

# Superlative model using word cloud for short answers evaluation in eLearning

Shailaja Jayashankar<sup>1</sup>  · R. Sridaran<sup>2</sup>

Published online: 25 October 2016

© Springer Science+Business Media New York 2016

**Abstract** Teachers are thrown open to abundance of free text answers which are very daunting to read and evaluate. Automatic assessments of open ended answers have been attempted in the past but none guarantees 100 % accuracy. In order to deal with the overload involved in this manual evaluation, a new tool becomes necessary. The unique superlative model discussed in this paper aims at providing improved accuracy by constructing word clouds. The model uses appropriate semantics with a visual appeal to partially automate free text evaluation. The model was applied at a K-12 school setup where the average human agreement rate was found to be 98 % and the accuracy score deviation from the mean was 2.82. This tool can be cast-off at any level starting from K-12 to higher education to evolve the way we view and evaluate answers.

**Keywords** Visual analytics · Word cloud · Relative cloud · Cohesion cloud · Automatic assessment

## 1 Introduction

Word clouds or tag clouds are a visual portrayal of words for any written material built on its frequency. Architect of Wordle (Feinberg 2010) declared people who create a word cloud take pride in creating something meaningful as it reflects and

---

✉ Shailaja Jayashankar  
shaila\_jm@rediff.com

R. Sridaran  
sridaran.rajagopal@gmail.com

<sup>1</sup> Faculty of Computer Science, National Academy for Learning, Bangalore, Karnataka, India

<sup>2</sup> Faculty of Computer Applications, Marwadi University, Rajkot, Gujarat, India

deepens the source text. Teachers are striving to include Bloom's higher order thinking by integrating new technologies in teaching and learning (Bloom 1956). But most of the technology is tough to learn and time consuming to formulate (Blin and Munro 2008). Word cloud is easy to comprehend and quick to embrace, it can be included into any class, subject and age. Word cloud is all about data mining and textual analysis still it does not need any of this to make one. Although traces of use of this powerful tool in education is scarce, when aptly used word cloud analysis offers a lot of benefits in e-learning. Limited works available are enumerated below.

### 1.1 Word cloud in education

Technology is swiftly progressing in our teaching and learning. Word cloud retains its plainness in use but sophistication of data analysis has made it powerful. Ramsden and Bate (2008) introduce the potential of using word clouds within teaching and learning and compared different word cloud software. To incorporate new teaching strategies (Huisman et al. 2011; Nickell 2012) have experimented with word cloud to promote engagement in online class settings. Perry (2012) has provided pragmatic confirmation for the same. (Baralt et al. 2011) converted students writing into word clouds which were hands on experience through discussion amongst peers they improved their vocabulary and grammatical tenses. Ahearn (2013) used word clouds in research and reflection, it can be effectually used in brainstorming sessions declares McNaught and Lam (2010).

A study by Miley and Read (2012) used word cloud as a tool to teach the four categories of students as per Kolb's learning theory. Kolb (2005) named the learners as accommodating, assimilating, converging or diverging as per their learning preferences. Whatever learning styles students fall into they all enjoyed it as it was engaging which points towards deep learning. Word cloud promotes active learning hence engaging, enjoyable and motivating. Engaging activities triggers prior knowledge and nurtures independent thinking. Jablon and Wilkinson (2006); Tsui's (2002) research has found that engagement is directly associated to critical and computational thinking (Wing 2008). Developing critical thinking (deNoyelles and Reyes-Foster 2015) skills at any education level still remains as an open challenge to educators. To support the above statement research by Kaptein et al. (2010); Behar-Horenstein and Niu 2011) established that visual representations of a document's text in the form of a word cloud nurtured both critical thinking and engagement within the framework of online discussions. Concetta et al. (2014) identified word cloud could be a formative assessment tool and provide feedback to students; they used it in its original form and left it to researchers to discover if it was an effective tool. Students and evaluators approved the visual appeal of the word cloud but not the quantitative approach for evaluation (DePaolo and Wilkinson 2014).

### 1.2 Visual analysis

Despite its beautiful visual representation and simplicity, the use of word cloud in education is very rare. That is mainly because prominence is more on quantity

rather than quality with respect to word cloud. A qualitative analysis of data is more important than just frequency count for it to penetrate into education as text and its processing is the crux. Natural Language Processing (NLP) is considered as a hard problem for computation (Collobert et al. 2011). Several machine learning approaches such as classifiers, latent semantic analysis and decision tree have been employed, thanks to the upsurge in processing power it has become a reality. Automatic evaluation of answer scripts have been attempted by several researchers but test result show only 80 to 85 % success rates in comparison to human evaluation. Answer script evaluation turns out to be a supervised Machine Learning (ML) classification problem (Maglogiannis 2007). This in turn needs a large training set, in reality questions and answers are not fixed so nearly impossible to train. Without which the model will create a lot of mistakes when pigeon-holing new documents.

Data mining is a research field which emerged from ML. Difference being ML concentrates on prediction based findings from the training data set whereas data mining emphasizes on the sighting of unknown assets in the data. To comprehend the features and developments in a large dataset, visualization can be a significant tool. Scientific and information visualization led to visual analytics that include knowledge management, statistical analysis, cognitive science and decision science as per Cook et al. (2007). Cook and Thomas (2005) defined “Visual Analytics (VA) as the skill for analytical reasoning which promotes decision making”. When data is enormous neither computational nor human power alone is not enough, VA comes into picture. The data involved in assessments is also enormous. Tags are becoming popular in social media to evaluate tweets and feeds (Bielenberg 2005); likewise keywords were always popular in answers. Keyword extraction is the main focus in this study to produce a visual cloud of keywords in the answer as a partial assessment tool. Instead of preprocessing data like the ML approach the tool employs post processing of the visual cloud. Open ended answers are unsuited for machine marking because of the struggle of coping with the innumerable possibilities in which a credit-worthy answer may be conveyed (Sukkarieh et al. 2003). When student’s marks/grades are concerned human discretion is a must to award the precise marks/credit which a machine cannot substitute.

