J Grid Computing (2019) 17:357–366
https://doi.org/10.1007/s10723-018-9452-4

Big Data Analytics, Text Mining and Modern English Language

Saqib Alam · Nianmin Yao

Received: 20 March 2018 / Accepted: 16 July 2018 / Published online: 4 August 2018

Abstract The modern English Language took centuries to convert from old English. The word *'hath'* of old English for example, has taken centuries to become 'have' in the modern English Language. If these changes had not been occurred there would not have been the possibility of modern words. A text written in fifteen century can be difficult to read and if we go back a couple of more centuries, it would be like reading a different language. In this paper, we have used the text mining techniques to analyze the old and modern English languages. We have introduced the Common-Words Counting algorithm that identifies common words of 15th century that diminishes gradually in the later centuries. We computed the speed of linguistic changes and identified the reasons behind them. For this purpose, 34000 text books were downloaded from Project Gutenberg of different authors, between 15th to 19th centuries. These books were categorized into five centuries in the range from 15th to 19th centuries. We selected most common words from the books of 15th century and calculated their frequencies in other centuries. We calculated the sum of Term Frequency-Inverse Document Frequency (TF-IDF) of these words and proved that frequencies of words were decreasing from 15th century to 19th century with some words even disappeared in other centuries, such as 'doth', 'hath', punt, guise and *'selfe'*. We calculated the speed of changing of words using the slope formula. We proved that the words were changing during each century with the speed of changing of words being the lowest during 16th – 17th centuries and the highest during 18th – 19th centuries which shows that the old words or their spellings were changed to the modern words during 18th – 19th centuries. The industrialization, modernization, and British Empire invasion were the key factors, which changed the old English language into modern English language.

Keywords Text mining · TF-IDF · English language · Speed of linguistic changes

1 Introduction

A language always changes, across space, social group, and time. It is difficult for a student of Arts subject to use the scientific methods to compute the linguistic changes. Books are available in a huge number and now with the introduction of social media and web 2.0, a huge amount of data is available online. Analyzing this huge amount of data is a challenge. Similarly applying quantitative methods on a qualitative text to search interesting patterns in it is a challenging

S. Alam (✉) · N. Yao
Department of Electronic Information and Electrical Engineering, Dalian University of Technology,
Black Building, Linggong Road No.2, Ganjingzi District,
Dalian, 116024, People's Republic of China
e-mail: alamsaqib@mail.dlut.edu.cn

N. Yao
e-mail: lucos@dlut.edu.cn

activity. This research analyzes linguistic changes in English during 15th to 19th centuries and identifies the reasons behind them by applying data mining techniques. The modern English Language took centuries to convert from old English. For example in the old English, the word '*hath*' has become '*have*' in modern English Language. If such changes would not have been occurred, we would not have the modern words. A text written in fifteenth century can be difficult to understand and if we go back a couple of more centuries, it would be like reading a different language. In this research, we focus on the books of different authors from different centuries and used text mining techniques to analyze the corpus. In recent years text analysis gain fame for evaluating text and extracting meaningful information from it. Recently, it has gained a huge attention because of the introduction of social web and availability of ebooks to analyze the textual data. We gathered the corpus from Project Gutenberg[1] which is a set of freely and publically available books. It was founded in 1971 and is the oldest free digital library.

As of 2015, its collection of items reached to 50,000. The books are available on Project Gutenberg in the form of plain text and other formats such as HTML, PDF, and EPUB. We used the plain text books in our research work. In this paper, we have introduced the Common-Words Counting algorithm that selects the most common words of 15th century and compares them with the words of other centuries. Our research has proved that many common words of the 15th century started diminishing gradually in the later centuries and new words were introduced because of the industrial and scientific revolutions. We calculated the speed of changing of words and proved that the speed of changing of words was the lowest during 16th – 17th centuries and it was the highest during 18th – 19th centuries. It is not like that an author was using the word '*hath*' and the next morning when he woke up

he started using '*have*'. There are some reasons which changed the old English into modern English.

2 Literature Review

Chou et al. [1] presented an automated methodology of classifying and clustering for the written verdict by Back-Propagation Network (BPN) and Self-Organization Map (SOM) methods. Based on neural networks, they developed a methodology classification and clustering of documents that helped law enforcement agencies to manage written judgment effectively. As an input of BPN, they select the keywords with the maximum occurrences, and selected seven criminal groups as output of it. The assessment were taken from the Judicial Yuan dynasty in Taiwan. Keywords were extracted and their frequencies as well as TF-IDF were calculated. The top keywords with the maximum frequencies were represented in the given written ruling. Relevant legal document were atomically generated by neural network which were pre-trained. They used SOM method for clustering of written criminal judgments. Their proposed work showed that their model helped finding relevant written judgments of criminal cases effectively.

