

A Quantitative Approach of Text Analysis for Team Formation

¹Akshay Chaturvedi

²Himanshi Khetal

³Manish Yadav

⁴Archit Saxena

⁵Mr. Amit Sinha,

^{1,2,3,4}B.Tech Scholar, ⁵Associate Professor,

Department of Information Technology

ABES Engineering College, Ghaziabad, India

Abstract: The text written by any person can reflect many outcomes. One of the important aspect of Text analysis is to identify the person with the similar knowledge on the same topic. This paper gives a method to form a team for the specific project. The incumbents are given a topic and asks them to write something. The texts collected are then passing through a defined data sets and calculated total points secured by each incumbent. The person obtained higher marks may be selected in the team.

I. INTRODUCTION

By analyzing the texts written by a person, we can predict whether the persons are like minded or not. This work aims at effectively using text analysis for the formation of teams in organisations. The text will be matched using a set of commonly used words previously stored in the database. Frequency matching will be the basis of analysis and a threshold frequency of the words will be set. Team formation using this technique will increase the productivity of an organization. The process of team formation in an organization can be more efficient with this technique such that positive workers are in one team. This process may decrease, up to some extent, the conflicts between team members as it would aim at keeping people with same thinking in one team. A good team always takes an organization to another level so making such a team would be beneficial for an organization. In advance we could predict the turnover of an organization with the help of positive team members.

II. OBJECTIVE

The proposed work has the following objectives:

- i. The plan is to prepare a set of data and define the unit of analysis.
- ii. After defining unit of analysis categorize the text.
- iii. Then a coding scheme will be developed to categorize the text and code for all text will be developed.

iv. Finally conclusions have to be drawn from the coded data and report the methods and findings.

III. METHODOLOGY

The work is performed in the following steps

- i. Collection of texts from different candidates on specific topic.
- ii. Design of standard words for three different categories.
- iii. Finding each word matching with standard words and calculation of its frequency.
- i. *Collection of texts from different candidates on specific topic*

A text will be selected to be used as the standard data for the classification. The candidates will be asked to submit their response on a topic into an automated system developed to input text by uploading a file (.txt, .docx).

The uploaded text is saved in the database and is analysed for the presence of previously defined words.

- ii. *Design of standard words for three different categories*

A set of words from the standard text will be identified on the basis of their occurrence in the text. The standard words will be categorized on the basis of their universal frequency of usage on a given topic.

There will be mainly 3 categories namely Category-1, Category-2 & Category-3. The words in each category will be assigned points and this point system is as follows:

Category-1: 1 point/word (Beginner)

Category-2: 2 points/word (Intermediate)

Category-3: 3 points/word (Expert)

- iii. *Finding each word matching with standard words and calculation of its frequency*

The numbers of standard words found in the text entered by the candidates are counted separately according to the categories in which they fall. The counted words are awarded points according to the categories i.e. Category-1, Category-2 or Category-3. The points awarded to a text and subsequently the user who submitted the text are calculated and the database is updated. On the basis of the total

awarded points and the number of vacancies, the top required candidates are selected for the post for which this test is conducted. The top 5 responses showing the standard word frequency closest to the threshold frequency will be selected as the basis for team formation.

IV. IMPLEMENTATION & RESULT

The work is implemented through a sample data. Here, five incumbents are invited and they have been given a topic “Web Design” and ask them to write something.

i. *Collection of the text*

The work can be implemented through some sample text on the same topic “Web Design” collected from different incumbents applied for the same post. These texts are collected, shown in figure-1;

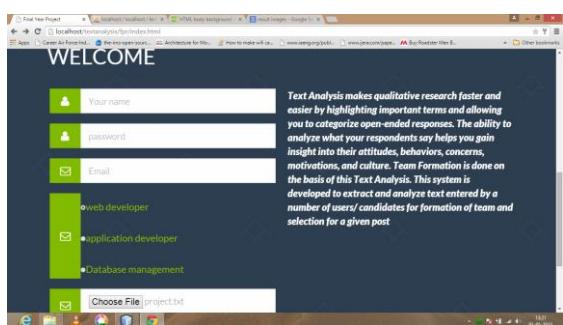


Figure 1: Input the details of the author

Incumbent-1: “Web design encompasses many different skills and disciplines in the production and maintenance of web. Web design books in a store interface web graphic design html css layout font Although web design has a fairly recent history, it can be linked to other areas such as graphic design raster vector responsive raster “

Incumbent-2: “Web design books in a store 1988—2001. Although web design has a fairly recent history, it can be linked to other areas such as graphic design. However web design can also be seen from a technological standpoint. It has become a large part of people’s everyday lives. It is hard to imagine the Internet without animated graphics, different styles of typography, background and music “