### 1.3 Problems faced by evaluators

Technology has reached classrooms and overcrowded at times too with several Learning Management Systems (LMS), personalization, blended learning and many more on offer. Students can find these overwhelming at times, not all are tech savvy (Jimoyiannis and Komis 2007). With differences in user interface it is difficult to navigate from one to the other. Technology has been incorporated into education to reduce the burden both on students and teachers, but we see the reverse trending. Faculty roles have changed from teachers to facilitators adding on to more work. However technology has bought in changes in teaching and learning, very little contribution to reduce teachers’ job of corrections which is a nightmare for most (Ingoley and Bakal 2012). Short answer evaluation (Burrows et al. 2015) approach has been attempted but the success rate is very meek as shown in Table 1.

**Table 1** Most popular method-based automatic short answering system with their performance

Serial no	Automatic Short answer evaluation tool	Success rate in percentage
1	Intelligent Essay Assessor	85 %
2	E-Rater	87 %
3	C-Rater	80 %
4	Bayesian Essay Test Scoring sYstem	80 %
5	Automark	93 %
6	e-Examiner	85 %
7	IndusMarker	92 %

Most of the e-learning tools have restored to Multiple Choice Questions (MCQ) as it is quick and does not need user intervention. This leads to deterioration of writing skills amongst students which is very essential. Hence at higher levels essay type answers are included but for corrections or evaluation they have to rely on peer reviews. This is not consistent enough; to attain these hiring evaluators is not a feasible solution. Not only is it time consuming, it can vary between a day to a week to get appropriate results and feedback. Short/essay type answering comes with a lot of advantages like students have a rich understanding, material-broad range of information upshots can be tested. Good for gauging ability to organize, integrate and express ideas against MCQs. Our word cloud based tool can be included into any LMS to attain easy and quick evaluation of short answers.

In this paper, a novel model for short answer evaluation using word cloud technique is employed. The model uses relative and cohesion cloud for evaluation. Section 2 provides an outline on the existing short answer evaluation approaches rummage-sale. Superlative model is elucidated in section 3. In Section 4, experimental results are accessible to certify the approach, before performance analysis comparison and conclusion.

## 2 Related work & background

Automatic short answer evaluation has been an active research since 1966. Many tools using different methodology and strategies have been developed over the years so Burrows et al. (2015) classified them into five eras namely: Concept mapping, Information extraction, Corpus-based methods, Machine learning, and Evaluation.

Concept mapping involves breaking the answer into concepts and checks the attendance or deficiency in the answer. Burstein et al. (1999) applied this at multiple reasoning levels. The c-rater (Leacock and Chodorow 2003) tried to map as many sentence level concepts with the teachers answer using four major concepts: model building, linguistic processing, recognizing the main points and scoring.

Information extraction practices pattern matching which was used in Automark (Mitchell et al. 2002) they castoff a large training set using bootstrapping. Three diverse machine learning approaches: Inductive Logic programming, decision tree learning and Naive Bayesian learning were used by Pulman and Sukkarieh (2005) for marking short free text responses automatically. They upgraded to Auto-Marking 2: An update on the UCLES-Oxford University research into using computational linguistics to score short, free text responses. CoSeC-DE (Comparing Semantics in Context) (Hahn and Meurers 2012) uses the lexical resource semantics by generating a graph to show equivalent meanings. Makers of IndusMarker, Siddiqi et al. (2010) used structure matching. It could evaluate only factual answers which have a clear discretion between yes and no. Unlike MCQ's the decision is not yes/no, automatic short answer assessment needs to compare with a correct answer which is a difficult task.

Corpus-based methods were applied on large text using statistical measures of text similarity. Atenea (Pérez et al. 2005) used BiLingual Evaluation and combined latent semantic analysis, Willow was its descendant. Mohler and Mihalcea (2009) used unsupervised method to find semantic similarities in text. Measurements from NLP gave rise to machine learning era, e-Examiner (Gütl 2007) used linear regression whereas k-nearest neighbour classifier was used in CAM (Content Assessment Module) (Bailey and Meurers 2008). However machine learning techniques require a large dataset to assess. This is not always available as most of the short assessment tools are developed by academicians who have used their own teaching dataset (Feinerer et al. 2016). To compare and for appraisal the era of evaluation a method-free competition emerged, hosting by Kaggle<sup>1</sup> company where data scientist all over the world compete for a prize money. Dzikovska et al. (2016) performed a joint student response analysis by bringing educational NLP and computational semantics researchers together and checked for completeness and correctness of the response. They concluded that additional research is required in the field.

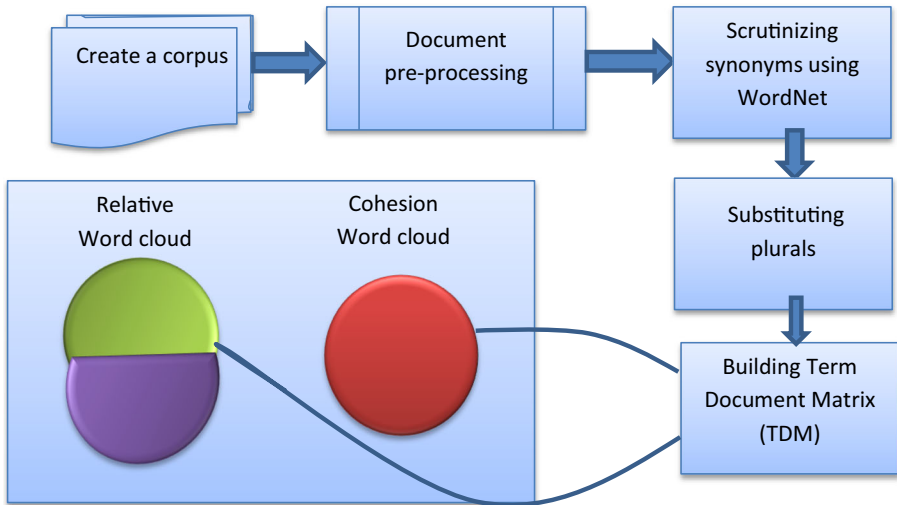
Our work is one of its kinds which integrate data mining, visualization/informatics and word cloud for short answer assessment. Short answer and essay type evaluation is perplexing especially if it comprises open ended answers. For more than 50 years now researchers have worked on automatic evaluation of answer scripts. The ongoing question concerns the quality of scores compared to human evaluation, faith and effectiveness in the process says Williamson et al. (2012); Butcher and Jordan (2010). In summary, we certain that our word cloud mapping approaches as a special case that benefits from crunching of data with the precision of humans that can be obtained as opposed to holistic marking. So there is benefit in commissioning this type of process.

