Text Mining: Twitter and the Problem With Spam Bots

By: Danny Cohen

Agenda

- Introduction
- Problem Description
- Background
- Methodology
- Assumptions
- Experimental Design
- Results
- Issues
- Conclusion

Introduction

- Senior
- Majoring in Computer Science
  - Emphasis in IT
- Minoring in
  - Math
  - Science
  - Theatre
- Interest in data mining stems from internship

Problem Description

- Spam bots in social media
- Estimated number of spam bots on twitter
    - 15% of total userbase
    - Almost 48 million "users"
- Different types of spambots
    - Useful
    - Harmless
    - Malicious

Background

- Text Mining in relation to Twitter
- Common methods of twitter text mining
  - Archives
  - Individual tweets
  - Account information
  - Searching tags or phrases

Methodology

- Analyzed 6 spam bots
  - 5 from different areas of interest
  - 1 which overlaps with another
  - Comparison of
    - Account info
    - Favorites
    - Recent tweets
- Analyzed 5 real people
  - 5 to be compared against each other
  - Comparison will be the same as spam bots

Assumptions

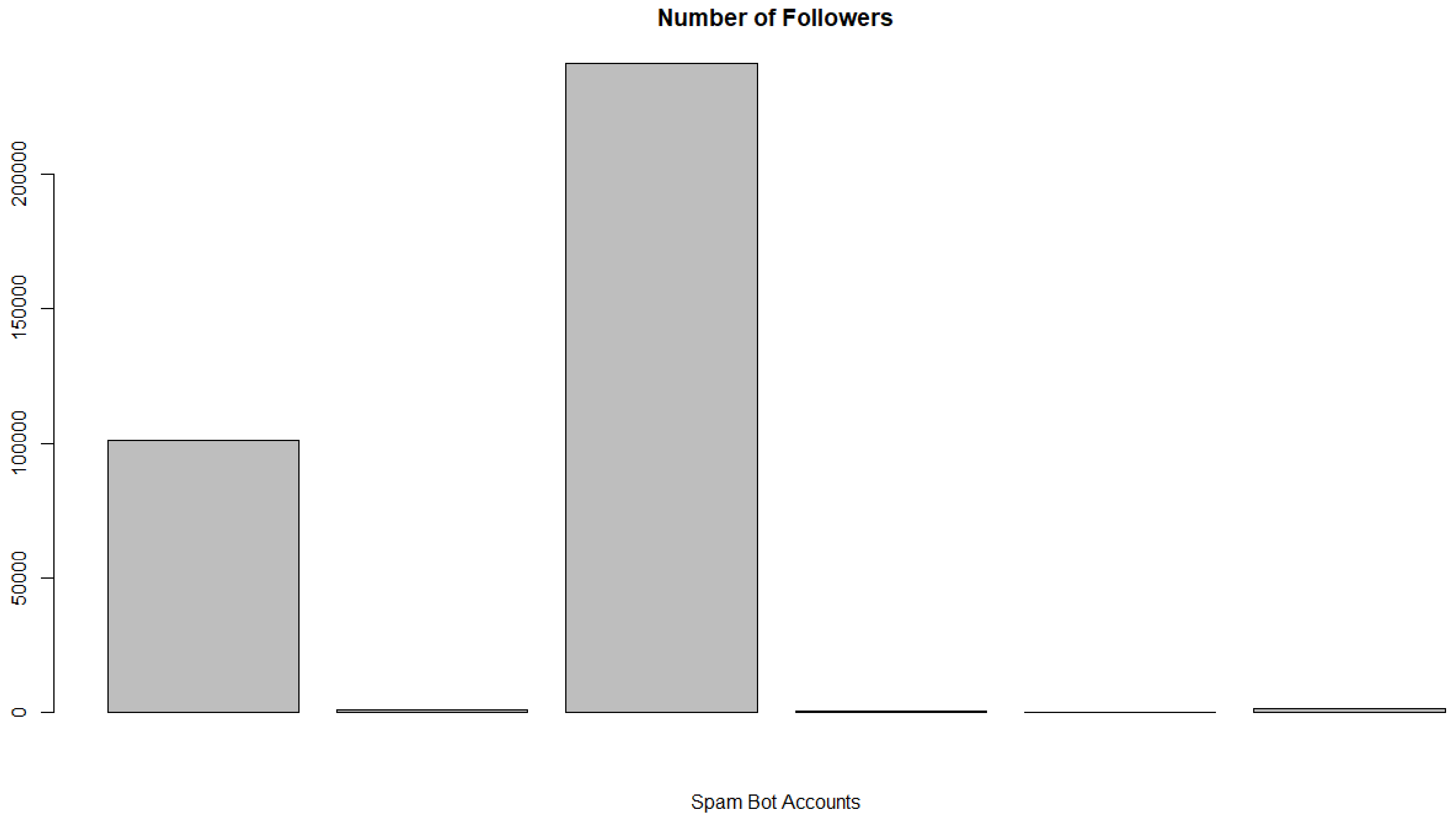
- Spam bots are easily identifiable
- All accounts are American
- Twitter must be functional

