INFORMATION EXTRACTION FROM WIKIPEDIA

TECHNICAL REPORT

eHonnef
Faculty of Information Technology
Brno University of technology
github.com/eHonnef/information-extraction

July 13, 2020

ABSTRACT

The information extraction field objective is to automatically extract useful information from, usually, text about any subject, i.e. reviews on a product, contact information, etc. and generate some practical statistics about it. This project aims to extract information from Wikipedia's Articles and its Wikidata item by parsing the content from the latest Wikipedia's database dump.

1 Introduction

1.1 Information Extraction

Information extraction (IE) is the ask of automatically extracting structured information from an unstructured text. Most of the cases it is necessary to process the text, because IE extracts information from the actual text of documents, by means of *natural language processing* (NLP) and also it is possible to process information from multimedia documents like images, audio and video [1]

The use of information extraction is to find contact information in a text, finding the proteins from journals and papers, classifying articles from Wikipedia, check if the reviews on a product are good or bad, and so on. [2]

Usually to be able to extract useful information from texts, there are a few sub tasks involved:

- *Pre-processing:* The text is prepared for processing where the text is cleared from any garbage (HTML tags, image links, etc) and with the help of computational linguistics tools such as tokenization, sentence splitting, morphological analysis, etc, the text is ready for processing.
- *Classifying concepts:* In this step the mentions of people, things, locations, events and other pre-specified types of concepts are detected and classified.
- *Connecting the concepts:* This is the task if identifying relationships between the extracted concepts.
- *Unifying:* Presenting the extracted data into a standard form.
- *Removing the noise:* Eliminate duplicate data.
- *Enriching the knowledge base:* The extracted information is added to a database or to an existing system, in other words, this is the step where you use the extracted data.

1.2 Natural language processing

The *Natural language processing* (NLP) is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language. This technology is rapidly advancing thanks to an increase interest in human-to-machine communication, availability of big data, powerful computing and enhanced algorithms. For example, NLP makes possible for computers to read text, hear speech, interpret it, measure sentiment and determine which parts are important. [3]

Natural language processing includes many different techniques for interpreting human language, ranging from statistical and machine learning methods to rules-basaed and algorithmic approaches. We need a broad array of approaches because the text and voice-based data varies widely, as do the practical applications. Basic NLP tasks include tokenization and parsing, lemmatization/stemming, part-of-speech tagging, language detection and identification of semantic relationships. [3]

1.3 Wikipedia

The Wikipedia is a multilingual online encyclopedia created and maintained as an open collaboration project by a community of volunteer editors using a wiki-based editing system. It is the largest and most popular general reference work on the World Wide Web, and is one of the most popular websites ranked by Alexa as of January 2020. It features exclusively free content and no commercial ads, and is owned and supported by the Wikimedia Foundation, a non-profit organization funded primarily through donations. [4]

1.3.1 Articles

A Wikipedia article or entry is a page on the Wikipedia website that has encyclopedic information on it, in other words, a Wikipedia article is a sort of summarization of the knowledge. Usually articles have an infobox on the side that contains a subset of information about the subject on the article, for example, if the article is about a person the infobox will provide the person's name, birth, nationality, occupations, and so on. Articles contains information about the subject's connections with another subjects, referred as wikilinks, the categories that the subject is part of and, also, it is possible to extract useful information about the subject, usually, in the first paragraph of the article. [5]

1.3.2 Wikidata

Wikidata is a collaboratively edited knowledge base hosted by the Wikimedia Foundation. It is a common source of open data that Wikimedia projects such as Wikipedia can use. Wikidata is a document-oriented database, focused on items, which represent topics, concepts, or objects. Examples of items include 1988 Summer Olympics (Q8470), love (Q316), Elvis Presley (Q303), and Gorilla (Q36611). Each item is identified by a unique number, prefixed with the letter Q, known as a "QID". This enables the basic information required to identify the topic the item covers to be translated without favoring any language. [6]

2 Methodology

2.1 Data from Wikipedia articles

The extracted data from Wikipedia articles includes the infobox (1.3.1) that have summarized information and it is possible to classify the subject using it, the first paragraph, more specifically the first sentence, contains defining words for the subject, take the *Abraham Lincoln* article for example, the first sentence is "Abraham Lincoln was an American statesman and lawyer who served as the 16th president of the United States from March 1861 until his assassination in April 1865." so we can extract the following defining words "American, statesman, lawyer".

It is possible to extract the subject's connections to another subject using the wikilinks in the article that later can be used to produce a graph of relationships, every article have a category, or multiple, so we can use it to assemble a set of common subjects.

2.2 Data from WikiData

There are plenty of data in the article's WikiData, a few examples are, the article name in other languages, the subject type (e.g.: human, book, series, tv show, etc.), the gender and so on. For this project it was only used the subject's type so we have a more narrow categorization for the articles, with that we can check the occurrence of the defining words for each categories and, possibly, feeding a neural network to predict the subject's type by giving a set of defining words.

2.3 Implementation

This sections will describe how the *Python* code works and the idea on how to parse and extract information from the files.

2.3.1 Data parsers

The code uses the latest split version of the Wikipedia dumps so it is easily paralleled (you can assign one core per file) or it is possible to split a single file to multiple cores since it is a multistream file. To reduce the memory usage the program feed the XML parser one line at time without the need to uncompress the file and after *N* articles readed it will save the parsed and processed content (described in the section 2.3.2) to a file and clear the stored data to continue the parsing.

The wikidata parser works in a similar way, but it is necessary to first parse the XML and then parse the JSON contained inside the <*text*> tag, the program uses the XML version of the wikidata dumps in this way the wikidata items are split in different files.