### 3 Superlative model

The employed ideal is innovative and is called the ‘Superlative Model’ (SM) for qualitative evaluation of answer scripts. In this research, we present a first-hand model to generate two types of word cloud called the cohesion and relative cloud. The broad-spectrum architecture is undertaken in Fig. 1 for the above SM to be possible a step wise prologue follows.

---

<sup>1</sup> <http://www.kaggle.com>



**Fig. 1** Overall architecture of the superlative model

### 3.1 Create a corpus

The input comprises of student response and answer key, they are in free text form which indeed is challenging to analyze, except one tells what to look for. Hence an answer key is aided for comparison. Patterns between the two have to be identified and sub-categorized. Simplest way is to guess the category which is not always right. So it is a typical text mining problem. Corpus can be created from any text file, PDF or a directory itself (Ingo and Kurt 2015). To create the corpus, text should be put in data frames. Answers are unstructured text data that needs to be transformed into structured data for it to be useful. Algorithm 1 is used to deconstruct words using spaces to prepare the data for text analysis.

```

Algorithm 1. To structure the data
> my.array <- dim ( ) //To hold the words an array is initialized
>while(fgetc(my.array[i])!=' ')//Read characters till the first space is searched
>myword[i]=currChar//The first word is stored till the position before the space
Repeat the step until all words are parsed.
  
```

### 3.2 Document pre-processing

Data cleansing is the most tiresome step, but utmost vital for text analysis (Liu and Curran 2006). The cleaning process involves removal of numbers, punctuation, whitespace and unnecessary words (stop words) to all elements of the corpus. Words which give only structure to the sentence like articles (a, an, the), conjunctions (and, or) are comprised as a Standard English list of stop words which can be removed. Words from the question can also be included into the stop word list.

Normally the next step in text mining is stemming where it chops the ends of words so that all forms of words are put together to increase the frequency of that word. For

example - words like sensing, sensors, sensed is reduced to sens and not sense. Problem with stemming algorithm is it axes the word to its root instead of the base. Stemming algorithm has not been used as it may cause problems and every word is important in an answer.

### 3.3 Scrutinizing synonyms using WordNet

Synonyms handled in a word cloud has never been attempted as per my survey. WordNet (Miller 1995; Feinerer et al. 2016) which is a lexical database of English is used for comparing similar words. The lemma of the word is determined, only noun is considered for comparison. Once installed the synonyms finding can happen offline, provided the dictionary environment is set to the current working directory. In text mining the “cosine” measure is used to search the similarity statistic (Hahsler et al. 2005). Its formula is:

$$\cos(\theta) = \frac{A*B}{\|A\|*\|B\|} \quad (1)$$

Where A and B are word frequencies.

### 3.4 Substituting plurals

In text mining the word and its plural is also considered as a separate word (Sag et al. 2002). This results in repetition of the same word in the relative cloud in multiple tenses. The word does not figure in the cohesion cloud due to the non-similarity which in turn may lead to miscalculation of marks.

### 3.5 Building Term Document Matrix (TDM)

The bag of words (Kao and Poteet 2007) is now ready after all the cleaning to be converted into a matrix. A matrix that lists all incidences of words are generated where in the terms (or words) are represented as rows and documents by columns. Hashing technique is used as it required less looping.

Algorithm 2. Explains the process.

**Algorithm 2.** To generate TDM  
 for (i = 1; i < 6; i++) // Outer loop is set for each record.  
   for (j=i; j <= 3; j++) // Inner loop goes through each word for each record.  
   If word\_in\_record= word\_in\_document  
   word\_count=word\_count+1  
 Loop until all words are compared in both the records.

### 3.6 Cohesion and comparative word cloud

The output of this model generates two word clouds, a cohesion and comparative cloud. Association rules are used to find the correlations between the two answers

(Leek et al. 2013). ‘A’ association ‘B’ are two disjoint sets, the three measures used for choosing intersecting rules are support, confidence and lift are as given in Eqs. 2, 3 and 4 respectively.

The equations (Brin et al. 1997) to calculate them are:

$$\text{support}(A \Rightarrow B) = P(A \cup B) \quad (2)$$

$$\begin{aligned} \text{Confidence } (A \Rightarrow B) &= P(B|A) \\ &= \frac{P(A \cup B)}{P(A)} \end{aligned} \quad (3)$$

$$\begin{aligned} \text{lift}(A \Rightarrow B) &= \frac{\text{confidence } (A \Rightarrow B)}{P(B)} \\ &= \frac{P(A \cup B)}{P(A)P(B)} \end{aligned} \quad (4)$$

Where  $P(A)$  is the probability in percentage of cases containing ‘A’. In the model two individual documents ‘A’ and ‘B’ (where A is Student Answer and B is Answer Key) are compared. Words that are common between the two documents form the cohesion cloud. The uncommon words form the relative cloud; common logic is as in Algorithm 3

**Algorithm 3.** To generate word clouds.

```
If A_W = weight_of_word_A   And   If B_W = weight_of_word_B
Then  word=cohesion_cloud(min(A_W, B_W))
Else  word=relative cloud(upper_half or lower_half)
```

Comprehensive algorithm to generate two separate word cloud; cohesion and relative clouds for the model is specified in Algorithm 4.

Let’s look at the two clouds in detail when generated it can be stored as a .png image for future reference.

