Creativity Metric for Written Text

Venkata Subramanian Mahalingam
Georgia Institute of Technology
venki@gatech.edu

ABSTRACT
In this paper we describe a novel creativity metric to assign a score specifying how creative a piece of written text is. This metric is based on the intuitive and common sense notion that the more varied the things referred by the text, the more creative it is and the fact that boring, droll text only talk about very few things and using common place words. Wikipedia is used as the primary source of knowledge of the words in the text by cross-referencing the categorization details given for that word in its page.

Author Keywords
Creativity, text, writing, wikipedia.

ACM Classification Keywords
H5.m. Algorithms, Measurement.

INTRODUCTION
Creativity in written text widely varies across different documents. The primary reasons for this is the fact that not all written text needs to be creative and the requirement and amount of creativity in the written document is dependent on the intended audience.

When the text is creative, it usually refers to multiple disparate things. We observed a clear demarcation in terms of the disparity of concepts that a document refers to for creative and boring texts. In our algorithm, we exploit this disparity property to construct the metric.

Though there has been previous work in this area [1], it has been constrained to using statistical machine learning on single sentences constrained based on keywords whereas our method can work on documents.

Measuring creativity using computers gives the advantage of speed and relative consistency across documents. Since the method is parameterizable, it allows the user to fine tune the parameters to fit his needs.

It should be noted that that creativity metric is only a heuristic constructed based on the study of various documents and trying to understand the basic property that has direct relevance to the level of creativity seen in the documents. And this basic property is the disparity seen in the words in terms of the concepts that they refer to and other characteristics like they type of the word (if it is the name of a location or a word not seen in the dictionary). Thus, it is not just the syntactic but the semantic knowledge of the document that is used in calculation of this metric.

To obtain the semantic knowledge about the words, we refer to Wikipedia. Wikipedia can be seen as not only the repository of information, but also the perspective that users have about the word. This perspective comes out by the way of categorization given by the users in the page of the word. The users of Wikipedia assign each word to a list of categories as deemed fit.

PARAMETERS
The algorithm basically calculates the values of a set of parameters calculated over the document. While the method of calculation will be explained in the next section, in this section we will explain the meaning and rationale behind the parameters.

Word count – this is used as the normalization baseline. All other parameters are in some way calculated as a ratio with respect to word count.

Word re-occurrence count – the more the repetition of words in the document, the less creative the document is.

Categories count – refers to the various categories the words of the document fall into. The more categories, the more creative the document is.

Categories re-occurrence count – the more the repetition of words in the document, the less creative the document is.

Count of words not in wiki (wiki fail) – refers to the words for which Wikipedia does not have a direct page entry. (example: proper nouns, made up words). The more such words, the more creative the document is.

Places count – refers to the words for which Wikipedia has the categorization as a location. It was observed that creativity in a document was directly proportional to the different locations referred in the document.

English words count – refers to the number of words that are from the English dictionary. The more such words, the less creative the document is.

ALGORITHM

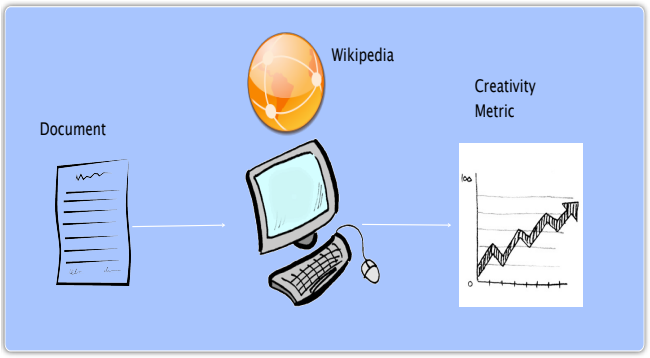

Figure 1: Information flow in the metric calculation

Figure 1 shows how the computer uses Wikipedia as the reference information to come up with the creativity metric by processing the document.

In this section, we describe the detailed method involved in starting out from the base document and ending up with the value for the creativity metric.

1. clean up the document by removing non-text characters.
2. Stem or singularize the words
3. Ignore words which are stop words – words like the, and, of etc.,
4. Update the word count.
5. If the word has been seen already, increase the word re-occurrence count.
6. If the word is present in the English dictionary, increase the English word count.
7. Try to access the Wikipedia page for the word. If Wikipedia does not have a unique page for the word, count it as wiki fail.
8. Try to access the categorization details of the page. Recurse to find the category of categories until it is 3 levels deep so as to get to the general category information of the page.
9. Calculate the ratios for all the above-calculated counts by using the word count as the base.
10. Calculate creativity index by the following formula. CREATIVITY INDEX = 10*(5*categories_ratio+wiki_fail_ratio+10*places_ratio)-(eng_word_ratio+word_recurrance_ratio+categories_reoccurance_ratio)

RESULTS

We tested the algorithm on paragraphs taken from a creative novel (Around the world in 80 days) and also from a non-creative document (software license) and in both cases the algorithm worked correctly where it gave a relatively much higher value to the creative document compared to the non-creative document.

Here were present snippets from two of the test cases and the scores assigned by the algorithm in both these cases.

(Snippet from the software license)

“

1.6 "Modifications" mean any addition to, deletion from, and/or change to, the substance and/or structure of the Original Code, any previous Modifications, the combination of Original Code and any previous Modifications, and/or any respective portions thereof. When code is released as a series of files, a

Modification is: (a) any addition to or deletion from the contents of a file containing Covered Code; and/or (b) any new file or other representation of computer program statements that contains any part of Covered Code.....

“

The score assigned in this case was **-0.38** thus branding this document as not creative.

(Snippet from the novel Around the World in 80 Days)

“

Meanwhile Mr. Fogg, after leaving the consulate, repaired to the quay, gave some orders to Passepartout, went off to the Mongolia in a boat, and descended to his cabin. He took up his note-book, which contained the following memoranda:

These dates were inscribed in an itinerary divided into columns, indicating the month, the day of the month, and the day for the stipulated and actual arrivals at each principal point Paris, Brindisi, Suez, Bombay, Calcutta, Singapore, Hong Kong, Yokohama, San Francisco,

“

The score assigned in this case was 3.4 thus branding this document as creative.

FUTURE WORK

The following works can make the algorithm better: Releasing the source code for public so other can use and give feedback on the heuristic so that it can be improved. Building a wikipedia category tree which can used to find relative distances in cateogory hierarchy between categories referred to in the document.

REFERENCES

1. Xiaojin Zhu, Zhiting Xu and Tushar Khot**, "**How Creative is Your Writing? A Linguistic Creativity Measure from Computer Science and Cognitive Psychology Perspectives"