2019

Is there a Correlation Between Wikidata Revisions and Trending Hashtags on Twitter?

Paula Dooley [Thesis]
Technological University Dublin.

Is there a correlation between Wikidata revisions and trending hashtags on Twitter?

Student Name

Paula Dooley

A dissertation submitted in partial fulfilment of the requirements of
Technological University Dublin for the degree of
M.Sc. in Computer Science (Advanced Software Development)

2019

I certify that this dissertation, which I now submit for examination for the award of MSc in Computing (Advanced Software Development), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Technological University Dublin and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: Paula Dooley

Date: **16th June 2019**

ABSTRACT

Twitter is a microblogging application used by its members to interact and stay socially connected by sharing instant messages called tweets that are up to 280 characters long. Within these tweets, users can add hashtags to relate the message to a topic that is shared among users. Wikidata is a central knowledge base of information relying on its members and machines bots to keeping its content up to date. The data is stored in a highly structured format with the added SPARQL protocol and RDF Query Language (SPARQL) endpoint to allow users to query its knowledge base.

This research, designs and implements a process to stream live Twitter tweets and to parse existing Wikidata revisions XML files provided by Wikidata to identify if a correlation exists between the top Twitter hashtags and Wikidata revisions over a seventy-seven-day period.
The statistical evaluation tools *'Jaccard Ratio'* and *'Kolmogorov-Smirnov'* have found that a significant statistical correlation does not exist between Twitter hashtags and Wikidata revisions over the studied period.

Key words: Wikidata, Twitter, Hashtags, SPARQL, Trending, Microblogging, Kolmogorov-Smirnov, Jaccard Ratio

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor Dr. Bojan Božić for his help, expertise and guidance throughout this project. He has always been on hand to suggest improvements, answer questions and provide guidance through to the completion of the project and without his continued expertise this could not have been completed.

I would like to thank Dr. Sarah Jane Delaney for the original idea and although it has changed and evolved it set me on the path to complete this challenging project and has allowed me to explore the discipline of research.

I would also like to thank my husband, Francis Mahon, for his continued support and encouragement throughout this course.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1 INTRODUCTION

1.1 Background

"*The World Wide Web is a large-scale digital compendium of information that covers practically every sphere of human interest and endeavour"* (Smart & Shadbolt, 2015). This information is available through home computers and mobile phones and, with continuous advancements in technology, people have become increasingly more electronically connected. Along with this information, there has come many powerful innovation services facilitating both how people access information and how they connect with one another. Social networking sites, such as Twitter and Facebook have evolved alongside wiki-sites containing huge amounts of information, such as Wikidata and Wikipedia.

Twitter, established in 2006, is a microblogging application (Small, 2011) allowing subscribers to share 280 characters in real-time data, referred to as a tweet (Doshi, Nadkarni, Ajmera, & Shah, 2017). Twitter is used for people to stay socially connected, where individuals express their views, share information and interact with others over the network (Doshi et al., 2017). Twitter data has become a significant research tool for analysis in areas such as, predicting stock behaviour (Li, Zhou, & Liu, 2016); book recommendations from twitter feeds (Arulselvi, Sendhilkumar, & Mahalakshmi, 2017); sentiment analysis (Ahuja & Dubey, 2017) (Haripriya & Kumari, 2017); burstiness (Al Tamime, Giordano, & Hall, 2018); longevity of trending topics with predictions (Sundar & Kankanala, 2015); as well as trend identification (Doshi et al., 2017).

Wikidata, launched in 2012, is a knowledge base, containing multilingual collections of structured data, (Vrandecic, 2013) maintained by voluntary individuals and machines also known as bots. The aim of Wikidata was connecting several Wikimedia projects, for example the knowledge source Wikipedia, Wikimedia, Commons containing media files and WikiSource consisting of historical documents (Ruttenberg, 2019). Wikidata is a centralized location which continuously catalogues and updates information,

providing access to the most accurate and consistent information across Wikipedia editors (Ruttenberg, 2019).

1.2 Research Focus

There are two main parts to this research project. The first part extracts the data from both Twitter and Wikidata. Twitter consists of tweets posted by individuals and consists of hashtags, URL's, plain text and user names. The focus of this study will look at Twitter hashtags for comparison. This data is cleaned and prepared for comparison with Wikidata revision article titles. Wikidata (Goldfarb & Merkl, 2018) like Wikipedia (Medelyan, Milne, Legg, & Witten, 2009) is an encyclopaedia of information which has evolved over time through authors continually revising the data to keep the information current. A revision is considered any one of insert, delete or substitution of data to an article (Jhandir, Tenvir, On, Lee, & Choi, 2017). The top Wikidata revision articles and Twitter hashtags are identified over a seventy-seven-day period.

The second part of this project compares the Wikidata revisions and Twitter hashtags to identify if a correlation exists between the hashtags posted and Wikidata revisions made. Statistical formulae, *Kolmogorov-Smirnov & Jaccard's Ratio*, will compare the text-ranked results from each group to determine if a statistically significant correlation exists. Visualisation analytics will be used to provide insight in to the results of the Twitter trends and Wikidata revisions over the studied period.

1.3 Research Problem

This project firstly looks to identify trending topics from Twitter over a seventy-seven-day period by extracting real-time data from Twitter. This approach is driven by the importance of real-time analysis of social media for organizations to identify actions and make decisions (Haripriya & Kumari, 2017). The data tweets are cleaned by extracting the hashtag and are ranked based on the number of occurrences. For the same period of the extracted Twitter data, the Wikidata revision articles are identified and the article title is extracted. The article title requires cleaning by removing any white space to allow for direct text comparison. The total number of revisions per article is recorded to determine the top edited Wikidata articles for the seventy-seven-day period. Statistical

tools will identify if a statistically significant correlation exists between the Twitter trending items and the top Wikidata page revisions. Visualisation techniques will be used for both the Wikidata revisions and the streamed Twitter data to provide insights in to the data.

1.4 Research Objectives

The aim of this research is to identify if a statistically significant correlation exists between Wikidata revisions and Twitter trending hashtags. "*The term correlation refers to a mutual relationship or association between quantities*" (Dalinina, 2017) where *'Jaccard Ratio'* and *'Kolmogorov-Smirnov'* are used to measure the correlation between both groups of data.

The main research objective is to determine if trending topics in the English language Wikidata, identified by the title of the most frequently edited pages, show a statistically significant correlation to the real-time streaming data top-trending hashtags on Twitter, over the seventy-seven-day period, using the statistical analysis tools *'Jaccard Ratio'* and *'Kolmogorov-Smirnov'*.

The research question and research hypothesis aim to support the objective defined as:

- Research Question: Is there a correlation between Wikidata revisions and trending topics hashtags on Twitter determined by *'Jaccard Ratio'* and *'Kolmogorov-Smirnov'*?
- Null hypothesis (H0): a correlation does not exist between Wikidata revisions and trending hashtags on Twitter determined by *'Jaccard Ratio'* and *'Kolmogorov-Smirnov'*.
- Alternative hypothesis (H1): a correlation exists between Wikidata revisions and trending hashtags on Twitter determined by *'Jaccard Ratio'* and *'Kolmogorov-Smirnov'*.

1.5 Research Methodologies

This research incorporates both primary and secondary research. Initially, secondary research was conducted on existing literature which examined studies focused on Wikidata and Twitter data processing and analysis. This secondary research provided insight on both the current techniques for processing and analysing data and on the statistical analysis methods for text comparisons.

Primary research was conducted through streaming live twitter data over a seventy-seven-day period, where the hashtag lists within each tweet were extracted for analysis. Secondary research also incorporated extracting revisions from Wikidata[1] downloads that were used for further analysis. An experimental research method has been used on both sets of data to quantify whether a statistically significant correlation exists between Wikidata revisions and trending hashtag topics in Twitter.

This project has four main objectives that will test the hypothesis:

- To retrieve streamed Twitter data, extracting its hashtag items per tweet. The data will be cleaned. Up to four n-grams will be applied and the data will then be ranked based on the volume of tweets over the study period.
- To extract Wikidata page details and revision data from Mediawiki data dumps and, using SPARQL Protocol and RDF Query Language (*SPARQL) API endpoint*, to retrieve the individual revision page titles. The data will then be cleaned for processing by removing all spaces before counting and ranking the number of page titles based on the number of revisions occurring per page title over the study period.
- To identify if a statistically significant correlation exists between both the top revised Wikidata pages and the top trending hashtags on Twitter, the statistical techniques to be used in identifying the presence of correlation are *Jaccard's Ratio* and *Kolmogorov-Smirnov*.
- To provide additional insights in to the data results, using visualisation techniques like word cloud and bar graphs.

[1] https://dumps.wikimedia.org/wikidatawiki/20190601/ Wikidata dumps

1.6 Scope and Limitations

Sourcing both Twitter streamed data and Wikidata revisions was met with a number of challenges. Using streamed Twitter data meant being confined to the API limit restrictions made available through the Twitter Streaming API. Twitter provides an enterprise Power Track API for paying customers. However, access to this resource was not made available, having contacted Twitter asking for a waiver of fees for student research. Additionally, Twitter quoted a cost of 12,500 US dollars for one million historic tweets that could also not be waivered for student research.

Steaming live Twitter data came with implementation challenges to ensure that a constant stream of data was available for analysis in this study. Having resolved these issues in the implementation, the data streaming starts from 15th of March 2019. The target for this research was three consecutive months of live streamed twitter data but due to the confines of the thesis deadline seventy-seven days of Twitter streamed data is available to analyse.

The source of Wikidata dumps changed during this process. Initial attempts at extracting all revisions, yielded only the latest revision per Wikidata page[2]. This resulted in a change of direction, where meta-data-history XML files were parsed to extract all revisions per page from the date the twitter streaming started. With the use of the *SPARQL*[3] *API endpoint* the additional information per Wikidata item page was sourced.

1.7 Document Outline

This section provides a summary of the five chapters of the document:

- Chapter 2 contains details of the Literature Review completed which examined existing research in the areas of Wikidata and Twitter data processing. This section begins by discussing trending and microblogging in a technologically changing society. An in-depth look is taken at the Wikidata structure and revisions that are the focus of this study. The Twitter tweets structure is also

[2] https://dumps.wikimedia.org/wikidatawiki/20190601/

[3] https://query.wikidata.org/

examined focusing on the hashtags' property list used in this study. Natural Language Processing (NPL) is examined and the statistical options to validate correlation between two string lists is also detailed.

- Chapter 3 summarises the three phases of the Design and Implementation process of this work. Phase one outlines how the Twitter data was retrieved and examines the data processing steps with details of the assumptions made as part of this phase. Phase two examines the Wikidata retrieval and processing, detailing assumptions made during this phase of the work. Finally, phase three details the experiment completed to test the hypothesis using statistical tools: *Jaccard's Ratio* and *Kolmogorov-Smirnov*. To complete this section an outline of the strengths and weaknesses of the design and implementation are documented.

- Chapter 4 discusses the Results and Evaluation of the experiment, testing the research hypothesis. The results are presented for both *Jaccard's Ratio* and *Kolmogorov-Smirnov* which outline if a correlation exists between Wikidata revisions and Trending twitter hashtags. Finally, the strengths and weaknesses of the results and evaluation approach are examined.

- Chapter 5 contains the Conclusion, summarising the results found and examining exciting areas of future work that could be completed.

2 LITERATURE REVIEW

Trending topics are the most popular talked about items at any point in time over a social media network (Sundar & Kankanala, 2015). As events are more frequently talked about, it becomes more popular for a period of time where it then peaks and falls. There are a number of areas to be considered when deciding on the approach to use for trend analysis. The data studied may be streamed or static data and may even be a combination of both. The data to be used in the study impacts which Natural Language Processing (NLP) techniques are selected, varying depending on whether the data is structured or unstructured. In addition, the data selected for analysis determines which statistical measures are best suited in identifying text similarity. The following section will examine previous research completed in these areas.

2.1 Twitter and Hashtags

Microblogging sites are a platform used by individuals to share information and voice opinions on any topic, like current events, products or services. To businesses, this information is invaluable with immediate feedback available on their products and services. Users often voice their likes through social networking sites but are just as likely to voice their dislikes opening an opportunity for businesses to respond quickly. It is becoming more frequent for organizations to use this information to gain insight in to their customers' views on their products and to help improve such products (Trupthi, Pabboju, & Narasimha, 2017). Real-time analysis of social media data is increasingly studied due to the use of social media in sharing information and connecting people, assisting companies to make decisions. (Haripriya & Kumari, 2017). There is a large amount of unstructured data available today on microblogging sites, like twitter hashtags, reviews and information articles. There are two hundred million members which produce approximately four hundred million tweets daily, (Tajalizadeh & Boostani, 2019) sharing their thoughts, views and opinions on a vast range of topics including products, services and events (Hao et al., 2011). In recent years there have been many studies completed on Twitter data for analysis in areas such as, predicting stock behaviour (Li et al., 2016); book recommendations from twitter feeds (Arulselvi et al., 2017); sentiment analysis (Ahuja & Dubey, 2017) (Haripriya & Kumari, 2017);

burstiness (Al Tamime et al., 2018); longevity of trending topic with predictions (Sundar & Kankanala, 2015); and trend identification (Doshi et al., 2017).

Twitter has been selected as the microblogging site to be used in this study because of its popularity among users where its hashtags represent popular topics. As such, it will be the focus of this study to identify trending items of interest to the public over time based on the number of the hashtag occurrences across a streamed tweet corpus.

A tweet contains a number of property attributes with specific attribute types as shown in table 2.1 below. The main groupings of this data include, the tweet data, retweet information and user details. The 'entities' property attribute contains the Twitter 'hashtags' list as a sub-property that is the focus of this study. The 'hashtag' was introduced by Twitter to assist individuals in joining conversations but has grown to become a way to broadcast information to the wider audience (Wang, Liu, & Gao, 2016). The importance and impact of Twitter hashtag use is supported by a study looking at the types of hashtags used during a movement on social media, finding that multiple hashtags in the one tweet coupled with reference to high-profile public individuals, resulted in it having a more viral impact across social media than a tweet without these qualities (Wang et al., 2016). The hashtag, contained within a tweet, is prefixed with the symbol '#' and is followed by a string of one or more characters, symbols or numbers. The structure of a full tweet in JSON format is shown in appendix A.

Property name	Property Type	Property Description
created_at	String	Tweet creation datetime.
id	Int64	Tweet unique identifier.
id_str	String	Tweet unique identifier as a string.
text	String	Tweet text content up to 280 characters in length.
source	String	Devide details used to post the tweet.
truncated	Boolean	Indicates if a the tweet text is truncated.
in_reply_to_status_id	Int64	Original tweet identifier in the cases where the tweet is a reply.
in_reply_to_status_id_str	String	Original tweet identifier as a string in the cases where the tweet is a reply.
in_reply_to_user_id	Int64	Original author identifier in the case where the tweet is a reply.
in_reply_to_user_id_str	String	Original author identifier as a string in the case where the tweet is a reply.
in_reply_to_screen_name	String	Original author screen name in the case where the tweet is a reply.
user	User object	User infromation posting a tweet including id, name, screen name, geo location, timezone, language etc.
coordinates	Coordinates	The location latitude and longitude provided by user or client application.
place	Place	Tthe place a tweet is associated with.
quoted_status_id	Int64	A quoted tweet identifier.
quoted_status_id_str	String	A quoted tweet identifier as a string.
is_quote_status	Boolean	Indicator if tweet is quoted.
quoted_status	Tweet	The original quoted tweet details.
retweet_status	Tweet	Represents the original tweet that was retweeted.
quote_count	Integer	Number of times the tweet has been quoted.
reply_count	Int	Number of times the tweet has been replied to.
retweet_count	Int	Number of times this tweet has been retweeted.
favorite_count	Integer	Number of times the tweet liked by other users.
entities	Entities	Entities taken from the text includes the hashtag list, url list, user mentions and symbol list.
extebded_entities	Extended Entities	Holds media data.
favorited	Boolean	Indicates if liked by authenticating user.
retweeted	Boolean	Indicates if retweeted by authenticating user.
filter_level	String	The filter levels required to stream this tweet.
lang	String	Language identifier.