### 3.6.1 Relative cloud

We consider two documents the top one is the student answer and bottom part being the words from the answer key as in Fig. 2. Different colors have been used for clear demarcation. In case of any doubts in marking and providing feedback this comes in handy.



**Algorithm 4.** Superlative algorithm

```

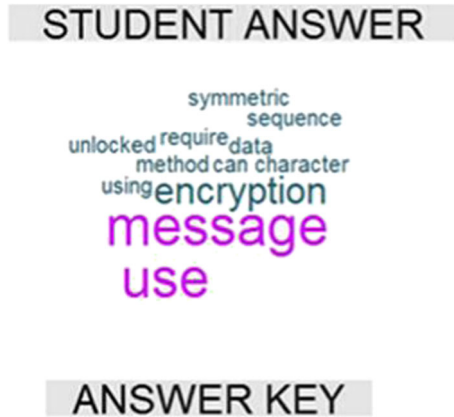
>tmpText1 = data.frame(c("student answer", "answer key"))
>ds <- DataframeSource(tmpText1)
>corp = Corpus(ds)
//Text cleaning- tm_map is an interface to apply transformation.
//The cleaning process involves removal of
Numbers ->corp = tm_map(corp,removeNumbers),
Punctuation ->corp = tm_map(corp,removePunctuation)
Whitespace -> corp <- tm_map(corp, stripWhitespace)
Stopword -> corp = tm_map(corp, function(x){removeWords(x,stopwords())})
//After all the cleaning a part of the corpus
>writeLines(as.character(corp[[1]]))
phishing creator sends legitimate looking email hope gathering personal
//Matrix is created
>str2<-matrix(as.character(corp[[1]]))
//Details of corp[1] i.e. student answer
corp[[1]]
<<PlainTextDocument>>
Metadata: 7
Content: chars: 187
//Plurals are managed
>str2 <- gsub("s\\b", "", str2)
//Managing synonyms using WordNet
str4.syn<-synonyms(str3.split[i], "NOUN")
//Term Document Matrix is computed
>term.matrix <- TermDocumentMatrix (corp)
>term.matrix
//Inspecting the term.matrix
<<TermDocumentMatrix (terms: 28, documents: 2)>>
Non-/sparse entries: 39/17
Sparsity : 30%
Maximal term length: 17
Weighting : term frequency (tf)
>term.matrix <- as.matrix(term.matrix)
//Creating the relative cloud
>comparison.cloud(term.matrix,max.words=300,random.order=FALSE,colors=c("#1F497D",
"#C0504D"), rot.per=0, main="Differences Between Answers")
//Creating the cohesion cloud
commonality.cloud(term.matrix,random.order=FALSE, rot.per=0, color="#F79646")

```

**Minutiae-** Let  $p_{i,j}$  be the rate at which word  $i$  occurs in document  $j$ , and  $p_j$  be the average athwart documents ( $\sum_i p_{i,j}/ndocs$ ). The size of each word is mapped to its maximum deviation ( $\max_i (p_{i,j} - p_j)$ ), and its angular position is determined by the document where that maximum occurs.

### 3.6.2 Cohesion cloud

Algorithm 6. plots a cloud of words common across documents, in this case two documents students answer and the answer key. The number of words will be directly



**Fig. 2** Relative cloud constructed by the superlative model

proportional to the marks obtained as shown in Fig. 3. For the question – “State what is meant by symmetric encryption” is considered to generate Figs. 2 and 3.

#### **4 Superlative model for qualitative evaluation of answer scripts– a case study**

In this section, the SM is described and its operation demonstrated through an illustrative example. Word cloud which is visualization and an infographic tool is used in this work. A relative and cohesion word cloud had been developed, where answers are crushed into keywords, for an easier and faster evaluation. Our work does not challenge nor try to replace the evaluators, as human intervention is required for effective evaluation when student’s marks are concerned. Our effort is to reduce the burden of going through the answer word by word and fishing out keywords to allot marks. The thankless job of not just one paper but going through each and every answer script is abridged. The tedious, repetitive job of reading data in the text form is made interesting as a colorful, visually appealing word cloud using the step below:



**Fig. 3** Cohesion cloud

#### 4.1 Word cloud generation

Word cloud employs a crude way of text analysis in comparison to Natural Language processing. Cohesive and relative cloud uses the same technique for drawing pictures of words

The steps for word cloud generation are:

1. Initially, the submitted document is in sentences which are broken into words
2. A numeric weight is assigned to each word according to its frequency. The formula is straight forward:  $\text{weight} = \text{word count}$ .
3. Proportional to its weight the font size is allocated considering magnitude of various constants.
4. Total area to be enclosed by the word cloud is estimated, centered on the factors: vaulting boxes for each word, summing and adjusting the areas for small and large words.
5. Words with higher frequency/weights are placed closer to the center, all in distinct rectangular boxes.
6. To station the words in the given space, randomized greedy algorithm is used. After its placement the word does not move.
7. Ever-increasing spiral is the easiest of all where in a rectangle is placed at the next best position without collision.
8. Efficient collision detection is an entire research field and a major problem in word cloud as words are converted into images and non-intersection of words has to be accomplished. Many algorithms like hierarchical bounding boxes, last-hit caching, and a quadtree spatial index are used which is beyond the scope of this paper due to its length.

#### 4.2 Datasets description

The purpose of this study is to use the Superlative Model for easy assessment of answers scripts and provide quick feedback to prepare the students for the board examination. The test was conducted at National Academy For Learning (NAFL)<sup>2</sup> a K-12 setup in Bangalore, India. NAFL is an International school, following the International General Certificate of Secondary Education (IGCSE) curriculum and a recognized Cambridge International Examination (CIE)<sup>3</sup> center.

All question and answers were chosen from the IGCSE syllabi for grade X. Students who took part in the study were grade X student (aged 14–16) taking their IGCSE board examination in March or May/June 2016. These questions were part of their continuous assessment/preparatory examination administered in Jan/Feb 2016 to prepare them well for their upcoming boards. Evaluators were subject experts teaching the curriculum for a minimum of 5 years at NAFL.

