WORD CLOUD GENERATION OF WEIGHTED WORDS FOR SENTIMENT ANALYSIS

Dr. U.M. Fernandes Dimlo¹, R. Harish Kumar², Chandanam Chandana³

¹ Professor and HOD, Department of CSE, Narsimha Reddy Engineering College, Maisammaguda, Dhulapally, Secunderabad, Telangana, India
² Assistant Professor, Department of CSE, Narsimha Reddy Engineering College, Maisammaguda, Dhulapally, Secunderabad, Telangana, India
³ M. Tech Student, Department of CSE, Narsimha Reddy Engineering College, Maisammaguda, Dhulapally, Secunderabad, Telangana, India.

Abstract - The broad utilization of area based administrations has prompted an expanding accessibility of direction information from urban situations. These information convey rich data that are helpful for enhancing urban communities through activity administration and city arranging. However, it likewise contains data about people which can imperil their security. In this examination, we work with the New York City (NYC) taxi trips informational index freely discharged by the Taxi and Limousine Commission (TLC). This informational collection contains data about each taxus taxicab ride that occurred in NYC. An awful hashing of the emblem numbers (the ID comparing to a taxi) permitted the recuperation of all the emblem numbers and prompted a security break for the drivers, whose wage could be effortlessly removed. In this work, we start an investigation to assess whether "culminate" namelessness is conceivable and if such a personality divulgence can be maintained a strategic distance from given the accessibility of different arrangements of outer informational collections through which the shrouded data can be recouped. This is refined through a spatio-fleeting join based assault which coordinates the taxi information with an outer emblem information that can be effortlessly assembled by an enemy. Utilizing a reproduction of the emblem information, we demonstrate that our assault can re-distinguish more than 91% of the cabs that employ in NYC notwithstanding when utilizing an ideal pseudonymization of emblem numbers. We likewise investigate the viability of direction anonymization procedures and show that our assault can at present recognize a critical division of the taxicabs in NYC. Given the limitations in distributing the taxi information by TLC, our outcomes show that unless the utility of the informational collection is altogether traded off, it won't be conceivable to keep up the security of taxi emblem proprietors and drivers.

Keywords - Big social data, Social set analysis, Social business, Visual analytics, geo-spatial, GIS, Taxi, Green cabs, Uber.

I INTRODUCTION

A word cloud (or label cloud) can be a helpful device when you have to feature the most generally referred to words in a content utilizing a snappy representation. Obviously, you can utilize one of the few on-line administrations, for example, wordle or tagxedo, exceptionally include rich and with a decent GUI. Being a R fan, I constantly needed to deliver this sort of pictures inside R and now, because of the as of
late discharged Ian Fellows' wordcloud bundle, at long last I can!
Keeping in mind the end goal to test the bundle I recovered the titles of the XKCD web funnies incorporated into my RXKCD bundle and created a word cloud in light of the titles' statement frequencies computed utilizing the ground-breaking tm bundle for content mining (I know, it resembles murdering a fly with a bazooka!).
Here in this report we will make a word cloud for a given archive, essentially in word cloud we demonstrate the word with higher recurrence with greater size and the words with bring down recurrence with littler size In this report we will create a word cloud from the PDF record.
Our corpus contain just PDF document so before stacking the corpus into memory it require to convert that into the txt record. make an index name corpus inside this registry make another catalog name pdf so full way resembles this our .R record and our corpus registry are in a similar index and pdf archive.

II RELATED WORK
Research on word clouds falls in one of two categories: 1) work that investigates the effectiveness and visual perception of word clouds, and 2) work that develops improvements and extensions to the word cloud visualization.
In addition, we define a third category for this work, consisting of text analysis systems that use word clouds as one of their components.

Effectiveness and Perception
There have been several attempts to investigate the effectiveness and perception of word clouds. Bateman et al. conducted a user study in which they systematically varied nine visual properties of word clouds. They found that the properties with the largest effect on the users’ attention are font size, weight, and color. Rivadeneira et al. also observed a strong effect of font size in their user study. Furthermore, Bateman et al. As well as Lohmann et al. found that terms in the middle of the cloud receive more attention on average than terms near the borders.

