Text Mining the Election Manifestos

Allan Brimicombe, University of East London

This year, as part of my teaching of predictive analytics, I decided to include text mining of the general election manifestos in order to inject some topical interest. Although the publication of the manifestos came late on in the term, text mining is relatively quick and I was able to slot it in. I present here the method as a generic way of analysing large amounts of qualitative data that complements conventional content analysis and some results that may be of more specific interest to criminologists. Since the Conservatives now have a majority in Parliament and their policies will dominate the next five years, their manifesto is picked out for particular attention.

The manifestos are published as PDF files available on-line. Adobe reader has an option to save any PDF file as a text file. These are messy files but contain all the text except for that which is embedded in photographs (e.g. photo of party leader with embedded quotable quote). We used the manifestos for all the parties included in the televised seven-party debate. Some cleaning of the text files is then necessary. After manually stripping off the contents and index pages, the rest can be carried out semi-automatically using 'find-replace' functions in a text editor. Each manifesto is reduced to a number of large paragraphs, one for each chapter in the manifesto. The word counts for the prepared texts are given in Table 1.

Table 1: Word counts for each party manifesto 2015

Conservative	Labour	Lib Dem	Green	UKIP	Plaid Cymru	SNP	Total
30,146	18,178	35,985	40,383	27,432	18,249	18,702	189,077

The software of choice for the analysis is open-source R (http://cran.r-project.org/) using mainly the tm text mining package (http://cran.r-project.org/web/packages/tm/). The first step is to load the text files and change their data type to a *corpus*, that is, a structured body of text suitable for statistical analysis. The corpus is then further cleaned to turn all the text to lower case (so we don't differentiated between 'Crime' and 'crime'),

remove all punctuation and numbers, remove unwanted words such as 'a', 'the' and 'we', as well as political party names. These are all automated functions that take just a few seconds of processing time. Table 2 shows an example of how a piece of raw data reads and how the finally prepared data looks in the corpus. One disadvantage of this type of text mining is that words are treated separately. So, for example (and as we will see below) 'female genital mutilation' is treated as three separate words. In Table 2, the stop word 'no' is removed (because we don't want a large number of 'no' to bias our analysis), leaving 'money' as neutral. The fact that the meaning is 'no money' rather than 'some money' is lost, but if the term 'money' turns out to be significant, then a return to the original documents with conventional content analysis would allow the correct interpretation.

Table 2: Example text before and after preparation as a corpus

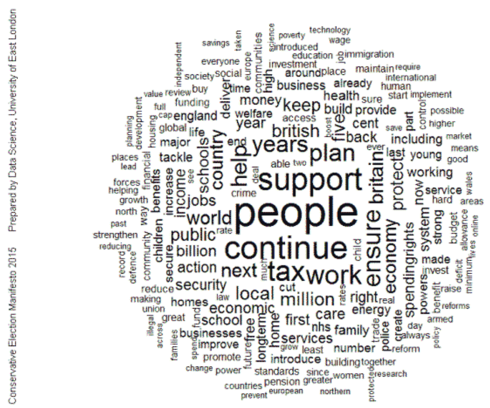
Before	FOREWORD Over the last five years, we have put our country back on the right track. Five years ago, Britain was on the brink. As the outgoing Labour Treasury Minister put it with brutal candour, 'there is no money'.
After	foreword last five years country back right track five years ago britain brink outgoing treasury minister brutal candour money

The next step is to turn the corpus into a term-document matrix (tdm). This is a very large matrix which records the number of times each unique term occurs in a manifesto chapter. This is quite a sparse matrix as many words only crop up once in one manifesto chapter leaving blanks in all the others. Thus the entire corpus becomes 9,438 unique terms with 95% of the tdm as blank. The Conservative manifesto (as corpus) includes 3,502 unique terms.

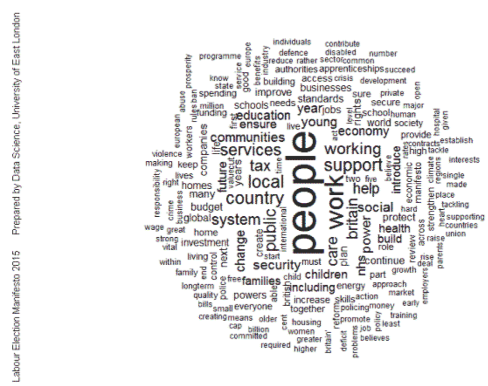
We can now visualise the main thrust of each manifesto graphically by text frequency as word clouds (e.g. Figure 1(a) and 1(b)). The word cloud for the Conservative Party can be compared with the top terms by word count in Table 3 to see how the word clouds work. We can also create more analytically interesting word clouds such as the difference between the 2010 and 2015 Conservative manifestos in Figure 1(c) and the differences between the 2015 Labour and Conservative manifestos in Figure 1(d). The first thing to note is that terms relating to crime and justice are hard to find – clearly not seen as vote winners. The main difference in Figure 1(c) is the introduction in 2015 of 'continue' (the second most frequent word) and 'plan' (ranked 7th). These two words feature again in Figure 1(d) as differentiating Conservative from Labour, with Labour

focusing much more on 'people' (in relation to manifesto word count) that do the Conservatives.

