Web Scraping & Text Analysis

Background

For this report, I selected two politicians for the analysis of their Twitter feeds: Stephen Morgan, the Labour MP for Portsmouth, and Royston Smith, the Conservative MP for Southampton. These politicians were chosen due to their opposing party affiliations and because they represent constituencies with several similarities. Both Portsmouth and Southampton are port cities located in Hampshire, boasting similar population sizes, universities, and strong ties to the British Armed Forces. This presents an intriguing opportunity to compare the Twitter usage of these MPs and ascertain if it reflects the resemblances between the cities.

Summary of Politicians tweets

I collected the most recent 3200 tweets (January 2021) from each politician's Twitter feed. Subsequently, I cleaned the metadata from the two datasets and transformed them into data framework models to facilitate further analysis.

Comparing Key-Word in context

Here are tables exported containing precedent and subsequent words surrounding the chosen keyword to explore “covid”, to assess similarities. Here the first 10 most recent have been copied from the exported table…

Royston Smith COVID tweets		
pre	keyword	post
have died from	#COVID19	STAY AT HOME
been vaccinated against	COVID-19	. That is
role in controlling	#COVID19	transmission in educational
in joining the	COVID-19	vaccination programme ,
out of the	#COVID19	vaccine in #Southampton
filling up with	Covid	patients and we
expected due to	#Covid	. . .
Government from the	Covid	Winter Grant Scheme
than ever .	#Covid19	means this year's
Your	Covid	Recovery is a

Stephen Morgan COVID tweets		
pre	keyword	post
new strain of	#COVID19	spreads very fast .
the spread of	#COVID19	https://t.co/b4wYwruC8c
part of the	Covid	Support Force mission
they're helping		response :
the update	#Covid19	https://t.co/4tNyGNFVw8
on border	Covid	security https://t.co/qsf1zwDtXh
, caring for	Covid-19	patients . She
moments of the	Covid-19	crisis . It
under strain before	Covid-19	struck . There
week due to	#COVID19	. Todayâ €
' to tackle	covid	. Their new

Stephen Morgan's COVID Tweets:

Stephen Morgan's Twitter feed contained 249 tweets that included the pattern "covid," while Royston Smith had just 25 such tweets. We conducted a keyword-in-context analysis by examining the text surrounding the keyword "covid" in both MPs' tweets. Notably, Stephen Morgan appeared to be more active in tweeting about COVID-19, with his tweets often referencing the new strain of the virus.

Royston Smith's COVID Tweets:

Royston Smith's tweets about COVID-19 were comparatively fewer. The context of his COVID-related tweets suggested a focus on topics such as vaccination, control measures, and government initiatives related to the pandemic.

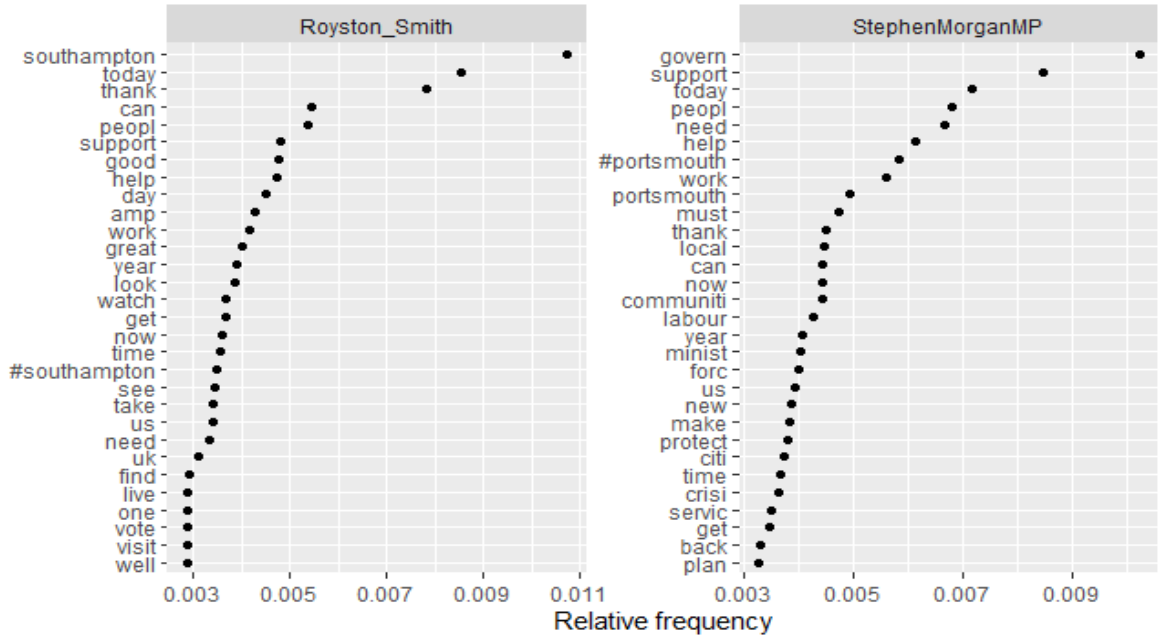
Term Frequency

I calculated term frequencies using Inverse Document Frequency Weighting, and further data cleaning was performed to remove symbols that had not been previously removed. To visually represent the most frequently occurring terms in the tweets of both MPs, I generated word clouds based on the term frequency models.

Royston Smith

Stephen Morgan

These word clouds provide insights into the most frequently used terms in the politicians' tweets. Royston Smith's word cloud may suggest a focus on topics related to his constituency, with "Southampton" being the most common term. In contrast, Stephen Morgan's word cloud suggests a greater emphasis on government-related discussions. To delve deeper into these patterns, a word frequency plot can be further explored.