<sup>2</sup> Official Website of National Academy For Learning, India - <http://www.nafl.in/>

<sup>3</sup> Official Website of Cambridge International Examinations UK, <http://www.cie.org.uk/programmes-and-qualifications/cambridge-secondary-2/cambridge-igcse/>

Teachers at NAFL have undergone subject specific training form CIE (either face-to-face or online) to teach and evaluate the IGCSE curriculum.

English is the medium of instruction at NAFL and the number of students taking the exam in March/June 2016 was 13. CIE offers over 70 subjects in any combination, at NAFL over 15 are on offer. SM assessments were applied on 8 subjects excluding languages like Hindi and French, Mathematics and its variants. Thus, the data set we worked with consists of a total of 13 student taking different subject combination (61 students x 69 answers/question = 4209). Human evaluation is not always consistent, hence two subject experts marked the papers and then the same were evaluated using cohesion cloud. To attain an average percentage (rounded off to one decimal place) the same assessors evaluated the subject papers using traditional correction methods and using the cohesion cloud method, Table 2 provides the details. A close look at the values reveals that both forms of evaluation have very close agreement rate.

The model has been evaluated on different subject datasets. To gauge its capability to help find commonalities and differences between multiple subjects, it has been tested on diverse source data sets. Most of the short answer evaluation approach has been tested on one particular subject/topic. Table 3 shows a list of sample questions from IGCSE curriculum used in different school subjects along with its mark weightage.

### 4.3 Experimental results

To demonstrate the working of the model question considered is—“Explain what is meant by Phishing.” is deliberated in algorithm 4. The model creates a cohesion cloud by comparing the student answer versus the answer key. The evaluator need not do any data mining on the student answer nor should the answer key be in keyword format. Natural language can be used; the tool even takes care of synonyms and plurals so that the student does not lose out. The evaluator need to simply count the number of words in the cohesion cloud and he/she arrives at the marks. Table 4. Comprises of the question, reference answer and two students answer evaluated by the subject expert and the cohesion cloud for the question “Explain what is meant by Phishing.”

### 4.4 Performance analysis

The evaluators regarded cohesive and relative cloud as a spontaneous and useful evaluation system. They indicated how tedious and methodical evaluation is brought to life using this tool. When asked about ambiguity and discrepancy in the answers they found it very clear to analyse the cohesive cloud. When in need the relative cloud was always there for rescue which could be visually compared keeping in mind the difference in colour.

Students found the relative cloud as an excellent means of feedback, customised for their answers. Students who used these clouds performed much better as they could visually see what was missing in their answers versus the key. This helped the students better their answers and tackle the CIE board exam with ease and performed exceptionally well.

**Table 2** Test data considered to evaluate the model

Subject	No of students	No of question	Average human evaluation	Average human evaluation using cohesion cloud
Computer Science	10	40	100 %	99.7 %
Physics	8	4	99.7 %	99.2 %
Biology	8	5	99.6 %	97.8 %
Business Studies	5	5	99.3 %	98.3 %
Economics	4	5	98.8 %	98.2 %
Geography	6	3	98.5 %	97.8 %
History	7	3	98.4 %	97.6 %
English	13	4	97.9 %	96.4 %
		Average	99.0 %	98.1 %

## 5 Comparison of superlative model with IndusMarker


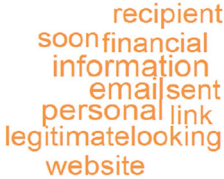

Major impact generated assessments tools have been discussed in section 2. In this section we closely examine IndusMarker (IM) vs Superlative Model (SM) as it is a recent development compared to the other models and accommodates different types of questions. IM uses structure-matching and SM practices text mining and word cloud approach. “Text mining could have been used for automated marking in IM but it was not selected because text mining is computationally quite

**Table 3** Sample set of questions for experiment

Subject	No	Question	Marks
Computer Science	1	Give one benefit of writing code in a high-level language.	1
	2	Describe two benefits of using USB connections between a computer and a device.	2
English	1	Using your own words, give two effects that the singing of the national anthem had on the writer.	2
	2	How did the piano player slightly spoil the musical effect of the national anthem?	2
Biology	1	Suggest why the chimpanzee population has decreased.	1
	2	State two methods, other than captive breeding of conserving endangered species.	2
Economics	1	Describe what is meant by a country’s inflation rate.	2
	2	Explain the factors that influence a country’s birth rate.	3
History	1	How significant was the failure of the Schlieffen Plan to the course of the First World War?	4
	2	How important were events at sea to the outcome of the First World War?	5

Source: Questions from (CIE) International General Certificate of Secondary Education (IGCSE) specimen paper. <http://www.cie.org.uk/programmes-and-qualifications/cambridge-secondary-2/cambridge-igcse/>

**Table 4** Algorithm result for the question

Type	Answer	Marks gained	Cohesive Word cloud
Answer key	Legitimate-looking emails sent to a user as soon as recipient opens clicks on link in the email attachment ... the user is directed to a fake website (without their knowledge). To obtain personal financial information data.	2	 <p>A word cloud containing terms like: without, recipient, legitimate, looking, website, information, personal, emails, directed, data, fake, USER, link, attachment, email, financial, knowledge, sent, opens, clicks.</p>
1 <sup>st</sup> candidate's answer	Phishing is where the creator sends out a legitimate-looking email in the hope of gathering personal and financial information from the recipient. As soon as the recipient clicks on the link in the email they are sent to a bogus website where they will be asked for personal information.	2	 <p>A word cloud containing terms like: recipient, soon, financial, information, email, sent, personal, link, legitimate, looking, website.</p>
2 <sup>nd</sup> candidate's answer	Emails sent to users from a criminal who is acting as an important organization. The criminal will ask for important and secure data from the customers by posing as the organization.	1	 <p>A word cloud containing terms like: emails, data, sent.</p>

Source: Answer key from Cambridge International Examination (CIE) International General Certificate of Secondary Education (IGCSE) Computer Science specimen paper. <http://www.cie.org.uk/images/166855-2015-paper-1-specimen-paper-markscheme.pdf>

expensive, it requires use of many tools/technologies for various stages of the text mining process and also requires considerably large training data set” Siddiqi (2010). SM takes the text mining approach as it proved to be computationally least expensive as it uses very few tools and training data set is not at all required. The only prerequisite being an answer key which is compulsory for any question paper.