Griffiths et al. [2] showed in their research that language is pass on to persons through generation by iterative mechanism. To understand the consequences of iterative learning is, therefore, an essential step in the advancement of linguistic change. They have established a structure for the analysis of iterative learning, which allows to distinguish the learning procedure from one learner to another. Their conclusions discovered that the part of iterative learning in the elucidation of linguistic universals and present a proper relationship between constraints on language acquisition and the languages that are spoken.

Reed et al. [3] proposed a model for unsupervised text clustering problem called Term Frequency-Inverse Corpus Frequency (TF-ICF). To calculated the effectiveness of the proposed model, they compared the five commonly used techniques for experimentation. According to their results TF-ICF can generate text clusters more faster and effetely as compared to other commonly used term weighting methods. Their experiment also showed that the performance of TF-ICF was above average and 11% below average in the worst case scenario. They found the performance of TF-ICF even better when the ICF.

Hills, Thomas et al. [4] used multiple language corpora which represents over 350 billion terms with more than 40,000 English lexis and demonstrated that in American English Language a systematic increase in real language over the last 200 years. Their results also found some evidence that the rise is effected by the increasing of population and may be associated

[1] www.gutenberg.org/ebooks/

with increasing number of second language learners or awareness in economic and technology in response to crowding in the language market. They also study the influence of gender and literacy. They demonstrated evolution in the psychological construction of American English Language, with a profound impact on cognitive processing, which is likely to permeate modern language use.

Ramos et al. [5] examined the results of applying TF-IDF to find out what terms in a corpus might be more favorable to use in a query. The implication of TF-IDF on terms evaluate the values for each term in a corpus through an inverse frequency of the word in a specific document to the total number of documents in which the word appears. The frequency of highest TF-IDF words means the relationship is more stronger with the document in which they appear. They suggested that if a word occurs in a query, the document could be of interest to the user. They provided evidence that their simple algorithm efficiently performs categorizations. They suggested that TF-IDF is a simple and effective algorithm for words to match it in a documents that are related to that query.

Laércio Dias et al. [6] investigated linguistic similarity and evolution of scientific fields. In their research they analyzed almost 20MB research papers from the past three decades. Their research indicated that the linguistic similarity is related but different from experts and citation-based classifications, leading to an improved view on the organization of science.

3 Proposed Common-Words Counting Algorithm

Lexicology is the study of words [7] and without words a language cannot be formed. A great deal of research still remains to be done into the semantic-contextual structure of the sentence and into its relation to word-order [8]. Many data mining techniques have been proposed for mining effective patterns in text documents. However, efficiently use and discovering the patterns are still an open research issue, specifically in the domain of text mining [9]. In this research we have proposed a common-words counting algorithm that counts the most frequent words of 15th century and comparers the frequency of those words with the same words appearing in other centuries in the range of 16th to 19th centuries. For this purpose, we developed a corpus of books from different centuries. We downloaded more than 34000 books from Project Gutenberg and created a single corpus. We categorized them by the centuries in which these books were written ranging from 15th to 19th centuries. We created a main corpus of total size 12.33MB that included books of all centuries. This main corpus was divided into five sub-corpora; one each for a century in the range of 15th to 19th. These sub-corpora of 15th, 16th, 17th, 18th, and 19th centuries are of the sizes 1.99MB, 1.99MB, 2.96MB, 3.53MB, and 1.86MB respectively. The reason of not having same sizes of the sub-corpora was the length of books in different centuries. Following are the steps of the proposed algorithm:

3.1 Stopwords Removal

Mostly, stopwords are highest at frequency in every corpus such as *the*, *to*, *and, also, a, an* and so on.These words do not have significant meaning in the text and it is necessary to filter out this noisy data from important textual data to enhance the quality [10, 16]. Stop words have lexical content and the presence of these words will fail the required results. We filtered out these words from our documents as a first step. The Python built-in library *nltk* was used for this purpose.

