The UK election manifestos over time: a topic modelling analysis

Stefano Costantini

3rd June 2015

Introduction

In the United Kingdom, before a general election each party issues a "manifesto". This is a document in which the party sets out the policies it will implement should it win. A governing party not delivering on its manifesto would be disadvantaged in the run-up to the following election. For this reason, a lot of work generally goes into preparing a manifesto, which becomes *the* document that "defines" a party before the electorate.

We analyse the text of the manifestos of the two main UK parties, the Labour Party and the Conservative Party over the years. The analysis covers the the period from the 1950 general election to the 2001 one (a total of fifteen elections). During this time, either of these two parties governed the country and no coalitions were formed.

We focus on the evolution of the main topics that are discussed in the party manifestos. In particular, we examine how the topics' popularity changes over time and whether it is related to macroeconomic conditions, as captured by three indicators: GDP growth, inflation and unemployment rate. We also consider the extent to which the parties differ from each other, both in terms of the topics discussed and in terms of the overall tone of their respective manifestos.

This paper is structured as follows. First, we describe the data we collected. Then we set out the main questions we seek to address. After that, we report on the text mining analysis we undertook to obtain a quantitative representation of each document in the corpus. Finally, we present the main analysis and we conclude by summarising our key findings, as well as by setting out some potential steps for further work.

Data

We use the text of the parties' manifestos as obtained from the websites labour-manifesto.com and conservativemanifesto.com. These websites provide the text of the original manifestos in `HTML` format.

The general elections we consider are those of 1950, 1951, 1955, 1959, 1964, 1966, 1970, 1974, 1979, 1983, 1987, 1992, 1997, 2001. Two elections were held in 1974, due to the first one delivering a hung parliament: for the purpose of this analysis we treat them as a single event.

Overall, we have 5,863 paragraphs of text, of which 2,577 comefrom Labour manifestos and 3,286 from Conservative manifestos. For each paragraph, we also have the following metadata: year, party name and party leader.

In addition to the text data, we have also obtained historic macroeconomic data (GDP growth, inflation and unemployment rate) from the UK Statistics office.

Question

We seek to address the following questions:

- whether the two main parties' policies follow similar popularity patterns;
- whether the topics' popularity is related to the surrounding economic environment, and
- the extent to which the parties differ from each other in their way of communicating with the electorate, both in terms of the choice of the individual topics addressed and in the overall tone of their manifesto.

While this is an initial descriptive analysis, we believe that it can provide interesting insights on these questions. Specifically, the analysis of the similarity between parties can also shed some light on whether they offer real alternatives or rather just differently nuanced views of the world.

We also believe that this analysis can provide a platform for further investigation by social scientists (economists, sociologists and political scientists). We develop this point in the concluding section.

Extracting content

We have obtained a quantitative representations of each paragraph in the corpus by carrying out LDA topic modelling, using the `topicmodels` library.

First, we have imported the raw text from the `manifestos.txt` file (obtained from the websites cited above) and pre-processed it using the `RawDocs` class provided in the `topicmodels` library. This has required the following steps:

- Tokenising the documents, that is separating each individual word. We have also used the `token_clean` method to remove all non-alphanumeric tokens, all numeric tokens and all symbols in the data.
- Removing "stopwords", that is frequent English language words that are grammatically necessary in written and spoken English but do not provide additional meaning. Removing the stopwords has reduced the number of tokens from 266,915 to 142,526.
- Stemming the tokens, using the Porter stemmer method provided by the `RawDocs` class. Stemming tokens helps group together words that carry the same meaning. For example the words *employee* and *employment* would both be turned into the stem *employ*, which, for the purpose of this analysis, carries the same meaning.
- Discarding stems with low *term-frequency inverse-document frequency (tf-idf)* score. The tf-idf score is higher for stems that occur many times in few documents, that is those stems that characterise specific documents and, for this reason, are more likely to carry meaningful information. Stems with low tf-idf score instead might add noise to the analysis, so removing them can improve the inference. We have carried out this step first by calculating the tf-idf score of all stems using the `tf-idf` method provided by the `RawDocs` class. Then we have plotted the ranking plot of the stems in the corpus, which is shown below.
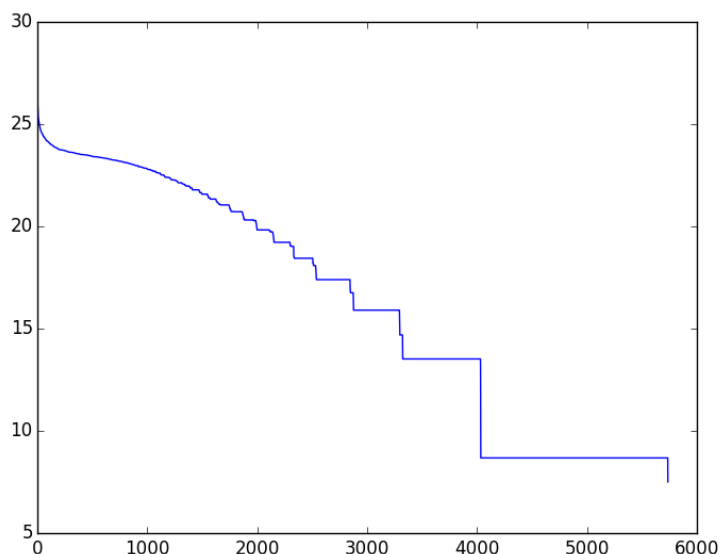
The chart shows that a likely cut-off point is reached at about 3,300 stems, so we have removed the stems featuring lower than that in the ranking.

The pre-processing has yielded a corpus of 131,532 total stems, with 3,279 unique stems. We have used this corpus to carry out LDA topic modelling.

