On Automating Basic Data Curation Tasks

Seyed-Mehdi-Reza Beheshti
University of New South Wales
Sydney, Australia
sbeheshti@cse.unsw.edu.au

Alireza Tabebordbar
University of New South Wales
Sydney, Australia
alirezat@cse.unsw.edu.au

Boualem Benatallah
University of New South Wales
Sydney, Australia
boualem@cse.unsw.edu.au

Reza Nouri
University of New South Wales
Sydney, Australia
snouri@cse.unsw.edu.au

ABSTRACT

Big data analytics is firmly recognized as a strategic priority for modern enterprises. At the heart of big data analytics lies the data curation process, consists of tasks that transform raw data (unstructured, semi-structured and structured data sources) into curated data, i.e. contextualized data and knowledge that is maintained and made available for use by end-users and applications. To achieve this, the data curation process may involve techniques and algorithms for extracting, classifying, linking, merging, enriching, sampling, and the summarization of data and knowledge. To facilitate the data curation process and enhance the productivity of researchers and developers, we identify and implement a set of basic data curation APIs and make them available as services to researchers and developers to assist them in transforming their raw data into curated data. The curation APIs enable developers to easily add features - such as extracting keyword, part of speech, and named entities such as Persons, Locations, Organizations, Companies, Products, Diseases, Drugs, etc.; providing synonyms and stems for extracted information items leveraging lexical knowledge bases for the English language such as WordNet; linking extracted entities to external knowledge bases such as Google Knowledge Graph and Wikidata; discovering similarity among the extracted information items, such as calculating similarity between string and numbers; classifying, sorting and categorizing data into various types, forms or any other distinct class; and indexing structured and unstructured data - into their data applications. These services can be accessed via a REST API, and the data is returned as a JSON file that can be integrated into data applications. The curation APIs are available as an open source project on GitHub.

Keywords

Data Curation, Big Data Analytics, Curation API

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3041021.3054726

1. INTRODUCTION

Understanding and analyzing big data is firmly recognized as a powerful and strategic priority [Chen et al. 2012, Beheshti et al. 2016b]. For deeper interpretation of and better intelligence with big data, it is important to transform raw data (unstructured, semi-structured and structured data sources, e.g., text, video, image data sets) into curated data: contextualized data and knowledge that is maintained and made available for use by end-users (e.g. data scientists and researchers) and applications (e.g. data and machine learning applications). In particular, data curation acts as the glue between raw data and analytics, providing an abstraction layer that relieves users from time consuming, tedious and error prone curation tasks. Data curation involves identifying relevant data sources, extracting data and knowledge, cleaning, maintaining, merging, enriching and linking data and knowledge. For example, consider a tweet in the Twitter [Kwak et al. 2010]: a microblogging service that enable users tweet about any topic within the 140-character limit and follow others to receive their tweets. It is possible to extract various information from a single tweet text such as keywords, part of speech, named entities, synonyms and stems [Gattani et al. 2013]. Then it is possible to link the extracted information to external knowledge graphs to enrich and annotate the raw data. Later, these information can be used to provide deeper interpretation of and better intelligence with the huge number of tweets in Twitter: every second, on average, around 6,000 tweets are tweeted on Twitter, which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year.

In particular, the data curation process enables extracting knowledge and deriving insights from the vastly growing amounts of local, external and open data. This task is vital for recent data analytics initiatives include: improving government analytical services, personalized advertisements in elections, and predicting intelligence activities [Tene and Polonetsky 2012, Beheshti et al. 2016a]. In this paper, we identify and implement a set of basic data curation APIs and make them available to researchers and developers *as services* to assist them in transforming their raw data into curated data. The curation services enable developers to easily add features - such as extracting keyword, part of speech, and named entities such as Persons, Locations, Organizations, Companies, Products, Diseases, Drugs, etc.; providing

synonyms and stems for extracted information items leveraging lexical knowledge bases for the English language such as WordNet[1]; linking extracted entities to external knowledge bases such as Google Knowledge Graph[2] and Wikidata[3]; discovering similarity among the extracted information items, such as calculating similarity between string, number, date and time data; classifying, sorting and categorizing data into various types, forms or any other distinct class; and indexing structured and unstructured data - into their data applications. These services can be accessed via a REST API, and the data is returned as a JSON file, an easy-to-parse structure, that can be integrated into (data and machine learning) applications. The basic data curation APIs are available as an open source project on GitHub[4]. The technical details for the curation APIs can be found in a technical report [Beheshti et al. 2016d]. The rest of the paper is organized as follows. In Section 2, we present an over view of the curation services, while in Section 3 we describe our demonstration scenario.