3.2 Tokenization

Tokenization is a process in text analysis which breaks up sequence of strings into words, keywords symbols or phrases or other meaningful elements. The main purpose of tokenization in text analysis is to identify the meaningful words. For this purpose, it converts each sentence into a tree form and discards symbolic characters, verbs, adverbs, and other irrelevant words. In various language processing functions some preliminary steps are, part of speech tagging, machine translation, spell checking, sentence boundary detection, information retrieval, and information extraction [11]. The tokenization step was applied on the corpus to identify meaningful words from the books of all centuries.

3.3 Case Converting

Books have words in different cases; some words are in lower case and others in upper case. We cannot identify and count words properly because of this issue. For example, if a word starts from upper case then it could be counted separately from the same word started in lower case. We converted all upper case keywords to lower case by using the built-in function of Python.

3.4 Filtering Words Greater than Three Characters

In this step of our proposed algorithm, it will filter out those words which will be less than three characters, such as *'all', 'zip', 'cup', 'joy'*. The purpose of this step is to drops those words which are not important to be considered for analysis purposes because the traditional stopwords libraries will not include all the words which are higher in frequency and not important for the analysis.

3.5 Counting of Common Words

The next step is the counting of common words in different centuries. Word count is significantly more accurate and simple the other methods [12, 17]. This step selects the 200 most commonly used words from the 15th century books and compares them with all words of other centuries to observe their occurrences in other centuries. We observed that words frequently used in 15th century were not commonly used in other centuries. We also observed that some words were not used even a single time in the books of other centuries of our corpus. Figure 1 shows the diagrammatic steps of the proposed common-words counting algorithm.

4 Features Selection

We used the TF-IDF weight as a feature selection metric in this research, to know the significance of the terms selected in the Common-Words Counting algorithm.

4.1 Term Frequency

To calculate that how frequently a term occurs in a document we used TF. As we know that the length of a document is usually different than other documents and the possibility is higher that a term will appear more times in lengthy document than a shorter document. Thus for normalization purposes, the TF is divided by the total number of words in a document. For the term $\boldsymbol{t_j}$ in a specific document $\boldsymbol{d_k}$, the term frequency is defined as bellow:

$$\mathrm{tf_{j,k}} = \frac{\mathrm{n_{j,k}}}{\sum_{\mathrm{j}}^{\mathrm{k}} \mathrm{n_{i,k}}} \tag{1}$$

In the above formula $\mathbf{n_{j,k}}$ is the occurrences of the desired term $\boldsymbol{t_j}$ in the document $\boldsymbol{d_k}$. The dominator is the total number of all words in the document $\boldsymbol{d_k}$. Where $\mathrm{tf_{j,k}}$ represents the term frequency for j to k, while in the Σ the initial values are not define.

4.2 Inverse Document Frequency

Inverse Document Frequency (IDF) was proposed in 1972, and has been commonly used by the researchers, and as generally is a part of TF*IDF function [13]. The IDF measures that how much important is a term in a desired corpus. After computing the TF it shows us that all the terms are important equally while IDF will weigh down the important and rare ones. Formula for IDF is as under:

$$idf_j = \log \frac{|T|}{|\{k : t_j \in d_k\}|} \tag{2}$$

In the above formula, |T| is the total number of documents in a corpus, while $|\{\boldsymbol{k} : \boldsymbol{t_j} \in \boldsymbol{d_k}|$ is the number of documents where the term $\mathbf{t_j}$ appears; such that $\boldsymbol{n_{j,k}} \neq \mathbf{0}$. The TF-IDF weight of the term $\boldsymbol{t_j}$ in a specific document $\boldsymbol{d_k}$ is the product of TF and IDF and its formula is as under:

$$(\mathrm{tf}-\mathrm{idf})_{\mathrm{j,k}} = \mathrm{tf_{j,k}} \times \mathrm{idf_j} \tag{3}$$

We have used the TF-IDF weight to calculate important words appeared in the documents of different centuries.

