Database Design & Web Implementation

Assignment 2 – Data Scrubbing and Excel
Hashim Hayat

Party Manifesto Analysis

Choose Data

For this assignment, I am looking at the manifestos published by political parties that present their political ideology and preferences to the general public.

Manifesto data fetched from: http://www.libdems.org.uk/read-the-full-manifesto

Part A: The Algorithm

Algorithm

The algorithm look at a party manifesto to calculate and analyze the political ideology of the party on the left-right scale using key words, sentences and phrases that left or right connotations. A similar has been used by many political scientists and data analyzers such as the *manifesto project* and Wagner and Meyer (2016). A third party sentiment analysis API has been used to judge the attitude in which each reference key word has been used in the manifesto.

INPUT
manifesto
ALGORITHM
OUTPUT
sentiment

The reference key words or sentences are divided into two categories: left and right. Left reference include words or phrases that are supported or rejected by the leftist parities and vice versa.

Step 1: **Fetching Data from the web**

The first step gets and downloads the manifesto in a text format directly from the web. Urllib, a python library to fetch the data. After the data is requested, it is converted into string format and then all the characters in the text are converted to lower case characters.

Step 2: **Removes old and temporary files**

This step removes any old and temporary data files from the working directory.

Step 3: **Writes the data acquired form Step 1 on file**

The manifesto text data fetched from the web is written on a file in this step for backup and offline usage.

Step 4: **Extracts data from reference files**

In order to analyze the manifestos, the user can include reference files that includes words, sentences or phrases that correspond to either political ideology (left or right). For instance, if leftist parties support the implementation of taxes, the user may include taxes into the left_ref.txt file.

The user may also include screen shots of words and phrases taken from published papers or anywhere. In order to fetch the words and phrases from the images, we use a python library that works as an OCR.

The words in the reference files are read from the text files or the image files and included into LEFT and RIGHT lists. Unimportant words, symbols and spaces such as and, for, that are removed from the reference lists.

Step 5: **Performs Left and Right Analysis**

In the final stage, the program goes through each word in the LEFT and RIGHT lists and find those words in the manifesto data. For each word in the list, a function called the word_freq counts the number of instances that word was repeated and returns the indexes at which the word is present.

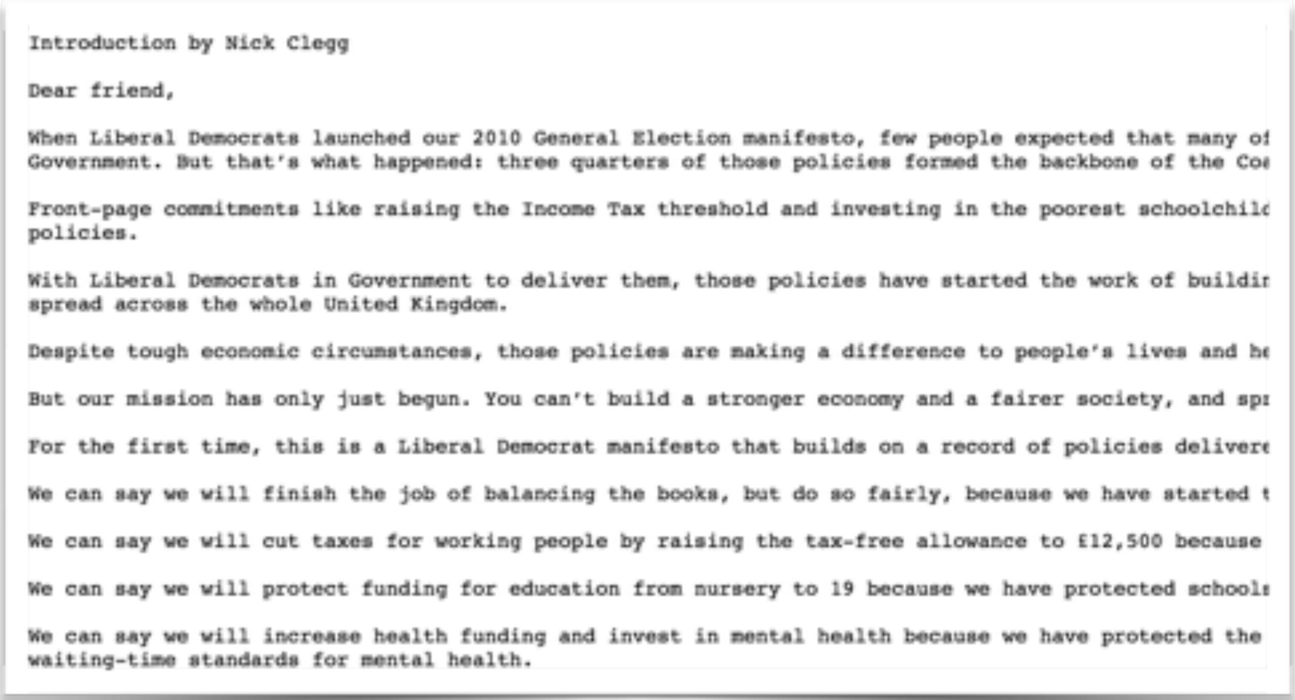
The indexes are later used to extract the sentences, one by one, in which the word was used using the function called extract_sentence.