We have chosen to approximate the inference on the topic posterior distribution by using Gibbs sampling, seeking to identify thirty topics. To improve the sampling accuracy, we have run three separate Gibbs sampling chains, allowing for a burn-in period of 500 iterations and then keeping 50 samples, picked using a thinning interval of 25. To select the sampling chain from which to extract the samples for the inference, we have considered the *perplexity* of each chain. *Perplexity* is a measure of goodness-of-fit.
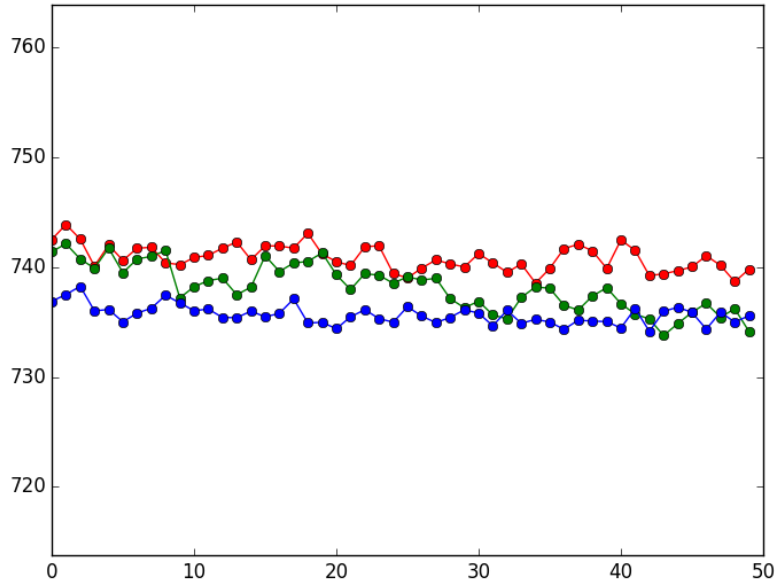

Figure 1: Perplexity of the three Gibbs sampling chains

As shown in the figure, the three chains appear to be generally equivalent in terms of their goodness-of-fit. Thus we have selected the chain with the lowest average perplexity across the sample. As the chain looks to have converged after the burn-in period, we have kept all 50 samples to estimate the topic distribution.

This phase of the analysis has allowed us to obtain a quantitative representation of each of the paragraphs in the corpus. More specifically, each paragraph in the corpus now can be represented as a vector in a 30-dimensional space, where each dimension corresponds to a specific topic.

Analysis

We now focus on the main analysis, seeking to answer the questions set out above. This section is structured as follows:

- First we present the main topics that the LDA analysis has identified.
- Then, for each main topic, we show its evolution over time in terms of its popularity. When relevant, we also considered its relationship with macroeconomic indicators, both in aggregate and separately for each party.
- Finally, we conclude by discussing the evolution of the similarity between the two main parties.

Main topics

We have carried out the LDA inference assuming 30 different topics. While each topic would carry a meaning, some may be more recognisable than others. To identify the main meaning of each topic, we have generated word clouds using the R package `wordcloud`. Following visual inspection of the word clouds, we have selected the following eight topics as their meaning seemed more clearly defined. The Appendix contains the word-clouds of each of the main topics, while the word-clouds of all 30 topics are provided as part of the supporting material to this paper.

Topic number	Meaning	Key stems
Topic 1	Welfare and benefits	Help, famili, need, support, give
Topic 3	Housing	Hous, home, build, council
Topic 7	Employment	Work, time, pay, employ
Topic 12	Education	School, educ, children, parent
Topic 24	Healthcare and the NHS	Servic, health, care, hospit
Topic 25	Taxation	Tax, rate, spend, cut, money
Topic 26	Inflation and cost of living	Price, increas, inflat, live, rise
Topic 30	Trade unions	Trade, union, right, member

We carry out the remainder of the analysis focusing on the topics above.

Popularity of main topics

For each of the main topics we have identified, we describe its evolution over time in terms of popularity. We have calculated the popularity of each topic by assigning a single topic to each paragraph in the corpus, selecting the topic with the highest probability. We have then calculated the frequency of each topic in each year.

As mentioned, for each topic we have considered how its popularity varies over time with respect to three key macro economic indicators. Instead of presenting every chart we have generated, we focus on reporting any notable patterns and differences. The full set of charts is provided as part of the supporting material to this paper.

Topic 1 - Welfare and benefits

The popularity of this topic appears to be linked to the rate of unemployment in the country. Unemployment started raising in the 70s, after a couple of decades of low unemployment.
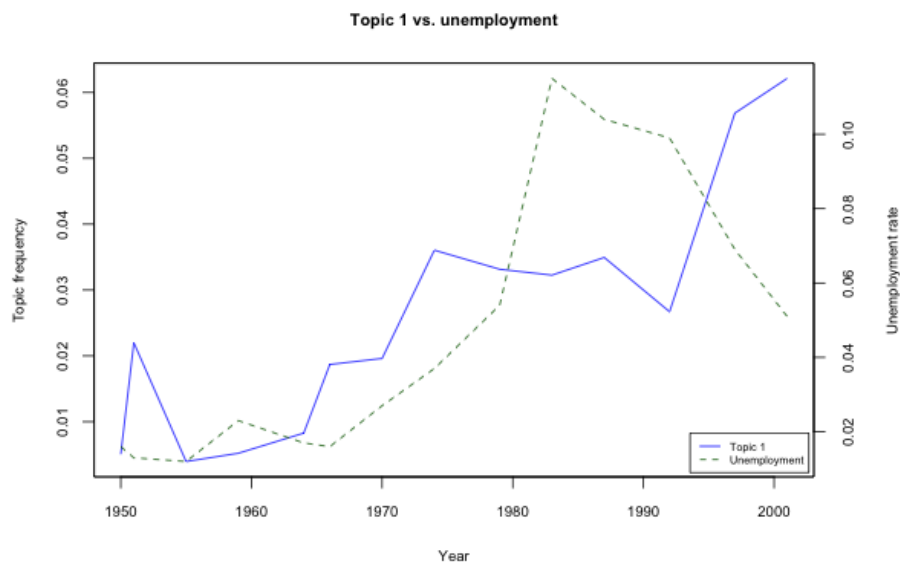
Figure 2: Topic 1 - Welfare and benefits (aggregate)

We note that this topic does not appear in the early Labour manifestos, while it does in the Conservative ones. However, its popularity increases faster in the Labour manifestos. Also, this topic becomes more prominent in recent years for both parties. Our analysis however does not allow us to say whether the attitude towards welfare and benefits is different between the two parties, as mentioning the same topic does not mean that the parties are necessarily in agreement with each other.
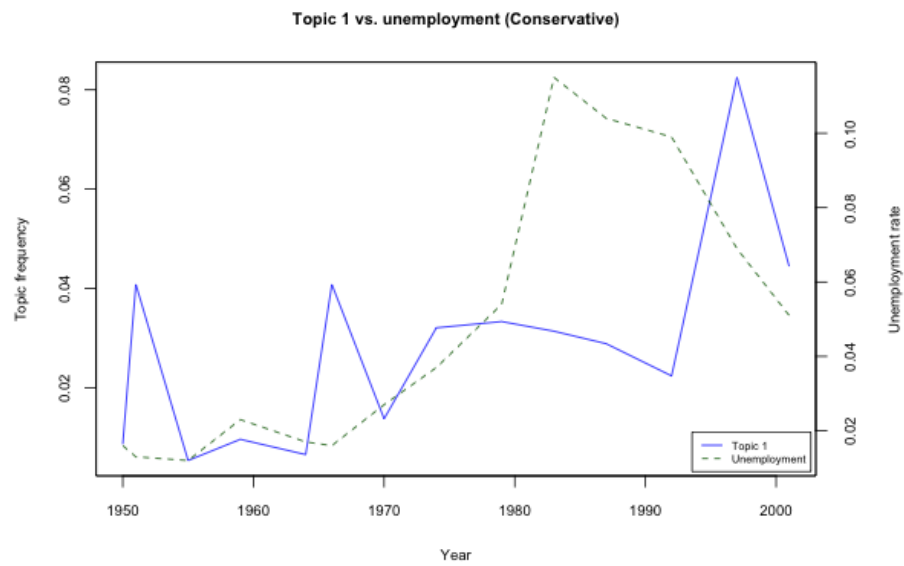
6

