Assignment 6 – Text Modeling and Analysis Using Quanteda

Course: EPPS 6326 / GISC 6323 – Data Collection and Production
Prepared by: Alfa Victor Lugard
PhD Student, University of Texas at Dallas

This report presents text modeling and analysis using the R package Quanteda, focusing on Twitter data surrounding the 2021 Biden–Xi Summit, and includes applications of Latent Semantic Analysis, Wordfish, and Correspondence Analysis.

1. Introduction

This assignment demonstrates computational text analysis using Quanteda in R. The dataset used consists of Twitter discussions on the Biden–Xi Summit in November 2021. The workflow includes preprocessing, Latent Semantic Analysis (LSA), and text network visualization of hashtags and mentions, followed by advanced scaling methods (Wordfish and Correspondence Analysis).

2. Tokenization and Document-Feature Matrix

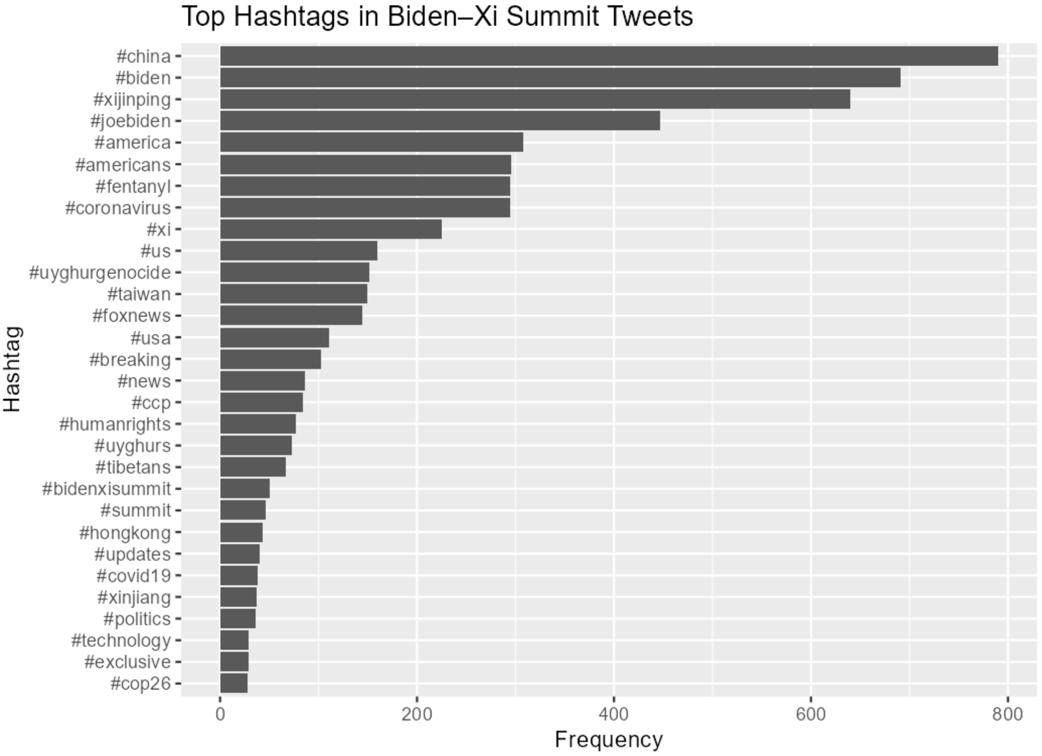
Tweets were tokenized, stopwords removed, and a document-feature matrix (DFM) created. The most frequent terms included 'biden', 'xi', 'summit', and 'china'. These tokens represent the main discussion topics during the summit.