2. CURATION SERVICES OVERVIEW

To enhance the curation process, we propose a framework for data curation feature engineering: this refers to characterizing variables that grasp and encode information from raw or curated data, thereby enabling to derive meaningful inferences from data. An example of a feature is 'mentions of a person in data items like tweets, news articles and social media posts'. We propose that features will be implemented and available as uniformly accessible data curation Micro-Services: functions or pipelines implementing features. In particular, we will support a concrete set of features [Anderson et al. 2013], organized in categories such as: Extracting, Classifying, Linking, and Indexing algorithms. The curation services use natural language processing technology and machine learning algorithms to transform raw data into contextualized data and knowledge, e.g. by extracting semantic meta-data from content, such as information on people, places, and companies and link them to knowledge graphs such as WikiData and Google KG - using similarity techniques - or classify the extracted entities using classification services. We provide curation API endpoints for performing content analysis on Internet-accessible web pages, posted HTML or text content. We have provided the technical details for the basic data curation APIs in a technical report [Beheshti et al. 2016d]. In the following we present an overview of the curation APIs.

2.1 Extraction Services

The majority of the digital information produced globally presented in the form of web pages, text documents, news articles, emails, and presentations expressed in natural language text. Collectively, such data is termed unstructured as opposed to structured data that is normalized and stored in a database. The domain of Information Extraction (IE) is concerned with identifying information in unstructured documents. In most cases, this activity concerns processing human language texts by means of Natural Language Processing (NLP) [Beheshti et al. 2016c]. Accordingly, analysts

[1]https://wordnet.princeton.edu
[2]https://developers.google.com/knowledge-graph/
[3]https://www.wikidata.org/
[4]https://github.com/unsw-cse-soc/Data-curation-API

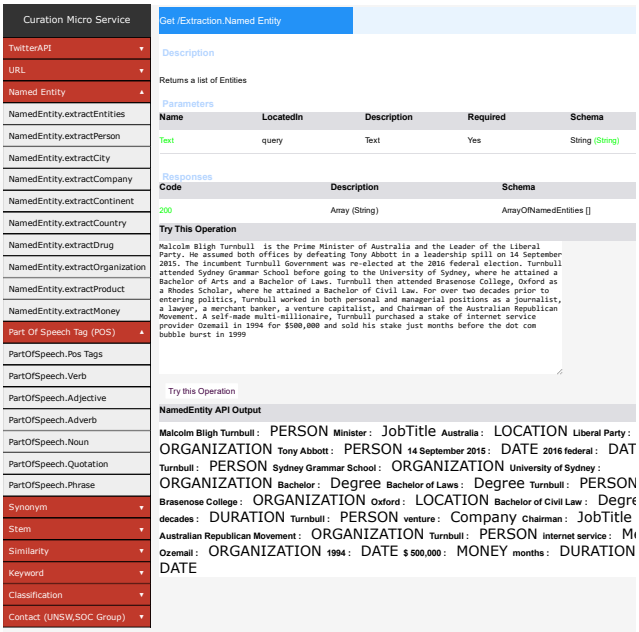

Figure 1: An example of entity extraction service. API endpoints are provided for performing content analysis on Internet-accessible web pages, posted HTML or text content.

may need a collection of natural language processing APIs to understand entities, Part of Speech (PoS), keywords, synonym, stems and more.

2.1.1 *Named Entity*

Named Entity Recognition (NER), also known as Entity Extraction (EE), techniques can be used to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, and percentages [Beheshti et al. 2016c]. In particular, named entities carry important information about the text itself, and thus are targets for extraction. Accordingly, NER is a key part of information extraction systems that supports robust handling of proper names essential for many applications, enables pre-processing for different classification levels, and facilitates information filtering and linking. State-of-the-art NER systems for English produce near-human performance [Beheshti et al. 2016c]. Figure 1 illustrates a screenshot of the basic data curation services presenting an example for the named entity extraction task.