Figure 3: Topic 1 - Welfare and benefits (Labour)

Figure 4: Topic 1 - Welfare and benefits (Conservative)

7

Topic 3 - Housing

Housing is a popular topic since the 50s. We do not find any relationship between the popularity of this topic and the macroeconomic indicators. The topic appears to lose popularity from the late 80s onwards, reaching its lowest level of popularity in 2001.
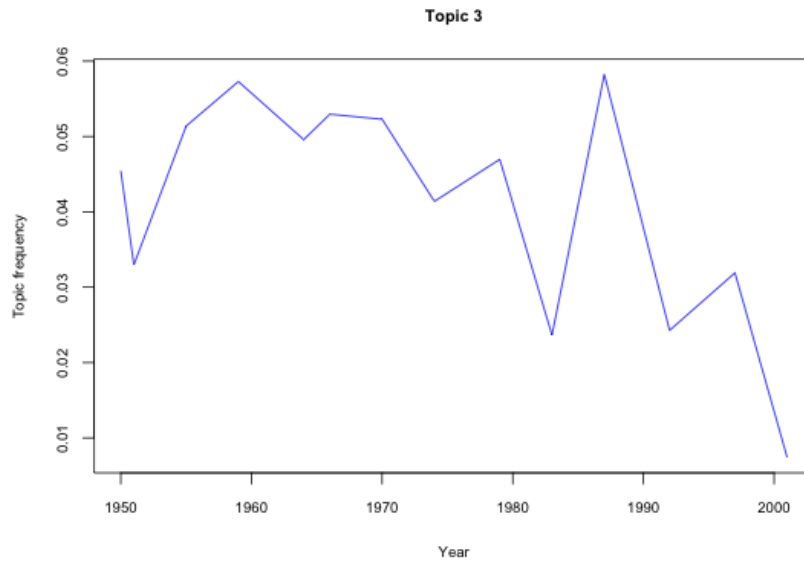

Figure 5: Topic 3 - Welfare and benefits (aggregate)

The pattern is similar for both parties, though the topic loses popularity first in Labour manifestos compared to Conservative ones. This might be due to the introduction of the 'right-to-buy' policy for council housing by the Conservative party in the 80s. This was one of the main policies of the Thatcher governments. The fact that they continue to feature in this party's manifestos for a while longer is not surprising.
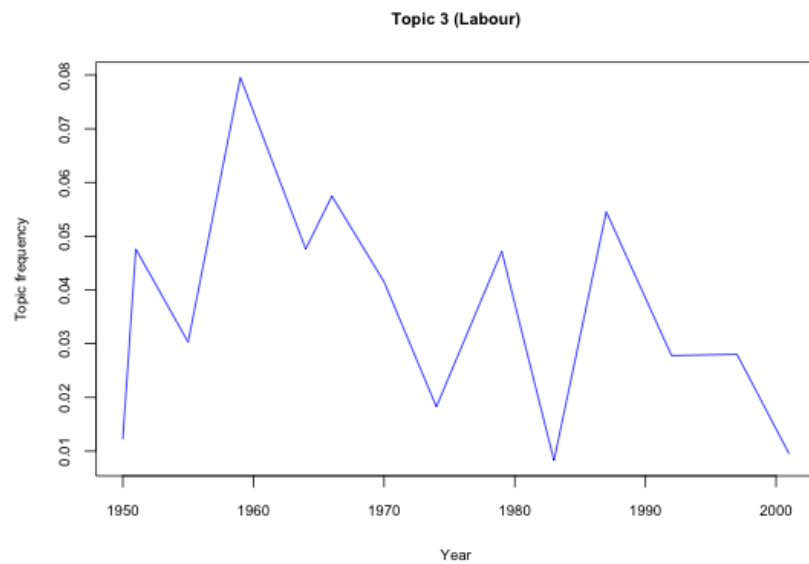
Topic 3 (Labour)

Figure 6: Topic 3 - Welfare and benefits (Labour)
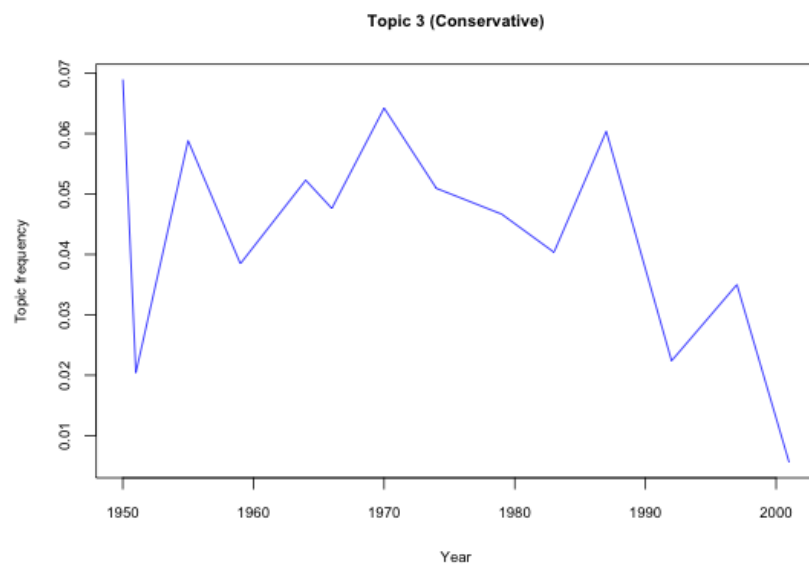
Topic 3 (Conservative)

Figure 7: Topic 3 - Welfare and benefits (Conservative)

