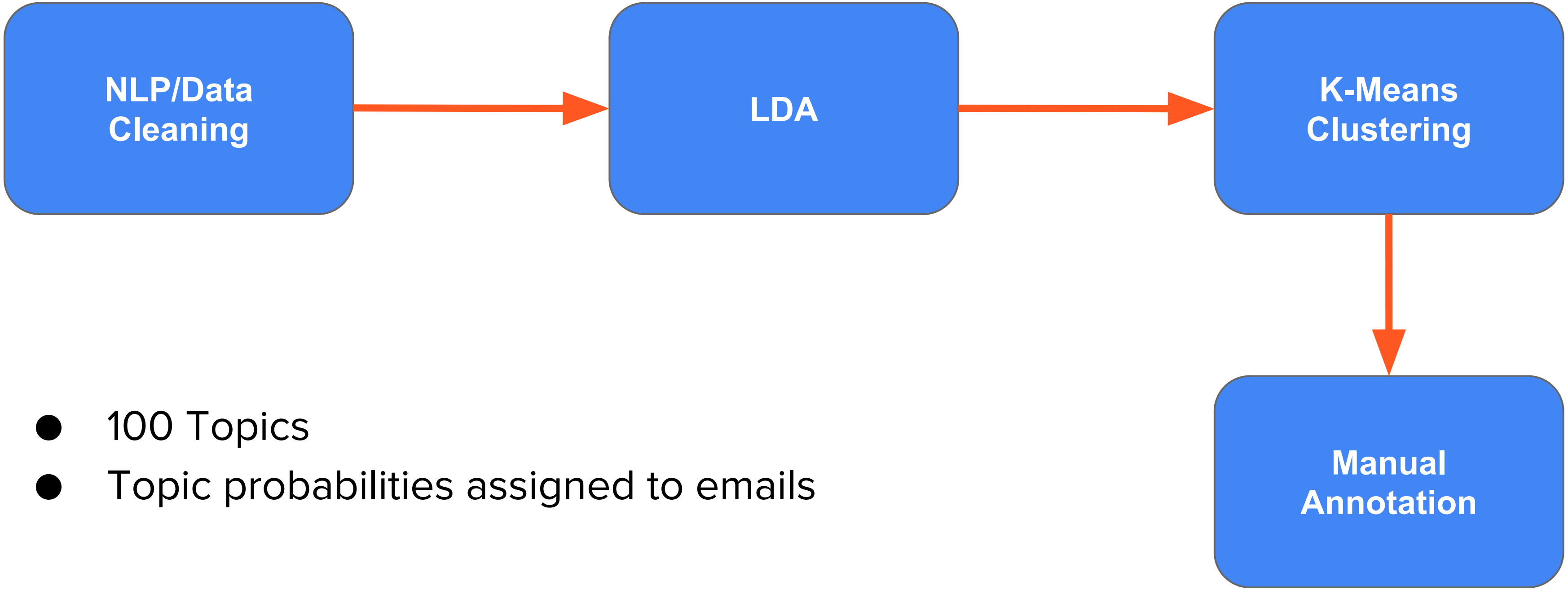


Approach: NLP/Data Cleaning



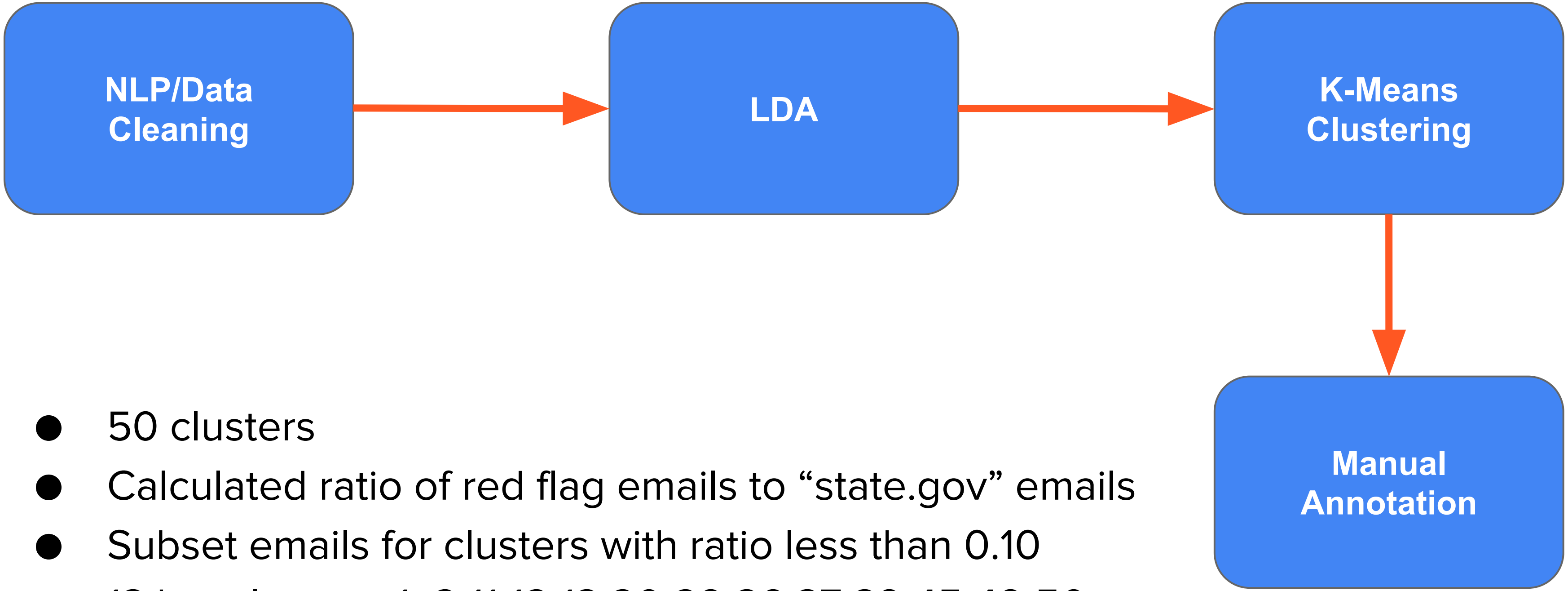
- Flagged “state.gov”, non “state.gov” and “ Sid” emails
- Removed “email words”
