Department of Electrical and computer Engineering

Part IV Research Project

Literature Review and Statement of Research Intent

Project Number: 84

Natural Language Processing and Text Analysis Report

Leighton Jonker

William Chao

Primary: Gill Dobbie
Secondary: Danielle Lottridge

30/04/1997

Declaration of Originality

This report is my own unaided work and was not copied from nor written in collaboration with any other person.

Leighton Jonker

Name: Leighton Jonker

ABSTRACT: This Literature Review and Statement of Research will be about the fourth-year engineering project that is being undertaken by Leighton Jonker and William Chao. The research that this project will be focused on will be the analyzing of text inputted by users. The text will be scanned for important and prominent words, which will then be put into an image search engine (such as Flickr) and retrieve the images that most suit the original inputted text to be then added to the text to enhance aesthetics by user choice. Various pieces of engineering literature will also be included, and will discuss the differences that these pieces of literature have compared to what we wish to research for our project.

1. Introduction

When using social media sites such as Facebook and Twitter, you can update your status with what you're currently feeling like; happy, sad, hungry or adventurous. Alongside the typing of a status users get the option to choose a colour to be put in the background of the status update to make it stand out and give it a little more of your personality. Currently this option is only limited to solid colours with the occasional shape of a star or moon. What we wish to achieve is a smart text analyser that will analyse the content of the status update and generate a list of photos that would be the most suited to making the status update stand out, which the user will then select to add into their status update. The application will allow users to customise their own status update to their liking with various fonts, text sizes and images to choose from.

This kind of technology is currently unused in popular social websites and helps develop the link between Natural Language Processing, Human-Machine Interfacing, machine learning and text analysis; meaning that this research has the potential to apply to various other applications rather than just use in social media. Having a feature such as automatically searching for the most suited image or perhaps different items will help quality of life for users of said system, increasing their engagement further, helping them retain interest and keeping them coming back time after time. A feature such as this may also have potential for advertisement, as results can be specially tailored for certain products or can feature professional photographers for further exposure to social media users.

Using Flickr APIs, you can send a search for a photo using specific tags, and categorise them based on upload dates, tags, photo taken dates, location and content type among other things. A web application would be created that allows users to type whatever they wish to be their status, and the application would analyse the text and return them with a list of the most compatible images that the users can then select and add to their status, alongside possible other formatting options (different fonts, sizes, special effects etc.) to

ultimately make the status stand out and gather attention to itself.

Ultimately, the purpose of this research is to create an algorithm that can discern what parts of a sentence in English are the most important and to then use those keywords as inputs to various other algorithms that are designed to use the internet (Flickr APIs) to find the most suited image for the keywords. All of this will be incorporated in an application which lets users input this sentence as well as pick from a list of returned images to add to this sentence to enhance the visual appeal of this sentence.

2. Natural Language Processing

Natural Language Processing is quite important when it comes to analysing text. Text may come in any form, whether it be digitally or analogue, old or new. The extent to which Natural Language Processing has progressed throughout the years has progressed to the electronic level. Every day there is millions of bytes of textual data being uploaded to the internet, and while most of this may not apply to your typical user, if someone wishes to access any of this information whether it may be from a news article or a status update, they need to look for it. Websites and news sources usually tag their articles with keywords which allow them to stand out when users wish to search the web for information but what happens in the case where tags are not created and there is nothing but a block of text that has no discernable way of being searched for? This is where computer science and natural language processing come into play; providing users with tools that analyse the internet in search for whatever a user may require, providing them with an organised list of what their search terms align with. The ability to analyse hundreds of thousands of documents is a very crucial ability to have when it comes to traversing the internet, as the internet grows every day. Technology has evolved to the point that search results that number in the millions can be returned to a user within a matter of milliseconds of searching for a term and can contain data about all aspects of the searched term; for example, searching the term noodles in the Google search engine will return upwards of 46 million results in just over half a second, providing information on what noodles are, how they are made, noodle shops that are close to you and nutritional information on noodles. In order to cut down the results gained users are required to refine their search term by adding various other keywords to narrow down what the search engine will be searching for, such as the word" recipes" when it comes to searching for noodle recipes etc., allowing the search engine to refine its large search and focus on a certain aspect of the major search term. The language processing that Google has been able to achieve allows users to have a seemingly unlimited amount of information available at the touch of their fingertips.