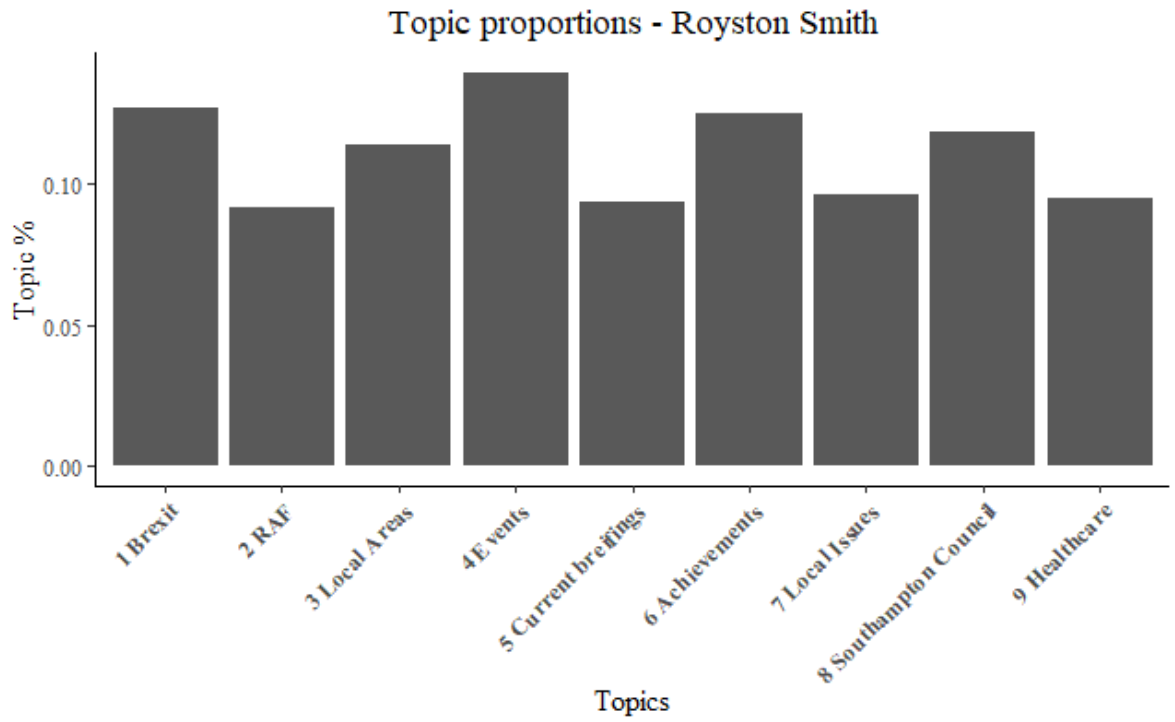


Text Analysis - Topic Modelling

I opted for topic modelling to analyse the politicians' tweets. Given the similarities in their constituencies, it was intriguing to investigate whether the topics they frequently tweeted about were comparable. Moreover, if common topics were identified, I sought to ascertain if the frequencies at which they tweeted about these topics were similar. Topic modelling utilizes a text mining algorithm based on the Latent Dirichlet Allocation (LDA) statistical model. This algorithm organizes words in the Twitter feed data to uncover prevalent themes. I applied this approach to the data frames (dfm's) of the politicians, and I assigned topic names based on the top words, cross-referencing them with the actual tweets to validate the naming choices.

Results – Royston Smith

In the topic analysis of Royston Smith's tweets, I identified nine distinct topics. The frequency of these topics is detailed in a bar chart, along with the top 20 words associated with each topic. Additionally, I provided example tweets that represent three of these topics:



Top 20 words for each topic

	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9
1	vote	spitfir	#southampton	thank	watch	good	treatment	thank	can
2	eu	year	amp	meet	evid	well	port	ve	gp
3	referendum	serv	join	cup	uk	thank	injuri	busi	find
4	leav	raf	morn	look	committe	congratul	minor	view	help
5	deal	sinc	bittern	great	take	money	head	council	need
6	brexit	arm	local	royston	read	great	urgent	thing	appoint
7	want	supermarin	#hampshir	#saintsfc	live	excel	hospit	itchen	practic
8	peopl	icon	lot	pleasur	affair	rais	litter	peopl	get
9	countri	forc	sunday	see	now	work	wait	level	day
10	prime	day	librari	fascin	intern	top	pick	colleagu	weekend
11	say	tribut	communiti	work	select	fantast	number	long	visit
12	poll	mark	nice	amp	report	luck	can	wonder	support
13	better	aircraft	fayr	forward	session	polic	help	email	even
14	govern	ago	us	togeth	2.45pm	achiev	collect	good	pharmaci
15	parliament	design	come	host	foreign	school	complet	small	#southampton
16	one	birthday	day	st	inquiri	award	week	run	close
17	promis	home	clean	visit	publish	citi	bin	mp	know
18	decis	men	cobbett	staff	relationship	worth	increas	walk	websit
19	agre	pilot	park	group	ask	fund	trade	much	cancer
20	democraci	#raf100	along	attend	polici	done	maritim	face	open

Example tweets of the 3 of the topics are presented.

Topic 4 (13%) – Expressions of appreciation

```
RSmithMP.csv.10
"A big thank you to @SMcPartland on @bbc5live today, with
@TVNaga01, saying categorically, victims should not be paying for
historic fire safety defects that are not our fault.\n\nWe urge
all MPs to please get behind him &amp;amp; @Royston_Smith.\nPlease
sign the McPartland Smith Amendment! https://t.co/GOYPyEUVIX"
                                RSmithMP.csv.23
"Proud to be British. https://t.co/ns32cVuBhr"
```