4.3 Term Frequency Inverse Document Frequency

Commonly the importance of a word is based on TF and IDF [14]. The TF-IDF is a numerical statistics which is used to reflect that how important is a word in a corpus. The TF-IDF weight is the composite of two factors; the first factor is the TF which normalizes the terms. It is equal to the number of terms appearing in a document divided by the total number of words in the same document. The second part, which is IDF, is the total number of documents in the corpus divided

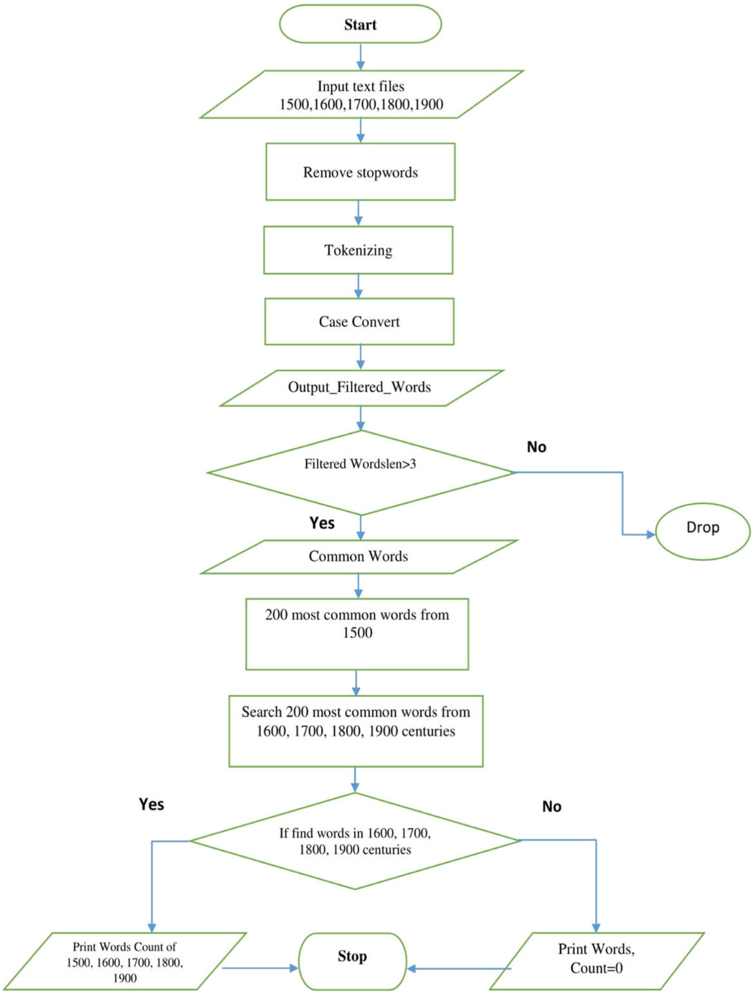
Start
Input text files 1500,1600,1700,1800,1900
Remove stopwords
Tokenizing
Case Convert
Output_Filtered_Words
Filtered Wordslen>3
No
Drop
Yes
Common Words
200 most common words from 1500
Search 200 most common words from 1600, 1700, 1800, 1900 centuries
Yes
If find words in 1600, 1700, 1800, 1900 centuries
No
Print Words Count of 1500, 1600, 1700, 1800, 1900
Stop
Print Words, Count=0

Fig. 1 Common-Web Counting Algorithm

by the number of documents where the specific word has appeared.

5 Experiment Design

This section discusses the experiment design, data used in the experiment, platform and tools used during the experiment.

5.1 Data Collection

We gathered data from Project Gutenberg which is an oldest digital library founded by Michal S. Hart in 1971. It is an open source digital library developed by volunteers. As of October 03, 2015, the number of books in Project Gutenberg was 50,000 in different formats like epub, pdf and text in different languages. It offers different methods to download books such as Download via BitTorrent, Edonkey/Emule, Jigdo, FTP and HTTP. We used wget to download all English

Table 1 The first 10 most frequently used words of 15th century

Words	15 th century
Good	913
Shall	855
Doth	811
Thou	672
Like	654
Well	649
Love	643
King	594
Would	584
Sweet	388

Table 2 Comparison of the first 10 most frequently used words of 15th century with other centuries on the basis of their frequencies

Words	15th century	16th century	17th century	18th century	19th century
Good	913	437	789	643	548
Shall	855	1116	605	570	404
Doth	811	92	52	83	1
Thou	672	1889	231	904	17
Like	654	399	743	1118	629
Well	649	386	793	815	579
Love	643	384	295	553	125
King	594	338	403	755	8
Would	584	1048	1448	1251	1197
Sweet	388	71	88	223	32

language books from Gutenberg using its link “url” to choose language “en”, filetype “txt”, encoding “flat”, Charset “ISO-8859, utf-8”. We found that the text books need to be renamed. As the number of books is more than 34,000 it was difficult to rename it manually. We created a *.batch* file and after running the file we got our required result, which was to rename files from source folder and the renamed files were placed in a separate folder. The code retrieve Author “A” and Publish Year “P” from the text file and sat as Title “T” of the text file. After renaming the text files we separated each era’s books, and then using command line to merge all books of each century in a single file.