Table 2.1 Twitter properties

2.2 Wikidata and Revision Structure

Wikidata launched in 2012 as a knowledge base of the Wikimedia foundation, storing its knowledge in the structured format of subject-predicate-object statements (Heindorf, Potthast, Engels, & Stein, 2017). The knowledge base is organized and structured in to pages (Erxleben, Günther, Krötzsch, Mendez, & Vrandečić, 2014) as shown below in figure 2.2 for the Technological University Dublin retrieved from Wikidata[4]. Wikidata content is language independent supporting four-hundred-and-ten languages (Kaffee & Simperl, 2018), where the item language displayed is determined by the user's language settings.

[4] https://www.wikidata.org/wiki/Q55619051

"The data model of Wikidata is based on a directed, labelled graph where entities are connected by edges that are labelled properties." (Bielefeldt, Gonsior, & Krötzsch, 2018). There are two types of entities including items and properties. Each item entity has a page relating to a subject area, for example, a city, person or a university as shown below in figure 2.1 where it's data can be entered, edited or viewed. (Erxleben et al., 2014).

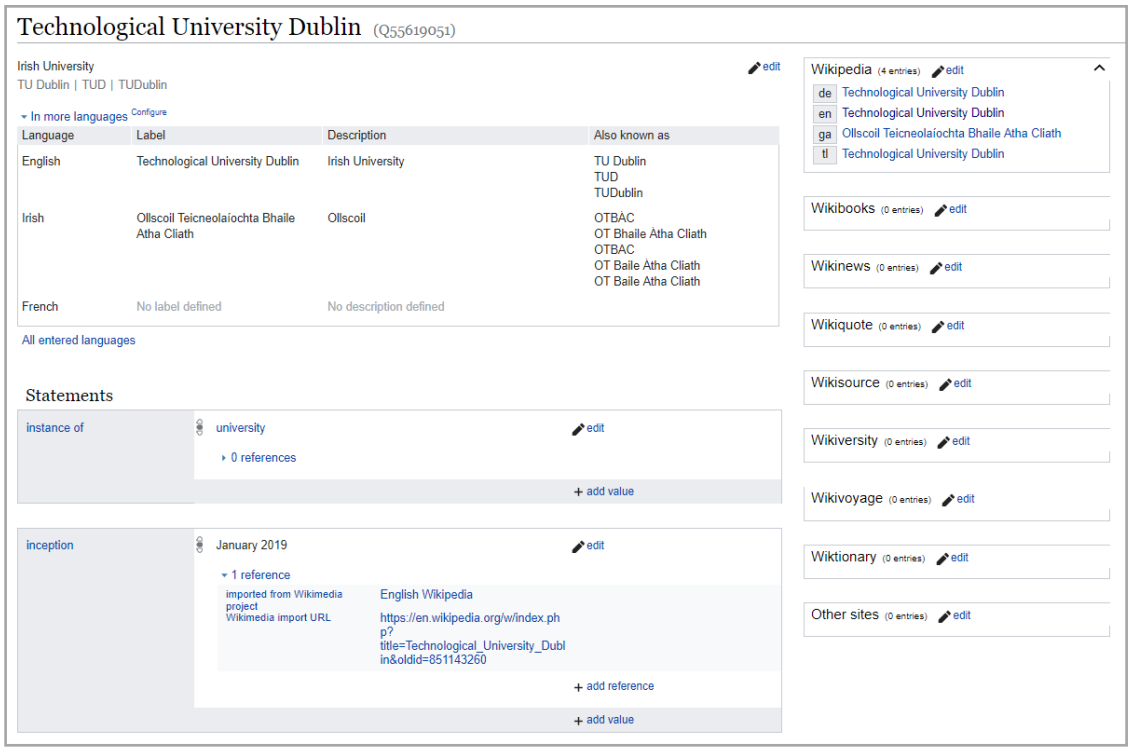


Figure 2.1 Wikidata page structure

The title is an opaque item identifier assigned automatically when the item is created beginning with the letter Q followed by a number (Erxleben et al., 2014). For example, 'Q5561905' is the title identifier for the Technological University Dublin. Its item head contains human-readable labels; descriptions and aliases; statements; and a set of site links supporting multiple languages codes (Heindorf et al., 2017). The items label, description and aliases, together referred to as terms, are used to display items in a natural language supported by the Wikidata (Erxleben et al., 2014). The site links consist of one link per site providing additional information, for example, links to Wikipedia articles.

The item body contains structured statements, also called sitelinks, encoding the structured knowledge of Wikidata in the form of subject-predicate-object triples where

the item is a subject, the property is a predicate and the value is an object (Heindorf, Potthast, Stein, & Engels, 2016). These property-value pairs are also referred to as claims (Heindorf et al., 2016). The statements are grouped by properties for example, ‘instance of’ or ‘inception’ as shown in figure 2.2, where each property is identified by an opaque identifier starting with the letter ‘P’ followed by digits.

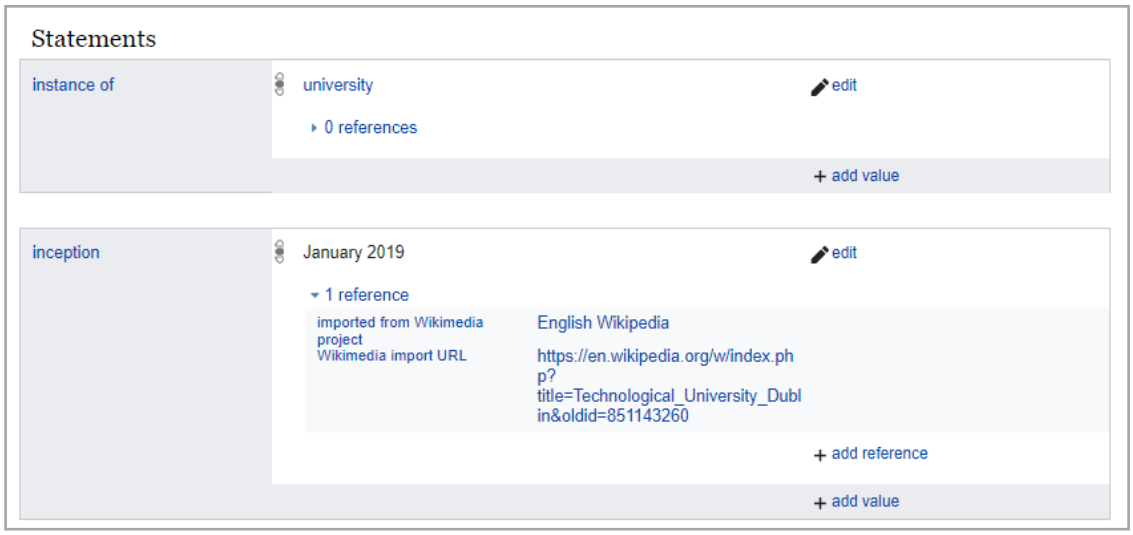


Figure 2.2 Wikidata statement structure

When a user edits the item, a new revision is created in the item revision history. Figure 2.3 shown below is the latest revisions for Technological University Dublin.

Figure 2.3 Wikidata page item revision history

Each Wikidata edit page contains the full revision history for the page by its Wikidata members in summary format. Each revision item can be selected to examine the changes

in more detail, with the facility to compare to the previous revision. Wikidata automated bot machines and its users keep the information up to date and accurate.

2.3 Data Retrieval

Full streaming of twitter data is used in studies, such as trend identification (Li et al., 2016), (Doshi et al., 2017), (Xie, Zhu, Jiang, Lim, & Wang, 2013) and sentiment analysis (Trupthi, Pabboju, & Narasimha, 2017), and will be used within this study. The approach to retrieving data from Twitter has varied across studies including examining historic data by topic (Sundar & Kankanala, 2015), (Ahuja & Dubey, 2017), as well as streaming the data by topic (Zangerle, Schmidhammer, & Specht, 2015), (Arulselvi et al., 2017). In one study, streaming twitter data by the topic was completed over a ten-month period to monitor the lifetime of trending topics over time finding, if a topic had six hundred or more tweets each day in the first week it would last a month, where positive and negative sentiment were impacted in tweets when determining if they would trend for more than one month (Sundar & Kankanala, 2015). Twitter provides a Streaming API that allows for the collection of publicly available tweets and this approach will be used to retrieve Twitter data. Wikidata dump files are made available through their website and come in a number of forms. The full Wikidata revision information can be downloaded which would rely on extracting the additional information via the SPARQL endpoint. SPARQL is a powerful API to access linked data collections that allow for retrieval of precise and insightful information in to the data. (Bielefeldt et al., 2018)

2.4 Natural Language Processing

"Computational linguistics, also known as natural language processing (NLP), is the subfield of computer science concerned with using computational techniques to learn, understand, and produce human language content."(Hirschberg & Manning, 2015)
There are a number of stages to natural language processing including tokenizing, stemming, stop-word removal, vector-space representation and similarity calculation (Runeson, Alexandersson, & Nyholm, 2007).

- Tokenizing is the process of changing the text to lower case and removing characters like brackets, hyphens and commas in the text so that the characters

are converted to a tokens (Runeson et al., 2007). These tokenized streams are often split in to words for further processing.

- The stemming process looks at the grammar meaning of the text and converting words that mean the same thing. An example of this would be working and work. (Runeson et al., 2007).
- Stop-word removal, involves the removal of common words for example ‘a’, ‘the’, ‘in’. These words do not contribute significantly to the statistical analysis of the data (Runeson et al., 2007). Dictionaries containing common stop words are available to compare the text under process against, and if found are removed. Inverse-frequency weighting to words is another approach that can be considered, where the most frequently occurring words in the full data set are considered for removal.

There are a number of Natural Language Processing (NLP) libraries that support NPL. The suite of libraries is used for text processing to clean the data before analysis. This process is taking unstructured data and applying a structure to the data (Trupthi, Pabboju, & Narasimha, 2017). The type of data under evaluation will vary the number of cleaning steps required to be completed. The NLP can include stop word removal, tokenization, stemming, classifying parsing and WordNet. (Trupthi, Pabboju, & Narasimha, 2017). Another technique when analysing text similarity is to split words in to n-grams to break up the sentences. This process can be completed at word-level or string-level as seen in the study examining duplication in text (Weissman, Ayhan, Bradley, & Lin, 2015).

2.5 Statistical Analysis Techniques for Text Correlation

There are a number of statistical analysis techniques to be considered when comparing text lists. When considering the statistical measures, the list characteristics are an important consideration. In the case of trend lists, in this study they are non-conjoined lists, where the lists may have different items within their lists. The lists are top-weighted; therefore, the top items of the list are more important than the lower ranked items and indefinite ranking will not be considered where a percentage of items will be examined. The following studies look at list similarity using statistical techniques:

- A study completed examining the correlations of search engine results URL's included *Jaccard Ratio* similarity distribution measure with different sizes for set similarity that included both with and without confidence levels, find a low overlap of two major search engines where 80% of queries had less than 3 search engine overlaps. (D'Alberto & Dasdan, 2011).
- In a study examining the likeness of Wikipedia pages for near duplicate detection Jaccard's similarity measure was used with a finding of a large amount of duplication within the Wikipedia page content (Weissman et al., 2015).
- Use of Jaccard Coefficient to determine the association between words was implemented in the language Python where it was found to be preforming well when measuring the similarity of words (Niwattanakul, Singthongchai, Naenudorn, & Wanapu, 2013).
- "*Weighted Kendall's Tahu is the number of swaps we would perform during the bubble sort in such a way to reduce one permutation to the other*"(D'Alberto & Dasdan, 2011), however this does not apply to this research as we do not have the same items in each list where an item may not exist in the second list.

2.6 Visualisation

Visualisation is a frequently used technique to display and explain results in a visual format and includes representation of data in formats such as a word cloud for visual representation of most frequent words, (Haripriya & Kumari, 2017); Time Series to show trends over time (Arulselvi et al., 2017), (Alsaadi, Almajmaie, & Mahmood, 2017); moving average to show the tweet rate (Arulselvi et al., 2017); and analysis bar graphs (Doshi et al., 2017).

3 DESIGN AND IMPLEMENTATION

This chapter details the design, implementation and statistical analysis performed to identify if a correlation exists between Twitter hashtags and Wikidata revisions. The overall process has been split in to three phases, where the details of each phase's implementation and processing details are outlined.

The implementation and processing work completed in this work consists of three distinct phases as outlined in figure 3.1. In phase one, data is streamed from Twitter and its hashtags are extracted and cleaned, applying n-grams before determining the top hashtags tweeted over a seventy-seven-day period. Secondly, for the same time-period, the Wikidata revisions are extracted from its available data dumps. The Wikidata titles are retrieved using SPARQL, identifying the top revision pages. Finally, statistical comparisons are completed on the top hashtags and Wikidata revisions to identify if a correlation exists. The edit-distance statistics will calculate the similarity between the text items in each list and a statistically significant correlation will be determined on the overall similarity of the text lists. The results are displayed through visualisation techniques.

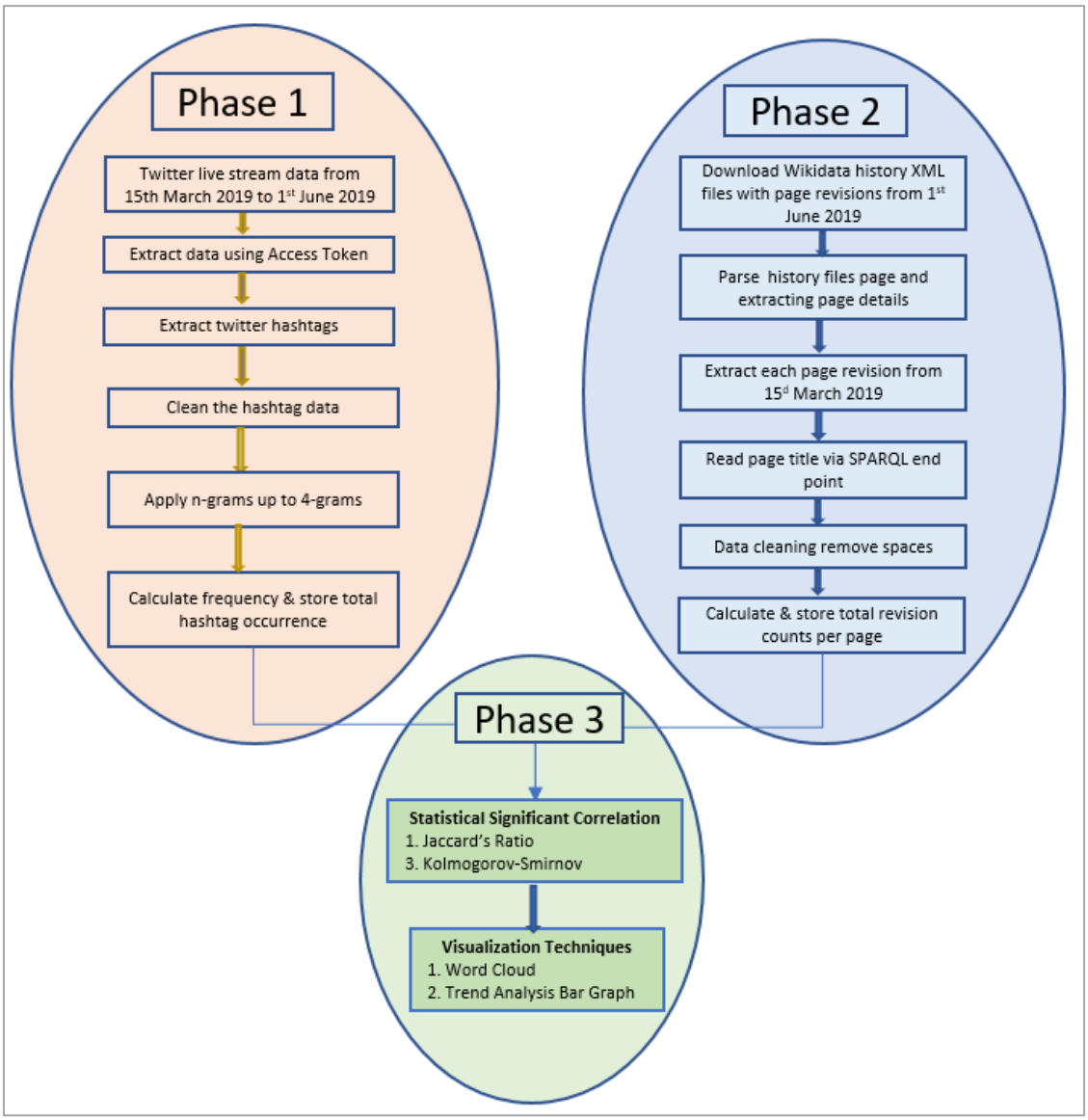


Figure 3.1 Three project phases of Wikidata and Twitter processing.