3. Human Machine Interface

Over the years human-machine interactions have been getting more and more streamlined, with continual improvements to human quality of life being made. Interaction between humans and machines are made to be as efficient and easy for the human as possible, in the terms of a smartphone; they are continually being improved to make the learning process easier and easier when it comes to using one. One of the benefits of continually improving human-machine interfaces is that humans are being exposed to technology at earlier ages, allowing for more effective learning methods at younger ages. Many children nowadays have access to some sort of iPad or smartphone with various educational applications that they use to learn as well as be entertained. The ability at which modern-era humans can process information in terms of memory, learning and problem solving are also attributed to the ever-increasing prowess of human-machine interfacing, allowing humans to easily access whatever information they require through the internet. Previously the curiosity of the human mind could not quite to satiated as information gathering required an individual to search for the information they required manually through reading books or researching things themselves, whereas now all one needs to do is search it up on the internet where detailed and specific information is available to them; all wrapped up in an easy to use interface that minimizes the problems that would typically occur when it comes to attempting to search for the information manually.

4. Text Analysis

Similar to Natural Language Processing, text analysis is the actual algorithms that are used to sift through the multitudes of information that the internet contains. Although our research project will involve the analysis of much smaller pieces of data the same concepts will apply. When it comes to text analysis there are multiple methods to choose from, each with their own benefits and drawbacks in certain situations. The main basic methods of text analysis as stated in reference [1] are Keyword Spotting, Lexical Affinity and Statistical NLP. Keyword spotting consists of scanning a document in search of designated keywords.

Lexical affinity is more complex than keyword spotting, as it assigns different affinities to words. For example, it would assign the word “Bake” an 80% probability that the document would be about food in some way shape or form. These probabilities would often be taken from various text corpora. Throughout the whole document these words with their affinities will be gathered and a result will be found by the end of the document, effectively categorizing the document.

Statistical NLP is one step further than keyword spotting and lexical affinity, as it introduces machine learning. Combining the keyword approach along with the arbitrary word analysis approach of lexical affinity;

statistical NLP allows for an even greater form of text analysis. Statistical NLP however has some drawbacks, as it requires very large documents to provide accurate findings, which will not work in our case.

Lemmatization is a form of text analysis which aims at breaking down sentences into much more basic forms in order to make the analysis easier for other algorithms. Lemmatization reads a text and converts the words it finds into their common base form, for example, trucks, truck's and trucks' all found in a text would simply be changed to truck to allow easier keyword recognition from other algorithms.

The Naïve Bayes algorithm is a family of probabilistic classifiers that are designed to categorize inputted words, and is based off the Bayes Theorem. This algorithm attempts to classify individual features of something to assume an object, whether this be colour, size or weight.

Currently text analysis algorithms are used to mass process documents, allowing for their efficient categorization. These categorizations are primarily used so that internet users can then efficiently search for any kind of document they wish using search engines. When a user uploads a blog or article to the internet and does not tag it in any way, normally it would be very difficult for others to find. With these systems being able to scan these untagged documents and generate tags for them, it makes it much easier for other users of the internet to find otherwise hidden websites.

Additionally, during the document tagging process a large amount of information can also be extracted from the document, and have summaries created for them, extract key parts of the information the text holds, and even compute a predicted bias of the author for certain subjects.

Text analysis is not just limited to analysing digital text, as it has been used in physical texts as well. Old books can be digitized by scanning them, ensuring that they last longer incase said book is destroyed, through giving them a digital form. Technology has come so far as to be able to have a camera look at and translate physical words into digital ones which can then have each word analysed, as opposed to just taking a picture of a page in a book and looking at it in the future. This means that these newly digitized text can be analysed for categorization as well.

5. Other Research

In the past there have been other text analysis and natural language processing projects which focus on social media, much like what we wish to do. These projects are like ours in the sense that they make use of text analysis algorithms to extract information from social media text such as from Twitter hashtags and status updates. One such research project, which focuses around microblogging and hashtags, analyses

social metrics [5]. After lemmatizing texts from Twitter by using a word dictionary consisting of the 186 most common English words, they would filter out the unnecessary words (such as and, to, from etc.). What they found was that most of the words that were left did not help in the description of the context of the text, these words were very common words such as happy, please, thank and video. On the contrary when they focused the filtration to focus on hashtags they found more defined results, as hashtags allow users to highlight certain defined topics, and are mainly the primary focus of most Twitter status updates.

Another research project used Facebook to study user data which contained the posts shared by the user [6]. They then used the Naïve Bayes algorithm with a word dictionary of 100 words split into 50 good and 50 bad, and extracted the good and bad words from the Facebook posts with the hope of being able to aid investigations by identifying potential digital lawbreakers who have posted negative things on Facebook. This project did not take into consideration the context of these posts as well as not define what being “Good” of “Bad” was, leaving the results they found quite vague and general.