2.1.2 *Part of Speech (PoS)*

A Part-of-Speech (PoS) is a category of words (or more generally, of lexical items) which have similar grammatical properties [Martin and Jurafsky 2000]. Words that are assigned to the same part of speech generally display similar behavior in terms of syntax - they play similar roles within the grammatical structure of sentences - and sometimes in terms of morphology, in that they undergo inflection for similar properties. Commonly listed English parts of speech are noun, verb, adjective, adverb, pronoun, preposition, conjunction, interjection, and sometimes numeral, article or determiner.

2.1.3 Keyword

In corpus linguistics a keyword is a word which occurs in a text more often than we would expect to occur by chance alone [Beheshti et al. 2016c]. Keywords are calculated by carrying out a statistical test which compares the word frequencies in a text against their expected frequencies derived in a much larger corpus, which acts as a reference for general language use. To assist analysts filtering and indexing open data, it will be important to extract keywords from unstructured data such as tweets text.

2.1.4 Synonym

A synonym is a word or phrase that means exactly or nearly the same as another word or phrase in the same language. An example of synonyms is the words begin, start, and commence. Words can be synonymous when meant in certain contexts, even if they are not synonymous within all contexts. For example, if we talk about a long time or an extended time, long and extended are synonymous within that context [Manning et al. 2014]. While analyzing the open data, it is important to extract the synonyms for the keywords and consider them in the analysis steps. For example, sometimes two tweets can be related if we include the synonyms of the keywords in the analysis: instead of only focusing on the exact keyword match. It is important as the synonym can be a word or phrase that means exactly or nearly the same as another word or phrase in the tweets.

2.1.5 Stem

A stem is a form to which affixes can be attached [Manning et al. 2014]. For example, the word friendships contains the stem friend, to which the derivational suffix -ship is attached to form a new stem friendship, to which the inflectional suffix $-s$ is attached. To assist analysts understand and analyze the textual context, it will be important to extract derived form of the words in the text. For example, considering the keyword 'health', using the Stem service, it is possible to identify derived forms such as healthy, healthier, healthiest, healthful, healthfully, healthfulness, etc; and more accurately identify the information items, e.g. tweets, that are related to health.

2.1.6 Information Extraction from a URL

A Uniform Resource Locator (URL), is a reference to a Web resource that specifies its location on a computer network and a mechanism for retrieving it. Considering a tweet that contains a URL link, it is possible to extract various types of information including: Web page title, paragraphs, sentences, keywords, phrases, and named entities. For example, consider a tweet which contains URL links. It is possible to extract further information from the link content and use them to analyze the tweets.

2.2 Linking Services

2.2.1 Similarity

Approximate data matching usually relies on the use of a similarity function, where a similarity function $f(v_1, v_2) \rightarrow s$ can be used to assign a score s to a pair of data values v_1 and v_2. These values are considered to be representing the same real world object if s is greater than a given threshold t. In the last four decades, a large number of similarity functions have been proposed for comparing [Beheshti et al. 2016c]: strings (e.g., edit distance and its variations, Jaccard similarity, and tf/idf based cosine functions), numeric values (e.g., Hamming distance and relative distance), images (e.g., Earth Mover Distance) and more. Accordingly, analysts may need a collection of similarity APIs to measure the Cosine similarity of two vectors of an inner product space and compares the angle between them, the Jaccard similarity of two sets of character sequence, the length of the longest common subsequence between two strings using an edit distance algorithm, the hamming distance between two strings of equal length and more.

2.2.2 Knowledge Bases

While extracting various features (e.g. named entities, keywords, synonyms, and stems) from text, it is important to go one step further and link the extracted information items into the entities in the existing Knowledge Graphs (e.g. Google KG and Wikidata). For example, consider that we have extracted 'M. Turnbull' from a tweet text. It is possible to identify a similar entity (e.g. 'Malcolm Turnbull'[5]) in the Wikidata. As discussed earlier, the similarity API supports several function such as Jaro, Soundex, QGram, Jaccard and more. For this pair, the Jaro function returns 0.74 and the Soundex function returns 1. To achieve this, we have leveraged the Google KG and Wikidata APIs to link the extracted entities from the text to the concepts and entities in these knowledge bases. For example, the Google API call will return a JSON file which may contain the url to Wikipedia[6].