The XML parser uses the library *xml.sax*, provided natively in the Python package, and the wikimedia parser uses the *MWParserFromHell* library [7], the wikidata uses the *JSON* package that is also provided natively.

2.3.2 Information extraction

To extract information from the articles its content is parsed using the *MWParserFromHell*, then the first paragraph is feeded to the Stanford NLP [8], to extract the defining words it is used the first sentence and looked for the lemma *"be"*, as the NLP create a dependency graph, the defining words are the nouns linked with the verb *"be"* until another verb is found.

There are some cases that is almost impossible to find the connection from the verb *"be"* to its "truly" defining words, we can use the *Apollo* [9] article as example, the first sentence is "Apollo is one of the most important and complex of the Olympian deities in classical Greek and Roman religion and Greek and Roman mythology.", so, currently, the defining words extracted are "one" instead of being "Olympian, deity". The reason is that sometimes the sentence structure is very complex and the dependency graph becomes very unpredictive, of course we could just filters the nouns of the sentence but then it'll have some nonsense, for example the Apollo [9] article, if we use only nouns it'll include the word "religion", and Apollo isn't a religion.

The information extraction includes the wikilinks, categories, the items from the infobox and the infobox's type.

3 Results

To show that the code works it was parsed one wikipedia dump, this file contains around 15 thousand articles excluding the "list of", redirect pages, category pages and others pages that are not considered articles, and one wikidata dump file, that contains around 22 thousand items excluding redirect items, items without label and some items that the type could not be read.

Figure 1 shows the search results of the defining word *"rock"*, as we can see it shows that almost every article that have "rock" probably is a rock band, but remember that it does not have all Wikipedia's article parsed and registered.

The reason that the entire Wikipedia dump is not processed for this work is the computational cost, it takes around 2 hours in a 3.38 GHz CPU to parse one articles dump file and around 20 minutes to parse one wikidata dump file. The difference in time is because the article needs to be XML parsed, parsed in the wikimedia parser and finally fed to the NLP to extract information. Also, it is considered the time to save to disk when the counter reach around 5000 articles read and the deletion of the variable data to reduce the memory usage.

4 Conclusions

This project showed that it is possible to retrieve information from the Wikipedia articles' text and their infoboxes, then draw some statistics as shown in the section 3 so in that way we could use to predict the type of the subject by the infobox type or by its defining words.

The project could possible extract more information from the articles and wikidata, for example, how strong is a connection between the subjects, and a creation of a neural network to predict the subject's type or category from given words. It has some flaws, as shown by the Apollo [9] article, so it needs more fine tunnings and adding more cases in the article parser.

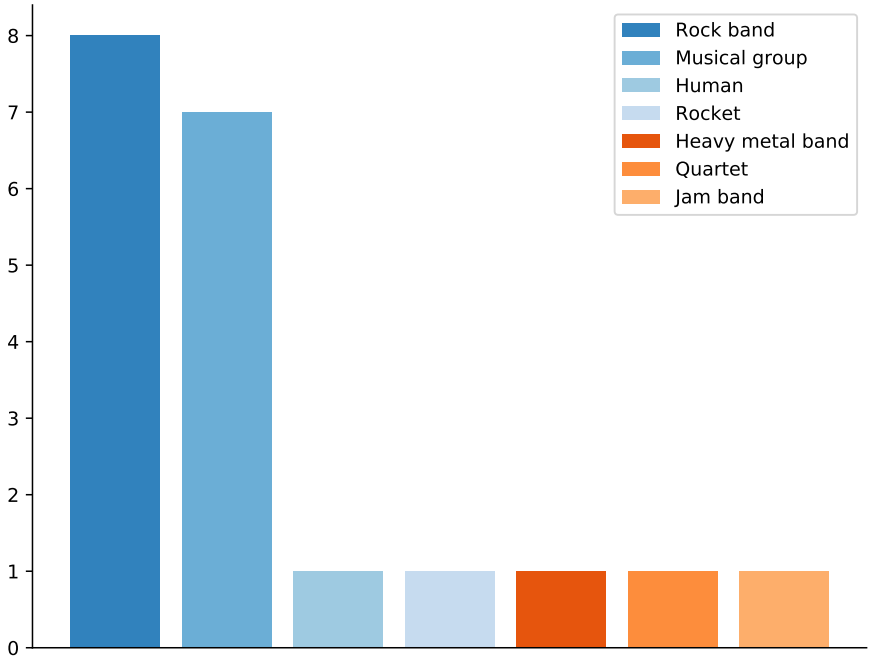

Figure 1: The results for the word *"rock"*

References

[1] Jim Cowie and Yorick Wilks. Information extraction, 1996.

[2] Jie Tang, Mingcai Hong, Duo Liang Zhang, and Juanzi Li. Information extraction: Methodologies and applications. In *Emerging Technologies of Text Mining: Techniques and Applications*, pages 1–33. IGI Global, 2008.

[3] SAS. What is natural language processing? `https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html`.

[4] Wikipedia contributors. Wikipedia — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Wikipedia&oldid=936603935`, 2020. [Online; accessed 20-January-2020].

[5] Wikipedia contributors. Wikipedia:what is an article? — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/wiki/Wikipedia:What_is_an_article`, 2020. [Online; accessed 20-January-2020].

[6] Wikipedia contributors. Wikidata — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Wikidata&oldid=935778700`, 2020. [Online; accessed 20-January-2020].

[7] Earwig, Legoktm, et al. Mwparserfromhell. `https://mwparserfromhell.readthedocs.io/en/latest/`.

[8] Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[9] Wikipedia contributors. Apollo — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Apollo&oldid=934274661`, 2020. [Online; accessed 20-January-2020].