Topic 1 (12%) - Brexit

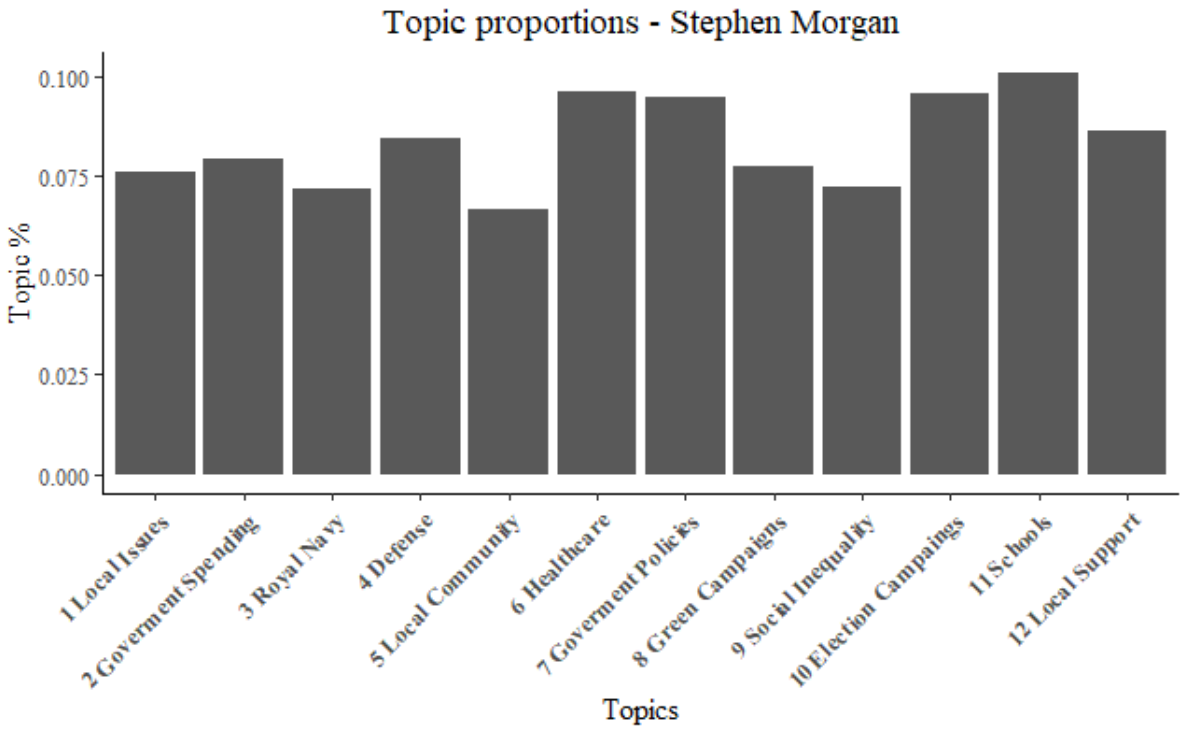
```
RSmithMP.csv.44
"Britain has just become a fully independent country again -
deciding our own affairs for ourselves.\n\nThank you to everyone
who worked with me &amp;amp; @BorisJohnson to get us here in the last
18 months.\n\nWe have a great future before us. Now we can build a
better country for us all."

RSmithMP.csv.75
"A year since GE19. After so much hard work in the dark and the
rain, we got @Royston_Smith elected and finally got #Brexit done!
\n\n2020 hasn't since gone quite as expected due to #Covid... but
I am always hugely proud that this honourable man is my
representative in Parliament. https://t.co/bDXdw1IKvD"
```

Topic 2 (9%) – Royal Air Force

```
RSmithMP.csv.33
"The Supermarine Southampton, designed by R. J. Mitchell and his
team at Supermarine in Woolston. They later designed and built the
Spitfire there.\n\nThe first flight was in 1925. They were used by
the RAF and civil operators, and they were sold to countries all
around the world. https://t.co/k2DwQbAhJc"

RSmithMP.csv.59
"The Royal Air Force has brought some Christmas cheer to a 98-
year-old Second World War Spitfire pilot after learning that
thieves had stolen from his home a cherished photo of him flying
the iconic aircraft.\n\nFull story: https://t.co/LMbU4wLcmU
https://t.co/23gM2QZ9Zy"
```

Results – Stephen Morgan

In the topic analysis for Stephen Morgan's tweets, I identified twelve topics. I presented a bar chart displaying the frequency of these topics and a table listing the top words associated with each topic. Moreover, I included example tweets that illustrate each topic:



	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9	topic10	topic11	topic12
1	join	cut	royal	forc	thank	nhs	busi	climat	equal	labour	school	communiti
2	view	crisi	ship	arm	#southsea	worker	job	citi	women	vote	test	good
3	link	council	sea	personnel	donat	protect	must	polic	action	portsmouth	minist	local
4	amp	year	hms	serv	fe0f	care	crisi	council	must	south	now	work
5	messag	fund	carrier	global	can	staff	economi	plan	day	back	trace	see
6	survey	care	base	bill	love	work	now	make	just	parti	must	team
7	festiv	pay	navi	year	communiti	keep	industri	tackl	state	elect	answer	great
8	can	servic	uk	veteran	year	us	self-employ	use	ban	leader	prime	fratton
9	comment	cost	year	nation	success	health	get	cycl	black	tori	public	portsmouth
10	websit	plan	aircraft	rememb	portsmouth	safe	coronavirus	emerg	stand	regist	meal	citi
11	petit	worri	first	war	local	social	small	transport	inequ	#ge19	child	new
12	onlin	local	deploy	countri	citi	frontlin	deal	travel	matter	starmer	children	morn
13	hous	less	naval	defenc	now	vaccin	grant	green	progress	mp	get	#supportloc
14	fan	conserv	oper	troop	big	covid-19	work	space	pension	beat	#pmqs	visit
15	question	caus	wish	tribut	famili	abus	minist	ocean	histori	say	level	thank
16	christma	struggl	queen	servic	27a1	servic	uk	local	found	stephen	poverti	busi
17	page	break	britain	contribut	st	home	sector	activ	amend	can	gt	enjoy
18	zoom	famili	time	protect	special	equip	end	fe0f	debat	trust	johnson	hear
19	veteran	credit	portsmouth	oper	team	ppe	review	200d	injustic	win	ask	along
20	check	due	capabl	day	kind	spread	left	action	speak	want	lockdown	award

Example tweets of each topic are presented. Community

Topic 11(10%) - Schools

```
SMorganMP.csv.139
"Digital poverty risks creating a generational disadvantage for
pupils who are unfortunately more likely to have been behind their
classmates prior to the pandemic. Iâ€™ve written to the Education
Secretary to urge him to commit to connect these children in our
city as a priority<U+0001F447><U+0001F3FB>
https://t.co/ogSI3tJZUv"

SMorganMP.csv.152
"This chaos over schools reopening was avoidable had the Education
Secretary listened on safety, testing, vaccinations, remote
learning and exams. It has forced unprecedented action by unions
and some local councils, putting many in an impossible
position<U+0001F447><U+0001F3FB>\nhttps://t.co/W7YsQODdPE"
```