2.3 Classification Services

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks. A classification task begins with a dataset in which the class assignments are known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time. In the terminology of machine learning, classification is considered as an instance of supervised learning while unsupervised procedure is known as clustering, and involves grouping data into categories based on some measure of inherent similarity or distance [Jajuga et al. 2012]. An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. Examples of classification algorithms include: Linear classifiers, Support vector machines, Quadratic classifiers, Kernel estimation and Decision trees. Figure 2 illustrates a screenshot of the basic data curation services presenting an example for the classification task.

2.4 Indexing Services

For the developers, it is important to expose the power of Elasticsearch [Gormley and Tong 2015] without the operational burden of managing it themselves. For example, it is important to automatically index entities/keywords for powerful, real-time Lucene (https://lucene.apache.org/) queries, e.g. while dealing with very large datasets such as Twitter data.

[5]https://en.wikipedia.org/wiki/_Turnbull
[6]https://en.wikipedia.org/

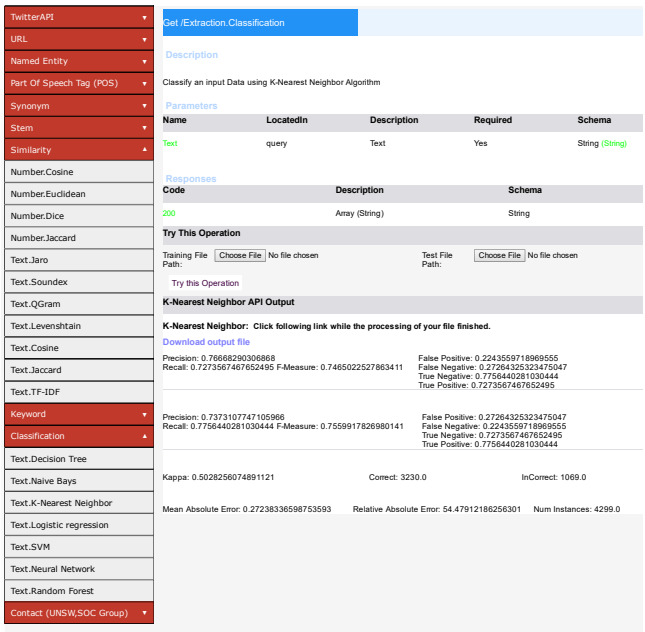

Figure 2: An example of the calssification service, supporting various algorithms such as NaiveBayes, SVM, DecisionTrees and K-nearest Neighbor.

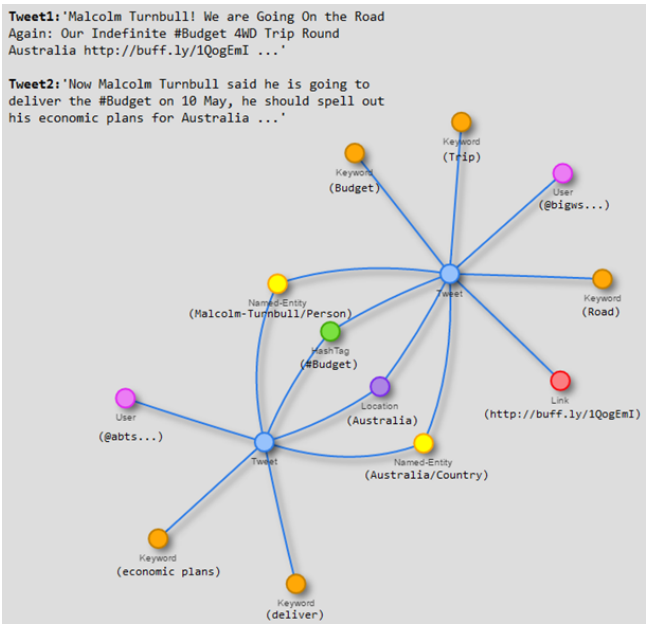

Figure 3: Use extracted features from Twitter to link related tweets.

2.5 Converter Services

The basic curation APIs may be applied to different data sources and file formats. To facilitate this task, the converter API can be used to convert PDF, Word, PowerPoint, XPS, and HTML documents into a text file to be fed to the basic curation APIs, where the result is returned as a JSON file, an easy-to-parse structure, and can be integrated into data applications. As an ongoing work, we are identifying various data sources and file formats to facilitate converting documents without user interaction.