5.2 Hardware and Software Tools

We used Windows 7 Ultimate Edition 32 bit operating system which is compatible with all the tools we used in our research work. The hardware we used in our research work is Intel®Core™2 Duo, 2.53GHz Processor with 2 MB L2 Cache, 4 GB of DDR2-400B RAM with Memory Clock 100MHZ and I/O Clock 200MHz.

Table 4 The Sum of TF-IDF of every word as per the century

Century	Sum of TF-IDF
15th	0.015013031
16th	0.010923866
17th	0.002617657
18th	0.000348641
19th	0.000631682

We used Python for Natural Language Processing (NLP), which is a widely used open source tool. Natural Language Toolkit (NLTK) is Python platform which makes human language data easy for programmers. NLTK is a wonderful library for linguistic computing and text analysis using Python. For selecting 200 most commonly used words from 15th century and comparing the existence of these words with other centuries, we used Python specifically its NLTK library.

KNIME is an open source data analytics and reporting tool that we used in our experiment. It has a Graphical User Interface (GUI) which makes the data processing (Extraction, Transformation, Loading), modeling, data analysis and visualization easier for the users. Similarly, we used OriginLab which is a scientific graphic and analysis tool that provides solutions for engineers and scientists for presenting their data graphically or analytically.

Table 3 Comparison of the first 10 most frequently used words of 15th century with other centuries on the basis of TF-IDF

Words	15th century TF-IDF	16th century TF-IDF	17th century TF-IDF	18th century TF-IDF	19th century TF-IDF
Good	0.002716	0.001215	0.001456	0.001029	0.001617
Shall	0.002544	0.003102	0.001117	0.000913	0.001192
Doth	0.002413	0.000256	0.000096	0.000133	0.000003
Thou	0.001999	0.00525	0.000426	0.001447	0.00005
Like	0.001946	0.001109	0.001371	0.00179	0.001856
Well	0.001931	0.001073	0.001464	0.001305	0.001708
Love	0.001913	0.001067	0.000544	0.000885	0.000369
King	0.001767	0.000939	0.000744	0.001209	0.000024
Would	0.001737	0.002913	0.002672	0.002003	0.003531
Sweet	0.001154	0.000197	0.000162	0.000357	0.000094

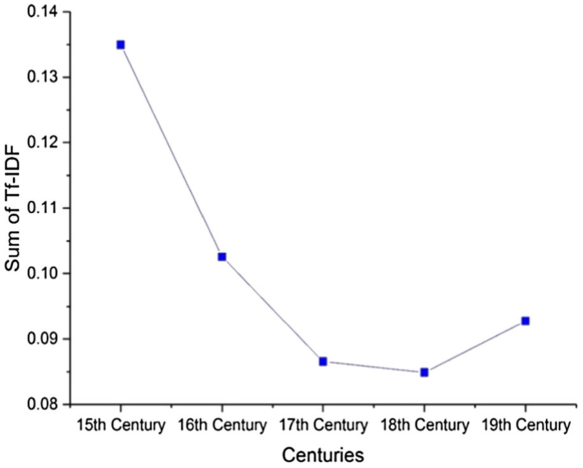

Fig. 2 Sum of TF-IDF from 15th Century to 19th Century

6 Results and Discussion

We downloaded 34,000 flat file (text format) books from Project Gutenberg to calculate the changes of words from old to modern language. We categorized the books from 15th to 19th centuries. From 15th century, we selected 200 most frequently used words and compared those words with other centuries.

Consider Table 1 that shows the 10 most frequently used words of 15th century in descending order. In Table 2, we compared the most frequently used words of 15th century with the words of other centuries by showing their frequencies in other centuries.

In the above Table 2 we can see that the frequency of some words decreased and some words even disappeared in other centuries, such as ‘*doth’, ‘thou’, punt, guise* and ‘*selfe’*. From Table 2 we calculated the TF-IDF of each word as shown in Table 3. We then summed up the TF-IDF of each century. Table 4 shows the sum of TF-IDF of all words in their respective century.