Topic 10 (9%) – Election Campaigns

```
SMorganMP.csv.555
"Congratulations from the other side of the pond @SeanCasten. A
brilliant win! Well done @VoteCasten <U+0001F1FA><U+0001F1F8>
<U+0001F1EC><U+0001F1E7> https://t.co/fvaxrLsBYE"

SMorganMP.csv.963
"How Keir Starmer's Labour intends to target rural and coastal
voters ahead of the next general election, in the words of shadow
cabinet members at the @LabourCC conference:
https://t.co/62ZK5WxyFr"
```

Topic 6 (9%) – Healthcare

```
SMorganMP.csv.28
"The Defence Secretary needs to set out a clear and credible plan
to protect troops on the ground, who are providing practical
support to roll out the vaccine and community testing. Good to
talk to @LauraReporting @ForcesNews about this call to
Government<U+0001F447><U+0001F3FB>\nhttps://t.co/uvGm8OxzAQ"

SMorganMP.csv.696
"Pregnancy<U+0001F930> can be challenging at the best of times,
but especially during #COVID19. Today I asked the Prime Minister
to urgently act to ensure hospitals can allow women to have a
birth partner with them from when they are admitted into hospital
for labour @MaternityAction https://t.co/1n4NMqz1KI"
```

Discussion

Social media offers politicians a direct channel to communicate with their constituents and the broader public. In this analysis, I delved into the Twitter feeds of two politicians, Stephen Morgan and Royston Smith, both representing constituencies in Hampshire, UK. Despite their similar regional backgrounds, their Twitter engagement and messaging appear to vary significantly.

One notable observation is the divergence in their engagement with the COVID-19 pandemic. Stephen Morgan's Twitter feed shows a significantly higher number of tweets mentioning "covid" compared to Royston Smith. This suggests that Stephen Morgan may have been more active in discussing the pandemic, which is an issue of great concern to the public. Furthermore, Stephen's tweets often refer to the new strain of the virus, indicating a keen interest in keeping the public informed about the latest developments. On the other hand, Royston Smith's tweets about COVID-19 were relatively fewer in number. His focus appears to be on topics such as vaccination, control measures, and government initiatives. This difference in approach may reflect varying priorities or strategies in addressing the pandemic. It's worth exploring whether these differences have had any impact on their constituents' perceptions and reactions.

In Royston Smith's word cloud, the prominence of "Southampton" suggests a strong emphasis on his constituency, reflecting his commitment to representing the interests of his local constituents. On the other hand, Stephen Morgan's word cloud highlights terms related to government, indicating a focus on national policies and issues. These word clouds hint at the different roles and responsibilities that come with being an MP. Royston Smith may place a high value on local representation and engagement, while Stephen Morgan appears to be engaged in national-level discussions and decision-making.

One intriguing common theme that emerged in the topics discussed by both MPs is their engagement with sectors of the British Armed Forces. This shared interest in matters related to the military is unsurprising, given the historical significance of the military presence in their respective constituencies. Southampton is renowned for its association with the Royal Air Force's Supermarine and its aviation heritage, while Portsmouth boasts a prominent naval base. It is clear that both MPs prioritize discussions related to defence and the Armed Forces, reflecting their constituencies' strong ties to these sectors. However, beyond this common ground, notable differences in their topic choices become evident. Stephen Morgan's Twitter feed exhibits a broader spectrum of political topics that are prevalent in his tweets. These encompass not only defence matters but also green campaigns and social inequality. This broader range of subjects reflects his active involvement in national and local issues, transcending the confines of traditional party lines. In contrast, Royston Smith's key political topic, prominently featured in his tweets, is Brexit. While there is overlap with some of the topics present in Stephen Morgan's tweets, such as discussions on local issues or healthcare, it's clear that Brexit remains a focal point of Smith's political communication. This could be attributed to his party affiliation and the ongoing implications of Brexit, highlighting his commitment to addressing this significant policy issue.

In the ever-evolving landscape of political communication, understanding the nuances of how politicians engage with their audience on platforms like Twitter provides valuable insights into their representation, priorities, and the dynamics of contemporary politics.

R Appendix

Assessment1.Rmd

- Scraping tweets - The last 3200 tweets from Stephen Morgan and Royston Smith

```
# # Stephen Morgan
# SM_tweets <- get_timelines(c("StephenMorganMP"), n = 3200, parse=T, toke
n=my_oauth)
#
# save_as_csv(SM_tweets, "SMorganMP.csv", prepend_ids = TRUE, na = "", fil
eEncoding = "UTF-8")
#
#
# # Royston Smith
# RS_tweets <- get_timelines(c("Royston_Smith"), n = 3200, parse=T, token=
my_oauth)
#
# save_as_csv(RS_tweets, "RSmithMP.csv", prepend_ids = TRUE, na = "", file
Encoding = "UTF-8")
```

- Setup Corpus' -

```
MorganTweets <- read.csv("SMorganMP.csv")
SmithTweets <- read.csv("RSmithMP.csv")
MorganTexts <- readtext("SMorganMP.csv", text_field = "text")
SmithTexts <- readtext("RSmithMP.csv", text_field = "text")

SMorganCorpus <- corpus(MorganTexts)
RSmithCorpus <- corpus(SmithTexts)
```

- Cleaning -

Removing meta data from Stephen Morgan

```
SMorganCorpus$retweets<-docvars(SMorganCorpus, "retweet_count")

library("lubridate")
Sys.setlocale("LC_TIME", "en_UK.UTF-8")

## Warning in Sys.setlocale("LC_TIME", "en_UK.UTF-8"): OS reports request
to set
## locale to "en_UK.UTF-8" cannot be honored

## [1] ""

SMorganCorpus$date_time <- docvars(SMorganCorpus, "created_at")
SMorganCorpus$time <- as.POSIXct(SMorganCorpus$date_time, tz = "UTC", form
at = "%a
%b %d %H:%M:%S %z %Y")

SMorganCorpus$date <- as.Date(SMorganCorpus$time)
SMorganCorpus$month <- as.numeric(format(SMorganCorpus$date, format="%m"))
SMorganCorpus$year <- as.numeric(format(SMorganCorpus$date, format="%Y"))
```

```
library("zoo")
SMorganCorpus$yr_m <- as.yearmon(paste(SMorganCorpus$year, SMorganCorpus$m
onth), "%Y %m")
```