This final piece of research relates much closer to our research over the previous pieces, as it focuses on keyword extraction from social media short text [7]. This project uses Word2vec and Textrank to capture semantic links between words, extract keywords and then rank them. Word2vec is a program that vectorizes the words in a text, which effectively creates a vector space that connects word frequency and proximity. Textrank is a graph-based ranking model, which uses ranking algorithms to determine the importance of words, with importance being determined by the linking of vertexes, with the more links contributing to word importance. They compare the algorithm that they created against the TF-IDF (Term Frequency – Inverse Document Frequency) and Textrank algorithms and find that the algorithm they created has similar results when it comes to long documents (such as Wikipedia) and is superior to TF-IDF / Textrank when it came to keyword extraction and ranking, while also being faster than Textrank but slower than TF-IDF.

6. Our Research

How all this research relates to our own research is by how we will be creating out own system that is very similar to the most common usage of text analysis algorithms on the internet to date. Normally text analysis systems would be created with the purpose of scanning and analyzing thousands of large documents, efficiently categorizing and extracting all the important information within them, but what we wish the achieve would be different and designed for a smaller scale, not analyzing pre-defined text, but text that is freshly input by a user, with the additional feature of providing immediate

feedback in relation to what the user had just input, along with the feature of allowing them to pick their own preference when it comes to the options presented to them by our system. Our final system will be a real-time system that reactively processes all live inputs given whilst returning a selection of most suited options for the user to choose from.

7. Current Aims

By the end of this project we aim to have created a web application that will allow users to input text, and from that text generate a list of suggested images that will relate to the initial text. Users will then be able to choose the image they like the most and it will appear alongside the text that they have written, making the text that they have written more aesthetically pleasing and stand out more. We wish to create the application using AngularJS, a JavaScript based framework. We have adequate experience with JavaScript and its various other modules, and believe that AngularJS is a well-established and supported framework that results in well-structured and high-quality applications. In the future if we wish to have our application work with other applications we believed that if we used a popular and easy-to-use framework such as AngularJS then it would make the merging process much easier and smoother.

8. Future Aims

In the future we wish to expand the application to be integrated with popular websites such as Facebook, Tumblr or Flickr as those websites do not currently have status analysis incorporated in their models. We saw that there was a lack of smart text-analysing software particularly in this region and believe that adding software such as this would make the websites more aesthetically pleasing and positively add to what they currently have. Additionally, we were hoping to expand the application to smartphones, as if we had already implemented the system on a web browser which is more complex it would be easier to port it over to the mobile scene for increased usage with major smartphone social media applications.

9. Conclusion

In conclusion, digital text analysis and natural language processing is mainly used when it comes to the analysis of digital documents. The purposes that encompass the use of text analysis contain but is not limited to information extraction, question answering and text summarization. Systems can be created to analyse thousands if not millions of documents extremely fast with relatively accurate results. Currently the main focus of this technology is around the preservation of data and analysis / keyword tagging of the internet. What we wish to do with this technology is contribute to its currently limited social media presence.

Currently no major social media website boasts an integrated system that allows users to input a string of words (status update) and have an algorithm search images dependent on those words that the user can then choose to add to their status. How we chose to approach this method was through the creation of a web application at first, with hopes to be able to transfer it to be able to work on a smart phone and with possible social media application interactions. We have decided to use AngularJS as it is popular within the web market which means it will have easy adaptability for newer applications and will have compatibility with the latest functions that AngularJS has to offer.

10. References

[1] E. Cambria and B. White, "Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]," in IEEE Computational Intelligence Magazine, vol. 9, no. 2, pp. 48-57, May 2014.

[2] E. Günther, T. Quandt. "Word counts and topic models: Automated text analysis methods for digital journalism research." in Digital Journalism 4.1 2016, pp. 75-88.

[3] S. K. Card “The psychology of human-computer interaction.” CRC Press, Xerox Palo Alto Research Center, 2017.

[4] K. Coyle "Mass digitization of books." *The Journal of Academic Librarianship 32.6*, 2006, pp. 641-645.

[5] T. Hachaj and M. R. Ogiela, "Clusters of Trends Detection in Microblogging: Simple Natural Language Processing vs Hashtags – Which is More Informative?," *2016 10th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS)*, Fukuoka, 2016, pp. 119-121.

[6] M. Ketcham, T. Ganokratanaa and S. Bansin, "The Forensic Algorithm on Facebook Using Natural Language Processing," *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Naples, 2016, pp. 624-627.

[7] D. Zhao, N. Du, Z. Chang and Y. Li, "Keyword Extraction for Social Media Short Text," *2017 14th Web Information Systems and Applications Conference (WISA)*, Liuzhou, Guangxi Province, China, 2017, pp. 251-256.