3.1 Twitter Data Gathering

During phase one, Twitter data is streamed to identity the top trending tweets by hashtag. The Twitter real-time data is accessed through its streaming Application Programming Interface (API) using tokens OAuth to ensure secure authorization data requests. The Streaming API returns the data and notifications in real-time from its public stream result in a JSON format (Li et al., 2016). Data from Twitter is streamed using the Twitter Streaming Application Program Interface (API) over a seventy-seven-day period.

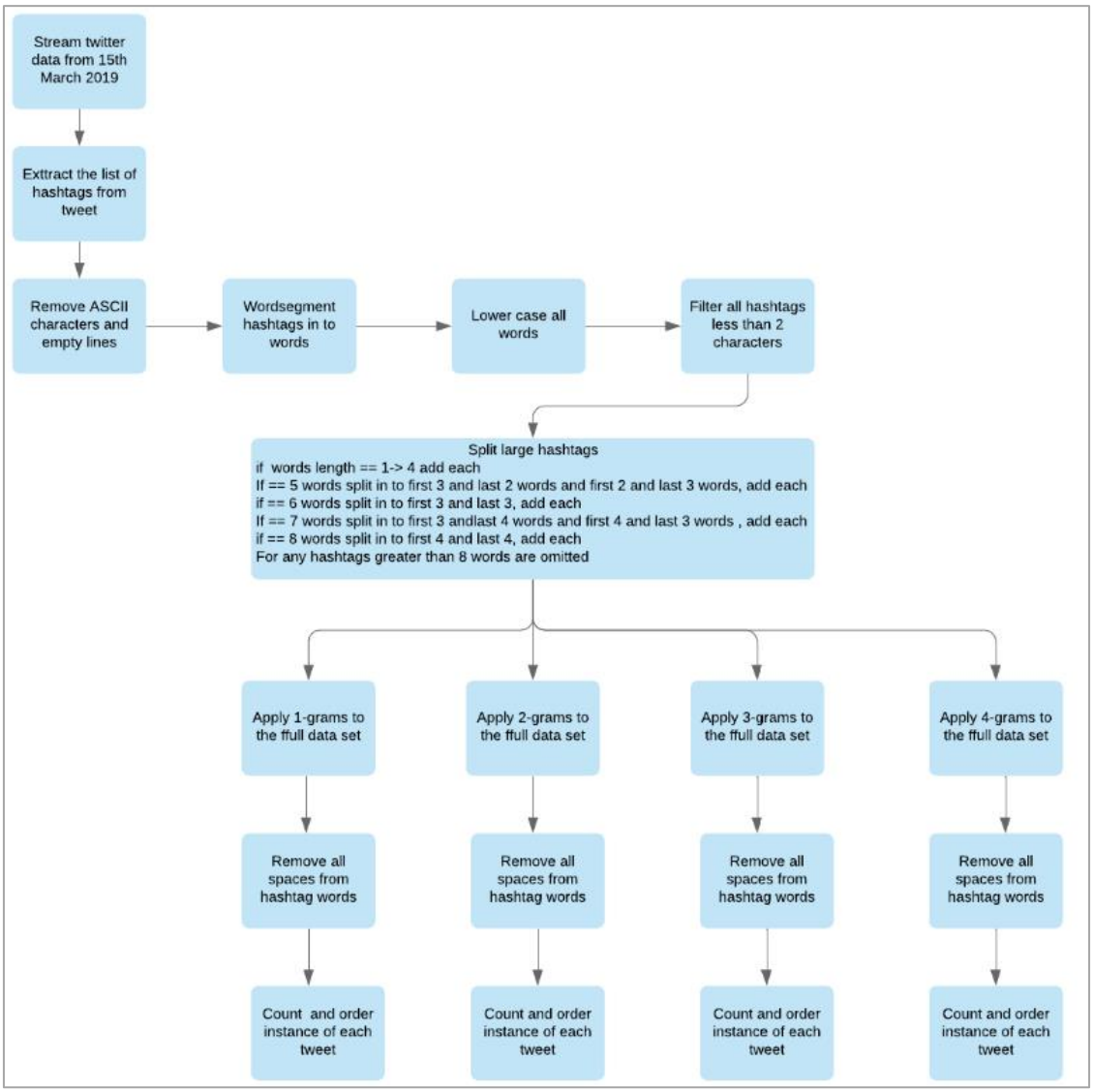


Figure 3.2 Twitter hashtag processing flow diagram

3.1.1 Create Twitter App

To access data through the twitter streaming APIs, a twitter developer account is set up on Twitter[5] with read and write access. An application is then created to generate the API credentials including API key; API secret; access token; and access secret token that will allow access to twitter from Python.

[5] https://apps.twitter.com/

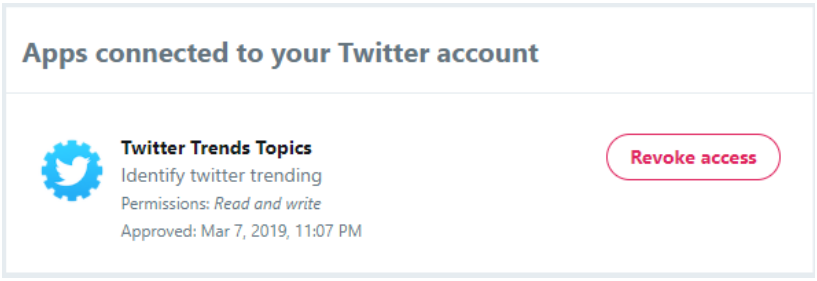


Figure 3.3 Twitter developer account

3.1.2 Accessing the Data

The selected language for processing twitter data is Python, an open source, cross platform programming language installed from Python[6]. A package manager tool, Pip[7], is installed to manage python packages and their installation. The IDE selected to retrieve and process the data is Visual Studio Code[8], a lightweight easy to use source code editor with rich support for the language Python. Twitter provides the streaming API that pushes messages to a persistent session, allowing the streaming API to download more data in real-time than could be completed using the REST API. Tweepy is an open source python library installed via Pip that allows python code to communicate with twitter using its Streaming API, providing access to twitter applications. In Tweepy, an instance of *'tweepy.Stream'*, establishes a streaming session and routes messages to a '*StreamListener'* instance. The *'StreamListener'* object monitors and catches the real-time tweets where its *'on_data'* method receives all messages and the *'on_status'* method receives status data from the *'on_data'* method returned in a JSON format that is stored locally (Doshi et al., 2017). The streaming API has three steps outlined below.

3.1.3 Tweepy OAuth Authentication

Authorising the app to access Twitter data requires the OAuth interface, where the Tweepy OAuthHandler method and the user configuration tokens are defined to provide access to Twitter. The authentication tokens include the customer_key; customer_secret;

[6] https://www.python.org/

[7] https://bootstrap.pypa.io/get-pip.py

[8] https://code.visualstudio.com/

access_token; and access_token_secret required to stream the Twitter data, as shown in figure 3.4 below.

```
# Create an authentication object
auth = tweepy.OAuthHandler(consumer_key,
consumer_secret)
# Set the user access token and consumer tokens
auth.set_access_token(access_token, access_token_secret)
# Create an API object passing the authentication information
api = tweepy.API(auth)
```

Figure 3.4 Python code to access twitter data

3.1.4 Create a StreamListener and a Stream

Tweepy's is a Python library providing access to the Twitter StramingAPI. Its '*StreamListener - on_data'* method, passes the data from 'statuses' to the *'on_status'* method. This method is inherited from *'StreamListener'* overriding its *'on_status'* method. The *'StreamListener'* stores the retrieved data in JSON file format. The tweets are stored in batches of five hundred tweets labelling each file based on date-time creation.

Once the API entry point to allow operations to be performed on twitter is available and the 'StreamListener' is available a stream object can be created as shown below in figure 3.5.

```
def main():
listen = SListener(api, 'myprefix')
stream = tweepy.Stream(auth, listen)
try:
stream.filter(track = '#', languages=['en'])
except:
stream.disconnect()
main()
if __name__ == '__main__':
main()
```

Figure 3.5 Python code to filter twitter data by # and language

3.1.5 Filter the Stream

Twitter provides limited options to filter real-time tweets. Option one, is the Enterprise PowerTrack[9] API with access to filter on the full Twitter data content. This is only available to Enterprise groups and therefore was not available for this project. The second option is a statuses/filter API, which returns public statuses that match one or more filter predicates. The filters applied were any tweet with a hashtag (#) that is in the English language.

```
stream.filter(track = '#', languages=['en'])
```

Figure 3.6 Python code to filter twitter data

3.1.6 Handling Errors

Error handling is an important part of twitter streaming with dangers of hitting rate limits or time-outs, where a restart of the process must be catered for.

3.1.7 Storing the Data

The data is stored in JSON format files. The full tweets are retrieved where they contain at least one hashtag (#) and are of locale English where they are stored in batches of five-hundred tweets, with file name labels based on date and time of file creation. When larger numbers of tweets were stored in files it was found the process slowed down, therefore files were created with five-hundred per file, which did not look to impact retrieval and storage.

3.1.8 Tweet Structure

The full tweet data is returned in a JSON format where the filter of a hashtag (#) exists in the tweet and where the locale is English. A full sample tweet is shown in Appendix A. The entity item hashtag list 'text' values are extracted from the tweet and stored in a

[9] https://developer.twitter.com/en/docs/tweets/filter-realtime/overview

.cvs of five-hundred tweet hashtags per file for further cleaning and processing. For example, the hashtag ‘Florida’ is extracted from the hashtag list as shown in figure 3.7


```
"entities":{
  "hashtags":[{
    "text":"Florida",
     "indices":[80,88]}],
     "urls":[{"url":"https:\/\/t.co\/Z98KvO6nhB",
     "expanded_url":"https:\/\/twitter.com\/i\/web\/status\/1112821872926777345",
      "display_url":"twitter.com\/i\/web\/status\/1\u2026",
      "indices":[117,140]}],
       "user_mentions":[],
        "symbols":[]},
```


Figure 3.7 Sample hashtag structure

3.1.9 Cleaning the Tweet

For each hashtag text extracted, all non-ASCII characters are removed, where only a-z characters remain. This includes removing foreign language characters, numerical data, punctuation etc. For example, hashtag like "text":”trump2020" is updated to “trump” removing the digits ‘2020’

3.1.10 Cleaning the Tweet

The tweet hashtags were split in to words for further processing. Two Python packages were examined to complete this process. The function *‘splitter.split’* was used to split the words of a hashtag initially but when the output was compared against the function *‘Wordsegment.segment’* it was found Wordsegment resulted in a better split of the words. The full twitter dataset was split based on Wordsegment and stored for further processing.

3.1.11 Removing Stop Words from the Tweet

The remaining tweet text is updated to lower case. Stop words are removed using '*ntlk.corpus*' of the English language. All tweets that are less than two characters are omitted from further processing with a maximum of five-hundred tweets stored per file.

3.1.12 Applying n-grams to the Tweet

Firstly, an n-gram pre-processing step was added to split large hashtags containing five or more words in to smaller groupings of words. For example, if a split hashtag contained five words it is split in to the first three words and last two words , then the first two words and the last three words where, as outlined in the next steps, n-grams are applied. In the case of an eight-word hashtag the words were split in to two groups of the first four and last four words.

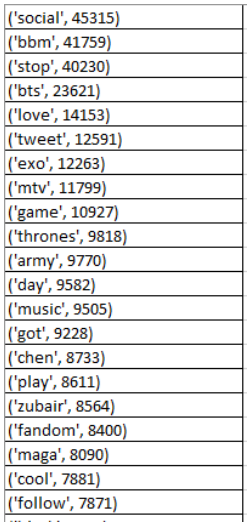
This process applied n-grams up to 4-grams to each of the extracted tweets as follows:

- The full hashtag has been split in to words where in the first sample 1-gram is applied to the full Twitter hashtag corpus. This involves taking any split hashtag with more than one word and splitting it in to individual words for processing.
- The process applies 2-grams to each of the applicable extracted tweets as follows. One-word hashtags are included, and two-word hashtags are included. For all hashtags greater than two, the hashtag is split and added for additional processing. This process required, in the case of a three-word hashtag, a twofold process. Firstly, that the first two words and the third word are extracted and added and secondly, that the first word and the last two words are extracted and added to the corpus for further processing. In the case of a four-word hashtag, the first two words and second two words were added.
- The process applies 3-grams to each of the applicable extracted tweets as follows. One-word up to three-word hashtags are included without change. For all hashtags greater than three, the hashtag is split and added for additional processing. This process required, in the case of a five-word hashtag, a twofold process. Firstly, that the first three words and the last two words are extracted and added and secondly, that the first two words and the last three words are

extracted and added to the corpus for further processing. In the case of a six-word hashtag, the first three words and the last three words were added.

3.1.13 Counting the Tweets

For all tweets collected a count is completed on each tweet occurring in the dataset and stored in a .cvs file for further processing.



('social', 45315)
('bbm', 41759)
('stop', 40230)
('bts', 23621)
('love', 14153)
('tweet', 12591)
('exo', 12263)
('mtv', 11799)
('game', 10927)
('thrones', 9818)
('army', 9770)
('day', 9582)
('music', 9505)
('got', 9228)
('chen', 8733)
('play', 8611)
('zubair', 8564)
('fandom', 8400)
('maga', 8090)
('cool', 7881)
('follow', 7871)

Figure 3.8 Cleaned counted and ordered hashtags 1-gram

3.2 Wikidata Mining and Understanding

In phase two, the English language Wikidata history files containing full revision history are downloaded and parsed for analysis. The figure 3.9 details the flow diagram of the overall process used to extract the revision data from Wikidata history revision files. Additional details are provided in the remaining sections of this chapter together with the additional processing required to prepare the data for analysis.