Removing meta data from Royston Smith

```
RSmithCorpus$retweets<-docvars(RSmithCorpus, "retweet_count")

library("lubridate")
Sys.setlocale("LC_TIME", "en_UK.UTF-8")

## Warning in Sys.setlocale("LC_TIME", "en_UK.UTF-8"): OS reports request
to set
## locale to "en_UK.UTF-8" cannot be honored

## [1] ""

RSmithCorpus$date_time <- docvars(RSmithCorpus, "created_at")
RSmithCorpus$time <- as.POSIXct(RSmithCorpus$date_time, tz = "UTC", format
= "%a
%b %d %H:%M:%S %z %Y")

RSmithCorpus$date <- as.Date(RSmithCorpus$time)
RSmithCorpus$month <- as.numeric(format(RSmithCorpus$date, format="%m"))
RSmithCorpus$year <- as.numeric(format(RSmithCorpus$date, format="%Y"))

library("zoo")
RSmithCorpus$yr_m <- as.yearmon(paste(RSmithCorpus$year, RSmithCorpus$mont
h), "%Y %m")
```

- Summary -

Smith - longest tweet

```
tokeninfoRS <- summary(RSmithCorpus, n=12916)
write.csv(tokeninfoRS, file="tokeninfoRS.csv", row.names=FALSE)
longesttweet <- tokeninfoRS[which.max(tokeninfoRS$Tokens), ]
texts(RSmithCorpus)[1453]

#
#
RSmithMP.csv.1453
## "You most certainly did make us proud <U+0001F3F4><U+000E0067><U+000E00
62><U+000E0065><U+000E006E><U+000E0067><U+000E007F> <U+0001F3F4><U+000E006
7><U+000E0062><U+000E0065><U+000E006E><U+000E0067><U+000E007F> <U+0001F3F4
><U+000E0067><U+000E0062><U+000E0065><U+000E006E><U+000E0067><U+000E007F>
<U+0001F3F4><U+000E0067><U+000E0062><U+000E0065><U+000E006E><U+000E0067><U
+000E007F> https://t.co/5wuLfbxnjl"
```

Morgan

```
tokeninfoSM <- summary(SMorganCorpus, n=12916)
write.csv(tokeninfoSM, file="tokeninfoRS.csv", row.names=FALSE)
longesttweet <- tokeninfoSM[which.max(tokeninfoSM$Tokens), ]
texts(SMorganCorpus)[longesttweet[1,1]]
```

```
#
#
SMorganMP.csv.681
## "October is â\200\230Domestic Violence Awareness Monthâ\200\231. Join K
irsty Mellor and: \n\n<U+0001F418> Gemma Greene - Parcs\n<U+0001F418> Dene
sha Rocastle - Parcs\n<U+0001F590><U+0001F3FC> Claire Lambon - Southern Do
mestic Abuse\n<U+0001F469><U+0001F3FD><U+200D><U+0001F91D><U+200D><U+0001F
469><U+0001F3FC> Sally Jackson - STADV\n<U+0001F469><U+0001F3FC><U+200D><U
+0001F3EB> Amanda Martin - NEU\n<U+0001F468><U+0001F3FB>Stephen Morgan MP\
n\nSign up here:\n\nhttps://t.co/dThnN2BVzB https://t.co/N6D9EvQKyQ"
```

Comparing Proflic-ness by date

```
library(dplyr)

tokeninfocollapsed <- tokeninfoRS %>%
  group_by(yr_m)%>%
  summarize(sum(Tokens)) %>%
  rename (Tokens = `sum(Tokens)`)

## `summarise()` ungrouping output (override with `.groups` argument)

if (require(ggplot2)) ggplot(data = tokeninfocollapsed, aes(x = yr_m, y =
Tokens)) +
  geom_line() + geom_point() + theme_bw()

## Warning: Removed 1 row(s) containing missing values (geom_path).

## Warning: Removed 1 rows containing missing values (geom_point).
```

Key-words in context - comparing use of tweets containing “covid”

```
covidtweetsRS <- kwic(RSmithCorpus, pattern = "*covid*", window = 3)
covidtweetsSM <- kwic(SMorganCorpus, pattern = "*covid*", window = 3)

write.csv(covidtweetsRS, file="covidtweetsRS.csv")
write.csv(covidtweetsSM, file="covidtweetsSM.csv")

 kwic(RSmithCorpus, pattern = "*covid*", window = 3)
```