- Removed non-English words
- DTM reduction: stemming, converting to lowercase

Approach: LDA



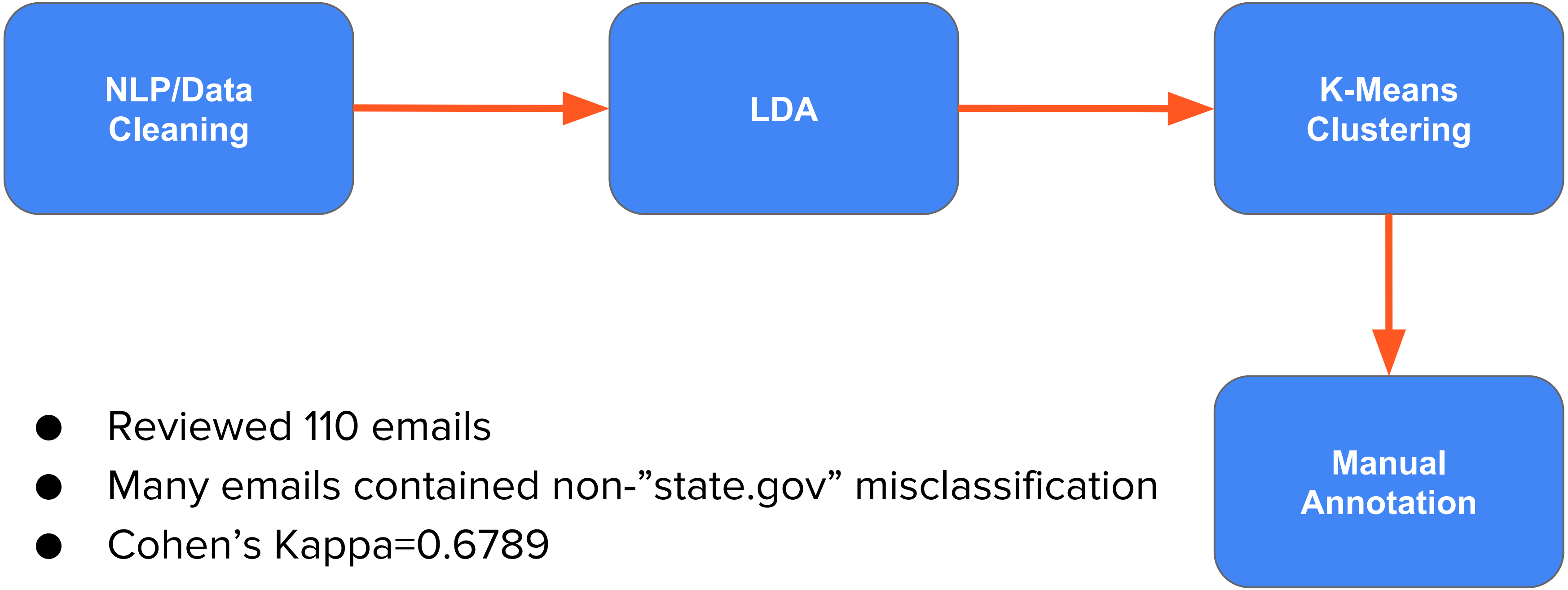
- 100 Topics
- Topic probabilities assigned to emails