Word clouds have been compared to unweighted lists and other user interfaces in a number of studies. The results indicate that users are on average more effective in spotting a specific term in an alphabetically ordered unweighted list than in an alphabetically ordered word cloud. However, frequently used terms are found more quickly in word clouds due to their larger font sizes.

Sinclair and Cardew-Hall [26] compared word clouds with a user interface simply consisting of a search box. While participants preferred the search box to enter specific terms, they favored the word cloud for more open-ended tasks. This finding is supported by Kuo et al. who used word clouds to summarize search results. Their results further indicate that word clouds are effective to give an impression of what information is present in a query result set. They draw the conclusion that word clouds are a good visualization technique to communicate an ‘overall picture’ of the text contents.

Improvements and Extensions
The above reported works mainly studied rectangular word clouds with a sequential line-by-line layout. However, several improvements and extensions to this basic layout have been proposed in the last couple of years.

Kaser and Lemire, for instance, use slicing trees, nested tables, and rectangle packing to optimize the distribution of space in HTML-based word
clouds. Seifert et al. present a related algorithm for white space optimization that can cope with differently shaped word clouds. It places terms in a circular fashion with the most frequent ones at the center and those with lower frequency towards the boundaries.

Other works use clustering techniques along with different kinds of word cloud layouts, ranging from line by-line layouts to force-directed layouts and topographical term landscapes. While the relatedness of the terms is indicated by their spatial distance in most of these works, some explicitly depict term relations, either by connecting the terms with arcs or by highlighting related terms in the word cloud.

Over the last years freely available word cloud generators, such as Wordle, Tagul, or Tagxedo, have been developed that produce visually appealing word clouds. These tools offer several options to customize the word cloud visualization by adapting typography, color, word orientation, or even the general shape of the word cloud. However, they are intended as ‘design tools’ rather than tools for text analytics. Consequently, the resulting word clouds are aesthetically pleasant but provide nearly no features to analyze the underlying text.

There are also attempts to include a temporal dimension in word clouds, for instance by using sparklines or histograms to depict changes in term use over time. Parallel Tag Clouds combine the ideas of word clouds and parallel coordinates to allow for a direct comparison of term frequencies at different points in time or from different data sources. Tree Clouds combine word clouds with trees to visualize the semantic relatedness of terms. Prefix Tag Clouds use prefix trees to group different word forms and visualize the subtrees as tag clouds. Finally, there are also 3D variants of word clouds, such as WP-Cumulus that provides a rotating, three-dimensional sphere of terms.

While a part of these extensions has been designed for specific application contexts, others can be used more generically. We adopted some of these ideas in our approach, such as the circular word cloud layout or the interactive highlighting of term relations.

III THE WORD CLOUD EXPLORER

Figure 1 shows a screenshot of the Word Cloud Explorer, generated with the popular Sherlock Holmes novel “The Hound of the Baskervilles” by Arthur Conan Doyle. The system consists of the central word cloud view and a number of other components providing additional information and functionality for the analysis.

The individual components are marked with letters in Figure 1. In the following, we will describe their functionality and explain how they support analysts.

Text Processing

After a text file is loaded, the system performs a linguistic analysis of its contents. We use the Stanford CoreNLP tools [28] for this purpose and perform several processing steps, consisting of tokenization, sentence splitting, part-of-speech tagging, lemmatization, and named-entity recognition. Based on the results of the part-of-speech tagger, we additionally implemented a detector for nominal multiword expressions. It joins all continuous sequences of proper nouns that occur in the same sentence. With this simple heuristic, we can detect most compound nominals and proper names in the
text (see for a comprehensive summary of different multiword phenomena).

The separate display of multiword expressions is important for many analysis tasks, especially those involving the identification of person or place names which are often multiwords (e.g. “Michael Jordan” or “New York”). Another benefit of considering multiwords is that the frequency counts of the individual terms are not artificially increased (e.g. “new” as part of “New York”).

**Word Cloud View**

The word cloud view (Figure 1a) is the main view of the system. It implements three different word cloud layouts that the users can choose from: two sequential line-by-line layouts, one ordered alphabetically, the other by frequency, and a circular layout showing the terms with the highest frequency in the center of the cloud and the lower frequency terms closer to the perimeter. The circular layout is shown in the screenshot of Figure 1; similar layouts have been presented.