Converting to dfm

```
dfmSmith <- dfm(RSmithCorpus)
dfmSmith

## Document-feature matrix of: 3,187 documents, 12,125 features (99.8% spa
rse) and 96 docvars.
##                 features
## docs             nice of bernie to visit joâ \200 \231 s bench
##   RSmithMP.csv.1    1  1      1  1     1   1 2 2 2     1
##   RSmithMP.csv.2    0  1      0  1     0   0 1 1 0     0
##   RSmithMP.csv.3    0  0      0  1     0   0 1 1 0     0
##   RSmithMP.csv.4    0  0      0  0     0   0 0 0 0     0
##   RSmithMP.csv.5    0  0      0  0     0   0 0 0 0     0
##   RSmithMP.csv.6    0  0      0  2     0   0 1 1 0     0
```

```
## [ reached max_ndoc ... 3,181 more documents, reached max_nfeat ... 12,1
15 more features ]

dfmMorgan<- dfm(SMorganCorpus)
dfmMorgan

## Document-feature matrix of: 3,195 documents, 13,040 features (99.7% spa
rse) and 96 docvars.
##                  features
## docs              did your mp commit to support the sewage bill ?
##   SMorganMP.csv.1   2    2  1      1  2       2   3      1    1 1
##   SMorganMP.csv.2   0    0  0      0  2       1   2      0    0 0
##   SMorganMP.csv.3   0    0  0      0  0       0   0      0    0 0
##   SMorganMP.csv.4   0    0  0      0  5       1   1      0    0 0
##   SMorganMP.csv.5   0    0  0      0  4       0   1      0    0 0
##   SMorganMP.csv.6   0    1  0      0  0       0   1      0    0 0
## [ reached max_ndoc ... 3,189 more documents, reached max_nfeat ... 13,0
30 more features ]

#top features of each
# topfeatures(dfmSmith, 10)

dfmSmith_trimmed<- dfm(RSmithCorpus, remove = c(stopwords("english"),"rt",
"@*", "+",
                                                "<","u",">","€","™","s" ,"0
001f3fb",
                                                "0001f3fb", "*â*","0001f447
", "t" ),
                       stem = TRUE, remove_punct = TRUE, tolower=T,
                       remove_symbols = T, remove_numbers = T, remove_url
= T)

# topfeatures(dfmSmith_trimmed, 50)

dfmMorgan_trimmed<- dfm(SMorganCorpus, remove = c(stopwords("english"),
                                                  "rt","@*","+","<","u",">","€
","™",
                                                  "s","0001f3fb","0001f3fb","
*â*",
                                                  "0001f447", "iâ" ),
    stem = TRUE, remove_punct = TRUE, tolower=T, remove_symbols = T,
    remove_numbers = T, remove_url = T)

# topfeatures(dfmMorgan_trimmed, 50)
```

term frequency - Inverse Document Frequency Weighting

```
dfmMorgan_tfidf<- SMorganCorpus %>%
  dfm(remove = c(stopwords("english"),"rt","@","rt", "@*", "+", "<", "u",
">", "€",
                 "™", "s","0001f3fb", "0001f3fb", "*â*","0001f447", "t"),
      stem = TRUE, remove_punct = TRUE, tolower=T) %>%
  dfm_tfidf( scheme_tf = "count", scheme_df = "inverse",base = 10)

dfmSmith_tfidf<- RSmithCorpus %>%
```

```
  dfm(remove = c(stopwords("english"),"rt","@","rt", "@*", "+", "<", "u",
">", "€",
                 "™", "s","0001f3fb", "0001f3fb", "*â*","0001f447", "t"),
      stem = TRUE, remove_punct = TRUE, tolower=T) %>%
  dfm_tfidf( scheme_tf = "count", scheme_df = "inverse",base = 10)
```

Wordclouds

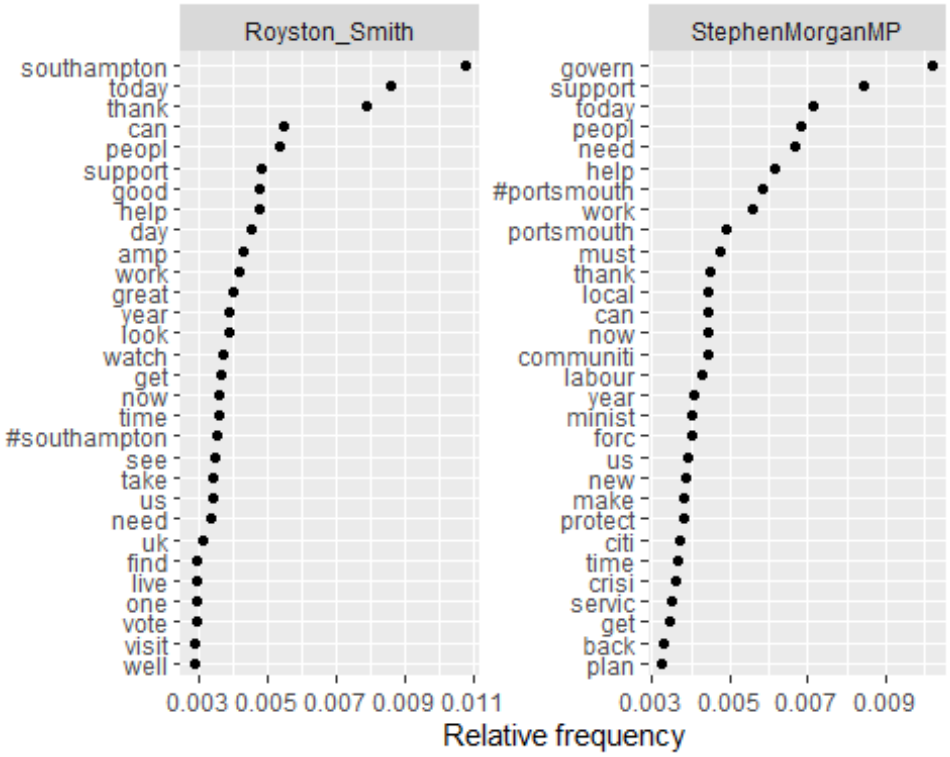
```
# textplot_wordcloud(dfmMorgan_tfidf, min_count = 6, random_order = FALSE,
rotation = 0.25,
#     color = RColorBrewer::brewer.pal(8, "Dark2"))
#
# textplot_wordcloud(dfmSmith_tfidf, min_count = 6, random_order = FALSE,
rotation = 0.25,
#     color = RColorBrewer::brewer.pal(8, "Dark2"))
```

Creating master corpus dfm

```
MasterCorpus <- RSmithCorpus+SMorganCorpus

MasterCorpus$author<-docvars(MasterCorpus, "screen_name")

dfm_Master<- MasterCorpus %>%
  dfm(remove = c(stopwords("english"),"rt", "@*", "+","<","u",">","€","™",
"s"
                 ,"0001f3fb","0001f3fb", "*â*","0001f447", "t"),
      stem = TRUE,
      remove_punct = TRUE, tolower=T) %>%
      dfm_group(groups = "author") %>%
      dfm_weight(scheme = "prop")

write.csv(convert(dfm_Master, to="data.frame"), file="dfm_group.csv", row.
names=FALSE)
```