Consider the following Fig. 2 created from Table 4.We can see that the frequencies of words in every century is decreasing which means that the words which were most frequently used in 15th century were diminished in 16th century and so on till 18th century. Figure shows a slight increase in the reuse of the most commonly used words of the 15th century in the 19th century. This slight increase shows that more authors of the 19th century have reused words of the 15th century in their writings as compared to the authors of 17th and 18th centuries.

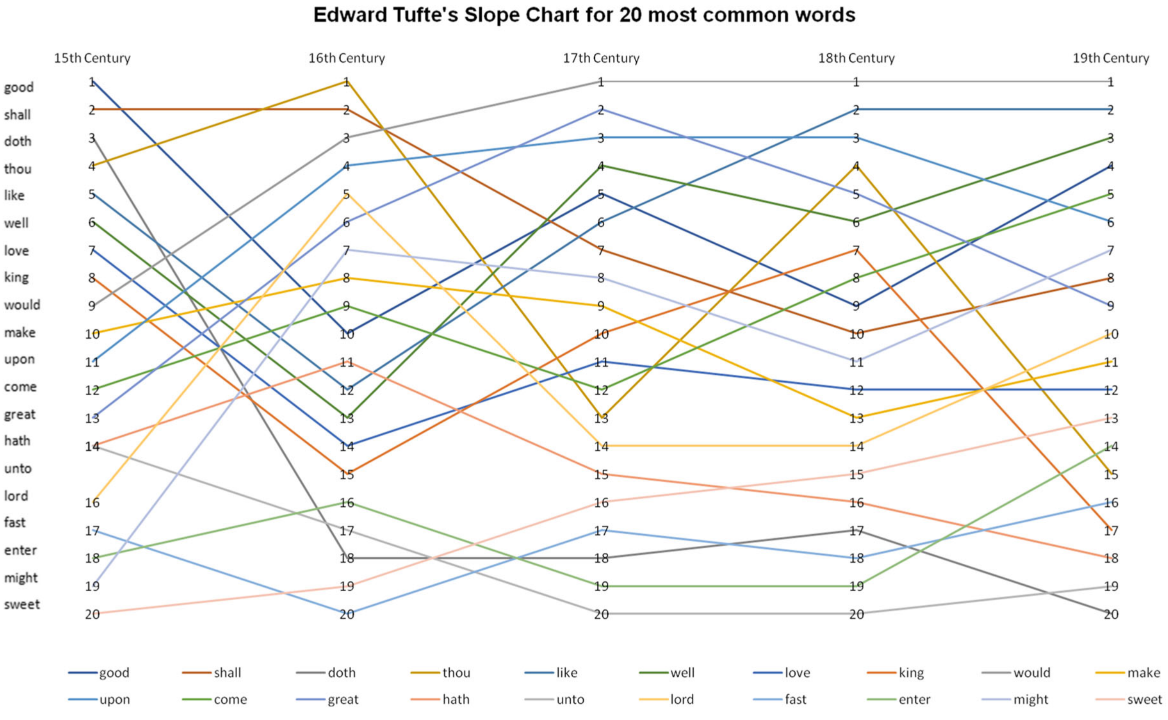

Fig. 3 Ranking of 20 most common words 15 to 19 centuries

Table 5 The Speed of Changing of Words

Centuries	Speed
15th – 16th	–4.08917E-05
16th – 17th	–8.30621E-05
17th – 18th	–2.26902E-05
18th – 19th	2.83041E-06

Figure 3 shows Edwards Tufte's Slope chart [15] of 20 most common words. The chart shows that most of the words of 15th century were diminished in other centuries.

To calculate the Edwards Tufte's slope chart [15], we ranked 20 most frequent words. The ranking is used in ascending order, the lower the rank the higher the frequency. In Fig. 3 we ranked the common words, in 15th century The word *"good"* ranked 1st as it is the most commonly used word in this century, while in 16th century the word *"thou"* become rank 1st. Similarly 17th, 18th and 19th centuries has one common word *"would"* which ranked 1st. Edwards Tufte's slopechart shows how a word ranked 1st in one century became less popular in other centuries. We can visually witness the changes in the words throughout the five centuries.