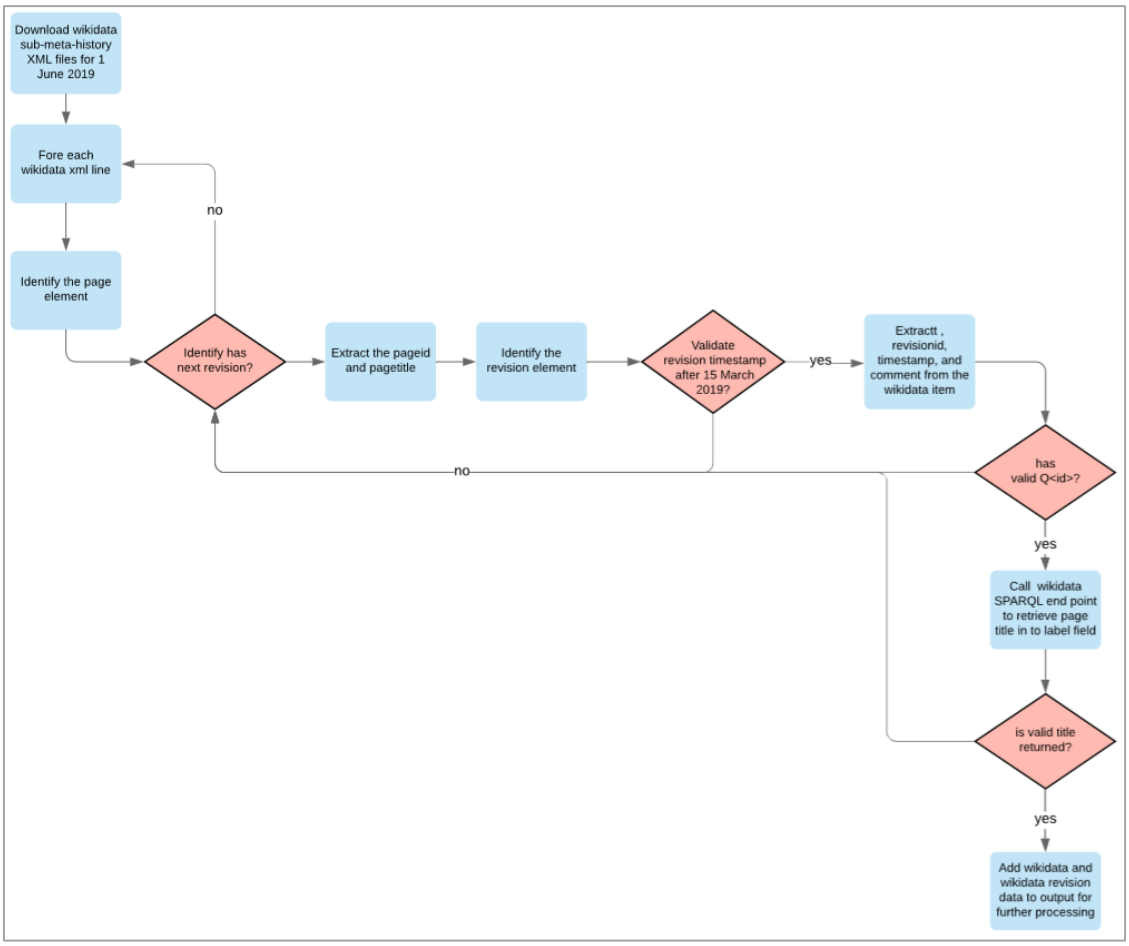


Figure 3.9 Wikidata revision data extraction process

3.2.1 Wikidata History Revision Files

In phase two the English language Wikidata history compressed files containing full revision history are downloaded and parsed for analysis with a name format 'Wikidata-{date}-stub-meta-history[num].xml'. These Wikidata dumps are released at regular intervals and available on the Wikidata site[10]. The selected revision files for this study contained the required revision information with minimal page data, for example wikidatawiki-20190601-stub-meta-history1.xml.gz. The twenty-seven Wikidata metadata history files from 1st of June 2019 were downloaded for revision analysis. These stub files contain the page and revision data without text content. These files contained the required revisions and were on average 1.8 GB each when compressed. When uncompressed these files were approximately 12 GB in size, except for the final

[10] https://dumps.wikimedia.org/wikidatawiki/

file wikidatawiki-20190601-stub-meta-history27.xml.gz, with a total size of 15.7 GB when compressed and approximately 78 GB when uncompressed. This final file, with a larger volume of data to the other twenty-six files, contains all the revisions since the previous release date of the wiki-media-history files. This is the intended design of revision output by Wikidata with this final file continuing to grow where other files should not (Wikimedia, 2018). Once the compressed files were decompressed the revision data per page in each xml file was extracted and this process is detailed in the next section.

3.2.2 Wikidata Download Process

The basic structure of a page revision is shown in figure 3.10 containing the page details and its related revisions outline.

```
<page>
  <title><Text></title>
  <id><Page Identifier</id>
  <revision>
      [First revision]
  </revision>
  <revision>
      [Second revision]
  </revision>
      [Additional revision information]
 </page>
```

Figure 3.10 Wikidata history file revision structure

The revision history metadata file consists of many page elements and revision elements of relevance in this study.

The page element <page> contains information about the Wikidata page with its sub elements revisions. This element is used to determine the start of the next page for its revisions to be considered. The sub elements of the page are as follows:

- The page title element <title> is the string representation of its identifier containing a number value. This is added to the output file as ‘pagetitle’.

- The element <id> represents the page identifier an is stored as ‘pageid’ in the output file.
- The <revision> list element contains each revision made to a page and many of its attributes are of relevance in this study to determine the total number of edits applied to a page.
    - The revision represents one revision item <revision> applied to a page.
    - This identifier relates to the revisions unique identifier and is stored as ‘revisionid’ in the output file.
    - The parent identifier is stored in the <parenteid> element linking the previous revision. This value is stored in the output as ‘parentid’.
    - The timestamp element is the date the revision occurred and is stored in the output file as ‘timestamp’.
    - The comment element contains the summary comment from the user when the revision was introduced and is stored as ‘comment’ in the output file.

Figure 3.11 shows a sample of revision data extracted from Wikidata history files where page elements ‘pageid’ and ‘pagetitle’ are extracted together with the revision element data. The revision element data includes its ‘datetime’ stamp if validated to be on or after 15nd March 2019 together with its ‘comment’, ‘parenteid’, and ‘revisionid’ all stored within .csv files for additional processing.

pageid	pagetitle	label	revisionid	timestamp	comment	parentid
20604	Q17758	Buttigliera d'Asti	38303	2019-03-17T00:44:01Z	b'/* wbsetreference-add:2| */ [[Property:P2046]]: 15.76 square kilometre, #quickstatements; [[:toollabs:quickstatements/#/batch/9360|batch #9360]] by [[User:Underlying lk]]'	885198340
20604	Q17758	Buttigliera d'Asti	38303	2019-03-16T12:32:09Z	b'/* wbcreateclaim-create:1| */ [[Property:P1082]]: 2,564, #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk]]'	804491948
20604	Q17758	Buttigliera d'Asti	38303	2019-03-16T12:32:11Z	b'/* wbsetqualifier-add:1| */ [[Property:P585]]: 1 January 2018, #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk]]'	884890338
20604	Q17758	Buttigliera d'Asti	38303	2019-03-16T12:32:13Z	b'/* wbsetreference-add:2| */ [[Property:P1082]]: 2,564, #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk]]'	884890367
20604	Q17758	Buttigliera d'Asti	38303	2019-03-16T12:32:15Z	b'/* wbsetqualifier-add:1| */ [[Property:P459]]: [[Q15911027]], #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk]]'	
20604	Q17758	Buttigliera d'Asti	38303	2019-03-17T00:44:03Z	b'/* wbsetqualifier-add:1| */ [[Property:P585]]: 9 October 2011, #quickstatements; [[:toollabs:quickstatements/#/batch/9360|batch #9360]] by [[User:Underlying lk]]'	884890423
20605	Q17759	Calamandrana	38303	2019-03-17T00:44:05Z	b'/* wbsetreference-add:2| */ [[Property:P2046]]: 19.16 square kilometre, #quickstatements; [[:toollabs:quickstatements/#/batch/9360|batch #9360]] by [[User:Underlying lk]]'	885198392
20605	Q17759	Calamandrana	38303	2019-03-16T12:32:18Z	b'/* wbcreateclaim-create:1| */ [[Property:P1082]]: 1,745, #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk]]'	804491936
20605	Q17759	Calamandrana	38303	2019-03-16T12:32:20Z	b'/* wbsetqualifier-add:1| */ [[Property:P585]]: 1 January 2018, #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk]]'	884890459
20605	Q17759	Calamandrana	38303	2019-03-16T12:32:22Z	b'/* wbsetreference-add:2| */ [[Property:P1082]]: 1,745, #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk]]'	884890485
20605	Q17759	Calamandrana	38303	2019-03-16T12:32:24Z	b'/* wbsetqualifier-add:1| */ [[Property:P459]]: [[Q15911027]], #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk]]'	884890512
20605	Q17759	Calamandrana	38303	2019-03-17T00:44:07Z	b'/* wbsetqualifier-add:1| */ [[Property:P585]]: 9 October 2011, #quickstatements; [[:toollabs:quickstatements/#/batch/9360|batch #9360]] by [[User:Underlying lk]]'	884890543
20606	Q17760	Genil	38303	2019-03-17T00:44:09Z	b'/* wbsetreference-add:2| */ [[Property:P2046]]: 12.79 square kilometre, #quickstatements; [[:toollabs:quickstatements/#/batch/9360|batch #9360]] by [[User:Underlying lk]]'	885198436
20609	Q17763	Calliano	38303	2019-03-16T12:32:26Z	b'/* wbcreateclaim-create:1| */ [[Property:P1082]]: 1,271, #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk]]'	804491922
20609	Q17763	Calliano	38303	2019-03-16T12:32:29Z	b'/* wbsetqualifier-add:1| */ [[Property:P585]]: 1 January 2018, #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk]]'	884890572
20609	Q17763	Calliano	38303	2019-03-16T12:32:31Z	b'/* wbsetreference-add:2| */ [[Property:P1082]]: 1,271, #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk]]'	884890598
20609	Q17763	Calliano	38303	2019-03-16T12:32:33Z	b'/* wbsetqualifier-add:1| */ [[Property:P459]]: [[Q15911027]], #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk]]'	884890629
20609	Q17763	Calliano	38303	2019-03-17T00:44:12Z	b'/* wbsetqualifier-add:1| */ [[Property:P585]]: 9 October 2011, #quickstatements; [[:toollabs:quickstatements/#/batch/9360|batch #9360]] by [[User:Underlying lk]]'	884890650

Figure 3.11 Wikidata revision with additional title information retrieved using SPARQL endpoint

The page title required for each revision is not available within the metadata revision history files and is required for processing in this work. However, each revision contains a ‘pageid’ in the format of Q<ID>, that is a unique identifier value item relating to its page article title. Using SPARQL, its value is read from the Wikidata SPARQL endpoint API and added to the field ‘label’ that is then added to the output file for later processing.

The edit titles are cleaned and the total number of edits per title is recorded during processing. The top edited article titles are identified using Python and its associated xml parsing libraries and stored in a related .cvs file for text comparison.

3.2.3 Wikidata Processing and Assumptions

Python has been used to parse the xml files to extract the Wikidata revision data in to individual records within a .csv file for additional processing. The attributes extracted per revision were ‘pageid’, ‘pagetitle’, ‘label’, ‘revisionid’, ‘timestamp’, ‘comment’ and ‘parented’ for each revision after the data 15th March 2019, from when twitter data was streamed. Extracting the Wikidata history revision files was a time-consuming process. The work to download these files was spread across four machines with ten instances running concurrently with each instance processing one XML uncompressed file per execution. The aim of this processing was to extract the revision history per page from the date twitter date began streaming 15th of March 2019.

For each item extracted a check was performed to validate the item has a Q<ID> relating to the page title. For all items that do not have a Q<ID> they are omitted from processing. Additionally, items with a Q<ID> recorded but do not have a valid title retrieved have also been omitted from the results as outlined in the assumptions below. The following assumptions have been made when processing this data:

3.2.3.1 Assumption 1 – Items without a page identifier are omitted

There are a number of references in the Wikidata history files that do have a Q<ID> defined but when retrieved via the SPARQL service from Wikidata, the page does not exist and returns an exception. For these values they are ignored and not included in the final result. It was confirmed these did not exist by running the SPARQL query from their provided service for a sample of those resulting in an exception. Within the code the exception is caught and passed over to continue processing the remainder of the document.

Example checking through the Wikidata SPARQL query service[11].

```
" SELECT DISTINCT * WHERE {wd:Q30 rdfs:label ?label . FILTER
(langMatches( lang(?label), "EN" ) ) } LIMIT 1
```

Figure 3.12 SPARQL query to retrieve page title

[11] https://query.wikidata.org/

3.2.3.2 Assumption 2 - User items and contacts omitted

Entries such as ‘user’ or ‘contact the developer’ pages as shown below in figure 3.13 have also been omitted from this study. These entries do not have a page ID that can be retrieved by SPARQL and therefore will be omitted from the final analysis result. Such entries would not have any relevance in the analysis as they relate to user main page updates and developer contacts.

179	User:Aschmidt	1787689	2019-03-1	b'/* wbse	8.49E+08
181	Wikidata:Contact the developr	88422	2019-03-1	b'/* {{P|P	8.84E+08
181	Wikidata:Contact the developr	5625	2019-03-1	b'Bot: Arcl	8.85E+08
181	Wikidata:Contact the developr	44949	2019-03-1	b'/* Unit h	8.86E+08
181	Wikidata:Contact the developr	2731518	2019-03-1	b'/* {{P|P	8.86E+08
181	Wikidata:Contact the developr	2814084	2019-03-1	b'/* Unit h	8.87E+08
181	Wikidata:Contact the developr	44949	2019-03-1	b'/* Unit h	8.87E+08
181	Wikidata:Contact the developr	44949	2019-03-1	b'/* Unit h	8.87E+08
181	Wikidata:Contact the developr	2814084	2019-03-1	b'/* Unit h	8.87E+08
181	Wikidata:Contact the developr	2814084	2019-03-1	b'/* Unit h	8.87E+08
181	Wikidata:Contact the developr	44949	2019-03-1	b'/* Unit h	8.87E+08
181	Wikidata:Contact the developr	887171808	2019-03-1	b'/* result	8.87E+08
182	MediaWiki:Common.css	3081030	2019-03-1	b'/* result	8.87E+08

Figure 3.13 Omitted revisions items

3.2.4 Retrieving the revision article title using SPARQL endpoint

SPARQL is a powerful API with which to access linked data collections that allow for retrieval of precise and insightful information in to the knowledge graph of Wikidata linked data. (Bielefeldt et al., 2018) The revision page title is retrieved and stored per revision item by querying the SPARQL endpoint as shown in figure 3.14

```
'SELECT DISTINCT * WHERE {wd:' + wiki_id + ' rdfs:label ?label .
FILTER (langMatches( lang(?label), "EN" ) ) } LIMIT 1'
```

Figure 3.14 SPARQL query structure to retrieve page title

The following example returned from the Wikidata revision xml files contained the Q<id> value of Q5561905, for the Technological University Dublin confirmed through the Wikidata SPARQL query service[12].

[12] https://query.wikidata.org/

```
SELECT DISTINCT * WHERE {wd:Q5561905 rdfs:label ?label .
FILTER (langMatches( lang(?label), "EN" ) ) } LIMIT 1
```

Figure 3.15 SPARQL query to retrieve the page title for Technological University Dublin

3.2.5 Additional Wikidata Processing

Once the Wikidata XML files were parsed, by extracting each revision that occurred on a page from 15th March 2019 to 1st June 2019, a number of cleaning steps were then required as shown in figure 3.16 below.

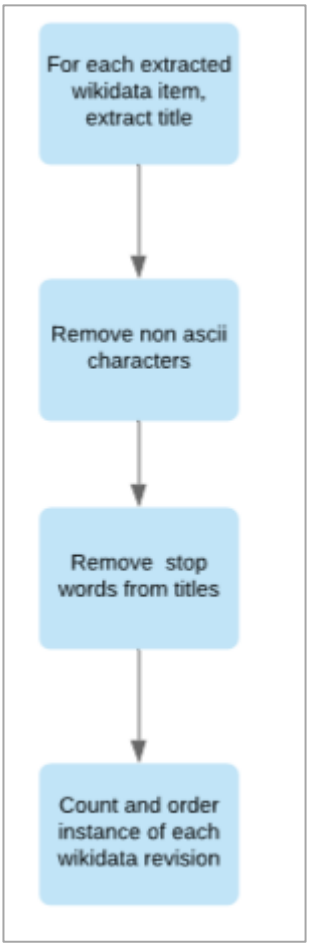


Figure 3.16 Wikidata additional processing flow diagram

The non-ASCII characters were extracted from the Wikidata page titles and stop words were removed. This used the same process, English language ‘ntlk’ stop word corpus,

that was applied to Twitter. To ensure the comparison with Twitter hashtag data was comparable, all spaces between words were removed from the Wikidata page titles. Finally, the Wikidata revisions per page were counted to make them available for statistical analysis.