Plot word frequency comparison

```
dfm_Master_Grouped<-dfm_sort(dfm_Master)

dfm_Master_Grouped2<-textstat_frequency(
  dfm_Master,
  n = 30,
  groups = "author")

#plot for comparison (using dfm with top 20th ranked words only)

ggplot(data = dfm_Master_Grouped2, aes(x = factor(nrow(dfm_Master_Grouped2
):1), y = frequency)) +
    geom_point() +
    facet_wrap(~ group, scales = "free") +
    coord_flip() +
    scale_x_discrete(breaks = nrow(dfm_Master_Grouped2):1,
                     labels = dfm_Master_Grouped2$feature) +
    labs(x = NULL, y = "Relative frequency")
```



Text analysis - Topic modelling

LDA probabilistic model

```
# tmod_ldaMS <- textmodel_lda(dfmMorgan_tfidf, k = 10)
# tmod_ldaRS <- textmodel_lda(dfmSmith_tfidf, k = 10)
```

Morgan top 20

```
# seededlda::terms(tmod_ldaMS, 20)
```

Royston top 20

```
# seededlda::terms(tmod_ldaRS, 20)
```

Further Cleaning & Stemming

```
dfmMorgan_tfidf_stem<- dfm(SMorganCorpus,
                           remove = c(stopwords("english"),"rt", "@*", "+"
, "<", "u",
                                      ">", "€","™","s", "0001f3fb", "0001f
3fb",
                                      "*â*","0001f447", "t","0*"),
                           stem=T,
  remove_punct = TRUE, remove_numbers = TRUE, remove_symbol = TRUE,tolower
=T) %>%
  dfm_tfidf(scheme_tf = "count", scheme_df = "inverse",base = 10)

dfmSmith_tfidf_stem<- dfm(RSmithCorpus,
                           remove = c(stopwords("english"),"rt", "@*", "+"
, "<", "u",
                                      ">", "€","™","s", "0001f3fb", "0001f
```

```
3fb",
                                          "*â*","0001f447", "t","0*","fe0f"),
                              stem=T,
  remove_punct = TRUE, remove_numbers = TRUE, remove_symbol = TRUE,tolower
=T) %>%
  dfm_tfidf(scheme_tf = "count", scheme_df = "inverse",base = 10)
```

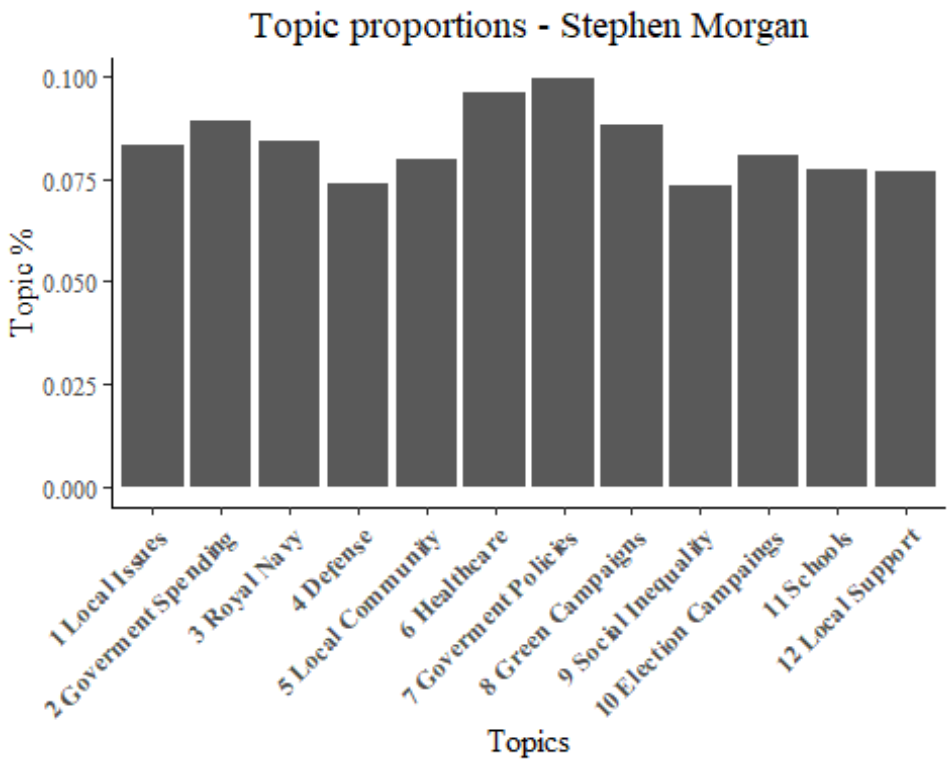
Re-apply model

```
require(quanteda)
require(readtext)
require(quanteda.corpora)
require(seededlda)
require(lubridate)
require(RColorBrewer)
require(dplyr)
require(ggplot2)
require(stm)
tmod_ldaMS <- textmodel_lda(dfmMorgan_tfidf_stem, k = 12)
tmod_ldaRS <- textmodel_lda(dfmSmith_tfidf_stem, k = 9)
```