3. DEMONSTRATION SCENARIO

The demonstration scenario consists of three parts. In the first part, we would like that the attendee appreciates the difficulties that one can encounter when dealing with the raw data. We start with a Twitter dump, three months tweets from May to July 2016. We illustrate how *extraction service* can be used to extract named entities, keywords, synonyms, PoS and stems from tweets. Then, in order to produce contextualized knowledge, we illustrate how we use the *linking service* to link extracted entities to knowledge bases such as Wikidata and Google KG. Next, we focus on a motivating scenario, understanding a Governments' Budget in the context of Urban Social Issues, to classify tweets that are related to Australian budget[7] 2016 into budget programs and categories. In particular, budget categories (e.g. 'Health', 'Social-Services', 'transport' and 'employment') defines a hierarchical set of programs (e.g. 'Medicare Benefits' in Health, and 'Aged Care' in Social-Services). These programs refers to a set of activities or services that meet specific policy objectives of the government. Afterward, we illustrate how we use the *indexing and classification services* to index the related tweets and group them in classes representing a budget program/category. It is possible to provide a set of entities as an input to the indexing API, and the API will index the set of related tweets containing those entities and/or keywords. We will demonstrate the attendee the classification result of the raw and conceptualized tweets.

In the second part, using the same Twitter data dump, we focus only on the health category of the government budget. We use the *extraction service* to extract keywords, phrases and named entities from tweets that are related to health. We illustrate how we trained this service, specifically for Australian budget, to be able to extract entities such as: (i) People, from GPs and Nurses to health ministers and hospital managers from Australian doctors directory[8]; (ii) Organizations, such as Hospitals, Pharmacies and Nursing Federation from myHospitals[9]; (iii) Locations, states, cities and suburbs in Australia from auspost[10]; (iv) Health funds, such as Medibank, Bupa and HCF from health-services[11]; (v) Drugs, such as Amoxicillin, Tramadol and Alprazolam from drug-index[12]; (vi) Diseases, such as Cancer, Influenza and Tuberculosis from medicine-net[13]; (vii) Medical Devices, such as Gas Control, Blood Tube and Needle from FDA[14]; (viii) Job titles, such as GP, Nurse, Hospital Manager, Secretary of NSW Health and NSW Health Minister from compdata[15]; and (ix) Keywords, such as healthcare, patient, virus, vaccine and drug from Australia national health and medical

[7]http://data.gov.au/dataset/budget-2015-16-tables-and-data

[8]https://www.ahpra.gov.au/
[9]https://www.myhospitals.gov.au/browse-hospitals/
[10]http://auspost.com.au/postcode/
[11]http://www.privatehealth.gov.au/
[12]http://www.rxlist.com/
[13]http://www.medicinenet.com/
[14]http://www.fda.gov/
[15]http://compdatasurveys.com/compensation/healthcare

research council[16]. We enrich these entities using KBs such as Wikidata, Google Knowledge Graph and Wordnet.

Then we use the *classification service* to identify the tweets with negative sentiment. Notice that, for the sentiment analysis, the classification API leverages the sentiment classifier implemented in the Apache PredictionIO (http://prediction.io). For example out of 2934 diabetes related tweets the algorithm identified 615 tweets with negative sentiment. As another example, we have identified 1549 tweets with negative sentiment in the Mental Health category. We propose to the attendee a scenario where she would be able to compare the classification result of the raw and conceptualized tweets.

In the third part, we discuss that, extracting all these features will be a great asset to summarize the large number of tweets. For example, 'entity summaries' of tweets containing the same named entity such as a person or organization; and 'keyword summaries' of tweets containing similar keywords. We may then analyze these related tweets to get valuable insights from the Twitter open data. For example, consider Figure 3 where two real tweets have been illustrated, it is possible to extract information (e.g. named entities, keywords, and hashtags) from the tweets and use them to generate a graph where nodes are the main artifacts and extracted information are the relationships among them. As illustrated in this figure, the tweets are linked through named entities and hashtags and this will generate an interesting graph which reveals the hidden information among the nodes in the graph: for example it is possible to see the path (transitive relationships among the nodes and edges) between user1 and user2 (in Twitter) which in turn reveals that these two users are interested in the same topics, and consequently may be part of some hidden micro-networks.