Experimental Design

- Locate all 6 spam bot accounts
- Collect their information through R
  - "twitteR" package
- Collect all 5 real individuals
- Obtain graphs and compare data

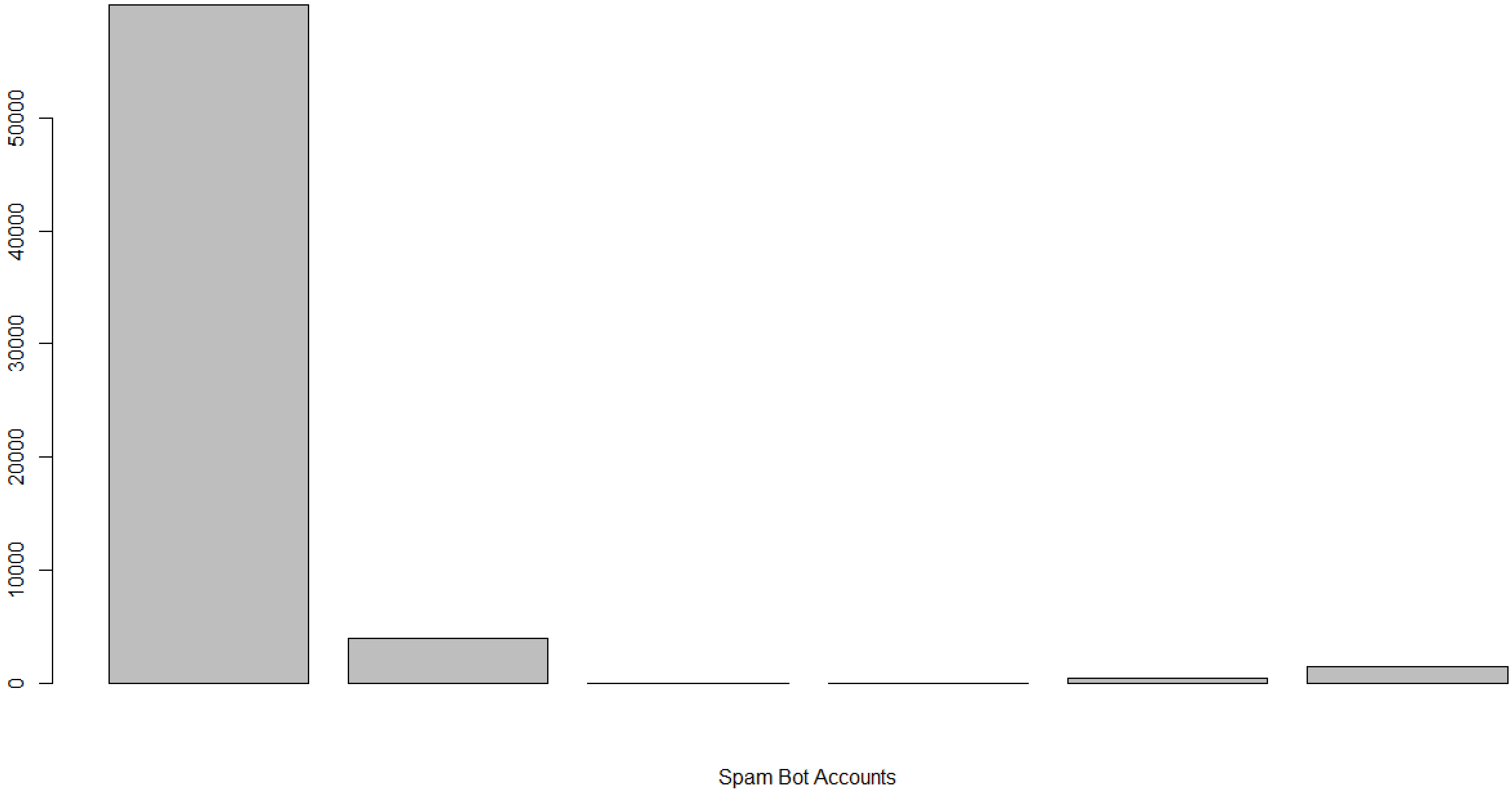
Results

- Spam Bot Followers



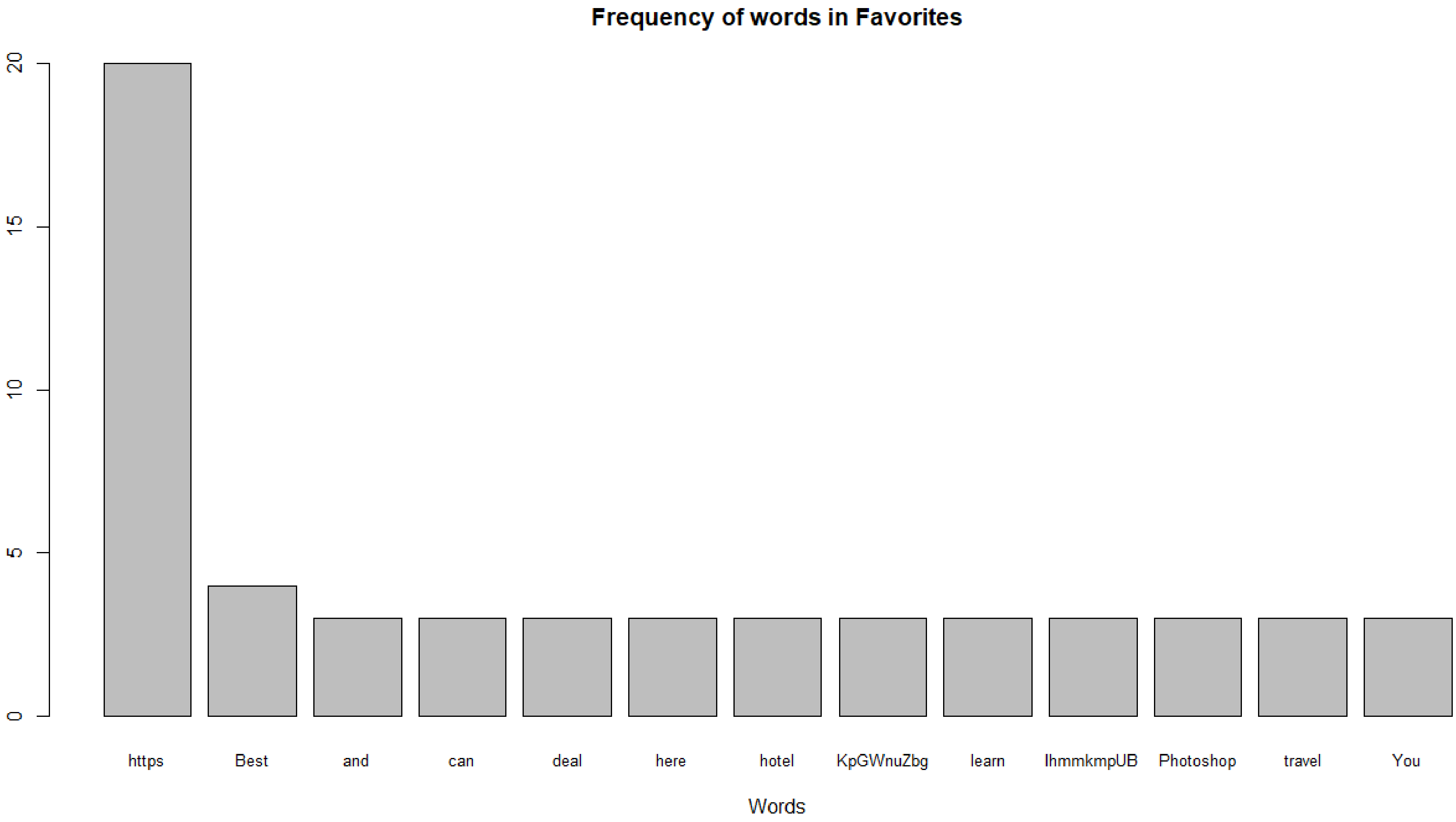
Results, cont.