9

Topic 7 - Employment

The popularity of this topic appears to be linked to the rate of unemployment. The aggregate popularity count shows that the topic is less popular in periods of low unemployment, that is until 1970, and then grows during the economic crisis of the 70s and the trade union disputes period of the early 80s. After a dip in the early 90s, its popularity grows again at the time of the 1997 election.
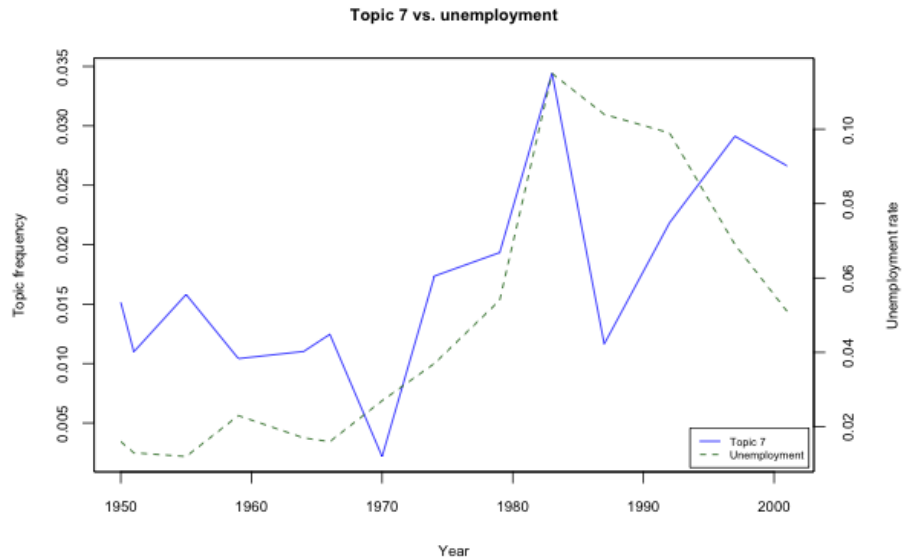

Figure 8: Topic 7 - Employment (aggregate)

The two parties show a different pattern. The topic is very popular in Labour manifestos of the late 50s, while, with the exception of the 1950 general election, it barely appears in Conservative manifestos until 1959. For both parties, the popularity of this topic rises during the 70s. During the 80s and 90s the topic remains popular for the Labour party, with the exception of a sharp reduction corresponding to the 1987 election. The topic loses importance in Conservative manifestos during the 80s, only to bounce at the time of the 1997 election.
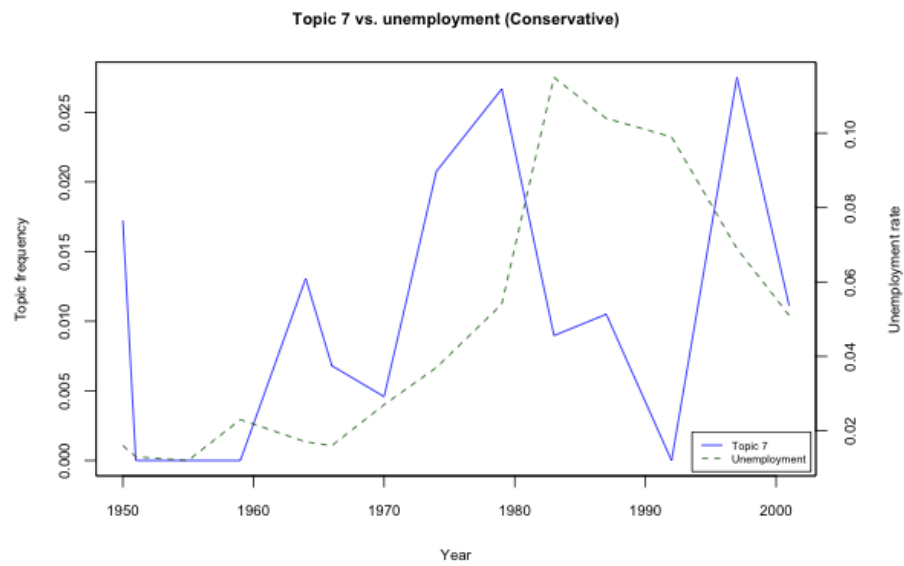
Figure 9: Topic 7 - Employment (Labour)

Figure 10: Topic 7 - Employment (Conservative)

11

Topic 12 - Education

The popularity of this topic does not appears to be related to any macroeconomic indicator in particular. Rather, the charts show that this topic has low-to-medium popularity for most of the period under analysis, until it jumps to prominence in 1997.
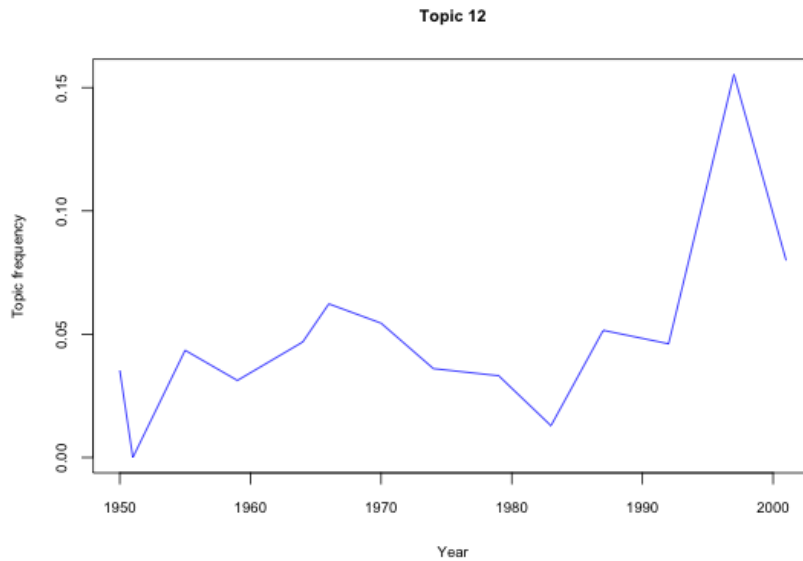

Figure 11: Topic 12 - Education (aggregate)