Approach: K-Means Clustering



- 50 clusters
- Calculated ratio of red flag emails to “state.gov” emails
- Subset emails for clusters with ratio less than 0.10
- 13 hot clusters: 1 2 11 12 13 20 29 36 37 39 45 49 50
- Identified 110 red flag emails requiring manual vetting

Approach: Manual Annotation



- Reviewed 110 emails
- Many emails contained non-”state.gov” misclassification
- Cohen’s Kappa=0.6789

Results

- Only 110 out of ~8,000 emails required annotation
- 18 emails identified as potentially non-compliant
  - Example email: 7668 - intelligence from former Foreign Minister/Vice Chancellor of Germany from 1998-2005

Source: http://www.thecommentator.com/system/articles/inner_pictures/000/006/105/original/Hillary_clinton_LATEST.jpg?1444470793

Email 7668

For: Hillary

From: Sid

Re: Iran, Saudi

Had dinner last night (Tuesday, February 16) with Joschka Fischer. We had an interesting conversation on Iran and Saudi Arabia, among other things. (As you know Fischer is now director of the Nabucco pipeline project.) On Iran, harsh, targeted sanctions are absolutely necessary, but are most effective diplomatically when always coupled with an offer to negotiate. The iron fist in the velvet glove approach achieves several objectives:

According to Fischer's intelligence, Ahmanijehad wished some negotiated settlement but was blocked. The regime has splits at the top. Perhaps true, perhaps not. But constantly pushing negotiations alongside sanctions puts additional pressure on internal divisions, whatever they are. Extending an open hand while brandishing a stick closes diplomatic and political room to maneuver for Iran: Its refusal to accept the open hand justifies application of the stick. Even when sanctions are enforced it always remains useful to say another way is open. The damage done to Iran is therefore the result of its own choice. This approach also aids the opposition. A purely condign sanctions strategy can contribute to the regime's will to punish and tighten repression. Talking of regime change, of course, undermines the cause of regime change. It is a gift to the regime. The opposition is a new factor in the Iran equation that must be taken into account on the political and moral level. Pushed to the wall, the regime may feel compelled to repress, which might involve thousands or tens of thousands of political killings. On Saudi Arabia, Fischer points out that if Iran develops nuclear weaponry the Saudis already have their own bomb. The Saudis invested in Pakistan's nuclear weaponry partly for this eventuality; that's their bomb in reserve.

Shiny Demo:

https://mdlenart.shinyapps.io/my1stApp

Future Actions

- Perform additional text cleaning
- Parallelize the data preprocessing and modeling steps
- Increase threshold ratio
- Attempt Author-Topic modeling

Questions?