While the alphabetical layout supports users in quickly spotting specific terms they are looking for, the frequency ordered layout lets them arrange terms according to how often they occur in the text. The circular layout complements the sequential line-by-line layouts as a space-efficient and visually appealing alternative.

Circular layouts have additionally shown to be most effective to spot high frequency terms in word clouds. Font size is scaled linearly with the occurrence frequency of the terms for all layouts. As the Word Cloud Explorer allows for a free placement of terms, additional word cloud layouts can easily be added to the system. Also, the mapping of the frequencies to the font size of the terms may be adapted. The word cloud view uses the information about different word forms provided by the lemmatization component to subsume them under one representative term in the word cloud. This means that, for example, not all inflections of a verb are shown as separate terms in the cloud but that their counts are added up and contribute to their most frequent representative within the given text.

Detected multiwords are displayed in camel case to make them easily recognizable as one entity. However, both features can be disabled in the menu if users do not want a special treatment of multiwords or the merging of different word forms.

**Co-occurrence Highlighting**

Users can hover over terms to highlight related terms. In the current implementation, two terms are related if they co-occur within the same sentence. Alternative implementations might compute the co-occurrences on larger text segments, such as paragraphs or whole documents. This feature of “co-occurrence highlighting” is an intuitive technique to show term relations in word clouds without producing visual clutter. Related terms are marked with a yellow box in our approach whose saturation corresponds to the relative co-occurrence frequency.

![Word Cloud Explorer](Image)
IV EXPERIMENT RESULTS

The study participants were surprised by the possibilities offered by such a simple and straightforward visualization as word clouds if enriched with context information and sophisticated interaction techniques. Overall, they performed very well in solving the analysis tasks with the Word Cloud Explorer. The paper with the hints was hardly ever used and the tasks were generally solved quickly and correctly without any help.

The participants rated the Word Cloud Explorer as an intuitive and useful text analysis system. They stated that they could imagine using it to analyze large bodies of text. However, they also stressed that they would only use it in combination with other tools complementing its functionality. For instance, while the ability to refer back to the actual text source was considered a crucial feature, some participants criticized the limited functionality of the text viewer component and proposed to integrate the approach with a powerful text editor. With respect to the word cloud layouts, an interesting finding was that all participants preferred the sequential layouts over the circular one, although they rated the circular one to be aesthetically most appealing. When asked about this apparent contradiction, most participants answered that they found it easier to visually compare relative word sizes using the line-by-line layout. This is because the lines could be used as visual anchors which facilitate to compare font height. Furthermore, we could observe the participants switching between the frequency and alphabetically ordered layouts according to whether they were interested in high frequency terms or searching for a specific term. This indicates that it is important to provide different word cloud layouts that users can choose from depending on the analysis task.

However, most participants preferred the search box over the word cloud when searching for a specific term. This result is not surprising, as using the search box is fast and also allows to find terms that are not part of the displayed word cloud. Furthermore, it is in line with the finding of Sinclair and Cardew-Hall that a search box is preferred for specific tasks while a word cloud for more general ones.

The participants were in disagreement about the usefulness of the part-of-speech coloring function. Some considered it a useful feature, while others found that it has little analytical value. It was argued that the part-of-speech of most words does not need to be visually communicated, as users normally know the part-of-speech of a word once they read it. Using the part-of-speech categories as a filter for the word cloud was considered more useful.

The named entity feature was unanimously found helpful and the participants used it frequently to solve the analysis tasks. The aggregation of multi words and different word forms was also positively evaluated by the participants. Overall, the participants assigned the linguistically and interactively improved word clouds many positive attributes (such as “tidy”, “clear”, “efficient”, “useful”). All this feedback indicates that word cloud based approaches have indeed some potential in text analytics if part of an advanced implementation like the Word Cloud Explorer. The user study revealed that their main advantages are flexibility and intuitiveness, which indicates that they might be particularly beneficial in environments were training times should be minimal.
V CONCLUSION & FUTURE WORK

The Basic Analysis for any data oriented application for any data set of unstructured data or semi-structured data is to extract the keywords from the data file using word cloud we can extract the words and form a word weighted graph which showcases how many times a word has been repeated how many times, this helps us to know exactly what data is speaking about.

The same application used to fetch keywords from streaming data which theoretically proves than even work on social media contents.

VI REFERENCES