3.2.6 Wikidata Processing Issues

Parsing of the Wikidata data dumps was completed using the language Python. This process was time consuming due to the large size of these files. This process could not be started until the cut-off date of Twitter collected data and required the data dumps to be made available on the same date. The date selected was 1st of June 2019. As a result, the time to process this data was short. During the parsing process two of the twenty-seven Wikidata dump XML files were fully parsed and eleven were partially parsed. This resulted in the collection of 1.8 GB of data revisions that occurred within the study period.

In addition, while processing the XML data there were many retrieval issues during parsing. This appeared to occur more in the later XML revision files. Errors occurred for a number of reasons, with the most frequently occurring causes being badly formatted XML. For example, incorrect values like commas in the element 'pageid' or an invalid 'pageid' caused a delay in trying to retrieve a page title via the SPARQL endpoint. Additionally, errors were encountered during parsing when the access limit was reached for SPARQL endpoint. When any of these errors occurred, the process continued to the next record.

3.3 Data Preparation for Statistical Analysis

The statistical analysis process included applying *Jaccard's Ratio* and *Kolmogorov-Smirnov* to a number of datasets, formed on a percentage total of the full datasets of Twitter hashtags in each n-grams and Wikidata page revisions. The language Python was used to implement the *Jaccard's Ratio* and *Kolmogorov-Smirnov* calculation functions, which were executed against these datasets. The percentage of data examined included 0.1%, 10%, 50% and 100% of these datasets. The results of the statistical analysis are detailed below in chapter 4 'Results, Evaluation and Discussion'.

The volume of revision data collected from Wikidata was 1.8 GB and resulted in out-of-memory exceptions when attempting to run the Kolmogorov-Smirnov against the full dataset. As a result, the lowest frequently occurring items were removed from the Wikidata dataset until a level was reached where this process could be successfully run. As outlined in figure 3.17, the total number of unique revisions, once ordered by the most frequent and counted in the full Wikidata dataset, is 1,867,281 unique pages. This number of pages was reduced to 270,135 unique pages, equating to 14.5% of the Wikidata unique revision pages, to allow for the Kolmogorov-Smirnov statistical formula to be run successfully. To determine this number, the lowest frequently occurring items with one-page revisions were firstly removed but the issue continued to occur. When Wikidata items containing three or less revisions were removed the Kolmogorov-Smirnov statistical formula could be run successfully. For all further references to 100% of Wikidata data this relates to the revised dataset containing 270,135 unique Wikidata pages.

Initially, the data was analysed using the statistical tool *Jaccard's Ratio* and *Kolmogorov-Smirnov* with 100% of the data but, when significant correlation was not found between Wikidata page revisions and Twitter hashtag frequencies, the lower percentage multiples of each data set were also examined. Figure 3.17 shows the breakdown of the number of both Wikidata items and Twitter hashtag for 100%, 50%, 10% and 0.1% of each dataset. Each counted item in the percentage groupings were counted based on frequency of occurrence. Therefore, each relate to unique references of both the Twitter hashtags and Wikidata pages.

wikidata	**Total**	**100%**	**50%**	**10%**	**0.1%**
	1867281	270135	135068	27014	270
Twitter	**Total**	**100%**	**50%**	**10%**	**0.01%**
1-ngram	N/A	52633	26317	5263	**53**
2-ngram	N/A	145133	72567	14513	**145**
3-ngram	N/A	132300	66150	13230	**132**
4-ngram	N/A	128791	64396	12879	**129**

Figure 3.17 Page numbers analysed for Wikidata revisions and Twitter hashtags

3.4 Jaccard's Ratio and Kolmogorov-Smirnov Statistical Measures processing

3.4.1 Kolmogorov-Smirnov

Kolmogorov-Smirnov is a measure of distribution similarity with a range of [0 – 2] where 2 indicates input distribution is equal (D'Alberto & Dasdan, 2011). This test is a statistical hypothesis test, determining if the two samples of Wikidata pages and Twitter hashtags follow the same distribution. To evaluate the samples with *Kolmogorov-Smirnov*, the null hypothesis H0 is defined where its output is unknown and used to validate if the two datasets come from the same distribution. Next, the data, in terms of probability, is examined to determine if the hypothesis is rejected. If the probability that the samples are from different distributions exceeds a confidence level the original null hypothesis H0 is rejected and so the two samples are from different distributions and thus accepting the alternative hypothesis H1. To evaluate this, a *statistic* value is calculated using both datasets.

The *Kolmogorov-Smirnov* p-value is the probability of the null hypothesis. Where the value is less than the significance level, the null hypothesis is rejected, and the alternative hypothesis is accepted. If the p-value is greater than the significance level of 5% (0.05) the null hypothesis is accepted. If the p-value is less than the significance level of 5% (0.05) the null hypothesis is rejected that both sets of data are from the same distribution.

3.4.2 Jaccard's Ratio

The statistical measure Jaccard's Similarity is a statistical hypothesis test used to evaluate the similarity between unordered sets containing a list of items. In this study the two sets of items are examined each containing string-lists of Wikidata page titles and Twitter hashtags. The *Jaccard's Ratio* (similarity) statistical measure was introduced in 1901 and is used determine set similarity between the two trend lists with a range of [0 - 1], where 0 represents no similarity and 1 indicates the same items exist in each list. (D'Alberto & Dasdan, 2011). The analysis for *Jaccard's Ratio* was completed for the full corpus of both datasets and run against the four datasets with n-gams applied.

Jaccard's similarity is the total of items shared (intersection) across both datasets, divided by the all the items in both datasets (union), to determine the similarity between the sample sets. The items in both lists are unique to the individual list. As a frequency count of both the Twitter hashtags and Wikidata revisions were completed as part of the data processing, all words in each dataset used to calculate Jaccard's similarity are unique.

An additional statistical measure Jaccard's distance is also used within the study to measure dissimilarity between sets. This value is calculated as 1 minus Jaccard's coefficient.

3.5 Visualisation Statistics

The data evaluation process takes an in-depth look at the results by examining visualisations of key areas in the data. Visualisations were implemented using the language R and Python 'mapplot'. The IDE RStudio with the R language was used to create word cloud charts for the most frequently used Twitter hashtags and Wikidata pages, based on revision frequencies for the studied period. The Python *'mapplot'* package was used to create bar charts, giving insight in to the frequency of top trending Twitter hashtags and Wikidata page revisions, as well as to create clusters showing statistical analysis output.

4 RESULTS AND EVALUATION

This chapter examines and discusses the results found from the statistical tools *Jaccard's Ratio* and *Kolmogorov-Smirnov,* which use quantitative techniques to identify if a significant correlation exists between the top Wikidata revisions and Twitter hashtag trends. Visualisation techniques will provide additional insight in to the data results and support identifying whether a correlation is found between both lists of data.

4.1 List Characteristics

When determining how to measure correlation between two lists of strings, the list characteristics must be considered. The Twitter hashtag words and Wikidata page lists both have the following characteristics:

- The lists contain string characters only. A cleaning process was completed on both Twitter data hashtags and Wikidata page titles. Cleaning the hashtags extracted from Twitter required removal of all non-ASCII characters; splitting the hashtags in to words; removal of stop words; applying n-grams up to 4-grams; and finally, removing the spaces between words resulting in the final hashtags that are ready for analysis. The hashtags were counted based on frequency and ordered from highest frequency to lowest frequency, at which point both lists are ready for the statistical analysis. This process is detailed in section 3.1.

  The Wikidata revisions details was extracted from its available data dumps, and its title retrieved via the SPARQL endpoint. The title was cleaned by removing spaces followed by a count on the number of edited titles and ordered to show the most frequently edited article. More in-depth details can be found in section 3.2.

- The trend lists are non-conjoined lists where one list does not cover all elements in the second list.

- The lists are top weighted where the top of the list is more important than the tail, ranked by the items occurring most frequently. For Twitter hashtags this

relates to the number of times the hashtag occurred in tweets and for Wikidata frequency relates to the number of revisions applied to a page.

- The top percentage of items from each list are then evaluated, therefore the evaluation will not consider indefinite ranking.

4.2 Visualisation of the Data

Data was collected and analysed by streaming data from the Twitter Streaming API from 15th March 2019 to 1st June 2019, and by downloading and parsing Wikidata data history dumps. This section examines views of the data through visualisation charts. Firstly, a bar graph outlined in figure 4.1 below, shows the total number of unique words and combined words tweets broken down by n-grams applied to hashtags once split. This gives an insight in to the volume of unique items processed per n-gram grouping without considering the frequency of each tweet item.

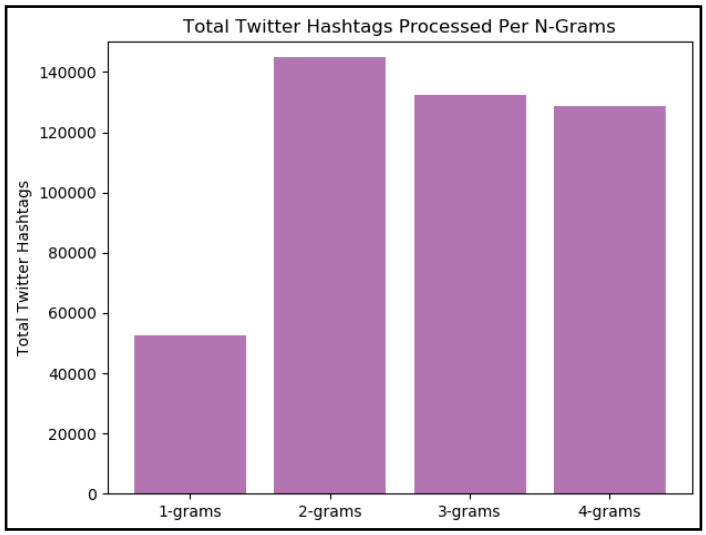


Figure 4.1 Total number of Twitter hashtags evaluated per n-gram

Figure 4.2 shows the total number of unique Wikidata articles collected based on the start date of Twitter data collection. This number of unique Wikidata revision pages processed is also shown, where 270,135 unique pages for the study together with their frequency were processed to allow for *Kolmogorov-Smirnov* statistical formula to be run successfully.

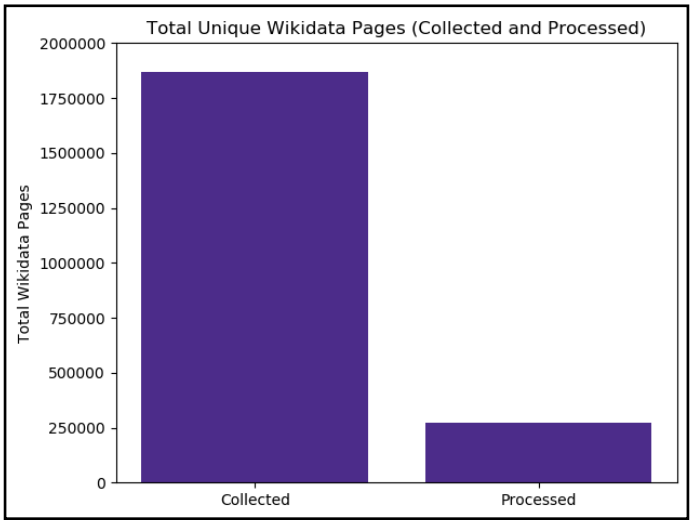


Figure 4.2 Total number of Twitter hashtags considered per n-gram

This equates to 14.5% of the total unique Wikidata pages collected without considering the frequency that were used in the study. For additional details see section 3.3 and, as stated there, all further references to 100% of Wikidata data will relate to the revised dataset containing 270,135 unique Wikidata pages.

4.2.1 Visualisation Word Cloud and Bar Graphs.

This section takes a look at some of the data through visualisation charts that show some insights in to the data collected from Twitter hashtags and Wikidata page revisions.

4.2.1.1 Wikidata visualisation

Firstly, examining the top Wikidata revision pages we can see some topical items appeared in the top twenty results. Item two ‘nursultan’ and item six ‘kleinerbriefkasten’ of the top twenty relate to renaming of the Kazakhstan capital city from Astana to Nursultan in honour of its outgoing leader a topical area at the end of March 2019. This gives a sense that the data is current and relevant to the time period the data was collected. What is surprising from the top twenty items, is the number of countries that appeared in the top twenty revised items in Wikidata where there have not been any major incidents occurring.

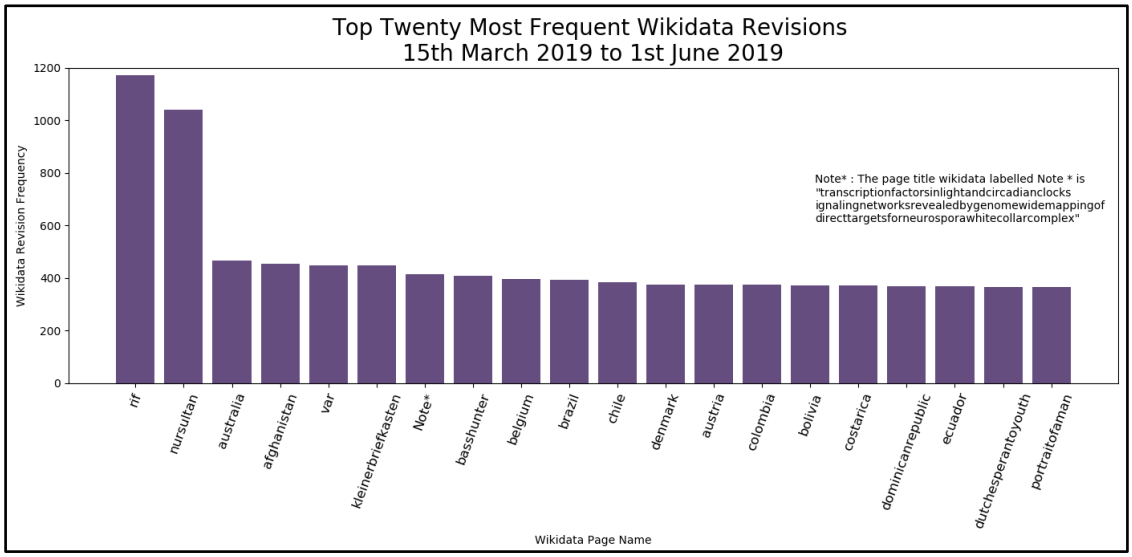


Figure 4.3 Top Wikidata Revision Pages

A word cloud has been generated for the top Wikidata page revisions for the period studied. This provides the opportunity to show more words on a visual with significant words highlighted. As discused above in the top Wikidata revision bar chart the number of Wikidata revisions relating to countries updated is more evident when examining the word cloud containing the top three hundred most revised pages over the study period as shown in Figure 4.4. This could be considered in further studies by creating a Wikidata bag of words to omit such items. However, within this word cloud countries are also included where major events have occurred, for example, Paris and Notre Dame are both included in the word cloud that would relate to Wikidata page updates in line with its world-famous cathedral being devastated by fire during the period of study.