7 Calculating the Speed of Changing of Words

The focus of this analysis was to determine how rapidly the words were changed and in which century these changes happened. For this purpose, we used the following slope formula:

$$S = \frac{c_1 - c_2}{f_1 - f_2}$$

Here S is the slope while $\boldsymbol{c_1}$ is the starting century and $\boldsymbol{c_2}$ is the ending century divided by $\boldsymbol{f_1}$which is the sum of $\boldsymbol{c_1}$ frequencies and $\boldsymbol{f_2}$ is the sum of $\boldsymbol{c_2}$ frequencies. For any two points on the line, we calculated the slope which is the rate of change. For example, we calculated the changes of a word *"hath"* which was commonly used in 1500 century, by calculating the slope between the first two points (centuries): (422-411)/(1500-1600) = 11/-100 = -0.11.

The fact that the slope is negative, because the line is going downward from left to right telling us

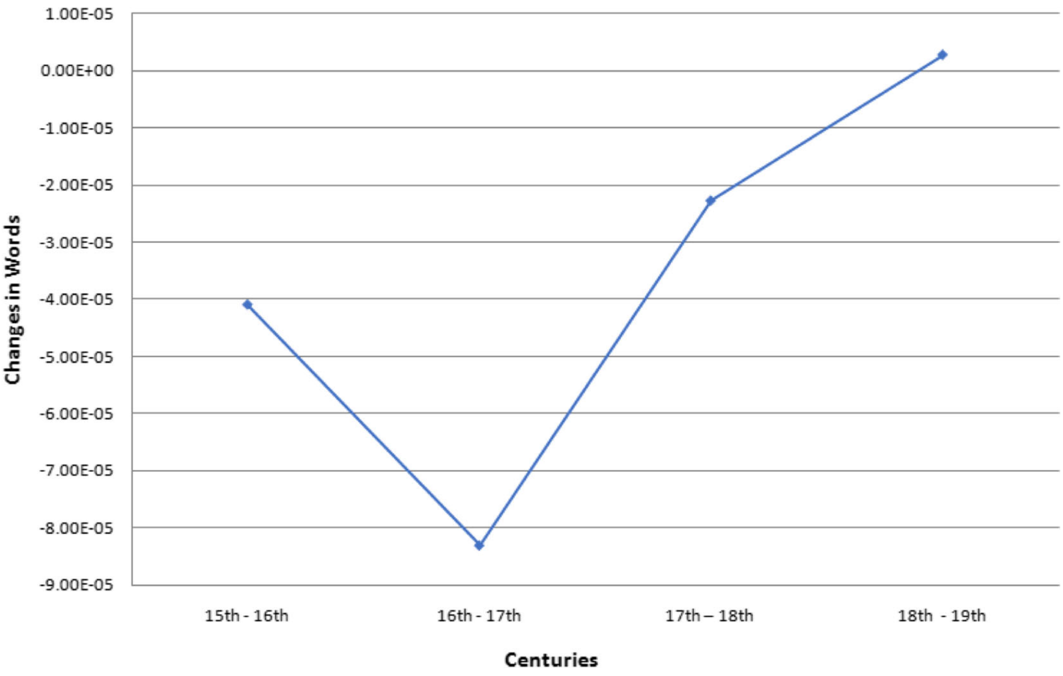

Fig. 4 A Visual View of the Speed of Changing of Words

that the frequency of the independent variable (incidence of *"hath"*) is diminishing. If the incidence was increasing, the line would go up, and the slope would be positive. The slope between the second and third points (1600 and 1700 centuries):

$$(411 - 182)/(1600 - 1700) = 229/-100 = -2.29.$$

Using the slope formula, the speed of changing of words was calculated as shown in Table 5. Figure 4 shows a visual view of the speed of changing of words. We can see that the speed of changing of words was the lowest during 16th – 17th centuries which means that less number of new words were introduced by authors during this span of time. On the other hand, the speed of changing of words is the highest during 18th – 19th centuries which shows that greater number of old words or their spellings were changed to the modern words.

In 15th century, mostly, people were not well educated; only children of the rich people could get education. The number of authors was also limited and they were writing plays for the common people which they could understand easily. The plays were mostly in slaying languages like Shakespeare's used slay language in his plays. In 16th century, the authors were mostly from 15th century so there was no big change in the English language. From 17th to 18th century, the number of authors increased and so as the education rate also increased. Hence we can see changes in the language. The 18th and 19th centuries were the age of modernization. Industries were setup, more people got educated, and scientific revolutions changed the world. New words were invented because of new inventions such as steamships and railways. British invasions were started at early 16th century but were at highest from 18th to 19th century, which also affected the English language because of the adaptation of words. Vowels were shifted which affected the spellings of words such as '*faire'* became '*fair'* '*owne'* became '*own'*.