This is especially true for the Labour party since 1992. This finding is interesting, as the focus on education was one of New Labour's flagship policies in the 1997 manifestos (recall, for example, Tony Blair's political mantra *"education, education, education"* often repeated a political rallies during that electoral campaign). The topic barely features in the Conservative 1992 manifesto, but becomes very popular in the 1997 one, probably as a reaction to Labour's focus on this theme.
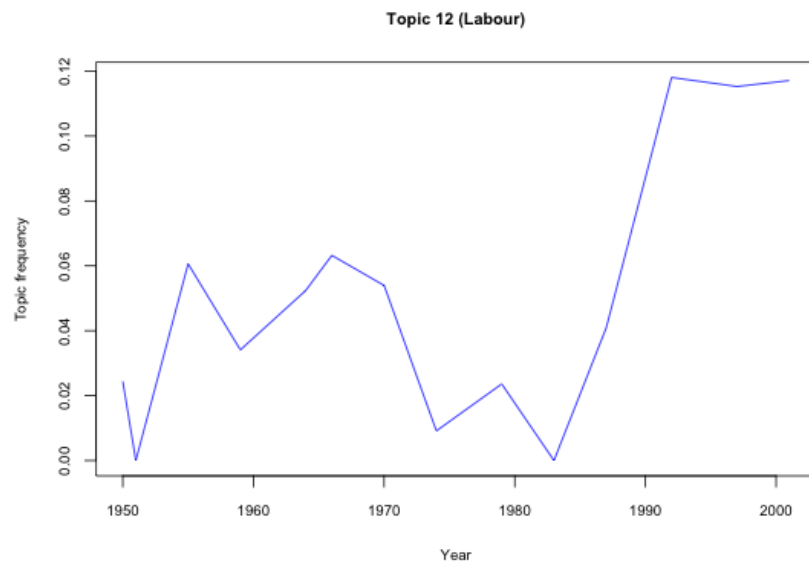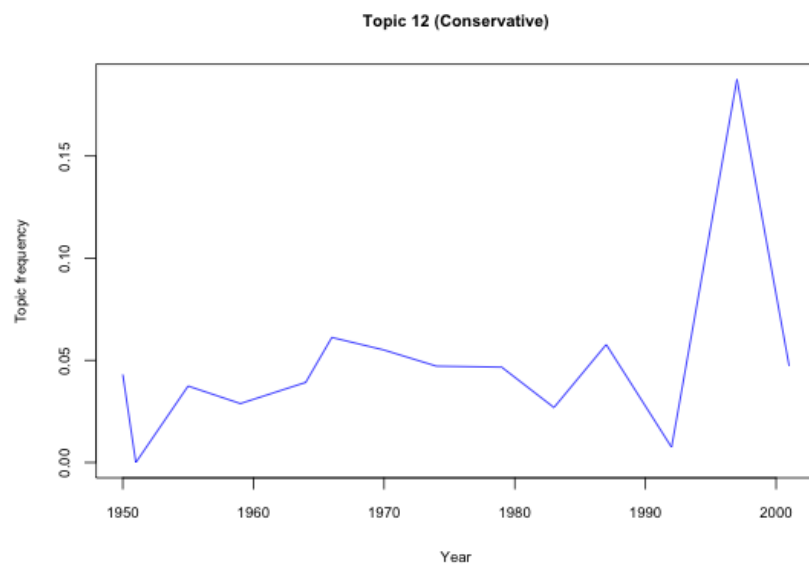
Figure 12: Topic 12 - Education (Labour)

Figure 13: Topic 12 - Education (Conservative)

Topic 24 - Healthcare and the NHS

The topic related to healthcare and the National Health Service (NHS) shows a
trend in popularity similar to that of Education. This topic does not feature
prominently in earlier manifestos, but its popularity increases significantly from
the 1979 election onwards. In terms of macroeconomic indicators, this topic
appears to be linked to the unemployment rate, at least until the 1983 election,
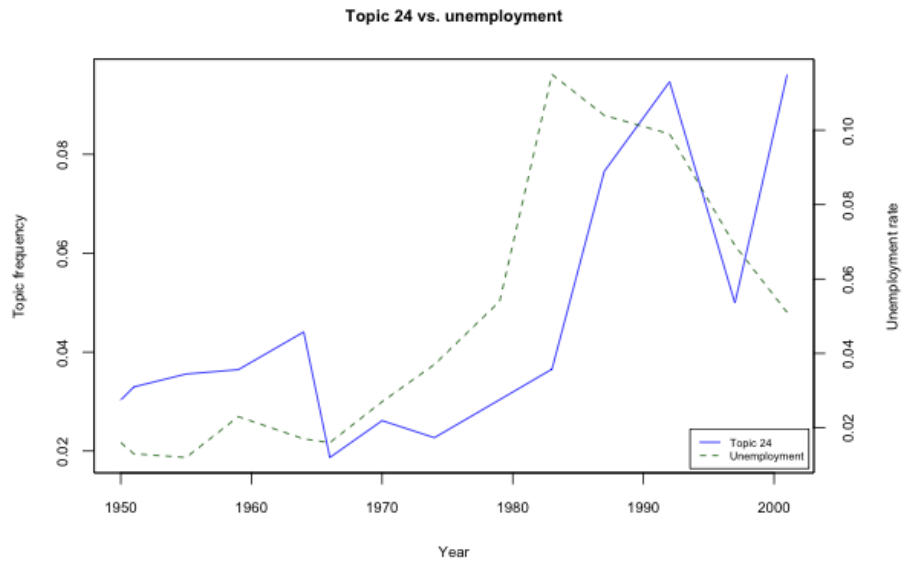as shown in the figure below.

Figure 14: Topic 24 - Healthcare and the NHS (aggregate)

The similarity between the two parties is remarkable, especially if we consider
that the Labour government introduced the NHS in 1948. For that reason, we
would expect this topic to be featured more prominently in Labour manifestos
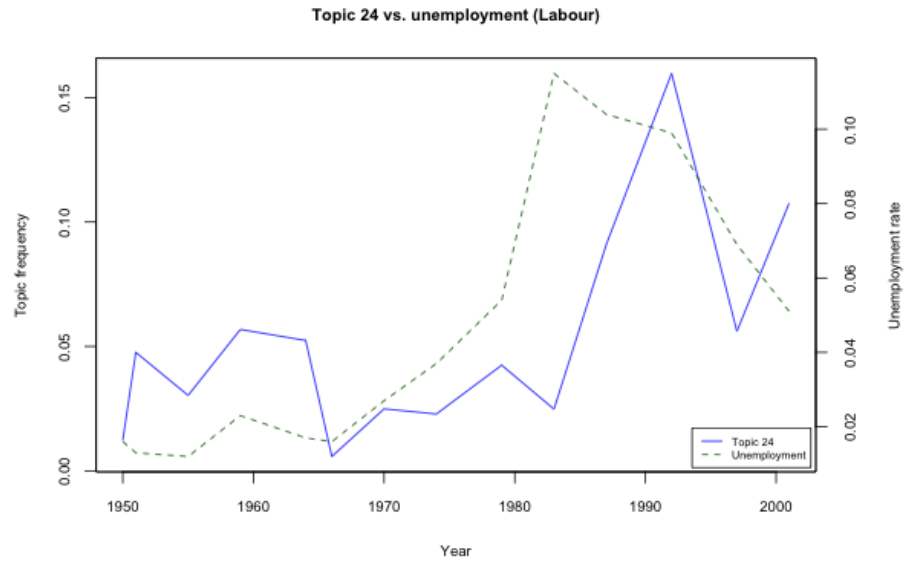during the 50s.