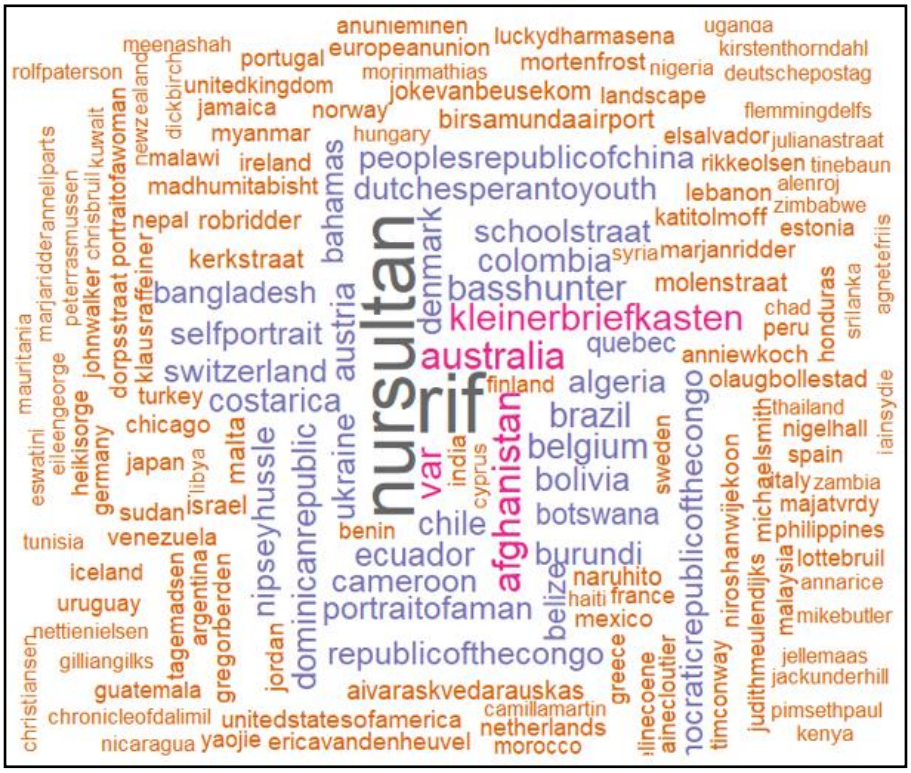


Figure 4.4 Wikidata word cloud top revisions

4.2.1.2 Twitter visualisation

The Twitter top twenty items are shown in bar graphs for each of the four n-gram analysed. The details of these are shown in figure 4.5 to figure 4.8 below. It can be seen from examining the top twenty hashtags across all 4-grams the ‘bbm stop social’ is the top hashtag. In the case of 1-grams as shown in figure 4.5 ‘bbm’, ‘social’ and ‘stop’ are the top trend words. This relates to the termination of the Blackberry messenger application for Android and iOS on May 31st, 2019 during the period of study. There is a number of trends in the top twenty items showing the data is relevant and topical to the period of study. This includes the popular television series ‘Game of Thrones’ appearing as the ninth ‘game’, the tenth ‘thrones’, and the fourteenth ‘got’ most popular item in 1-grams. This television show aired its season eight during the period of study.

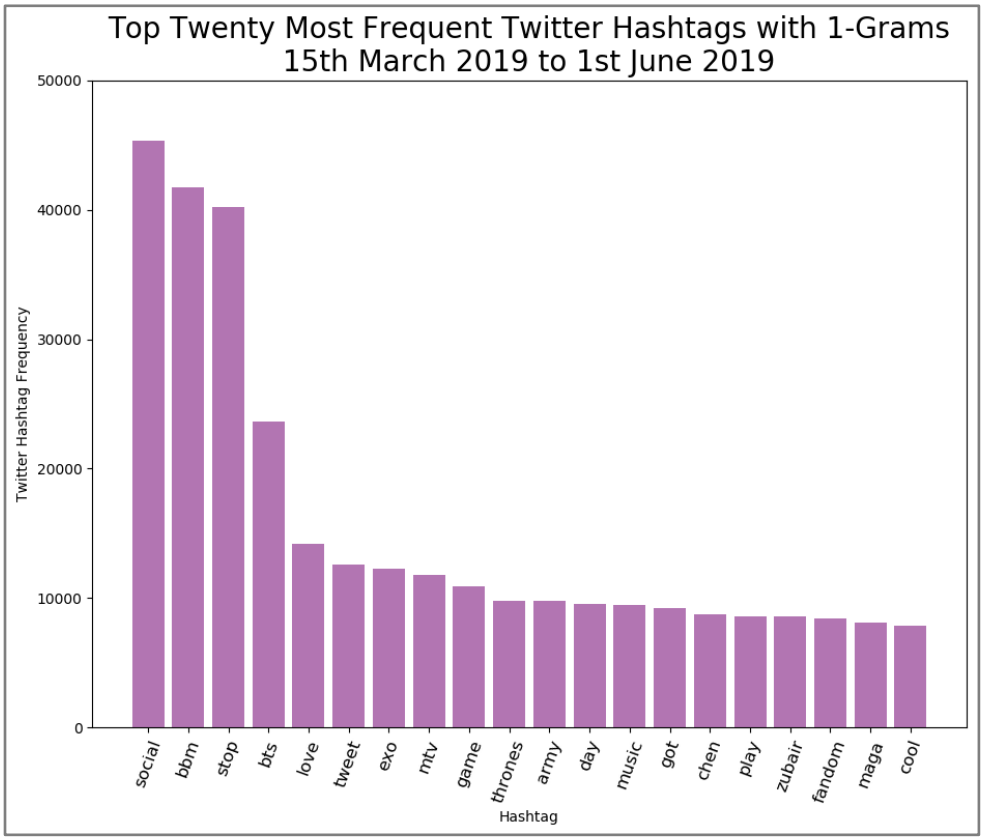


Figure 4.5 Twitter top twenty hashtags of 1-grams

When the top twenty hashtags for 2-grams was examined the results shown in n-gram one is reflected. The Blackberry messenger application termination hashtags ‘bbmstop’ and ‘stopsocial’ feature as the top two Twitter hashtag items with the television show ‘gamethrones’ ranked at number three together with the related hashtag ‘got’ at rank six. Like in 1-grams top twenty hashtag occurrences, there are a number of general language words also included like ‘cool’, ‘play’ and ‘fashion’ which could be omitted from the study by the introduction of a bespoke bag of words during cleaning.

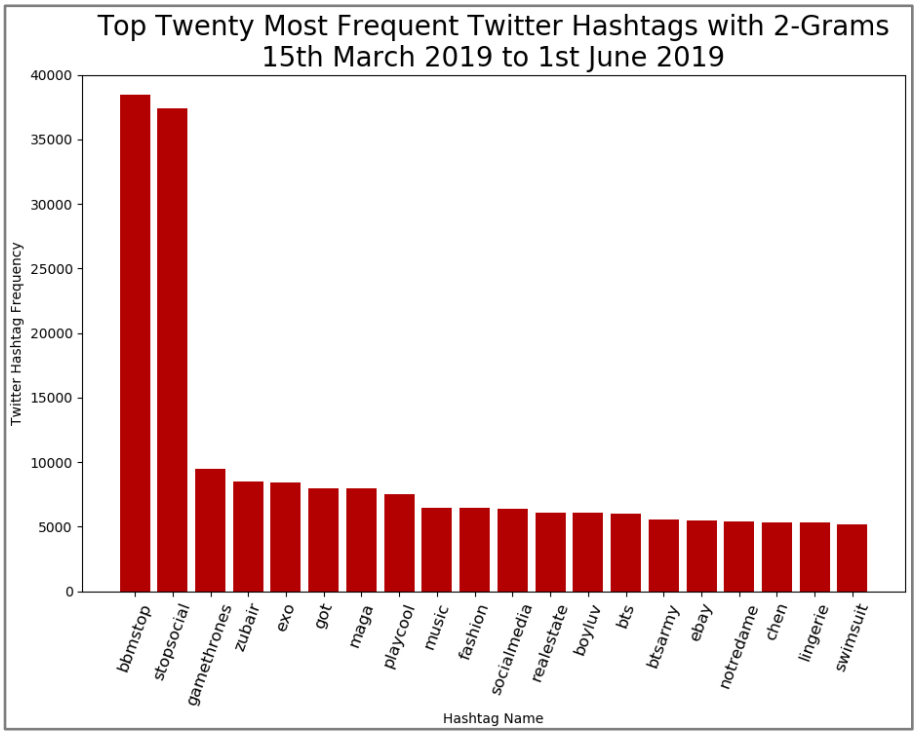


Figure 4.6 Twitter top twenty hashtag of 2-grams

As shown in figure 4.7 the top twenty, 3-gram results are again reflective of the previous n-gram results with 37,302 tweets relating to the termination of the Blackberry messenger app and the TV show 'Game of Thrones' related tweets ranked as the third and fifth most popular hashtags over the studied period.

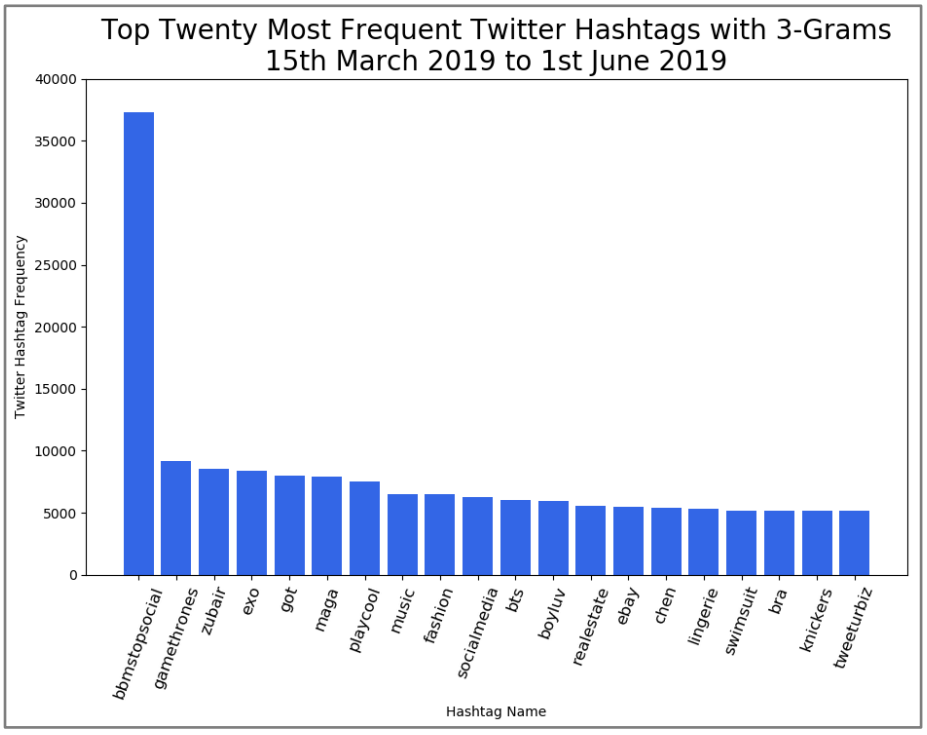


Figure 4.7 Twitter top twenty hashtags 3-grams

When examining 4-grams top ranked list, there is no difference in the top twenty output results where again termination of the Blackberry messenger app and the TV show ‘Game of Thrones’ related tweets ranked as the third and fifth most popular hashtags over the time period. This shows that the top trending hashtags were never greater than three words.

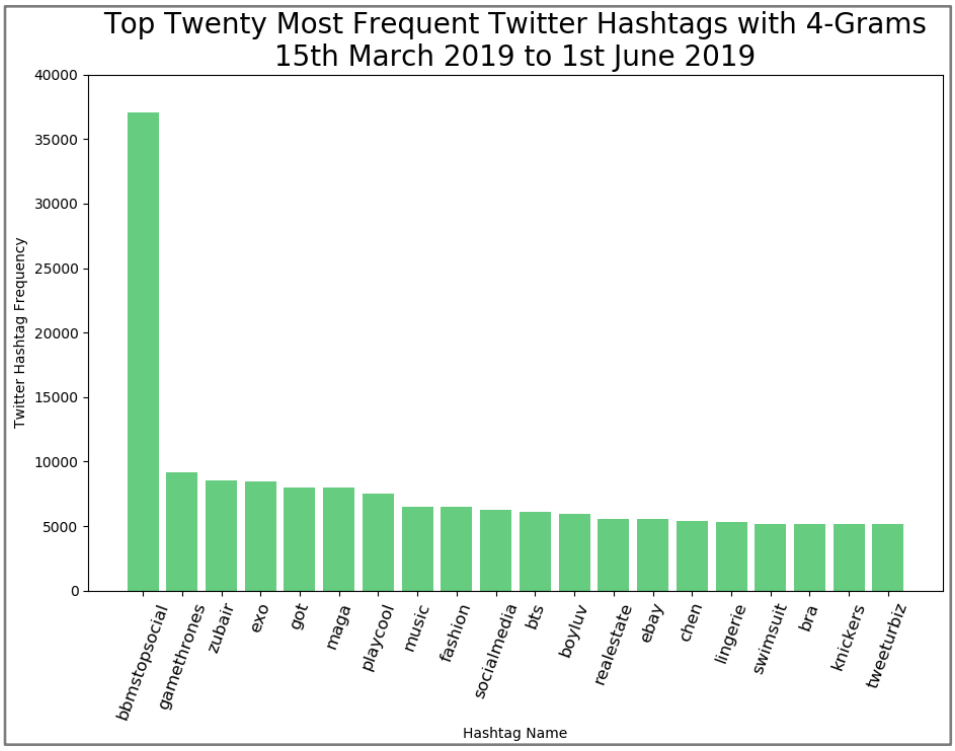


Figure 4.8 Twitter top twenty hashtags 4-ngrams

The n-grams visualizations show a consistency across all 4-grams where the termination of the Blackberry messenger application was the most tweeted hashtag across all n-grams. Also, consistently the television show ‘Game of Thrones’ is always high on the frequency list and is spread across a number of hashtag entries. This supports the possibility of introduction a bespoke bag of words to allow combining of related tweets like ‘gameofthrones’ occurring 9145 times and ‘got’ occurring 8016 times as shown in figure 4.5, in to one related hashtag item because they relate to the same topic. Similarly, a bespoke translator could convert ‘bbm’ to ‘Blackberry messenger’ for better comparison to Wikidata. A number of general words also included like ‘music’ and ‘fashion’ could be omitted from the study by the bespoke bag of words during cleaning for the twitter data.

When the Wikidata page items list was examined for ‘Game of Thrones’ related pages, three items were identified from the data extracted. These included seven revisions on the page ‘listofgameofthronescharacters’, seventy-five revisions on the page ‘gameofthrones’ and nine revisions on ‘agameofthrones’. Similarly, the data retrieved from Wikidata pages was examined for references to blackberry with twenty-five revisions on the page ‘blackberry’.

Word clouds were generated for the most frequently occurring words within each n-gram up to a maximum of three-hundred as detailed below in figure 4.9 to figure 4.12 for the studied period. These visualisations provides the opportunity to show more words on a visual with significant words highlihgted by size. As shown in figure 4.9 the size of the word on the word cloud visualisation represents the greater frequency of occurrence of each hashtag for the period studied. Examining the word cloud shows a number of improvements can be made to the visualisation results by having supplementary bag of words to ommit general day to day words like ‘find’ or ‘make’ that were not considered for removal during the stop word cleaning phase. What is very clear from examining the visualization is a need for a process step to remove slang word used on Twitter and rude words which are very common within the Twitter hashtag word clouds.

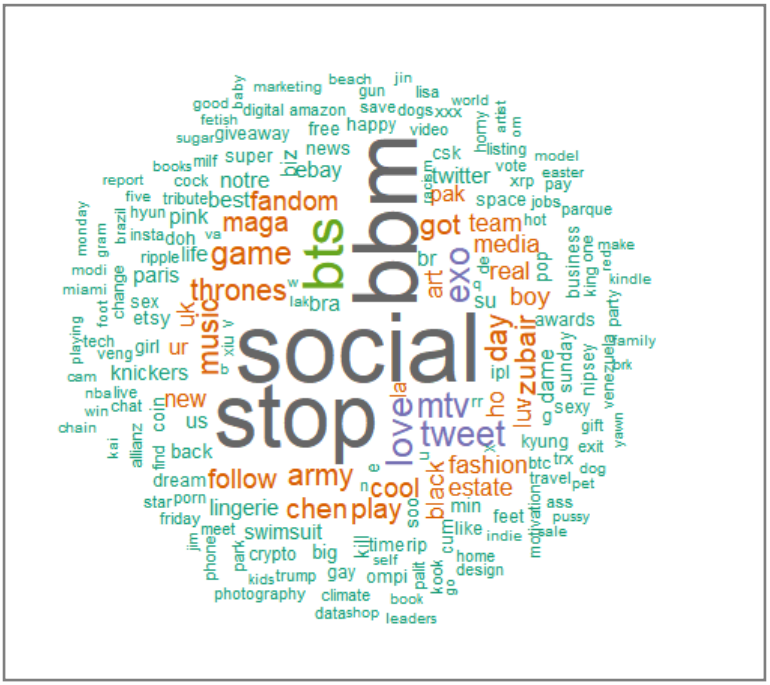


Figure 4.9 Word cloud for twitter hashtags of 1-grams

By examining the less frequent hashtag words within each of the word clouds it is clear there are many occurrences of topical issues and major events represented that have occurred during the study period and that cross over with Wikidata edits including ‘paris’ and ‘notredam’ which are both included in the word cloud that would relate to Wikidata page revision where the Paris’ world-famous cathedral Notre Dame was devastated by fire during the period of study. Additionally, high profile figures words like ‘trump’ relating to the president of the United States are included as well as climate change, a topical issue of the time. While these words appear lower down in the number of Wikidata revision ordered lists we can see some of these words are represented in both datasets studied.