- Spam Bot Followed accounts



Results, cont.

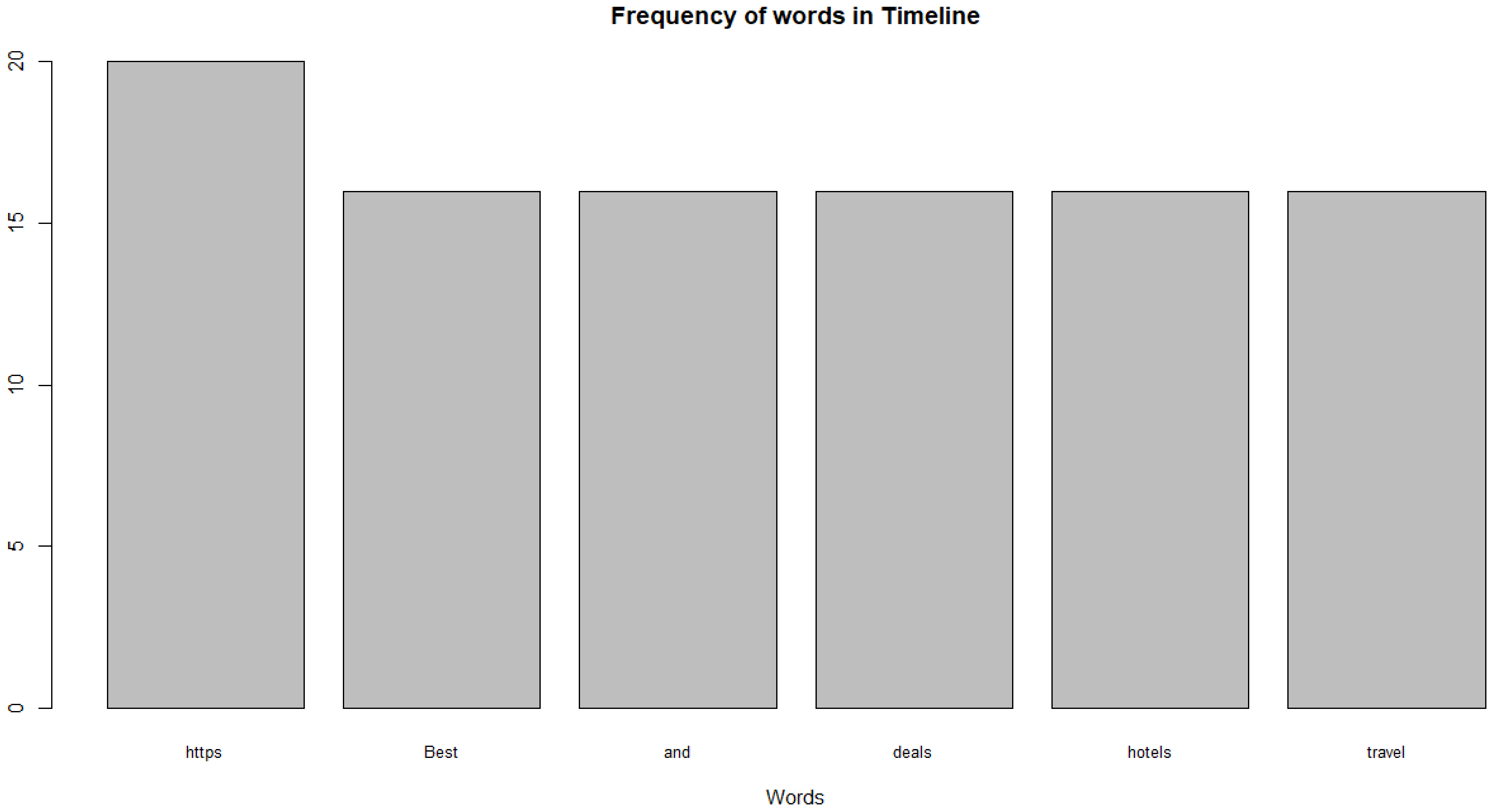
- Spam Bot Favorites
  - More conclusive
  - Patterns of favorites
  - No overlap amongst each account’s favorites