Figure 15: Topic 24 - Healthcare and the NHS (Labour)

Figure 16: Topic 24 - Healthcare and the NHS (Conservative)

15

Topic 25 - Taxation

The popularity of this topic appears to be linked the unemployment rate. Thus, taxation does not feature prominently in the early phase of our analysis, but its popularity increases from the 1970s onwards, as shown in the figure.
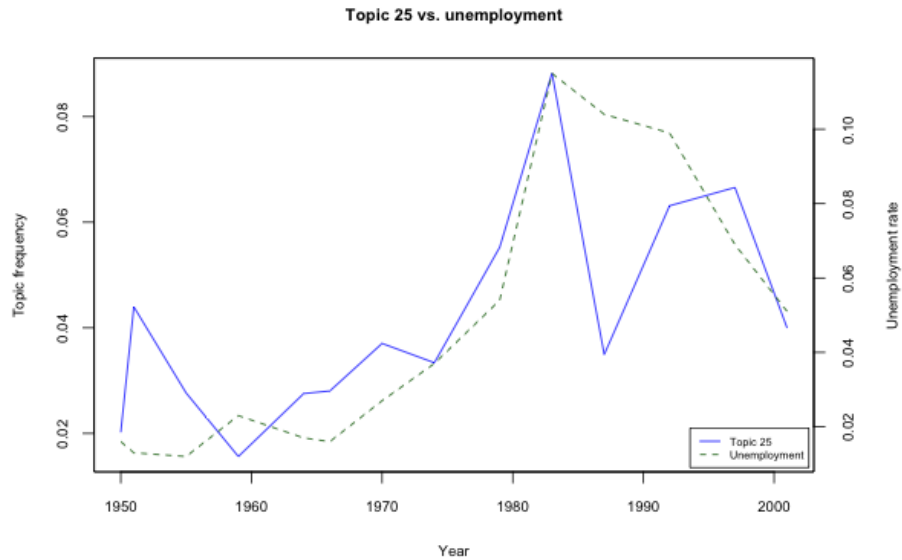

Figure 17: Topic 25 - Taxation vs. unemployment (aggregate)

The popularity pattern appears to be different for the two parties. For Labour, taxation is a popular topic early on but it declines for most of the period considered, with the exception of a peak at the time of the 1983 election (a period of very high unemployment). For the Conservative party, taxation becomes a popular topic from the mid-70s onwards, albeit with some peaks and troughs.
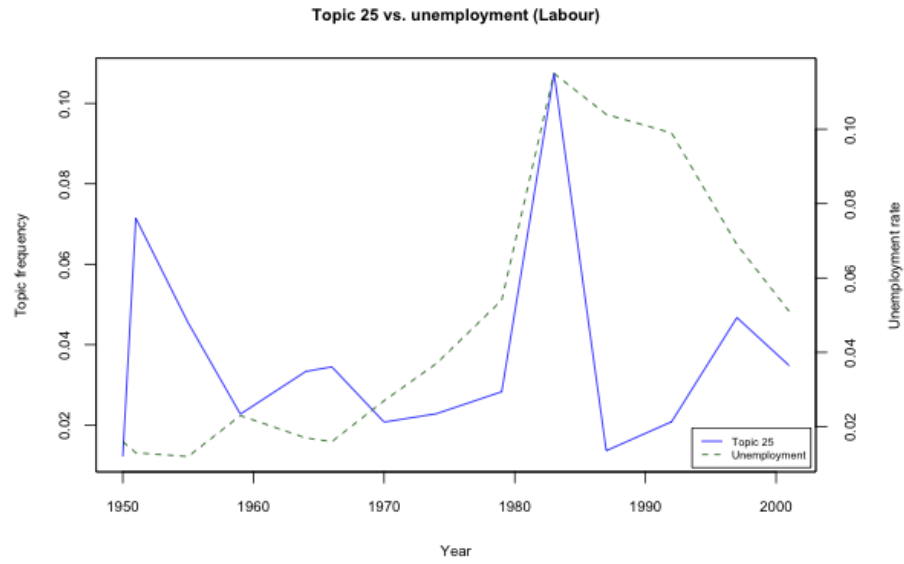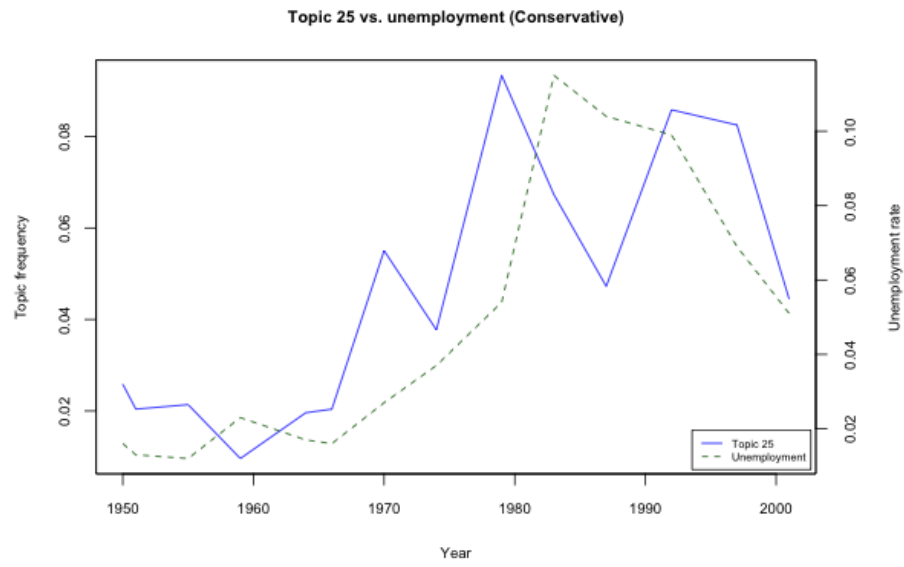
Figure 18: Topic 25 - Taxation (Labour)

Figure 19: Topic 25 - Taxation (Labour)

Topic 26 - Inflation and the cost of living

Not surprisingly, in aggregate terms the popularity of this topic is related to the level of inflation over the years, as shown in the figure.

Figure 20: Topic 26 - Inflation (aggregate)

However, the aggregate measure hides a different pattern for the two parties. Specifically, while this is a popular topic for both the Labour party and the Conservative party in 1950, its popularity in Labour manifestos declines gradually over time, while in the Conservative manifestos it tracks the rate of inflation more closely. This would suggest that the Conservative party paid more attention to this macroeconomic indicator when shaping its policies.