Figure 4.10 Word cloud for Twitter hashtags 2-grams

Figure 4.11 Word cloud for Twitter hashtags 3-grams

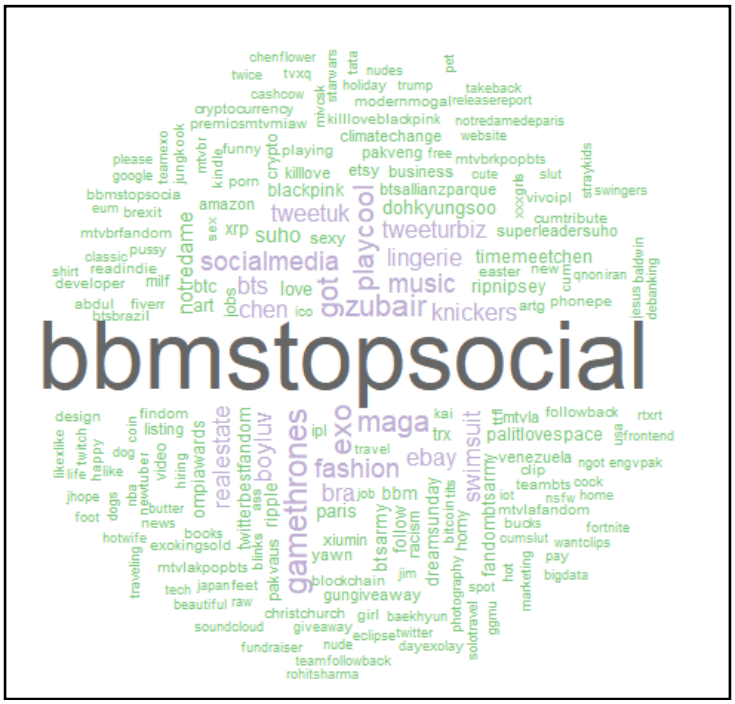


Figure 4.12 Word cloud for Twitter hashtags 4-grams

4.3 Jaccard's Ratio and Kolmogorov-Smirnov Statistical Measures Results and Evaluation

This analysis was completed by firstly separating the Twitter hashtags retrieved by its StreamingAPI and created n-gram up to 4-gram grouping of the split hashtag words. For further details on the retrieval and processing steps see section 3.1. Initially the data was analysed using the statistical tool Kolmogorov-Smirnov with 100% of the data made up of 1,867,281 unique pages, but the number of pages included in the calculations was reduced to 14.5% of the overall data with 270,135 unique pages because of performance issues in running the calculation across the full Wikidata page revisions as detailed in section 3.3. Within each n-gram groupings the data was grouped by the percentage of data to be analysed. For each n-gram the following coverage split was completed 0.1%, 10%, 50% and 100% of the Twitter per n-grams. The same split percentage was also applied to the Wikidata sets within each grouping. The data was evaluated using the statistical tools *Kolmogorov-Smirnov* and *Jaccard's Similarity*, to identify if a correlation exists between Wikidata page revisions and Twitter hashtags. The number of unique Twitter hashtags and Wikidata pages are detailed in Table 4.1 below.

wikidata	**Total**	**100%**	**50%**	**10%**	**0.1%**
	1867281	270135	135068	27014	270
Twitter	**Total**	**100%**	**50%**	**10%**	**0.01%**
1-ngram	N/A	52633	26317	5263	**53**
2-ngram	N/A	145133	72567	14513	**145**
3-ngram	N/A	132300	66150	13230	**132**
4-ngram	N/A	128791	64396	12879	**129**

Table 4.1 Page numbers analysed for Wikidata revisions and Twitter hashtags

Based on the list characteristics of the Twitter hashtags and Wikidata pages the *Jaccard's Ratio* and *Kolmogorov-Smirnov* statistical measures were used to evaluate the Wikidata revision and Trending Twitter hashtags to determine if a correlation strength existed between the two sets of variables. The finding has accepted the null hypothesis and rejected the alternative hypothesis indicating a statistically significant correlation was not found between Wikidata page revisions and Twitter hashtags for the studied period when applied across a number of percentages of the datasets including Wikidata items

and Twitter hashtag for 100%, 50%, 10% and 0.1% of each dataset. The following section discusses and evaluates the results.

4.3.1 Jaccard's Ratio statistical measure

The *Jaccard's Ratio* (similarity) statistical measure was used to determine set similarity between the two trend lists with a range of [0 - 1] where 0 represents no similarity and 1 indicates the same items exist in each list. (D'Alberto & Dasdan, 2011). Jaccard's Similarity is a statistical hypothesis test evaluating the similarity between unordered sets containing a list of items. In this study the two sets of items are examined each containing string-lists of Wikidata page titles and Twitter hashtags. The analysis for *Jaccard's Ratio* was completed for the full corpus of both datasets and run against the four datasets with n-gams applied. Additionally, analysis was completed for Jaccard's Ratio against 0.1%, 10%, 50% and 100% of both datasets. An additional statistical measure Jaccard's distance is also computed against both list of text-strings used within the study to measure dissimilarity between sets. This value is calculated as 1 minus Jaccard's coefficient. The results are shown below in Table 4.1.

Test & % of data	**1-grams (100%)**	**2-grams (100%)**	**3-grams (100%)**	**4-grams (100%)**
Jaccard's Similarity (100%)	0.04171830622256648	0.032609560564164794	0.03312188246891775	0.03350060752397064
Jaccard's Distance (100%)	0.9582816937774336	0.9673904394358352	0.9668781175310822	0.9664993924760293
Test & % of data	**1-grams (50%)**	**2-grams (50%)**	**3-grams (50%)**	**4-grams (50%)**
Jaccard's Similarity (top 50%)	**0.056381942920177175**	0.03804319638424777	0.03952566096423017	0.03989906625862186
Jaccard's Distance (top 50%)	0.9436180570798228	0.9619568036157522	0.9604743390357698	0.9601009337413782
Test & % of data	**1-grams (10%)**	**2-grams (10%)**	**3-grams (10%)**	**4-grams (10%)**
Jaccard's Similarity (top 10%)	0.03921884567045857	0.023691581282223585	0.02560272958444652	0.02622825564315872
Jaccard's Distance (top 10%)	0.9607811543295415	0.9763084187177764	0.9743972704155535	0.9737717443568413
Test & % of data	**1-grams (0.1%)**	**2-grams (0.1%)**	**3-grams (0.1%)**	**4-grams (0.1%)**
Jaccard's Similarity (0.1%)	0.0	0.0024271844660194173	0.002506265664160401	0.0
Jaccard's Distance (0.1%)	1.0	0.9975728155339806	0.9974937343358397	1.0

Table 4.2 Jaccard's Similarity and Jaccard's Distance statistical results

Interpreting Jaccard Similarity results will have values in the range of 0-1 where 0 represents no similarity and 1 represents an exact match. Firstly, looking at the results in Table 4.1 for 1-grams across 0.1%, 10%, 50% and 100%, we can see there is no similarity of words when similarity was calculated on 0.1% of the datasets with a result of 0. This 0.1% of the dataset equated to top 53 unique hashtags from Twitter and the

top 270 Wikidata pages ranked by most revisions. This value is also reflected in the Jaccard's distance where the calculated value is 1 indicating the greatest distance. By increasing the size of the datasets to 10% for 1-grams this equates to 145 Twitter hashtags and 27,014 Wikidata pages, we can see an increase in similarity to 0.3921 and a reduction in distance with a value of 0.96078. An increase in the similarity continues to occur up to 50% of the 1-grams data sample and reduces again as the dataset is analysed at 100% of the sample.

This is an interesting pattern that is reflected across each of the n-grams where the similarity is low on 0.1% of the data in all n-grams datasets analysed and increases in similarity when 50% of the data is analysed, but after 50% the similarity decreases again when 100% of the data was analysed but that 100% distance value is always greater that the recorded 10% n-gram value. Similarly, the pattern established for Jaccard's Distance as outlined for 1-grams above is consistent across all n-grams with a decrease in distance up to 50% of the sample and an increase again when 100% of the data is analysed for each of the n-grams.

The lowest possible similarity was calculated for 1-grams and 4-grams with a value of 0 showing no similarity. The highest similarity was recorded for 1-grams when 50% of the data was examined. This equates to 26,317 unique top Twitter hashtags and 135,068 ordered unique Wikidata pages. A value of 0.05638 was recorded for similarity and a value of 0.9436 recorded for distance with this value being the only one that reached above the 0.05 threshold. The next closest similarity value measured for similarity was also identified within the 1-grams analysis a value of .04171 was calculated when 100% of the data was analysed. For remaining distance values calculated they were all less than 0.04

4.3.2 Kolmogorov-Smirnov statistical measure

Kolmogorov-Smirnov is a measure of distribution similarity with a range of [0 – 2] where 2 indicates input distribution are equal (D'Alberto & Dasdan, 2011). This test *Kolmogorov-Smirnov* is a statistical hypothesis test, determining if the two samples of Wikidata pages and Twitter hashtags come from the same distribution. To evaluate the samples with *Kolmogorov-Smirnov*, the null hypothesis H0 and the alternative

hypothesis H1 are defined without knowledge of its result. The null hypothesis and alternative hypothesis were defined in this study as follows:

- Null hypothesis (H0): a correlation does not exist between Wikidata revisions and trending hashtags on Twitter determined by *'Jaccard Ratio'* and *'Kolmogorov-Smirnov'*.

- Alternative hypothesis (H1): a correlation exists between Wikidata revisions and trending hashtags on Twitter determined by *'Jaccard Ratio'* and *'Kolmogorov-Smirnov'*.

Next, the data, in terms of probability, is examined to determine if the hypothesis is rejected. A number closer to 0 indicates a likelihood the two samples are coming from the same distribution. If the probability that the samples are from different distributions exceeds a confidence level the original null hypothesis H0 is rejected and so the two samples are from different distributions and thus accepting the alternative hypothesis H1. To evaluate this, a *statistic* value is calculated using both datasets.

The *Kolmogorov-Smirnov* p-value was also calculated as part of this study used to determine the probability of the null hypothesis. If the p-value is greater than the significance level of 5% (0.05) the null hypothesis is accepted. If the p-value is less than the significance level of 5% (0.05) the null hypothesis is rejected. A low p-values means that the two samples are significantly different. The results for the Kolmogorov-Smirnov statistic and p-value are shown below in table 4.2.

Test & % of data	**1-grams (100%)**	**2-grams (100%)**	**3-grams (100%)**	**4-grams (100%)**
Kolmogorov-Smirnov p-value (100%)	5.726436890827359e-181	0.0	2.4486e-320	3.579683e-318
Kolmogorov-Smirnov statistic (100%)	0.06869303067890309	0.06606143443961077	0.06440017803089293	0.06476615112259088
Test & % of data	**1-grams (50%)**	**2-grams (50%)**	**3-grams (50%)**	**4-grams (50%)**
Kolmogorov-Smirnov p-value (top 50%)	1.1769474555258024e-102	8.234367674015068e-172	4.6452529163793994e-154	3.0203599339841 34e-103
Kolmogorov-Smirnov statistic (top 50%)	0.05574604004972983	0.05310761900520611	0.05144423653098529	0.052103042190194904
Test & % of data	**1-grams (10%)**	**2-grams (10%)**	**3-grams (10%)**	**4-grams (10%)**
Kolmogorov-Smirnov p-value (top 10%)	1.305210408847932e-25	3.9472506178244786e-83	7.845116134062559e-78	2.8082545624747394e-78
Kolmogorov-Smirnov statistic (top 10%)	0.08131552634938832	0.06466022948979733	0.06302907334652164	0.06335630137067927
Test & % of data	**1-grams (0.1%)**	**2-grams (0.1%)**	**3-grams (0.1%)**	**4-grams (0.1%)**
Kolmogorov-Smirnov p-value (top 0 .1%)	0.4183080902726968	0.4268711788289691	0.15201927607963006	0.4183080902726968
Kolmogorov-Smirnov statistic (top 0.1%)	0.12939662567915355	0.08826414704667493	0.11853344306024576	0.12939662567915355

Table 4.3 Kolmogorov-Smirnov static and p-value results

When the *statistic* value and *p-value* from the *Kolmogorov-Smirnov* test are examined together where a small *statistic* value together with a high *p-value* then the hypothesis that the distributions of the two samples are the same cannot be rejected. From the results we can see a high p-value across the majority of tested samples where its value is always greater than the 5% threshold of 0.05 as a result this supports the acceptance of the null hypothesis that there is not a statistically significant correlation between Wikidata page revision frequencies and Twitter hashtags for the period and data evaluated. There is one exception to this when datasets of 2-ngrams when tested with 100% of the data resulted in a p-value of 0 that is slightly higher than the 0.06606 score calculated for the dataset. The *Kolmogorov-Smirnov* statistic p-values contained very high levels across all datasets examined. An additional test was completed against a sample of the data by reducing the dataset lists to be of the same length where the *Kolmogorov-Smirnov* was calculated but it was found reducing the lists to be the same size did not impact the p-value result significantly.

While the outcome of this study rejects the alternative hypothesis that a correlation exists between the data sets examined, improvements identified during this study may have a positive impact on the result. These main suggested improvements include:

- Increased processing power to allow statistical analysis calculations to be run over large datasets. In this study the Wikidata sample was reduced to 14% of the collected sample to run the calculation *Kolmogorov-Smirnov* without memory errors.
- Introduction of a bespoke bag of words may also improve the results by removing slang words, noisy data words and identifying similar meaning words so that they are combined.

4.4 Hypothesis outcome

Having analysed Wikidata page titles of the most revised items against Twitter trending hashtags using the statistical tools *Kolmogorov-Smirnov and Jaccard's Ratio,* the null hypothesis (H0) is accepted, and the alternative hypothesis (H1) has been rejected. This result is based on having identified a high Jaccard's distance value, and a low Jaccard's

similarity value between both lists across all data tests completed in the data. Additionally, when the data was examined with the *Kolmogorov-Smirnov* a high p-value was found together with a low statistic value across supporting acceptance of the null hypothesis.

5 CONLUSTION AND FUTURE WORK

5.1 Conclusion

This study has examined Wikidata revisions page titles and streamed Twitter trending hashtags over a seventy-seven-day period to identify if a correlation exists between both sets of data. The results from this study have accepted the null hypothesis that a correlation does not exist between Wikidata revisions and trending hashtags on Twitter validated by the results from the statistical measures *'Jaccard Ratio'* and *'Kolmogorov-Smirnov'*. This work has included the mining of live streamed data for a seventy-seven-day period and parsing of Wikidata history revision XML files.

5.2 Future Work

There are many interesting areas where this work could either be extended or improved upon, that were not examined in this study because of limited access to data and time constraints. These are discussed below.

Improvements Through Data Availability

The volume of tweets studied relied on the available downloaded tweets through its publicly available Twitter StreamingAPI. However, if access was available to the enterprise Power Track API that is currently only available for paying customers this would allow access to a larger volume of steamed tweets to be used in the research.

With access to historical tweets in large volumes this could also provide additional insights in to the study but was but was outside the scope of this research due to high quoted costs of acquiring this data from Twitter as detailed in section 1.6.