Frequency of words in Favorites

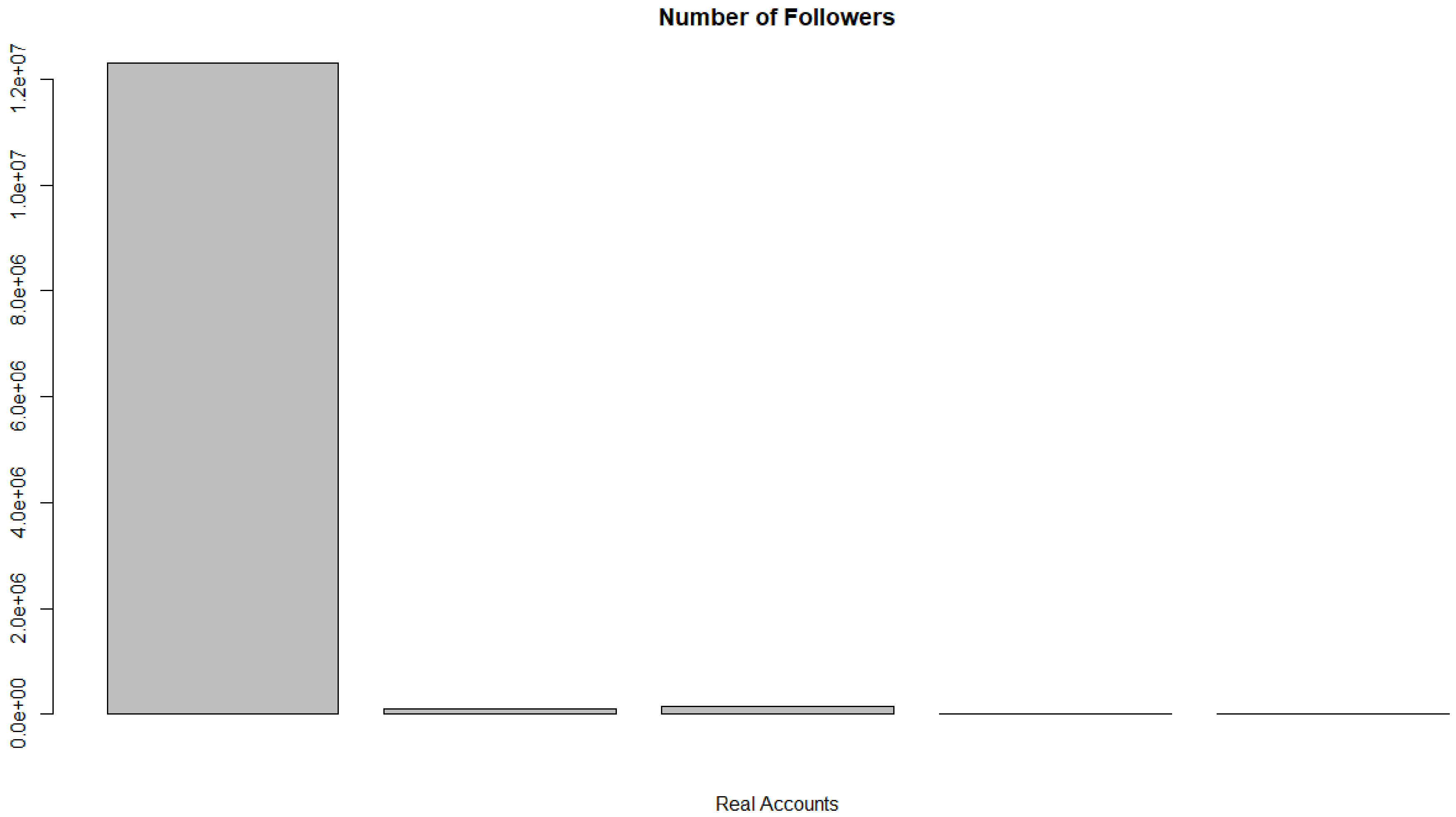
Results, cont.