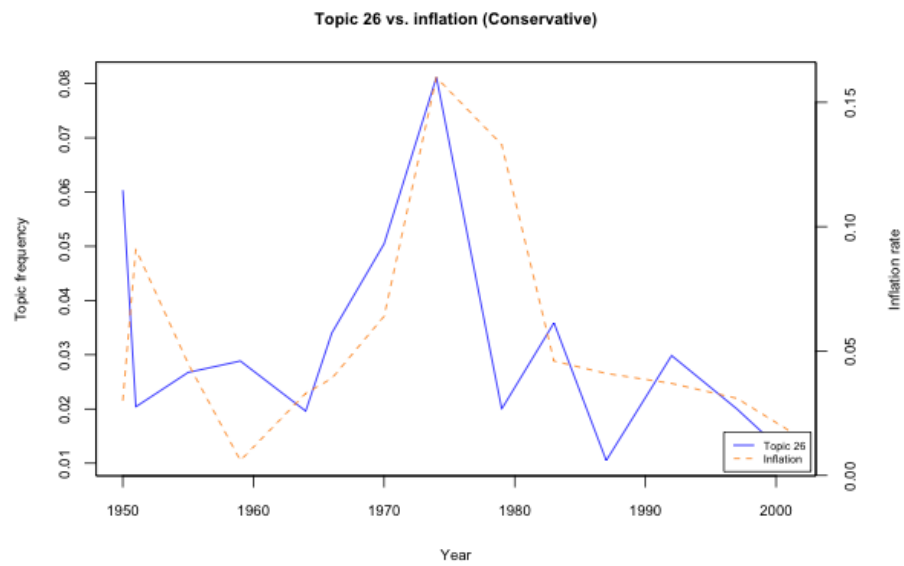18

Figure 21: Topic 26 - Inflation (Labour)

Figure 22: Topic 26 - Inflation (Conservative)

19

Topic 30 - Trade unions

This topic does not appear to be closely linked to any of the macroeconomic indicators we considered. Its popularity is high at the beginning of the analysis period. It declines in the 50s and 60s, to then rise again to prominence during the turbulent 1970s. It peaks at the time of the 1979 election, which took place after the so-called *'winter of discontent'*, a period characterised by widespread industrial action. After this time, the topically gradually loses popularity, with the exception of a further peak in 1992.
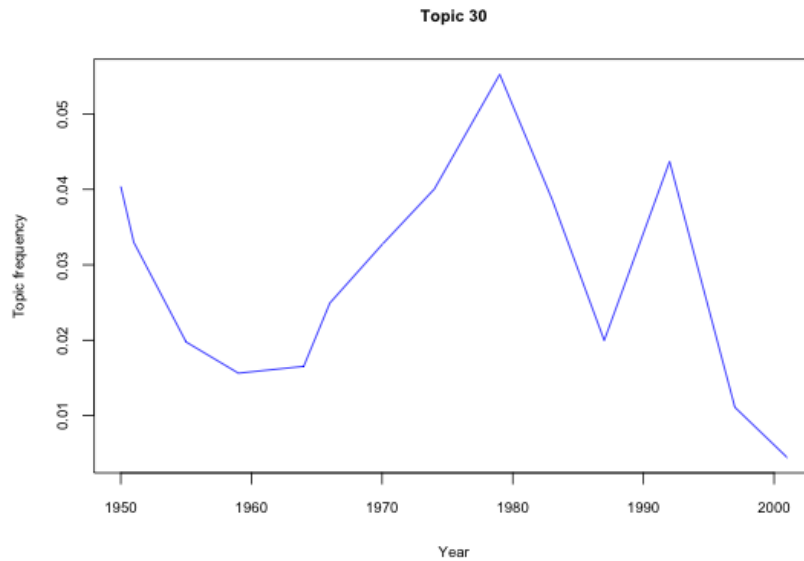

Figure 23: Topic 30 - Trade unions (aggregate)

The popularity patterns for the two parties are similar in the first part of the analysis period, though we can observe that the topic acquires popularity in the Labour manifestos earlier than in the Conservative ones. After 1983, the topic's popularity decline in the manifestos of both parties, with an exception of a peak in 1992 for the Conservative party. This may be related to trade union issue associated with privatisations.
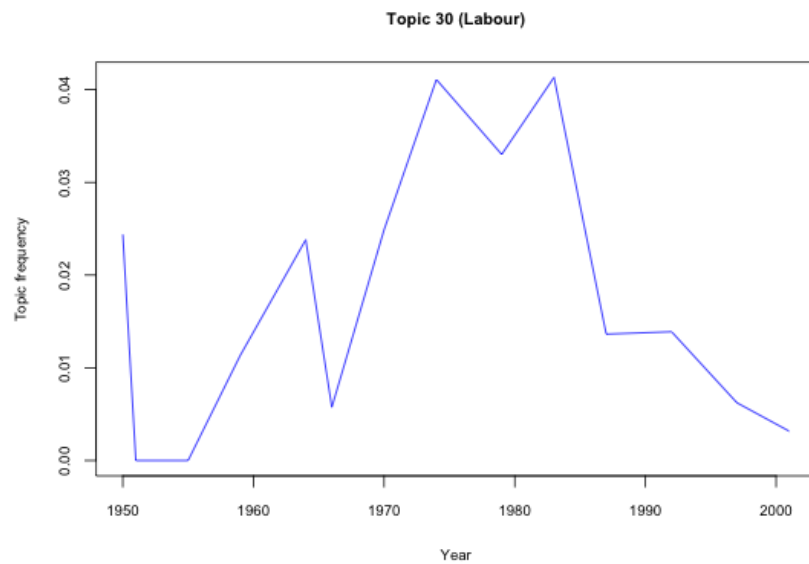
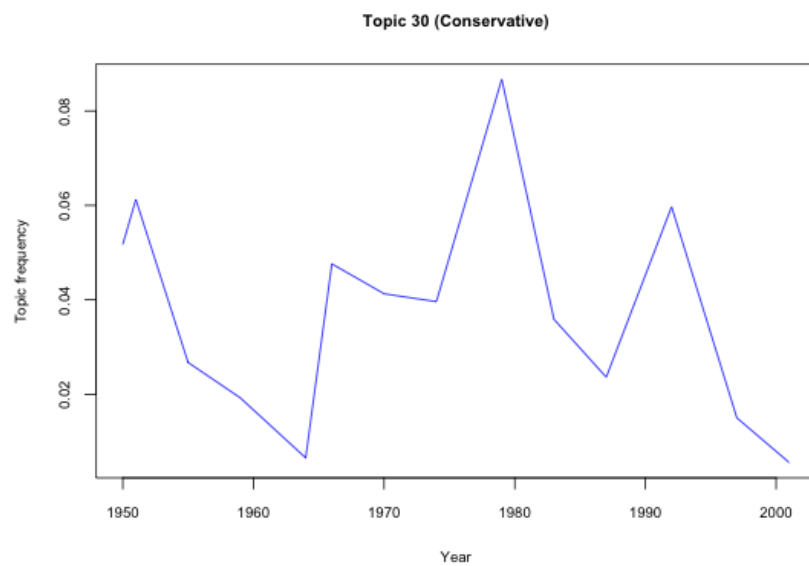Figure 24: Topic 30 - Trade unions (labour)

Figure 25: Topic 30 - Trade unions (conservative)

21

Similarity between parties

The analysis presented so far has identified a degree of similarity in the popularity of the topics covered in the election manifestos of both parties. To better explore this aspect, we consider the evolution of the entire topics distribution for the two parties over time.

