The Title of Your Dissertation

Jacob Barrow - 40337360

Submitted in partial fulfilment of
the requirements of Edinburgh Napier University
for the Degree of
BSc (Hons) Computing Science

School of Computing

October 3, 2020

Authorship Declaration

I, Jacob Barrow, confirm that this dissertation and the work presented in it are my own achievement.

Where I have consulted the published work of others this is always clearly attributed;

Where I have quoted from the work of others the source is always given. With the exception of such quotations this dissertation is entirely my own work;

I have acknowledged all main sources of help;

If my research follows on from previous work or is part of a larger collaborative research project I have made clear exactly what was done by others and what I have contributed myself;

I have read and understand the penalties associated with Academic Misconduct.

I also confirm that I have obtained informed consent from all people I have involved in the work in this dissertation following the School's ethical guidelines.

Signed:

Date:

Matriculation no:

General Data Protection Regulation Declaration

Under the General Data Protection Regulation (GDPR) (EU) 2016/679, the University cannot disclose your grade to an unauthorised person. However, other students benefit from studying dissertations that have their grades attached.

Please sign your name below one of the options below to state your preference.

The University may make this dissertation, with indicative grade, available to others.

The University may make this dissertation available to others, but the grade may not be disclosed.

The University may not make this dissertation available to others.

Abstract

Contents

List of Tables

List of Figures

Acknowledgements

Insert acknowledgements here

I would like to thank...

1 Introduction

You can fill out sections as you please.

1.1 Overview Of Project Milestones

This is a sub sub section with a list of bullet points.

- A working X, that will be used for this investigation.
- Investigation of current tools and their potential use during an investigation of X .
- Programming of X with related frameworks Y and Z.
- That is all.

2 Background Research

2.1 Existing Approaches

A range of studies have already considered several aspects of this project.

Manjesh, Kanakagiri, Vaishak, Chettiar, and Shobha (2017) used Natural Language Processing (NLP) to identify 'clickbait' headlines, however they did not consider the headline's relationship to the article.

)

2.2 Possible Approaches

2.2.1 Neural Networks and Deep Learning

2.2.2 Natural Language Processing

3 Data Collection

3.1 Attributes

Before collecting the data, it's important to decide what form it'll take and what attributes will be stored.

As the aim of the project is to identify incongruence between an articles headline and body, these two attributes will be included in the dataset. In order to identify trends and allow for further analysis, the article's date of publication and the publisher (e.g. BBC, The Guardian, etc.) will also be stored.

Collection could have gone further and retained the articles category (e.g. 'politics', 'sport' etc.), but different publishers categorise articles in different ways - for instance, the BBC has a combined 'Science and Environment' category, whereas The Guardian splits these into two distinct categories. Additionally, similar news articles can be filed under different categories, depending on the publisher. As this project's focus is on the article's content, and not categorisation, it can be considered out of scope to investigate the interplay between different publisher's approach to categorising articles.

3.2 Sources

The Independent is one of the only online publishers to make available their entire archives. Using the methods mentioned in Section 3.3, XXX articles were collected, from 2011 to 2020. This 9-year period should prove a useful

Publisher	Earliest	Latest	Total
BBC (Archive)	1950-01-21	2005-12-11	1857
The Independent (Archive)	2011-01-01	XXX	XXX
BBC (Daily)	2019-04-18	XXX	XXX
Daily Mail (Daily)	2019-09-24	XXX	XXX
The Guardian (Daily)	2020-06-25	XXX	XXX

Table 1: Extents of the data sources collected

dataset to analyse a potentially changing landscape in the congruity of news headlines.

The BBC has an 'On This Day' page[1] that has a very select archive from 1950-2005, and analysing these articles could produce some interesting results. However, each of these articles will have been hand-picked (as evidenced by the 'In Context' notes alongside each article), and only represent historic world news events. Therefore, these articles will not be a suitable representation for the average of the time period they are from.

As well as archives, current news was also collected from a range of publishers. Table 1 shows the full list of data sources collected, as well as the time range they cover and the total records obtained.

3.3 Obtaining the Data

To collect the data, several Python scripts were created. For the daily news, the publishers' various RSS feeds were consulted, and for the archives a more customised approach was taken.

These scripts utilise the BeautifulSoup library to parse each article's webpage and scrape them for the headline, date and body text. As each publisher builds their websites using different design patterns and with different technologies, each script had to be tailor made to fit the page structure.

In addition, some sites implemented a strict rate-limit on requests - to make a copy of The Independent's archive took around XXX days to complete, scraping one article every 15 seconds.

3.4 Ethics

Across a variety of datasets, XXX articles were collected for analysis. This is a substantial amount of data, and represents the work of many individual journalists and news publishers.

[1] `http://news.bbc.co.uk/onthisday`

While automated techniques were used to collect the data, everything collected was publically accessiable. In addition, it is legal to make a digital copy of copyrighted data for non-commercial research (Intellectual Property Office, 2014). Even so, care still needs to be taken in the obtainment of the data in order to avoid overloading or altering the regular service of these archives. As mentioned above, requests were rate-limited to avoid inadvertant denial of service attack, and spread out over a long period of time.

References

BBC. (2013, Feb). *On this day.* http://news.bbc.co.uk/onthisday. (Accessed: 2020-09-29)

Intellectual Property Office. (2014, Jun). *Exceptions to copyright.* https://www.gov.uk/guidance/exceptions-to-copyright#text-and-data-mining-for-non-commercial-research. (Accessed: 2020-10-02)

Manjesh, S., Kanakagiri, T., Vaishak, P., Chettiar, V., & Shobha, G. (2017). Clickbait pattern detection and classification of news headlines using natural language processing. In *2017 2nd international conference on computational systems and information technology for sustainable solution (csitss)* (p. 1-5).

Appendices

A Project Overview

A.A Example sub appendices

...

B Second Formal Review Output

Insert a copy of the project review form you were given at the end of the review by the second marker

C Diary Sheets (or other project management evidence)

Insert diary sheets here together with any project management plan you have

D Appendix 4 and following

insert content here and for each of the other appendices, the title may be just on a page by itself, the pages of the appendices are not numbered, unless an included document such as a user manual or design document is itself pager numbered.