- Spam Bot Timelines



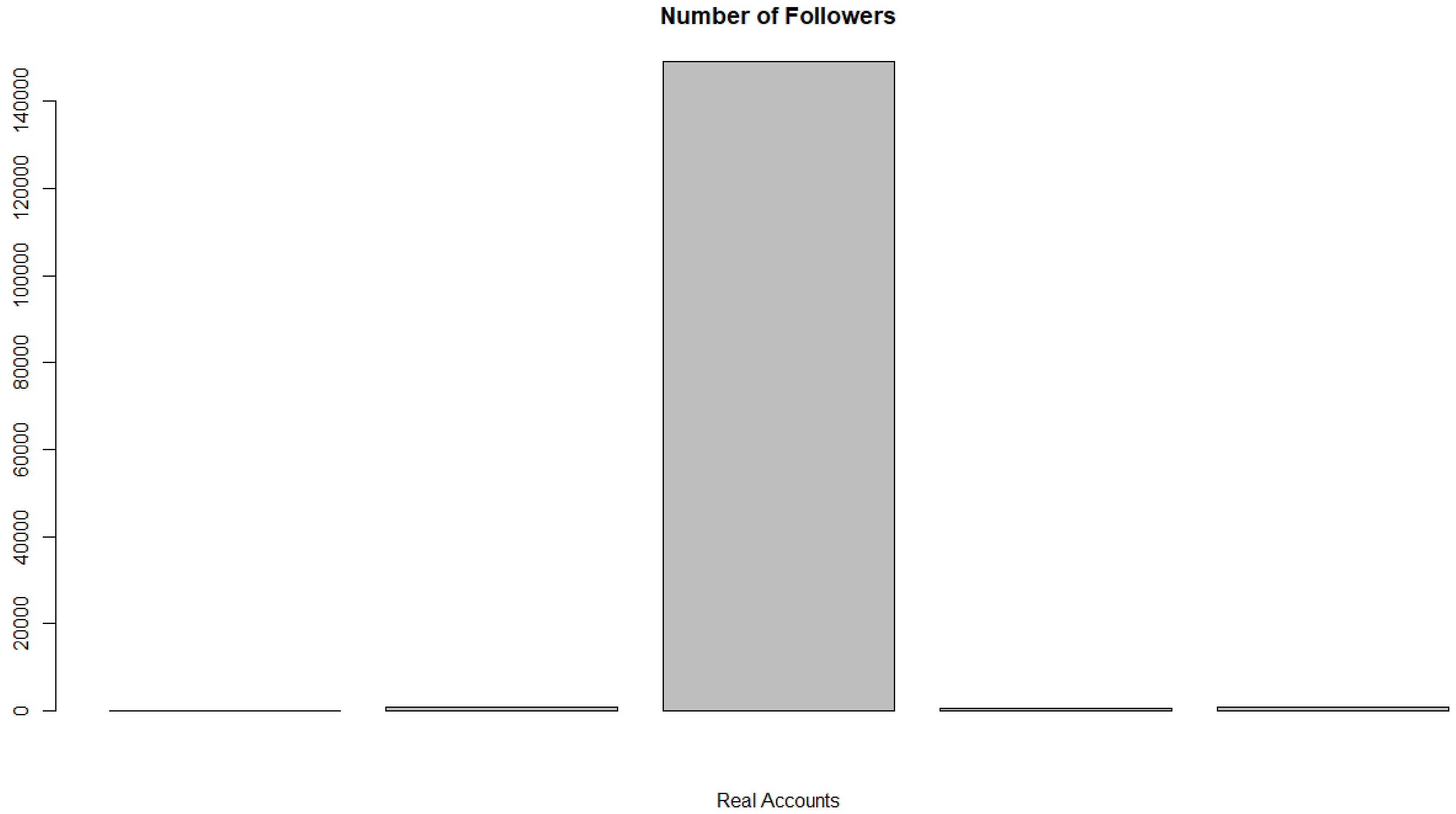
Results, cont.