Though there are differences in the approach finding solution to ease out short answer evaluation is unchanged. Both models compare student’s answers with the answer key. IM relies on structure matching which is why a lot of pre-processing is required to structure the answers. In SM there is no such requirement; both student and answer key can be in its natural language. To evaluate SM same questions were executed as in IM and Oxford-UCLES system’s evaluation. The subject chosen was biology: answers, word length, type of questions and marks allotment were identical. Both human evaluation and model based evaluations were engaged. Empirical methods to evaluate the system are human-system agreement rates and time taken to formulate and validate are given in Table 5. To compare the performances of IM and SM same computing parameters and equations were used. The equations used and as specified in IM to computer average answer length, human-system agreement and time are represented by Eq. 5, 6 and 7 respectively. The equations and their explanation follow:

**Table 5** System performance to compare the models IM and SM

Sl. No.	Question type	IM Average answer length (words)	SM Average answer length (words)	IM Average human-system agreement rate z1	SM Average human-system agreement rate z2	IM Average time taken in minutes a1	SM Average time taken in minutes a2
1	Sentence completion	1.4	1.3	99.52 %	100 %	2.0	1.0
2	Single term generation	2.2	1.2	100 %	100 %	2.20	1.10
3	“Quantity” required	2.9	2.5	100 %	100 %	2.20	1.20
4	“Numerical value” generation	2.3	2.1	99.04 %	100 %	3.0	1.10
5	“Location” or “source” required	3.4	2.5	97.61 %	100 %	4.40	1.30
6	“Example” required	2.8	2.4	97.14 %	100 %	4.20	1.20
7	List	7.3	7.2	93.80 %	97.03 %	7.50	1.60
8	“Ordering / rearrangement” required	10.5	10.4	93.33 %	94.23 %	7.50	2.10
9	Short explanation / description	8.7	9.1	91.90 %	95.83 %	11.00	2.0
10	“Reason” or “justification” required	10.1	11.2	90.47 %	93.54 %	12.40	2.30
11	“Way of doing something” required	9.6	10.3	90.95 %	94.78 %	10.50	2.10
12	Definition	13.6	13.8	88.57 %	95.03 %	12.40	2.50
13	Compare	10.7	11.5	95.23 %	96.47 %	8.00	2.50
14	Contrast	19.6	18.5	88.09 %	92.98 %	16.40	3.40
15	Composite	15.9	16.1	85.71 %	93.67 %	14.00	3.20

The average answer length ‘y’ is calculated using summation of words ‘w<sub>i</sub>’ for the ‘i’ question, divided by the product of questions ‘q’ and answers ‘a’ per question.

$$y = \frac{\sum_{i=1}^q w_i}{q * a} \quad (5)$$

To arrive at human system agreement an answer script is evaluated by a subject expert manually and using the tool. Number of correct judgment ‘c’ divided by total number of judgments ‘t’ is equal to human-system agreement  $r = c/t * 100$ . The average human-system agreement rate ‘z’ is computed as summation of human-system agreement rate ‘r<sub>i</sub>’ for the ‘i’ question, divided by the total number of questions ‘q’.

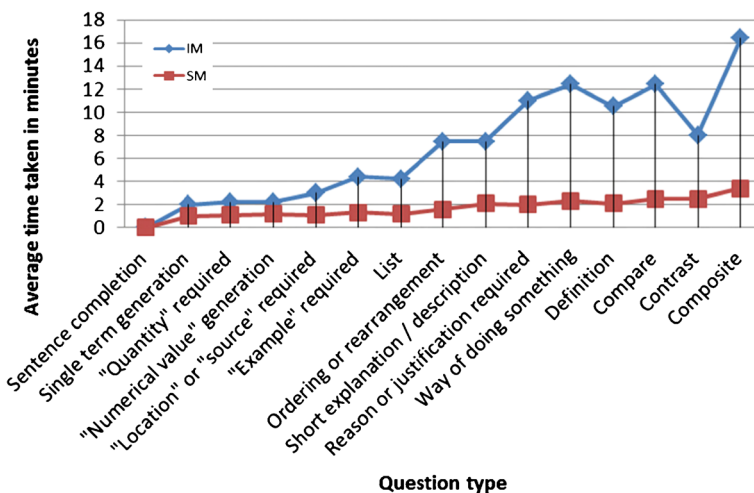
$$z = \frac{\sum_{i=1}^q r_i}{q} \tag{6}$$

On the same lines average time taken to formulate the required structures in IM for each question type is given as ‘a’ by summing ‘t<sub>i</sub>’ time taken to express the required structure for ‘i’ question, for all ‘n’ the total number of questions used for the question type.

$$a = \frac{\sum_{i=1}^n t_i}{n} \tag{7}$$

Using these equations, IM and SM is compared for the same set of parameters, questions and types. The results are tabulated in Table 5 where average human-system agreement of IM is denoted as z1 and SM is represented as z2. Average time taken by IM is indicated as a1 and SM symbolises as a2.

The efficiency of any automatic short answer analysis tool is assessed as per accuracy, cost and time taken. Machine learning techniques for evaluating short answers have produced 80 to 85 % results, the drawback being large datasets for training which is not always available in education. Both supervised and unsupervised ML techniques are an expensive affair. Cohesive and relative cloud approach being cheaper, easy and quick yields around 96 to 99 % accuracy in evaluation. R (Fellows 2013) is used in this tool as it is classic statistical software moreover a free open source software available for Windows too. R’s TM package (Meyer et al. 2008) was best suited as any other statistical analysis software’s are vendor specific and very expensive. SM proved to be time efficient as training data set is not involved which consumes time. While SM’s performance is much better than IM in terms of average time and human-system agreement as show in Table 5. A line graph comparing IM and SM average time is shown in Chart 1.