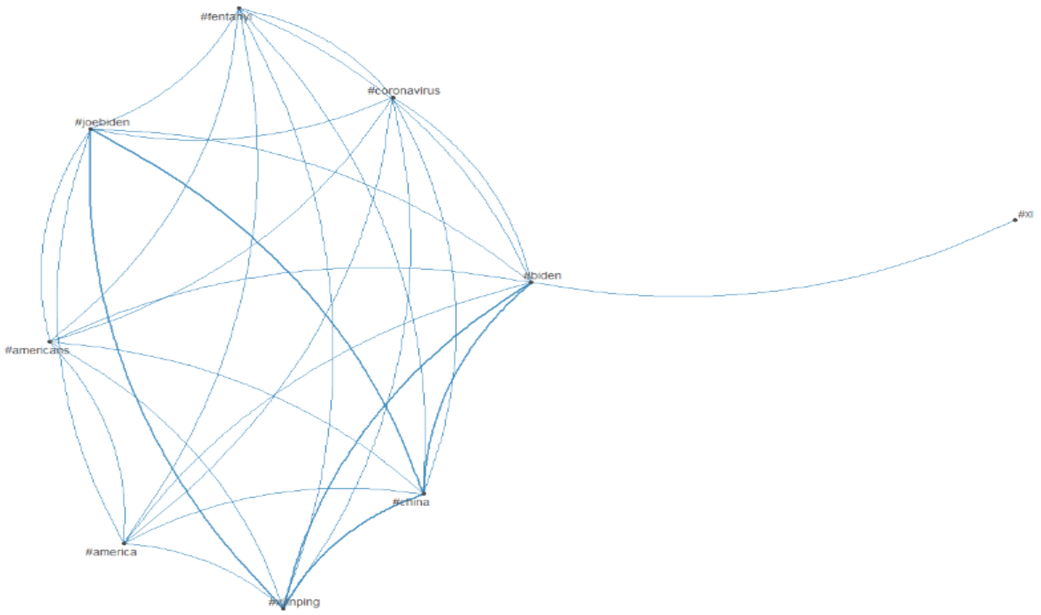
3. Latent Semantic Analysis (LSA)

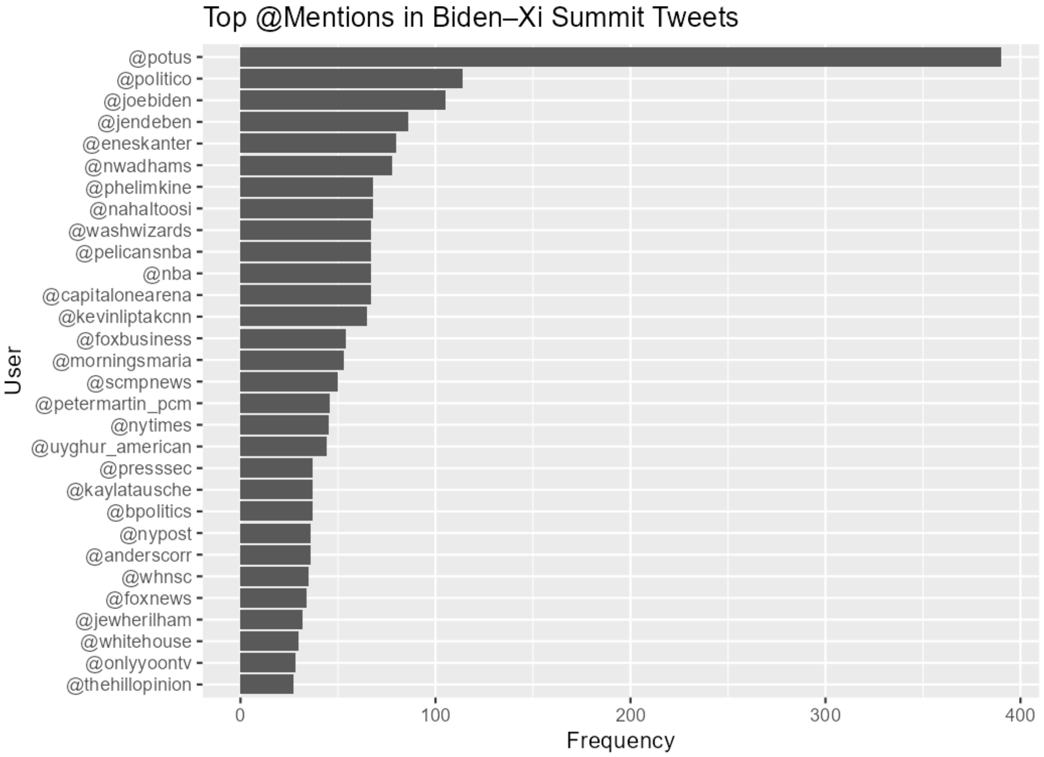
LSA was used to uncover latent semantic structures among words by reducing dimensionality. This helps identify underlying contexts and co-occurrence relationships between terms that may not appear directly connected in raw frequency counts.