- Real accounts' number of followers



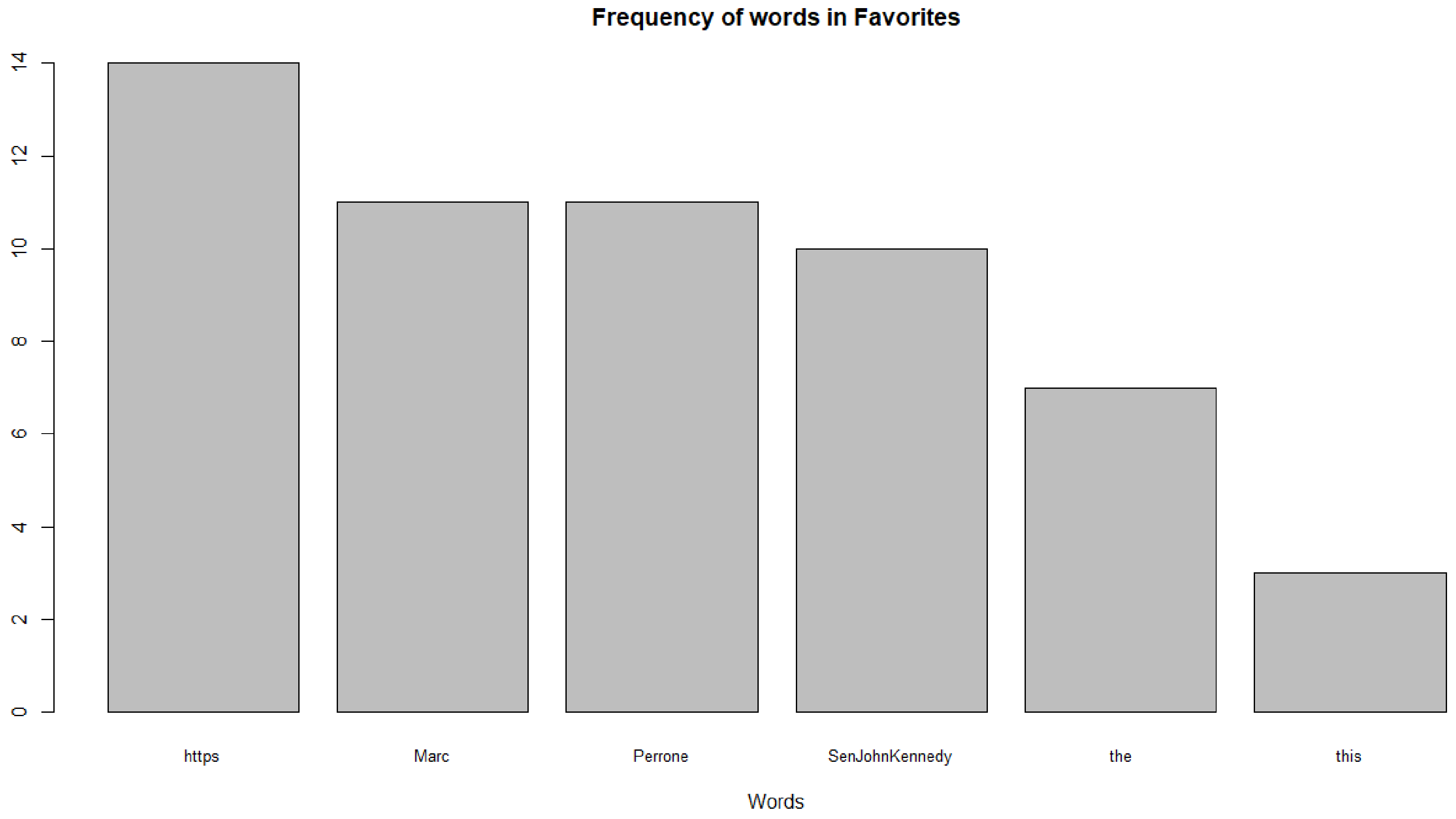
Results, cont.