**Chart 1** Graph depicting the difference in average for different short-answer question types



As IM employed structure matching using Stanford Parser it come to be computationally affluent and expended a lot of time. “Another problem with the Stanford Parser is that there is no guarantee that it will produce an accurate parse of student’s answer text” (Leek et al. 2013) hence accuracy reduces for long answers. As seen in Table 5 both average time and accuracy is directly proportional to the average number of words. In SM the same holds good but the average time has drastically decreased and accuracy has improved, as SM practices post human analysis were as IM encompasses pre analysis. To compare two data sets effectively Standard Deviation (Kragten 1994) of time values  $a_1$ ,  $a_2$  and accuracy  $z_1$  and  $z_2$  are calculated using Eq. 8:

Sample Standard Deviation:

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (8)$$

Where  $\bar{x}$  is the sample mean and  $n$  is the sample size. Standard Deviation has been computed for accuracy and time for both the models and the results are as tabulated in Table 6.

Accuracy scores deviation from the mean is only 2.82 in SM compared to 4.7 in IM which is almost 50 % cut down. In terms of time SM takes 4 epochs less time to compute any type of question to be evaluation. Therefore SM has proved its enhancement in terms of both accuracy and time efficiency.

Central errors other than human-system agreement found in IM were of two types: false positives and misses. If an answer gets extra marks then it deserves to be reflected as false positive. It is considered a miss when the answer gets less than it justifies. In IM the misses contributed at 69 % as it is impossible to list all possibilities and false positive arrives at 31 % as stated by the author. SM eliminates misses and false positives as human evaluators assess the end product; in case of disagreement they refer the relative cloud.

## 6 Conclusion and future work

The tool has mainly been designed to reduce the burden on teachers for a speedy and fair evaluation. This word cloud technique has been tested on different subject answers script evaluation including English as a subject. The lone prerequisite being the presence of an answer key to compare with. Students and teachers alike have given favorable acceptance of this tool.

Stemming is not used by this tool as it provided undesirable results and could not recall back to the base word rather it arrives at the root word. More refined techniques

**Table 6** Computed values of standard deviation for accuracy and time

Model	Accuracy	Time
IndusMarker	4.720417	4.753625
Superlative Model	2.818505	0.775948

like lemmatization which handles grammatical context by determining the lemma of a word using part of speech can be future extensions. WordNet has been used to look up synonyms. Only noun has been implemented, verb and adjectives can be tackled. It is not reflected in this tool due to time complexity. A total of 4209 questions were considered across 8 subjects to test the model, a larger data set can produce better results. This study included questions from Grade X, testing in higher education is a forthcoming extension. Future work can include all or one of these and a quicker method can be developed for essay type of answer evaluation.

As the saying goes “a picture is worth a thousand words”. Ten words in the form of a picture can be worth ten thousand is an inclusion. Therefore SM using the cohesion and relative word cloud works as a good visual analytics and evaluation tool in e-learning.

## References

- Ahearn, L. M. (2013). Commentary: keywords as a literacy practice in the history of anthropological theory. *American Ethnologist*, 40(1), 6–12.
- Bailey, S. & Meurers, D., (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 107–115). Association for Computational Linguistics.
- Baralt, M., Pennestri, S., & Selvandin, M. (2011). Action research: using wordles to teach foreign language writing. *Language Learning & Technology*, 15(2), 12–22.
- Behar-Horenstein, L. S., & Niu, L. (2011). Teaching critical thinking skills in higher education: a review of the literature. *Journal of College Teaching and Learning*, 8(2), 25.
- Bielenberg, K., (2005). *Groups in social software: Utilizing tagging to integrate individual contexts for social navigation* (Doctoral dissertation, Universität Bremen).
- Blin, F., & Munro, M. (2008). Why hasn't technology disrupted academics' teaching practices? understanding resistance to change through the lens of activity theory. *Computers & Education*, 50(2), 475–490.
- Bloom, B.S. (1956). *Taxonomy of educational objectives: The classification of educational goals: Cognitive Domain*. Longman.
- Brin, S., Motwani, R., Ullman, J.D. & Tsur, S., (1997). Dynamic itemset counting and implication rules for market basket data. In *ACM SIGMOD Record* (Vol. 26, No. 2, pp. 255–264). ACM.
- Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1), 60–117.
- Burstein, J., Wolff, S. & Lu, C., (1999). Using lexical semantic techniques to classify free-responses. In *Breadth and depth of semantic lexicons* (pp. 227–244). Springer Netherlands.
- Butcher, P. G., & Jordan, S. E. (2010). A comparison of human and computer marking of short free-text student responses. *Computers & Education*, 55(2), 489–499.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug), 2493–2537.
- Concetta, A. DePaolo, Wilkinson, K. (2014). get your head into the clouds: using word clouds for analyzing qualitative assessment data. *TechTrends-May/June 2014*, Springer, Volume 58, Number 3.
- Cook, K. A. & Thomas, J. J. (2005). *Illuminating the path: The research and development agenda for visual analytics* (No. PNNL-SA-45230). Pacific Northwest National Laboratory (PNNL), Richland, WA (US).
- Cook, K., Earnshaw, R., & Stasko, J. (2007). Guest editors' introduction: Discovering the unexpected. *IEEE Computer Graphics and Applications*, 27(5), 15–19.
- deNoyelles, A., & Reyes-Foster, B. (2015). Using word clouds in online discussions to support critical thinking and engagement. *Online Learning*, 19(4), n4.
- DePaolo, C. A., & Wilkinson, K. (2014). Get your head into the clouds: using word clouds for analyzing qualitative assessment data. *TechTrends*, 58(3), 38–44.
- Dzikovska, M. O., Nielsen, R. D., & Leacock, C. (2016). The joint student response analysis and recognizing textual entailment challenge: making sense of student responses in educational applications. *Language Resources and Evaluation*, 50(1), 67–93.
- Feinberg, J. (2010). Wordle. In J. Steele & N. Iliinsky (Eds.), *Beautiful visualization* (pp. 37–58). Sebastopol: O'Reilly.