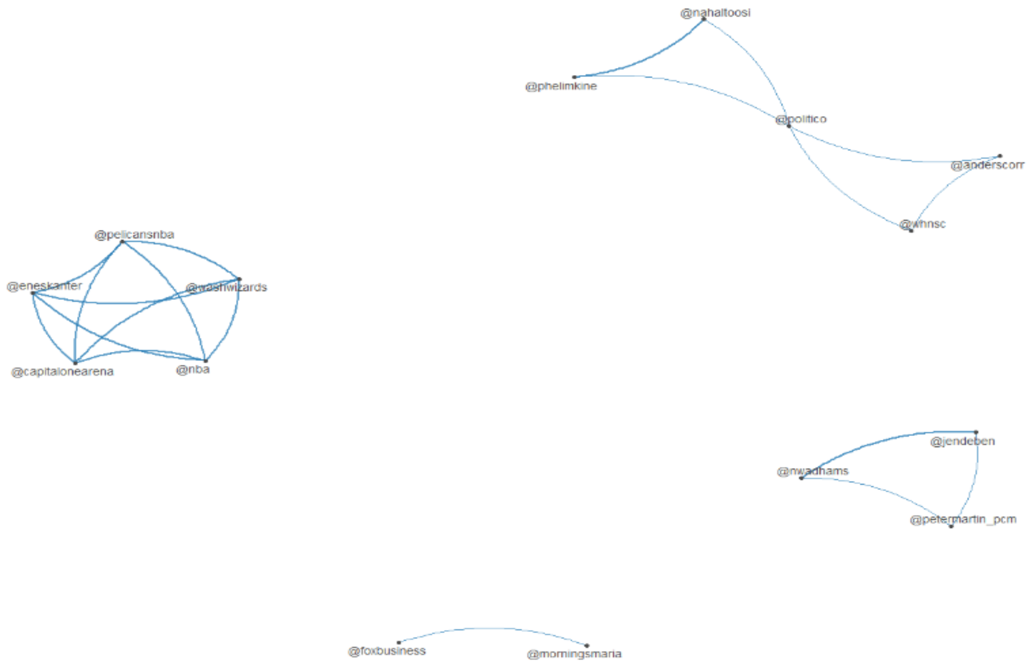
4. Hashtag and Mention Analysis

This section highlights the most common hashtags and user mentions. The visualizations below show both bar charts and co-occurrence networks, demonstrating communication clusters around political figures, media outlets, and discussion themes.