The sentence is then fed into the sentiment analyzer function that uses an API call to a third part sentiment analyzer on the web to analyze whether the sentence has a positive sentiment or negative.
Sentiment Analysis API: http://text-processing.com/api/sentiment/

For instance, if the word is “military”, the algorithm extracts all the sentences in which the words military has been used. Then each sentence is fed into the sentence analyzer to find out if the word military is used in a positive reference or a negative reference.

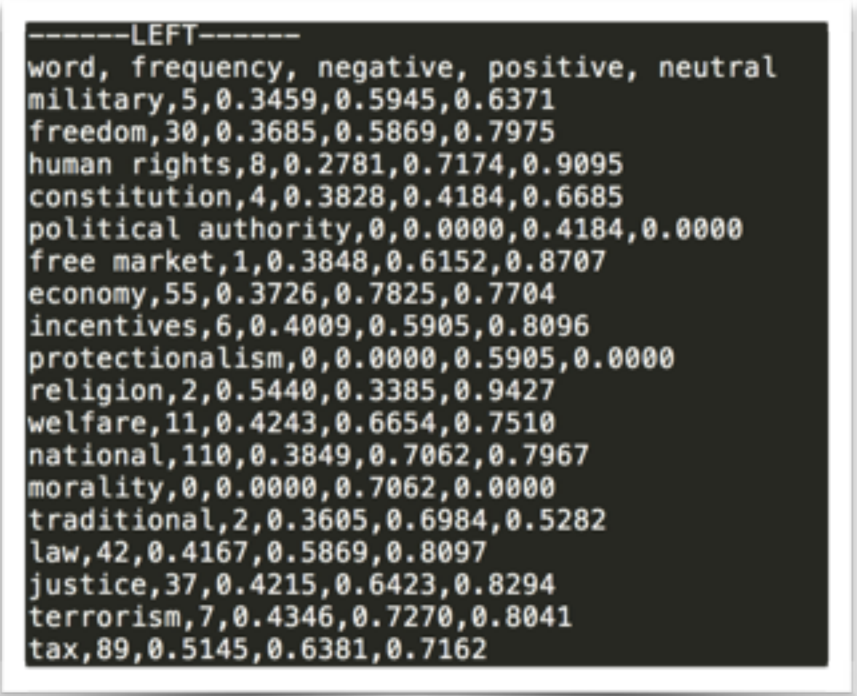
If a word is used more than once, the average ratio of sentiment is returned. The words along with their frequency, negative and positive sentiments are written on the two csv-files: left_analyis and right_analyis in the result folder.

Input File (web-view):



```
Introduction by Nick Clegg

Dear friend,

When Liberal Democrats launched our 2010 General Election manifesto, few people expected that many of
Government. But that's what happened: three quarters of those policies formed the backbone of the Coa

Front-page commitments like raising the Income Tax threshold and investing in the poorest schoolchild
policies.

With Liberal Democrats in Government to deliver them, those policies have started the work of buildin
spread across the whole United Kingdom.

Despite tough economic circumstances, those policies are making a difference to people's lives and he

But our mission has only just begun. You can't build a stronger economy and a fairer society, and spr

For the first time, this is a Liberal Democrat manifesto that builds on a record of policies delivere

We can say we will finish the job of balancing the books, but do so fairly, because we have started t

We can say we will cut taxes for working people by raising the tax-free allowance to £12,500 because

We can say we will protect funding for education from nursery to 19 because we have protected schools

We can say we will increase health funding and invest in mental health because we have protected the
waiting-time standards for mental health.
```

Output File:

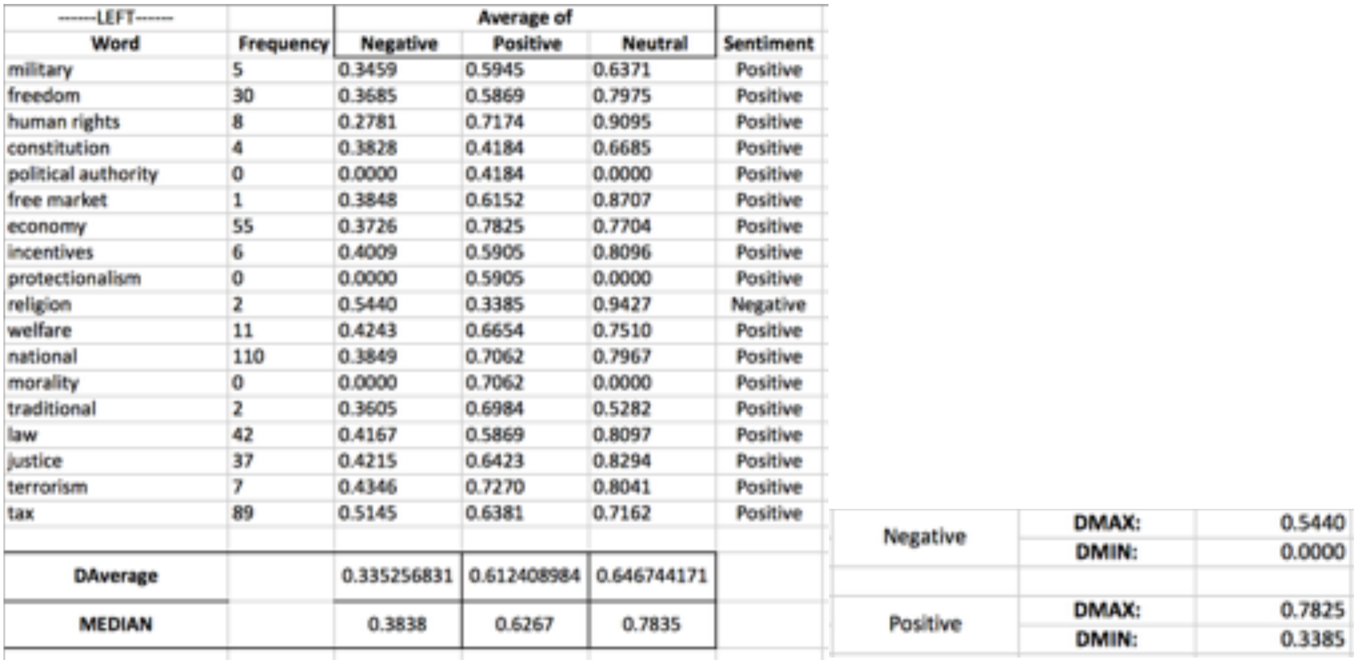


```
------LEFT------
word, frequency, negative, positive, neutral
military,5,0.3459,0.5945,0.6371
freedom,30,0.3685,0.5869,0.7975
human rights,8,0.2781,0.7174,0.9095
constitution,4,0.3828,0.4184,0.6685
political authority,0,0.0000,0.4184,0.0000
free market,1,0.3848,0.6152,0.8707
economy,55,0.3726,0.7825,0.7704
incentives,6,0.4009,0.5905,0.8096
protectionalism,0,0.0000,0.5905,0.0000
religion,2,0.5440,0.3385,0.9427
welfare,11,0.4243,0.6654,0.7510
national,110,0.3849,0.7062,0.7967
morality,0,0.0000,0.7062,0.0000
traditional,2,0.3605,0.6984,0.5282
law,42,0.4167,0.5869,0.8097
justice,37,0.4215,0.6423,0.8294
terrorism,7,0.4346,0.7270,0.8041
tax,89,0.5145,0.6381,0.7162
```

Part B: Microsoft Excel

Version: Mac 2011

The comma separated files, left_analyis.csv and right_analyis.csv are imported into the excel spread sheet for further analysis.



------LEFT------		Average of			
Word	**Frequency**	**Negative**	**Positive**	**Neutral**	**Sentiment**
military	5	0.3459	0.5945	0.6371	Positive
freedom	30	0.3685	0.5869	0.7975	Positive
human rights	8	0.2781	0.7174	0.9095	Positive
constitution	4	0.3828	0.4184	0.6685	Positive
political authority	0	0.0000	0.4184	0.0000	Positive
free market	1	0.3848	0.6152	0.8707	Positive
economy	55	0.3726	0.7825	0.7704	Positive
incentives	6	0.4009	0.5905	0.8096	Positive
protectionalism	0	0.0000	0.5905	0.0000	Positive
religion	2	0.5440	0.3385	0.9427	Negative
welfare	11	0.4243	0.6654	0.7510	Positive
national	110	0.3849	0.7062	0.7967	Positive
morality	0	0.0000	0.7062	0.0000	Positive
traditional	2	0.3605	0.6984	0.5282	Positive
law	42	0.4167	0.5869	0.8097	Positive
justice	37	0.4215	0.6423	0.8294	Positive
terrorism	7	0.4346	0.7270	0.8041	Positive
tax	89	0.5145	0.6381	0.7162	Positive
DAverage		0.335256831	0.612408984	0.646744171	
MEDIAN		0.3838	0.6267	0.7835	

Negative	**DMAX:**	0.5440
	DMIN:	0.0000
Positive	**DMAX:**	0.7825
	DMIN:	0.3385

AVERAGE and **Median** values are is calculated for the Negative, Positive and Neutral average values of each key word. The spreadsheet also calculates the **DMAX** and **DMIN** for Positive and Negative sentiments for each file. Moreover, according to the Negative and Positive sentiment of each key word, an average sentiment is calculated in the **Sentiment** column.

EXTRA CREDIT

- Data fetched from the Web
- Unusual form of data used. Not regular statistical data with columns and rows.
- Use of web API to analyze data.
- Python pytesseract library used to read data from images.
- Proper file management, cleaning up of temporary and old files.
- Program made in a reusable manner.