The table below shows, for each general election year and for each party, the top three topics in terms of popularity.

Party and ranking	1950	1951	1955	1959	1964	1966
Labour - Top 1	27	26	6	29	5	13
Labour - Top 2	23	29	26	3	13	10
Labour - Top 3	21	6	5	5	22	12
Conservative - Top 1	6	22	5	5	5	9
Conservative - Top 2	2	23	2	6	13	13
Conservative - Top 3	9	2	6	18	22	4

Party and ranking	1970	1974	1979	1983	1987	1992
Labour - Top 1	4	2	20	10	24	24
Labour - Top 2	9	16	27	25	5	12
Labour - Top 3	13	27	6	15	13	14
Conservative - Top 1	29	20	25	5	24	23
Conservative - Top 2	3	26	30	21	3	5
Conservative - Top 3	9	3	2	25	11	28

For the table above, we can see that in 7 out of 14 election years, the Labour and Conservative parties had a least one topic in common at the top their respective topic popularity ranking. This suggests that the similarity between the manifestos is potentially high even if it tends to fluctuate over time.

To quantify the extent of the similarity over time, for each year and for each party we have calculated the average topic distribution across all 30 topics. This has provided us with an average vector space representation of the content of each manifesto in each year, which we could interpret as a quantitative representation of the general "tone" of each manifesto. Using this information, we have calculated the cosine similarity between each vector pair (i.e. Labour vs. Conservative in each year). The following chart shows how the cosine similarity varies over time.
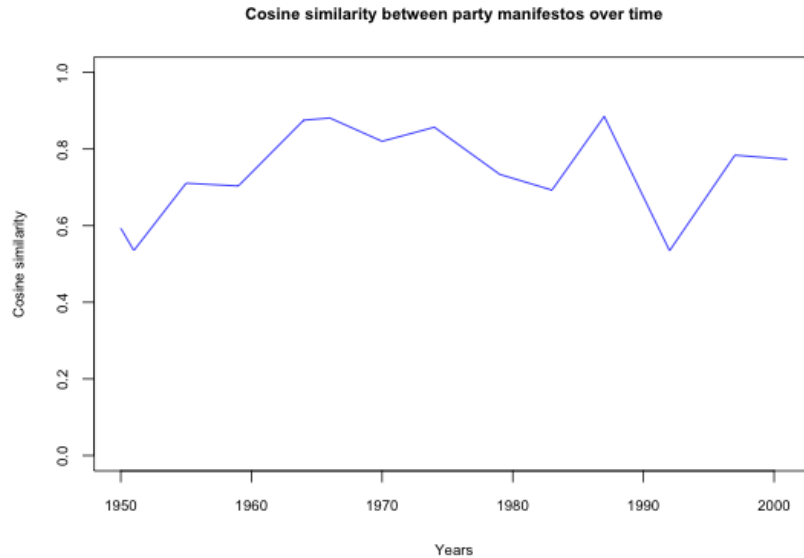
Figure 26: Cosine similarity over time

First of all, we note that cosine similarity is generally high, dipping below 0.6 only on a couple of occasions. This indicates a high degree of similarity, even though this type of analysis does not allow us to determine the degree of agreement on each topic between the two parties.

The analysis also shows that the two parties appear more dissimilar earlier on in the analysis period, but they converge during the 50s, 60s and the early 70s. From about 1974 onwards, there is a slight decline in similarity for a couple of elections, which is however counterbalanced by a sharp increase in 1987.

Cosine similarity dips again in 1992, a year in which Labour was seeking to position itself as a dramatic alternative to more than a decade of Conservative governments.

Cosine similarity increases again in 1997. It could be argued that Tony Blair's New Labour won the 1997 election by conquering the middle-class electorate. This might have been achieved by shifting the position of his party towards the political centre, making it more similar to the Conservative party.

Conclusion

In this analysis we set out to address the following questions:

- whether the two main parties' policies follow similar popularity patterns;

23

- whether the topics' popularity is related to the surrounding economic environment, and
- the extent to which the parties differ from each other in their way of communicating with the electorate, both in terms of the choice of the individual topics addressed and in the overall tone of their manifesto.

While the analysis has been carried out at relatively high level, we can say that some topics appear to be directly linked to the unemployment rate and, to a lesser extent, to inflation. We do not notice any relationship with GDP growth.

We also observe that the two main parties tended to show a degree of similarity over the years. The similarity in the "tone" of their respective manifestos tend to be relatively high over time, albeit with some peaks and troughs.

We believe that this analysis, while mainly descriptive, could provide an interesting platform for further investigation in various disciplines such as economics, sociology and political science. Possible developments could include:

- the analysis of how different topics may be blended together
- the extent to which parties, while discussing similar topic, show a degree of agreement over them
- a focus on specific topics which, paired with the collection of more metadata and socio-economic indicators, could shed more light on the policies put forward in these areas by the two main parties;
- the collection of additional text information, for example from official parliament records (e.g. Hansard), which would provide a corpus covering a continuous period of time.

Appendix - Word clouds of main topics

Topic 1 - Welfare and benefits

Figure 27: Topic 1 - Welfare and benefits

Topic 3 - Housing

Figure 28: Topic 3 - Housing

Topic 7 - Employment

Figure 29: Topic 7 - Employment

Topic 12 - Education

Figure 30: Topic 12 - Education

Topic 24 - Healthcare and the NHS

Figure 31: Topic 24 - Healthcare and the NHS

Topic 25 - Taxation

Figure 32: Topic 25 - Taxation

Topic 26 - Inflation and cost of living

Figure 33: Topic 26 - Inflation and cost of living

Topic 30 - Trade unions

Figure 34: Topic 30 - Trade unions