4. CONCLUSION AND FUTURE WORK

In this paper, we identified and implemented a set of curation services to make them available to researchers/developers to assist them in transforming their raw data into curated data. We have provided the technical details for the curation APIs in a technical report [Beheshti et al. 2016d]. As an ongoing work, we are identifying and implementing more services to support enriching, annotating, summarizing and organizing raw data.

5. ACKNOWLEDGMENTS

We Acknowledge the Data to Decisions CRC (D2D CRC) and the Cooperative Research Centers Programme for funding this research.

6. REFERENCES

[Anderson et al. 2013] Michael R Anderson, Dolan Antenucci, Victor Bittorf, Matthew Burgess, Michael J Cafarella, Arun Kumar, Feng Niu, Yongjoo Park, Christopher Ré, and Ce Zhang. 2013. Brainwash: A Data System for Feature Engineering.. In *CIDR*.

[Beheshti et al. 2016a] Seyed-Mehdi-Reza Beheshti, Boualem Benatallah, and Hamid Reza Motahari-Nezhad. 2016a. Scalable graph-based OLAP analytics over process execution data. *Distributed and Parallel Databases* 34, 3 (2016), 379–423. DOI: http://dx.doi.org/10.1007/s10619-014-7171-9

[Beheshti et al. 2016b] Seyed-Mehdi-Reza Beheshti, Boualem Benatallah, Sherif Sakr, Daniela Grigori, Hamid Reza Motahari-Nezhad, Moshe Chai Barukh, Ahmed Gater, and Seung Hwan Ryu. 2016b. *Process Analytics - Concepts and Techniques for Querying and Analyzing Process Data*. Springer. DOI: http://dx.doi.org/10.1007/978-3-319-25037-3

[Beheshti et al. 2016d] Seyed-Mehdi-Reza Beheshti, Alireza Tabebordbar, Boualem Benatallah, and Reza Nouri. 2016d. Data Curation APIs. *CoRR* abs/1612.03277 (2016). http://arxiv.org/abs/1612.03277

[Beheshti et al. 2016c] Seyed-Mehdi-Reza Beheshti, Boualem Benatallah, Srikumar Venugopal, Seung Hwan Ryu, Hamid Reza Motahari-Nezhad, and Wei Wang. 2016c. A systematic review and comparative analysis of cross-document coreference resolution methods and tools. *Computing* (2016), 1–37. DOI: http://dx.doi.org/10.1007/s00607-016-0490-0

[Chen et al. 2012] Hsinchun Chen, Roger HL Chiang, and Veda C Storey. 2012. Business intelligence and analytics: From big data to big impact. *MIS quarterly* 36, 4 (2012), 1165–1188.

[Gattani et al. 2013] Abhishek Gattani, Digvijay S. Lamba, Nikesh Garera, Mitul Tiwari, Xiaoyong Chai, Sanjib Das, Sri Subramaniam, Anand Rajaraman, Venky Harinarayan, and AnHai Doan. 2013. Entity Extraction, Linking, Classification, and Tagging for Social Media: A Wikipedia-Based Approach. *PVLDB* 6, 11 (2013), 1126–1137. http://www.vldb.org/pvldb/vol6/p1126-gattani.pdf

[Gormley and Tong 2015] Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: The Definitive Guide.* " O'Reilly Media, Inc.".

[Jajuga et al. 2012] Krzystof Jajuga, Andrzej Sokolowski, and Hans-Hermann Bock. 2012. *Classification, clustering, and data analysis: recent advances and applications*. Springer Science & Business Media.

[Kwak et al. 2010] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue B. Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*. 591–600. DOI: http://dx.doi.org/10.1145/1772690.1772751

[Manning et al. 2014] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*. 55–60. http://aclweb.org/anthology/P/P14/P14-5010.pdf

[Martin and Jurafsky 2000] James H Martin and Daniel Jurafsky. 2000. Speech and language processing. *International Edition* 710 (2000).

[Tene and Polonetsky 2012] Omer Tene and Jules Polonetsky. 2012. Big data for all: Privacy and user control in the age of analytics. *Nw. J. Tech. & Intell. Prop.* 11 (2012), xxvii.

[16]https://www.nhmrc.gov.au/