- Feinerer, I., Hornik, K., Wallace, M. & Hornik, M.K., (2016). Package ‘wordnet’. <https://CRAN.R-project.org/package=wordnet>, Accessed 21 January 2016.
- Fellows (2013). <https://cran.r-project.org/web/packages/wordcloud/index.html>, Accessed 22 January 2016.
- Gütl, C. (2007). e-Examiner: towards a fully-automatic knowledge assessment tool applicable in adaptive e-learning systems. In *Proceedings of the 2nd international conference on interactive mobile and computer aided learning* (pp. 1–10).
- Hahn, M. & Meurers, D. (2012). Evaluating the meaning of answers to reading comprehension questions a semantics-based approach. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 326–336). Association for Computational Linguistics.
- Hahsler, M., Grün, B. & Hornik, K. (2005). A computational environment for mining association rules and frequent item sets.
- Huisman, R., Miller, W., & Trinoskey, J. (2011). We’ve Wordled, have you? digital images in the library classroom. *College & Research Libraries News*, 72(9), 522–526.
- Ingo, F & Kurt, H. (2015). tm: Text Mining Package. R package version 0.6-2. <http://CRAN.R-project.org/package=tm>, Accessed 21 January 2016.
- Ingoley, S., & Bakal, J. W. (2012). Use of fuzzy logic in evaluating students’ learning achievement. *International Journal on Advanced Computer Engineering and Communication Technology (IJACECT)*, 1(2), 47–54.
- Jablon, J.R. & Wilkinson, M. (2006). Using engagement strategies to facilitate children’s learning and success. *YC Young Children* 12-16.
- Jimoyiannis, A., & Komis, V. (2007). Examining teachers’ beliefs about ICT in education: implications of a teacher preparation programme. *Teacher Development*, 11(2), 149–173.
- Kao, A. & Potet, S.R. eds. (2007). Natural language processing and text mining. Springer Science & Business Media.
- Kaptein, R., Hiemstra, D. & Kamps, J., (2010). How different are language models and word clouds?. In *European Conference on Information Retrieval* (pp. 556–568). Springer Berlin Heidelberg.
- Kolb, A. Y. (2005). The Kolb Learning Style Inventory—Version 3.1 2005 Technical Specifications. *Boston, MA: Hay Resource Direct*, 200.
- Kragten, J. (1994). Tutorial review. calculating standard deviations and confidence intervals with a universally applicable spreadsheet technique. *Analyst*, 119(10), 2161–2165.
- Leacock, C., & Chodorow, M. (2003). C-rater: automated scoring of short-answer questions. *Computers and the Humanities*, 37(4), 389–405.
- Leek, J, Peng, R. & Irizarry, R. (2013). Simply statistics”, <http://www.r-bloggers.com/a-wordcloud-comparison-of-the-2011-and-2012-sotu/#>, Accessed 21 January 2016.
- Liu, V. & Curran, J. R. (2006). Web Text Corpus for Natural Language Processing. In EACL.
- Maglogiannis, I. G. (2007). *Emerging artificial intelligence applications in computer engineering: real world AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies* (Vol. 160). Ios Press.
- McNaught, C., & Lam, P. (2010). Using Wordle as a supplementary research tool. *The Qualitative Report*, 15(3), 630.
- Meyer, D., Hornik, K. & Feinerer, I. (2008). Text mining infrastructure in R. *Journal of Statistical Software* 25(5), pp.1-54. ISSN 1548–7660. URL <http://www.jstatsoft.org/v25/i05>, Accessed 21 April 2016.
- Miley, F., & Read, A. (2012). Using word clouds to develop proactive learners. *Journal of the Scholarship of Teaching and Learning*, 11(2), 91–110.
- Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39–41.
- Mitchell, T., Russell, T., Broomhead, P. & Aldridge, N. (2002). Towards robust computerised marking of free-text responses.
- Mohler, M. & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 567–575). Association for Computational Linguistics.
- Nickell, J. (2012). Word clouds in math classrooms. *Mathematics Teaching in the Middle School*, 17(9), 564–566.
- Pérez, D., Gliozzo, A. M., Strapparava, C., Alfonseca, E., Rodríguez, P. & Magnini, B. (2005). Automatic assessment of students’ free-text answers underpinned by the combination of a BLEU-inspired algorithm and latent semantic analysis. In *FLAIRS conference* (pp. 358–363).
- Perry, L. (2012). Using word clouds to teach about speaking style. *Communication Teacher*, 26(4), 220–223.
- Pulman, S.G. & Sukkarieh, J. Z. (2005). Automatic short answer marking. In *Proceedings of the second workshop on Building Educational Applications Using NLP* (pp. 9–16). Association for Computational Linguistics.

- Ramsden, A. & Bate, A. (2008). Using word clouds in teaching and learning.
- Sag I. A., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword expressions: a pain in the neck for NLP. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 1–15). Springer Berlin Heidelberg.
- Siddiqi, R. (2010). Improving learning and teaching through automated short answer marking, complete thesis, <https://www.escholar.manchester.ac.uk/api/datastream?publicationPid=uk-ac-manscw:90867&datastreamId=FULL-TEXT.PDF>, Accessed 28 July 2016.
- Siddiqi, R., Harrison, C. J., & Siddiqi, R. (2010). Improving teaching and learning through automated short-answer marking. *IEEE Transactions on Learning Technologies*, 3(3), 237–249.
- Sukkarieh, J. Z., Pulman, S.G. & Raikes, N. (2003). Auto-marking: using computational linguistics to score short, free text responses. In the annual conference of the International Association for Educational Assessment (IAEA), Manchester, UK.
- Tsui, L. (2002). Fostering critical thinking through effective pedagogy: evidence from four institutional case studies. *The Journal of Higher Education*, 73(6), 740–763.
- Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1), 2–13.
- Wing, J. M. (2008). Computational thinking and thinking about computing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 366(1881), 3717–3725.