Incumbent-3: “Web designers use a variety of different tools depending on what part of the production process they are involved in. These tools are updated over time by newer standards and software but the principles behind them remain the same. Web graphic designers use vector and raster graphics packages to create web-formatted imagery or design prototypes. Technologies used to create websites include standardised mark-up, which can be hand-coded or generated by WYSIWYG editing software. “

Incumbent-4: “A raster image is made of up pixels, each a different color, arranged to display an image. A vector image is made up of paths; each with a mathematical formula (vector) that tells the path how it is shaped and what color it is bordered with or filled by. The W3C has released new standards of HTML (HTML5) and CSS (CSS3), as well as new JavaScript API's, each as a new but individual standard.

Incumbent-5: “Web designers use a variety of different tools depending on what part of the production process they are involved in. These tools are updated over time by newer standards and software but the principles behind them remain the same. Web graphic designers use vector and raster graphics packages to create web-formatted imagery or design prototypes interface web graphic design html css layout font “

(i) The standard words for “Web Design”

The development of data sets containing standard words is decided on the basis several Literature survey and going through hundreds of written text on the same topic. On sample data set for “Web Design” is divided among different categories such as:

Category-1 (giving 1-point for each word) =
{ Web, Design, Graphic, HTML, CSS }

Category-2 (giving 2-points for each word) = {Interface, Layout, Frontend, Font, Pixels}

Category-3 {giving 3-points for each word) = {Raster, Vector, Responsive}

These are shown in figure 2:



Figure 2: Extraction of words

(ii) *Finding each word matching with standard words and calculation of its frequency*

On analysing all the text collected above and passing through the same data sets, we observed the following.

Table-I: Quantitative calculation of texts

Standard Words	Text-1	Total Points gain	Text-2	Total Points Gain	Text-3	Total Points gain	Text-4	Total Points gain	Text-5	Total Points gain
Web	5	5	3	3	3	3	0	0	4	4
Design	5	5	4	4	1	1	0	0	2	2
Graphic	2	2	1	1	1	1	0	0	2	2
Html	2	2	0	0	0	0	2	2	1	1
CSS	1	1	0	0	0	0	2	2	1	1
Interface	1	2	0	0	0	0	0	0	1	2
Frontend	0	0	0	0	0	0	0	0	0	0
Font	1	2	0	0	0	0	0	0	1	2
Pixels	0	0	0	0	0	0	1	2	0	0
Layout	1	2	0	0	0	0	0	0	1	2
Raster	2	6	0	0	1	3	1	3	1	3
Vector	1	3	0	0	1	3	2	6	1	3
Responsive	1	3	0	0	0	0	0	0	0	0
TOTAL POINTS		33		8		11		15		22

The above analysis shows that the Incumbent-I may be considered for selection.

V. CONCLUSION

The text analysis may be used for various purposes. One of the important applications is to form a team of persons for any specific project. The incumbents may be given a topic that is useful and mandatory for the project and asks them to write latest development over that topic. The texts collected will then be passed through a set of predefined standard data sets and calculate their frequency of occurring. The standard data sets are developed keeping in mind that it includes all the conceptual and latest technological words. According to the frequency calculated, a best person for a particular project will be selected. This work can also be an aid to final selection list.

REFERENCES

- I. Haytham Mohtasseb and Amr Ahmed, "More Blogging Feature for Author Identification", in ACM 2007.
- II. Amit Sinha and Ashok K. Sinha, " An Improved Human Trait Modeling using Fuzzy Inference System", in IJERT Vol. 1-Issue 6 pg.1-6.
- III. J. Oberlander and S. Nowson, " Whose thumb is it anyway? Classifying author personality from weblog text", in Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics, Sydney, Australia, 2006.
- IV. Scott Nowson and Jon Oberlander, "Identifying More Bloggers", in ICWSM 2007 USA.
- V. J.M. Dewaele and A. Furnham, "Extraversion: The unloved variable in applied linguistic research", Language Learning, 49:509–544, 1999.
- VI. S. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker, "Lexical predictors of personality type", in Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America, 2005.
- VII. J. W. Pennebaker and L. King, "Linguistic styles: Language use as an individual difference", Journal of Personality and Social Psychology, 77:1296–1312, 1999.
- VIII. K. Scherer, "Personality markers in speech", in K. R. Scherer and H. Giles, editors, Social Markers in Speech, pages 147–209. Cambridge University Press, Cambridge, 1979.
- IX. Yla R. Tausczik and James W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods" published in "Language Style Matching Predicts Relationship Initiation and Stability" Psychological Science January 1, 2011 22: 39-44.