8 Conclusion and Future Work

In this research, we have analyzed books of different centuries using the text mining techniques. The main objective of this research was to study changes in words of English language from 15th century to 19th century. For this purpose, we downloaded 34,000 flat file (text format) books from Project Gutenberg to calculate the changes of words from old to modern language. These books were categorized into five centuries in the range from 15th to 19th centuries. We selected the 200 most common words used from the books of 15th century and calculated their frequencies in the 16th, 17th, 18th, and 19th centuries. We calculated the sum of TF-IDF of these words and proved that frequencies of words were decreasing from 15th century to 19th century with some words even disappeared in other centuries, such as '*doth', 'hath', punt, guise* and '*selfe'*.

We calculated the speed of changing of words using the slope formula. We proved that the words were changing during each century with the speed of changing of words being the lowest during 16th – 17th centuries and the highest during 18th – 19th centuries which shows that the old words or their spellings were changed to the modern words during 18th – 19th centuries. The reason of this quick change was that it was the age of modernization. Industries were setup, more people got educated, and scientific revolutions changed the world. New words were invented because of new inventions such as steamships and railways. In the future work, we will include books of 20th and 21st centuries to our corpus and analyze changes in the words. Similarly, we will identify the most common words in the 15th century and analyze them with respect to the words of other centuries. We will repeat the same process for the rest of centuries in our future work.

References

1. Chou, S., Hsing, T.P.: Text mining technique for Chinese written judgment of criminal case. In: Pacific-Asia workshop on intelligence and security informatics, pp. 113–125. Springer, Berlin (2010)
2. Griffiths, T.L., Kalish, M.L.: Language evolution by iterated learning with Bayesian agents. Cogn. Sci. **31**(3), 441–480 (2007)
3. Reed, J.W., Jiao, Y., Potok, T.E., Klump, B.A., Elmore, M.T., Hurson, A.R.: TF-ICF: A new term weighting scheme for clustering dynamic data streams. In: 5th international conference on machine learning and applications, 2006. ICMLA'06. (pp. 258-263). IEEE (2006)
4. Hills, T.T., Adelman, J.S.: Recent evolution of learnability in American English from 1800 to 2000. Cognition **143**, 87–92 (2015)
5. Ramos, J.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional

conference on machine learning, vol. 242, pp. 133–142 (2003)

6. Dias, L., Gerlach, M., Scharloth, J., Altmann, E.G.: Using text analysis to quantify the similarity and evolution of scientific disciplines. R. Soc. Open Sci. **5**(1), 171545 (2018)
7. Grzega, J., Schoener, M.: English and general historical lexicology. Eichstätt-Ingolstadt, Katholische Universität (2007)
8. Firbas, J.A.N.: De Vordre des mots dans les langues anciennes com- pares aux langues modernes Question de grammaire generate (1844)
9. Zhong, N., Li, Y., Wu, S.T.: Effective pattern discovery for text mining. IEEE Trans. Knowl. Data Eng. **24**(1), 30–44 (2012)
10. Munková, D., Munk, M., Vozár, M.: Influence of stop-words removal on sequence patterns identification within comparable corpora. In: ICT innovations 2013, pp. 67–76. Springer, Heidelberg (2014)
11. Rehman, Z., Anwar, W., Bajwa, U.I., Xuan, W., Chaoying, Z.: Morpheme matching based text tokenization for a scarce resourced language. PloS one **8**(8), e68178 (2013)
12. Blumenstock, J.E.: Size matters: word count as a measure of quality on wikipedia. In: Proceedings of the 17th international conference on World Wide Web (pp. 1095–1096). ACM (2008)
13. Robertson, S.: Understanding inverse document frequency: on theoretical arguments for IDF. J. Doc. **60**(5), 503–520 (2004)
14. Azam, N., Yao, J.: Comparison of term frequency and document frequency based feature selection metrics in text categorization. Expert Syst. Appl. **39**(5), 4760–4768 (2012)
15. Tufte, E.R.: Beautiful evidence, vol. 1. Graphics Press, Cheshire (2006)
16. Hofmann, T.: August. Probabilistic latent semantic indexing. In: ACM SIGIR Forum (Vol. 51, No. 2, pp. 211-218). ACM (2017)
17. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: ACM SIGIR Forum (Vol. 51, No. 2, pp. 268-276). ACM (2017)