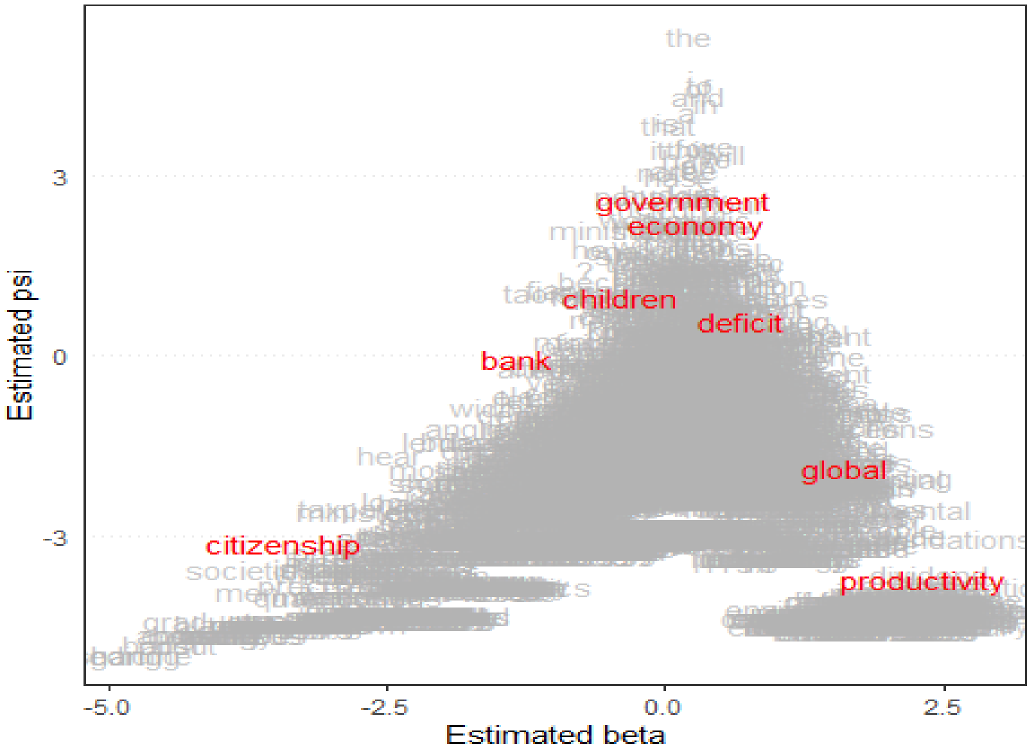


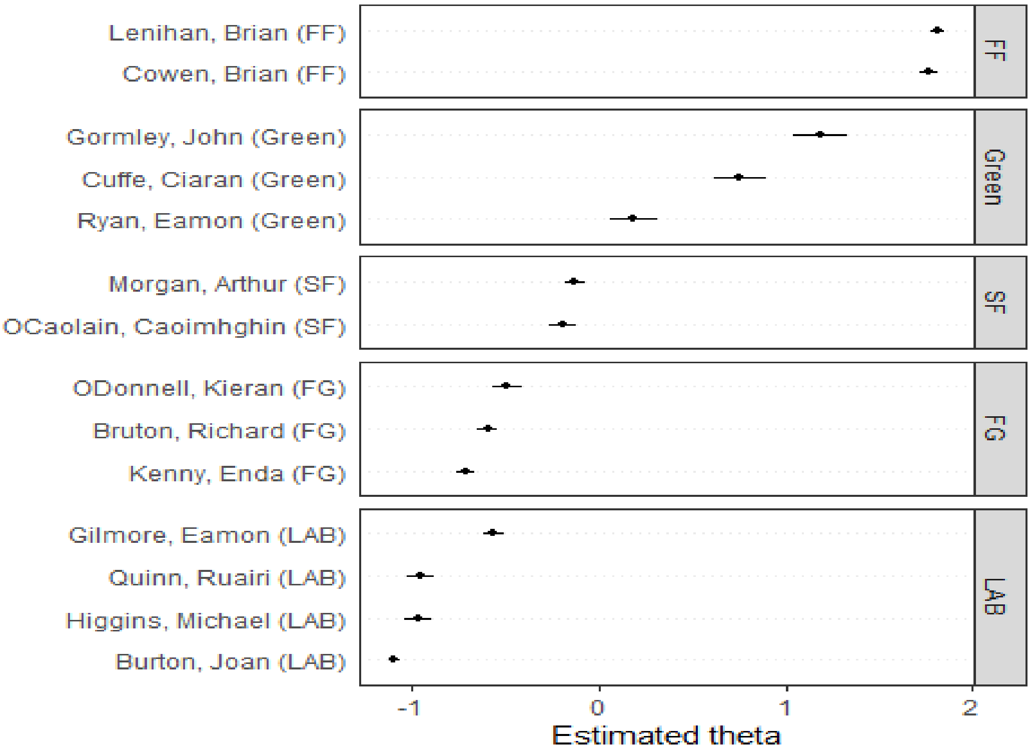

Top @Mentions in Biden–Xi Summit Tweets

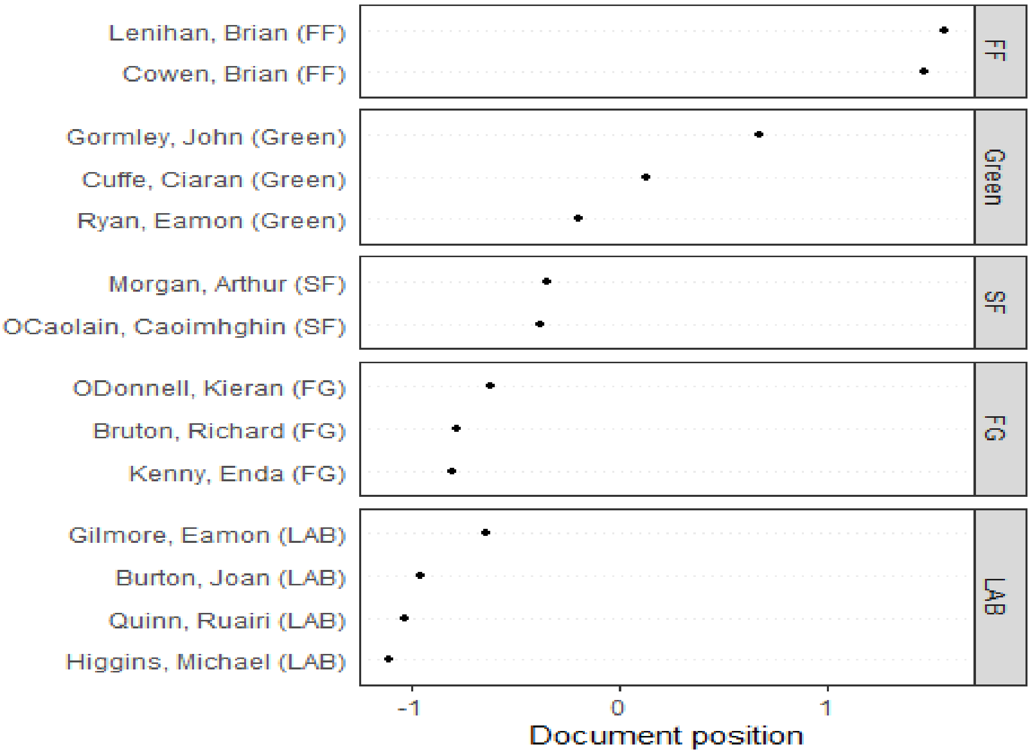

@nahaltoosi
@phelimkine
@politico
@anderscorr
@whnsc
@pelicansnba
@eneskanter
@washwizards
@capitalonearena
@nba
@jendeben
@nwadhams
@petermartin_pcm
@foxbusiness
@morningsmaria

5. Text Scaling Models: Wordfish and Correspondence Analysis

Using the built-in dataset `data_corpus_irishbudget2010`, we estimated document and word positions with the Wordfish model. The analysis reveals ideological separation between political parties—Fianna Fáil (FF) members lean more positively, while Labour (LAB) members lean negative. The figures below visualize estimated word positions, document positions by party, and a correspondence analysis (CA) projection.




Lenihan, Brian (FF)
Cowen, Brian (FF)
Gormley, John (Green)
Cuffe, Ciaran (Green)
Ryan, Eamon (Green)
Morgan, Arthur (SF)
OCaolain, Caoimhghin (SF)
ODonnell, Kieran (FG)
Bruton, Richard (FG)
Kenny, Enda (FG)
Gilmore, Eamon (LAB)
Quinn, Ruairi (LAB)
Higgins, Michael (LAB)
Burton, Joan (LAB)
FF
Green
SF
FG
LAB
-1
0
1
2
Estimated theta


Lenihan, Brian (FF)
Cowen, Brian (FF)
Gormley, John (Green)
Cuffe, Ciaran (Green)
Ryan, Eamon (Green)
Morgan, Arthur (SF)
OCaolain, Caoimhghin (SF)
ODonnell, Kieran (FG)
Bruton, Richard (FG)
Kenny, Enda (FG)
Gilmore, Eamon (LAB)
Burton, Joan (LAB)
Quinn, Ruairi (LAB)
Higgins, Michael (LAB)
FF
Green
SF
FG
LAB
-1
0
1
Document position

6. Discussion and Conclusion

This assignment demonstrates how Quanteda enables efficient text preprocessing, modeling, and visualization. The hashtag and mention analyses revealed dominant actors and topics in global political discourse, while Wordfish and Correspondence Analysis provided quantitative insights into ideological and thematic dimensions of text corpora. Such methods bridge qualitative political analysis with reproducible quantitative techniques.