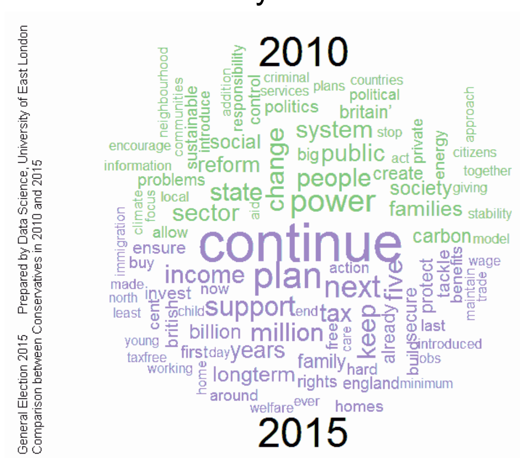
Figure 1: Selected word clouds (for full set for all parties, see: http://www.uel.ac.uk/geo-information/GeneralElection2015.htm)



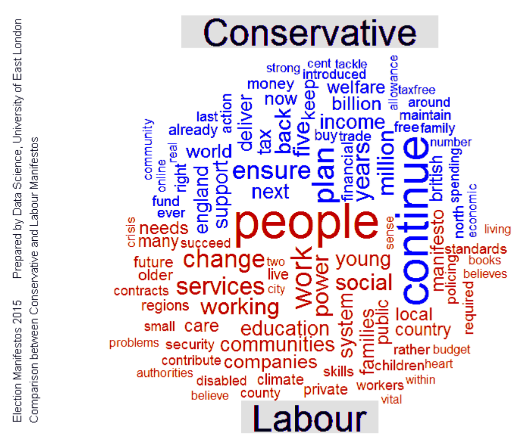
Conservative Party Manifesto 2015



Labour Party Manifesto 2015



Difference between 2010 & 2015 Conservative Manifestos



Difference between Conservative and Labour Manifestos 2015

Table 3: Top terms in Conservative manifesto

Conservative Manifesto		
Term	Count	Rank
people	159	1
continue	115	2
support	110	3
tax	99	4
work	89	5
ensure	86	6
plan	81	7

Looking a bit further at word counts and rankings, Table 4 summarises the top 10 terms using all the party manifestos that relate to 'key issues' facing society at the election. This is then compared with the Conservative manifesto. The ranks provide the main comparison. Thus 'work' ranks highly but 'education' in all manifestos ranks 20th but for the Conservatives ranks 119th, noting however 'schools' ranks higher. The NHS is not ranked as highly as we might have perceived from the debates. Crime and justice still does not feature. So Table 5 picks out some terms of interest to criminologists. The term 'security' is highest but relates mostly to external threats; 'violence' is ranked higher by the Conservatives though 'policing' and 'justice' come way down.

Table 4: Comparison of ranking of 'key issues'

All Party Manifestos			Conservative Manifesto	
Term	Count	Rank	Count	Rank
work	597	3	89	5
health	363	9	35	50
economy	296	18	58	14
education	284	20	20	119
energy	279	21	32	55
children	246	24	35	50
nhs	242	25	36	45
schools	217	33	47	22
jobs	212	35	43	26
housing	208	36	29	64

Table 5: Comparison of ranking of terms of interest to criminology (singular and plural terms have been added together, though the ranking given is for the term with the highest rank)

All Party Manifestos			Conservative Manifesto	
Term	Count	Rank	Count	Rank
security	184	49	40	31
crime(s)	171	82	26	96
police	127	108	28	73
justice	106	145	4	756
policing	59	319	6	552
criminal(s)	83	345	8	448
prison(s)	69	369	11	374
violence	49	397	11	268

Finally we look at some word associations. Associations are where two words occur within the same manifesto chapter. A pair of words may also occur singly elsewhere in the corpus, so a measure of the association is a positive correlation (0, 1) where 1 indicates that two terms always occur together. Table 6 summarises the correlations >=0.8 for ‘crime’, ‘justice’, ‘police’ and ‘violence’ using all manifestos. The crime agenda appears to be most focussed on ‘prison(s)’, ‘offenders’ and ‘re-offending’. Under ‘crime’ so too are two terms ‘genital’ and ‘mutilation’ showing FGM on the crime agenda as well as ‘trafficking’. Correlations with ‘justice’ shows ‘cannabis’ on the agenda. For my own research into domestic violence and abuse (DVA), ‘girls’ and ‘sexual’ are most highly correlated with ‘violence’ reflecting the VAWG agenda, with ‘domestic’ not far behind.

Table 6: Word associations with correlations >=0.8 using all manifestos

crime		justice		police		violence	
term	correlation	term	correlation	term	correlation	term	correlation
prisons	0.92	criminal	0.92	victims	0.85	girls	0.89
prison	0.90	courts	0.87	court	0.81	sexual	0.88
offenders	0.88	diverting	0.86	officers	0.80	crimes	0.86
reoffending	0.88	reoffending	0.85			domestic	0.80
victims	0.87	crime	0.83				
behaviour	0.86	police	0.82				
custody	0.86	sentences	0.81				
hurt	0.86	cannabis	0.80				
policing	0.86						
possession	0.85						
sentences	0.85						
commissioners	0.82						
arrested	0.81						
confidence	0.81						
genital	0.80						
mutilation	0.80						
trafficking	0.80						

To conclude, text mining is a rapid approach to handling and analysing large amounts of qualitative data. Whilst it has been a bit of fun to look at the 2015 election manifestos, my main use of this technique is, for example, in analysing the free text fields in which the modus operandi are written up in police recorded crime thereby being able to tap into a rich seam for research.