- Real accounts’ followed accounts



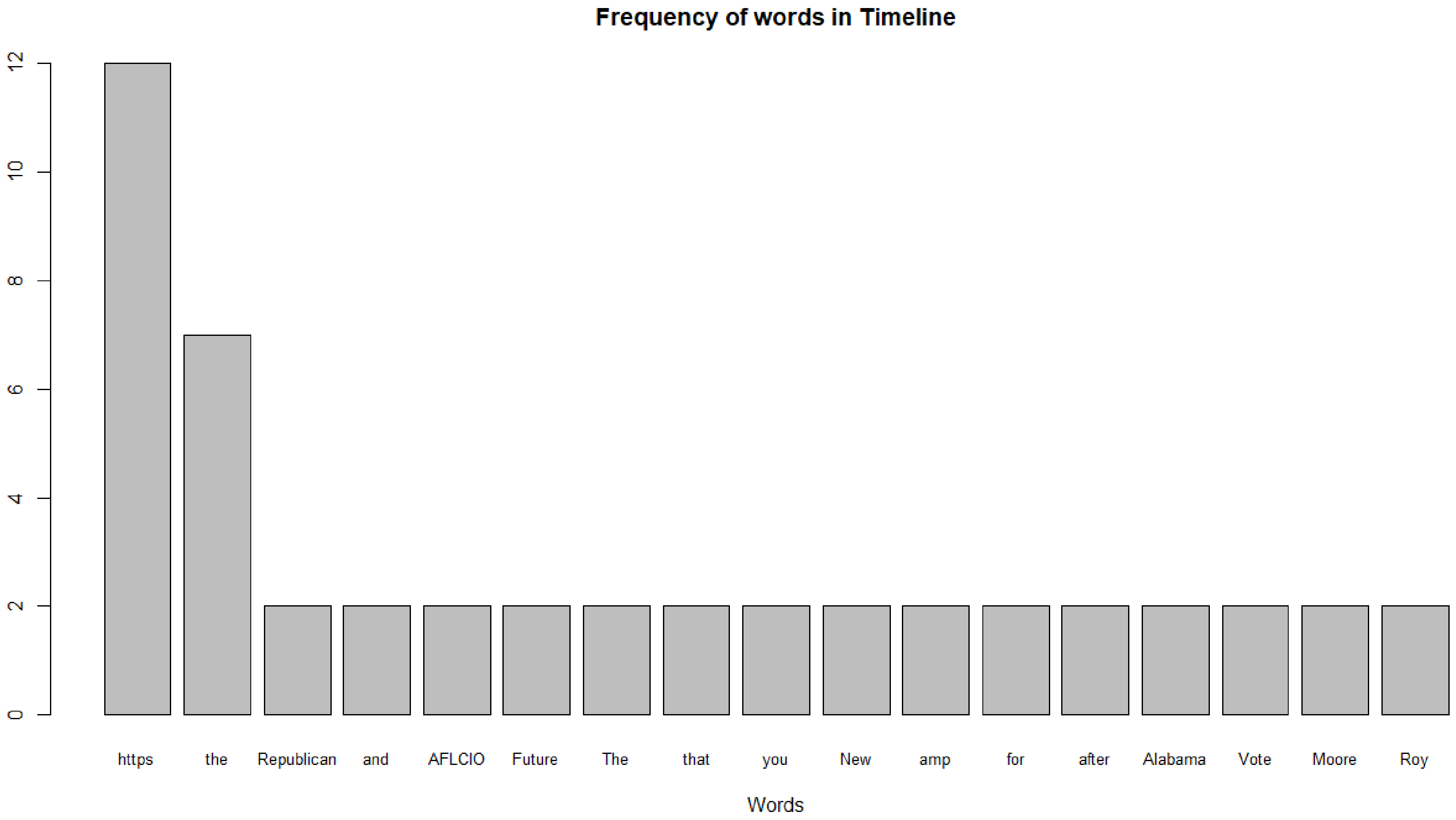
Results, cont.

- Real Individual's favorites & word frequency



Results, cont.

- Real individual's timeline



Issues

- Certain users lacked original tweets
- Data from popular accounts could have potentially shifted
- Age of spam bot accounts

Conclusion

- Largest differences
  - Constant tweeting of links
  - Heavy use of specific words such as “free” or “cheap”
- Recommended approach
  - Identification of Spam Bots
  - Avoid automated deletion
  - Potential for logging associated accounts
- Future Work
  - “Useful” spam bots

Bibliography

- Newberg, M. (2017, March 10). As many as 48 million Twitter accounts aren't people, says study. Retrieved December 5, 2017, from https://www.cnbc.com/2017/03/10/nearly-48-million-twitter-accounts-could-be-bots-says-study.html
- Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017, March 27). Online Human-Bot Interactions: Detection, Estimation, and Characterization. Retrieved December 5, 2017, from https://arxiv.org/pdf/1703.03107.pdf