csv table

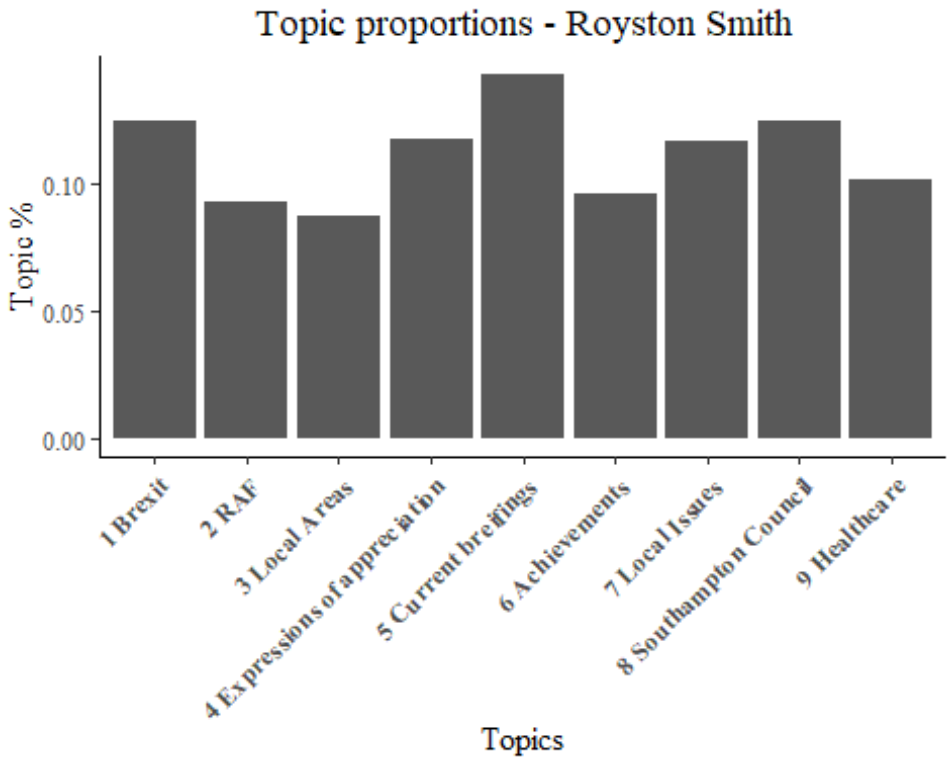
```
topwords<-as.data.frame(seededlda::terms(tmod_ldaMS, 20))
write.csv(topwords, file="top_wordsSM.csv")
# View(topwords)


topwords<-as.data.frame(seededlda::terms(tmod_ldaRS, 20))
write.csv(topwords, file="top_wordsRS.csv")
# View(topwords)

MorganTweets$topic <- seededlda::topics(tmod_ldaMS)
View(MorganTweets[,c(5,91)])

SmithTweets$topic <- seededlda::topics(tmod_ldaRS)
View(SmithTweets[,c(5,91)])

#attach the topic variable to the corpus object as well:
SMorganCorpus$topic <- seededlda::topics(tmod_ldaMS)
RSmithCorpus$topic <- seededlda::topics(tmod_ldaRS)

topics_tableM<-ftable(MorganTweets$topic)
# View(topics_tableM)
topicsprop_tableM<-as.data.frame(prop.table(topics_tableM))
# View(topicsprop_tableM)

topics_tableS<-ftable(SmithTweets$topic)
# View(topics_tableS)
topicsprop_tableS<-as.data.frame(prop.table(topics_tableS))
# View(topicsprop_tableS)


#visualise the topic frequencies
```

```
ggplot(data=topicsprop_tableM, aes(x=Var1, y=Freq)) +
  geom_bar(stat = "identity") +
  theme_classic()+
  labs (x= "Topics", y = "Topic %")+
  labs(title = "Topic proportions - Stephen Morgan") +
  scale_x_discrete(labels=c("topic1" = "1 Local Issues", "topic2" = "2 Gov
erment Spending",
                            "topic3" = "3 Royal Navy",
                            "topic4" = "4 Defense", "topic5"="5 Local Comm
unity",
                            "topic6" = "6 Healthcare", "topic7" = "7 Gover
ment Policies",
                            "topic8" = "8 Green Campaigns",
                            "topic9"="9 Social Inequality",
                            "topic10" = "10 Election Campaings","topic11"=
"11 Schools",
                            "topic12" = "12 Local Support")) +
  theme(axis.text.x = element_text(face="bold",
                           size=10, angle=45,hjust = 1)) +
  theme(text=element_text(size=12,  family="serif")) +
  theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(data=topicsprop_tableS, aes(x=Var1, y=Freq)) +
  geom_bar(stat = "identity") +
  theme_classic()+
  labs (x= "Topics", y = "Topic %")+
  labs(title = "Topic proportions - Royston Smith") +
  scale_x_discrete(labels=c("topic1" = "1 Brexit", "topic2" = "2 RAF", "to
pic3" = "3 Local Areas",
                            "topic4" = "4 Expressions of appreciation", "t
opic5"="5 Current breifings",
```

```
                              "topic6" = "6 Achievements", "topic7"="7 Local
Issues",
                              "topic8" = "8 Southampton Council", "topic9"="
9 Healthcare")) +
  theme(axis.text.x = element_text(face="bold",
                            size=10, angle=45,hjust = 1)) +
  theme(text=element_text(size=12,  family="serif")) +
  theme(plot.title = element_text(hjust = 0.5))
```




```
texts(SMorganCorpus)[28]

##
SMorganMP.csv.28
## "The Defence Secretary needs to set out a clear and credible plan to pr
otect troops on the ground, who are providing practical support to roll ou
t the vaccine and community testing. Good to talk to @LauraReporting @Forc
esNews about this call to Government<U+0001F447><U+0001F3FB>\nhttps://t.co
/uvGm8OxzAQ"

texts(SMorganCorpus)[696]

#
#
SMorganMP.csv.696
## "Pregnancy<U+0001F930> can be challenging at the best of times, but esp
ecially during #COVID19. Today I asked the Prime Minister to urgently act
to ensure hospitals can allow women to have a birth partner with them from
when they are admitted into hospital for labour @MaternityAction https://t
.co/1n4NMqz1KI"

texts(RSmithCorpus)[33]
```

```
##
RSmithMP.csv.33
## "The Supermarine Southampton, designed by R. J. Mitchell and his team a
t Supermarine in Woolston. They later designed and built the Spitfire ther
e.\n\nThe first flight was in 1925. They were used by the RAF and civil op
erators, and they were sold to countries all around the world. https://t.c
o/k2DwQbAhJc"
```

```
texts(RSmithCorpus)[59]
```

```
#
#
RSmithMP.csv.59
## "The Royal Air Force has brought some Christmas cheer to a 98-year-old
Second World War Spitfire pilot after learning that thieves had stolen fro
m his home a cherished photo of him flying the iconic aircraft.\n\nFull st
ory: https://t.co/LMbU4wLcmU https://t.co/23gM2QZ9Zy"
```