Improvements Through Extending the Period Analysed

While the initial aim of this study was to download streamed data over a three-month period, the final study examined the tweet downloads over a seventy-seven-day period. Extending the corpus of tweets to the intended three-month period may increase the

accuracy of this study; allow for improvement and alternative analysis with Wikidata; or analysis of other sources of available data, for example Wikipedia.

Extending the Techniques of Data Analysis

This work could be extended to include ‘likes’ and ‘retweets’ per Twitter item. The impact of a trending hashtag can increase when a tweet is liked or retweeted by high profile individuals and could better identify correlations between trending hashtags and Wikipedia revisions.

Creation of a bespoke bag of words to handle individual tweet parts containing slang words or abbreviations for example may also be added to the study to improve results accuracy.

Improvements on the Horizon (due to technology)

An interesting area to consider for future work, is in the area of the semantic web. Technologies like Word Net an and Context that would provide additional insights in to the data.

6 BIBLIOGRAPHY

Ahuja, S., & Dubey, G. (2017). Clustering and sentiment analysis on Twitter data. *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*, 1–5. https://doi.org/10.1109/TEL-NET.2017.8343568

Al Tamime, R., Giordano, R., & Hall, W. (2018). Observing Burstiness in Wikipedia Articles during New Disease Outbreaks. *Proceedings of the 10th ACM Conference on Web Science - WebSci '18*, 117–126. https://doi.org/10.1145/3201064.3201080

Alsaadi, H. I., Almajmaie, L. K., & Mahmood, W. A. (2017). Forecasting of Twitter hashtag temporal dynamics using locally weighted projection regression. *2017 International Conference on Engineering and Technology (ICET)*, 1–4. https://doi.org/10.1109/ICEngTechnol.2017.8308166

Arulselvi, A. C., Sendhilkumar, S., & Mahalakshmi, S. (2017). Classification of tweets for sentiment and trend analysis. *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, 566–573. https://doi.org/10.1109/ICCONS.2017.8250525

Bielefeldt, A., Gonsior, J., & Krötzsch, M. (2018). *Practical Linked Data Access via SPARQL: The Case of Wikidata*. 10.

D'Alberto, P., & Dasdan, A. (2011). On the Weakenesses of Correlation Measures used for Search Engines' Results (Unsupervised Comparison of Search Engine Rankings). *ArXiv:1107.2691 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1107.2691

Dalinina, R. (2017). Introduction to Correlation. Retrieved June 7, 2019, from Introduction to Correlation website: https://www.datascience.com/learn-data-science/fundamentals/introduction-to-correlation-python-data-science

Doshi, Z., Nadkarni, S., Ajmera, K., & Shah, N. (2017). TweerAnalyzer: Twitter Trend Detection and Visualization. *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 1–6. https://doi.org/10.1109/ICCUBEA.2017.8463951

Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., & Vrandečić, D. (2014). Introducing Wikidata to the Linked Data Web. In P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, … C. Goble (Eds.), *The Semantic Web – ISWC 2014* (Vol. 8796, pp. 50–65). https://doi.org/10.1007/978-3-319-11964-9_4

Goldfarb, D., & Merkl, D. (2018). Visualizing Art Historical Developments Using the Getty ULAN, Wikipedia and Wikidata. *2018 22nd International Conference Information Visualisation (IV)*, 459–466. https://doi.org/10.1109/iV.2018.00086

Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Keim, D. A., Haug, L., & Hsu, M. (2011). Visual sentiment analysis on twitter data streams. *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 277–278. https://doi.org/10.1109/VAST.2011.6102472

Haripriya, A., & Kumari, S. (2017). Real time analysis of top trending event on Twitter: Lexicon based approach. *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–4. https://doi.org/10.1109/ICCCNT.2017.8204123

Heindorf, S., Potthast, M., Engels, G., & Stein, B. (2017). Overview of the Wikidata Vandalism Detection Task at WSDM Cup 2017. *ArXiv:1712.05956 [Cs]*. Retrieved from http://arxiv.org/abs/1712.05956

Heindorf, S., Potthast, M., Stein, B., & Engels, G. (2016). Vandalism Detection in Wikidata. *Proceedings of the 25th ACM International on Conference on*

Information and Knowledge Management - CIKM '16, 327–336. https://doi.org/10.1145/2983323.2983740

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *ARTIFICIAL INTELLIGENCE*, *349*(6245), 7. https://doi.org/10.1126/science.aaa8685

Jhandir, M. Z., Tenvir, A., On, B.-W., Lee, I., & Choi, G. S. (2017). Controversy detection in Wikipedia using semantic dissimilarity. *Information Sciences*, *418–419*, 581–600. https://doi.org/10.1016/j.ins.2017.08.037

Kaffee, L.-A., & Simperl, E. (2018). Analysis of Editors' Languages in Wikidata. *Proceedings of the 14th International Symposium on Open Collaboration - OpenSym '18*, 1–5. https://doi.org/10.1145/3233391.3233965

Li, Q., Zhou, B., & Liu, Q. (2016). Can twitter posts predict stock behavior?: A study of stock market with twitter social emotion. *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, 359–364. https://doi.org/10.1109/ICCCBDA.2016.7529584

Medelyan, O., Milne, D., Legg, C., & Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, *67*(9), 716–754. https://doi.org/10.1016/j.ijhcs.2009.05.004

Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013). *Using of Jaccard Coefficient for Keywords Similarity*.

Runeson, P., Alexandersson, M., & Nyholm, O. (2007). Detection of Duplicate Defect Reports Using Natural Language Processing. *29th International Conference on Software Engineering (ICSE'07)*, 499–510. https://doi.org/10.1109/ICSE.2007.32

Ruttenberg, J. (2019). *ARL White Paper on Wikidata: Opportunities and Recommendations*. 60.

Small, T. A. (2011). WHAT THE HASHTAG?: A content analysis of Canadian politics on Twitter. *Information, Communication & Society*, *14*(6), 872–895. https://doi.org/10.1080/1369118X.2011.554572

Smart, P. R., & Shadbolt, N. R. (2015). *Personalized Search: Epistemic Boon or Burden*.

Sundar, D. S., & Kankanala, M. (2015). Analyzing and predicting Lifetime of trends using social networks. *2015 International Conference on Computer Communication and Informatics (ICCCI)*, 1–7. https://doi.org/10.1109/ICCCI.2015.7218090

Tajalizadeh, H., & Boostani, R. (2019). A Novel Stream Clustering Framework for Spam Detection in Twitter. *IEEE Transactions on Computational Social Systems*, 1–10. https://doi.org/10.1109/TCSS.2019.2910818

Vrandecic, D. (2013). The Rise of Wikidata. *IEEE Intelligent Systems*, *28*(4), 90–95. https://doi.org/10.1109/MIS.2013.119

Wang, R., Liu, W., & Gao, S. (2016). Hashtags and information virality in networked social movement: Examining hashtag co-occurrence patterns. *Online Information Review*, *40*(7), 850–866. https://doi.org/10.1108/OIR-12-2015-0378

Weissman, S., Ayhan, S., Bradley, J., & Lin, J. (2015). Identifying Duplicate and Contradictory Information in Wikipedia. *Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries - JCDL '15*, 57–60. https://doi.org/10.1145/2756406.2756947

Xie, W., Zhu, F., Jiang, J., Lim, E., & Wang, K. (2013). TopicSketch: Real-Time Bursty Topic Detection from Twitter. *2013 IEEE 13th International Conference on Data Mining*, 837–846. https://doi.org/10.1109/ICDM.2013.86

Zangerle, E., Schmidhammer, G., & Specht, G. (2015). #Wikipedia on Twitter: Analyzing Tweets About Wikipedia. *Proceedings of the 11th International Symposium on Open Collaboration*, 14:1–14:8. https://doi.org/10.1145/2788993.2789845

APPENDIX A


```
{
"created_at":"Tue Apr 02 00:47:20 +0000 2019",
"id":1112879150061293568,
"id_str":"1112879150061293568",
"text":"RT @TruckersVote: . - - - GOP Corruption - - - \n # 536\nSenator Rick Scott of #Florida \nwas convicted of defrauding Med\u2026",
"source":"\u003ca href=\"http:\/\/twitter.com\/download\/android\" rel=\"nofollow\"\u003eTwitter for Android\u003c\/a\u003e",
"truncated":false,
"in_reply_to_status_id":null,
"in_reply_to_status_id_str":null,
"in_reply_to_user_id":null,
"in_reply_to_user_id_str":null,
"in_reply_to_screen_name":null,
"user":{
        "id":394984800,
        "id_str":"394984800",
        "name":"Bruce Balemian",
        "screen_name":"BruceBalemian",
        "location":"Warwick, RI",
        "url":"http:\/\/alternativeheatingsolution.com",
        "description":"I am the owner of Expert auto repair, and Alternative Heating solutions",
        "translator_type":"none",
        "protected":false,
        "verified":false,
        "followers_count":442,
        "friends_count":762,
        "listed_count":4,
        "favourites_count":140046,
        "statuses_count":7539,
        "created_at":"Thu Oct 20 23:09:55 +0000 2011",
        "utc_offset":null,
        "time_zone":null,
        "geo_enabled":true,
        "lang":"en",
        "contributors_enabled":false,
        "is_translator":false,
        "profile_background_color":"C0DEED",
        "profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png",
        "profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png",
        "profile_background_tile":false,
        "profile_link_color":"1DA1F2",
        "profile_sidebar_border_color":"C0DEED",
        "profile_sidebar_fill_color":"DDEEF6",
        "profile_text_color":"333333",
        "profile_use_background_image":true,
        "profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/378800000534919536\/2b32c905e10fff2a2f2a60f39f9e72df_normal.jpeg",
        "profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/378800000534919536\/2b32c905e10fff2a2f2a60f39f9e72df_normal.jpeg",
        "profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/394984800\/1430435771",
        "default_profile":true,
        "default_profile_image":false,
        "following":null,
        "follow_request_sent":null,
        "notifications":null},
"geo":null,
"coordinates":null,
"place":null,
"contributors":null,
"retweeted_status":{
        "created_at":"Mon Apr 01 20:59:44 +0000 2019",
        "id":1112821872926777345,
        "id_str":"1112821872926777345",
        "text":".- - - GOP Corruption - - - \n # 536\nSenator Rick Scott of #Florida \nwas convicted of defraudi\u2026 https:\/\/t.co\/Z98KvO6nhB",
        "display_text_range":[0,140],
```

```
"source":"\u003ca href=\"http:\/\/twitter.com\/download\/android\" rel=\"nofollow\"\u003eTwitter for Android\u003c\/a\u003e",
"truncated":true,
"in_reply_to_status_id":null,
"in_reply_to_status_id_str":null,
"in_reply_to_user_id":null,
"in_reply_to_user_id_str":null,
"in_reply_to_screen_name":null,
"user":{
        "id":573173793,
        "id_str":"573173793",
        "name":"The Trucker Vote",
        "screen_name":"TruckersVote",
        "location":"On the Road U.S.A.",
        "url":null,
        "description":"- - - Defending the American dream one tweet at a time. - - - Known to encourage perfect strangers to be reliable Democratic voters - - -",
        "translator_type":"none",
        "protected":false,
        "verified":false,
        "followers_count":34617,
        "friends_count":32921,
        "listed_count":115,
        "favourites_count":52584,
        "statuses_count":13983,
        "created_at":"Sun May 06 22:59:13 +0000 2012",
        "utc_offset":null,
        "time_zone":null,
        "geo_enabled":true,
        "lang":"en",
        "contributors_enabled":false,
        "is_translator":false,
        "profile_background_color":"C0DEED",
        "profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png",
        "profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png",
        "profile_background_tile":false,
        "profile_link_color":"1DA1F2",
        "profile_sidebar_border_color":"C0DEED",
        "profile_sidebar_fill_color":"DDEEF6",
        "profile_text_color":"333333",
        "profile_use_background_image":true,
        "profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/883569789183991808\/Osadv4Pe_normal.jpg",
        "profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/883569789183991808\/Osadv4Pe_normal.jpg",
        "profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/573173793\/1434410969",
        "default_profile":true,
        "default_profile_image":false,
        "following":null,"follow_request_sent":null,"notifications":null},
 "geo":null,
"coordinates":null,
"place":null,
"contributors":null,
"is_quote_status":false,
"extended_tweet":{
        "full_text":".    - - - GOP Corruption - - - \n          # 536\nSenator Rick Scott of #Florida \nwas convicted of defrauding Medicare for $1.8 billion and took the 5th 75 times. https:\/\/t.co\/f9Su9kDlAu",
         "display_text_range":[0,170],
         "entities":{
         "hashtags":[{
                  "text":"Florida",
                  "indices":[80,88]}],
                 "urls":[],
                "user_mentions":[],
                "symbols":[],
        "media":[{
                   "id":1112821849350533121,
                "id_str":"1112821849350533121"
                "indices":[171,194],
                "media_url":"http:\/\/pbs.twimg.com\/media\/D3GJk3TU8AEn11z.jpg",
"media_url_https":"https:\/\/pbs.twimg.com\/media\/D3GJk3TU8AEn11z.jpg",
```

```
            "url":"https:\/\/t.co\/f9Su9kDlAu",
            "display_url":"pic.twitter.com\/f9Su9kDlAu",
            "expanded_url":"https:\/\/twitter.com\/TruckersVote\/status\/1112821872926777345\/photo\/1",
            "type":"photo",
            "sizes":{   "thumb":{"w":150,"h":150,"resize":"crop"},
                        "small":{"w":680,"h":680,"resize":"fit"},
                        "medium":{"w":1200,"h":1200,"resize":"fit"},
                        "large":{"w":2048,"h":2048,"resize":"fit"}}}]},
"extended_entities":{
            "media":[{
                        "id":1112821849350533121,
                        "id_str":"1112821849350533121",
                        "indices":[171,194],
                        "media_url":"http:\/\/pbs.twimg.com\/media\/D3GJk3TU8AEn11z.jpg",
                        "media_url_https":"https:\/\/pbs.twimg.com\/media\/D3GJk3TU8AEn11z.jpg",
                        "url":"https:\/\/t.co\/f9Su9kDlAu",
                        "display_url":"pic.twitter.com\/f9Su9kDlAu",
                        "expanded_url":"https:\/\/twitter.com\/TruckersVote\/status\/1112821872926777345\/photo\/1",
                        "type":"photo",
                        "sizes":{
                                    "thumb":{"w":150,"h":150,"resize":"crop"},
                                    "small":{"w":680,"h":680,"resize":"fit"},
                                    "medium":{"w":1200,"h":1200,"resize":"fit"},
                                    "large":{"w":2048,"h":2048,"resize":"fit"}}}]},
"quote_count":1,
"reply_count":3,
"retweet_count":33,
"favorite_count":33,
“entities":{"hashtags":[{
            "text":"Florida",
            "indices":[80,88]}],
            "urls":[{"url":"https:\/\/t.co\/Z98KvO6nhB",
            "expanded_url":"https:\/\/twitter.com\/i\/web\/status\/1112821872926777345",
            "display_url":"twitter.com\/i\/web\/status\/1\u2026",
            "indices":[117,140]}],
            "user_mentions":[],
             "symbols":[]},
            "favorited":false,
            "retweeted":false,
            "possibly_sensitive":false,
            "filter_level":"low",
            "lang":"en"},
"is_quote_status":false,
"quote_count":0,
"reply_count":0,
"retweet_count":0,
"favorite_count":0,
"entities":{
            "hashtags":[{
                        "text":"Florida",
                        "indices":[98,106]}],
                         "urls":[],
                        "user_mentions":[{
                                    "screen_name":"TruckersVote",
                                     "name":"The Trucker Vote",
                                      "id":573173793,
                                     "id_str":"573173793",
                                     "indices":[3,16]}],
 "symbols":[]},
"favorited":false,
"retweeted":false,
"filter_level":"low",
"lang":"en",
"timestamp_ms":"1554166040326"
}
```