Better Results in Automatic Arabic Text Summarization System Using Deep Learning based RBM than by Using Clustering Algorithm based LSA

Marwa Elgamal, Prof. Dr Salwa Hamada, Prof. Dr Reda Aboelezz and Dr Mohamed Abou-Kreisha

Abstract— The fast growing amount of Arabic interactive books and electronic texts makes the research in automatic Arabic documents summarization and clustering an important field in the area of natural language processing (NLP), text mining and Deep Neural Networks (DNN). In this Paper we investigate extractive text summarization that give the most accurate and robust results in Automatic Arabic text Summarization using different approaches. Then we evaluate the outputs from these approaches to decide which one resolve better results. We focus on extractive text summarization using clustering algorithm (CA) with Latent Semantic Analysis (LSA) method, and deep learning approach using Deep Restricted Boltzmann Machine (DRBM). Arabic documents Datasets are used from Arabic stories and Arabic educational book to convert the text into a summarized storyline. The experimental results show that the proposed models with (CA-LSA) clustering algorithm and DRBM deep learning Approach performing an effective summarization from the Arabic document. After using manual evaluation by human judges we found that text summarization using deep learning approach with DRBM method give better results in the Arabic text summarization.

Index Terms— Clustering Algorithm (CA), Latent Semantic Analysis (LSA), Deep Learning, Extractive text Summarization, Deep Restricted Boltzmann Machine (DRBM), Arabic Text Summarization, Manual Evaluation.

1 INTRODUCTION

The rise in information technology improves the quantity and the difficulty of information sources, especially that related to online Arabic text documents. Like online scientific papers, blogs, e-books, e-learning educational books and Arabic stories. The automatic document summarization becomes very important because the increasing in online Arabic documents with the increase in Arabic readers and the need to have a summary that covers important information from original document to the extracted text [1]. The key point behind different summarization techniques is to find a representative subset of the original document such that the main principle of the document which is contained in this subset from the semantic and conceptual standpoints [2].

Achieving this is very challenging because it is difficult to determine the quality of the summary, since it depends on many factors such as the user's requirements and the compression ratio among others. However, there is some success to tackle these problems where the scientists have been able to reach a level in which the machine is able to generate a human readable summary, especially for English documents. Unfortunately, Arabic document summarization is still receiving a little attention [3].

Recently, significant research efforts have been dedicated to solve the Arabic documents clustering and summarization problem.

Many of these systems address the summarization problem depending on the required application, based on input output and content types and the user's needs.

There are two common types of automatic text summarization, extractive and abstractive [4]. In this paper we mainly focus on extractive summarization, we chose two different extractive approaches, one is using clustering algorithm with LSA approach and the second with deep learning using RBM approach. In text summarization problem, developing feature reduction techniques are important for efficient representation of textual features. This concern has generated by using CA-LSA and DRBM to solve the summarization problem. CA-LSA is a powerful unsupervised analytical tool. It groups the documents to a meaningful cluster according to its semantic relations between words which is considered in this method. The performance of CA-LSA-based summarization algorithms depends on the quality of the document representation [3, 5]

- *Marwa Elgamal is currently pursuing Doctoral degree program in Systems and Computer Department at Faculty of Engineering, Al-Azhar University. E-mail: eng.marwa.m.m@gmail.com*
- *Prof. Dr Reda Aboelezz, Systems and Computer Department at Faculty o Engineering, Al-Azhar University. E-mail: Reda2018aboelezz@gmail.com*
- *Prof. Dr Salwa Hamada, National Research Institute, Cairo, Egypt. E-mail hesalwa@hotmail.com.*
- *Dr Mohamed Abou-Kreisha, Al-Azhar University, Mathematical Depart ment at Faculty of Science, Cairo, Egypt.E-mail:drkresha@gmail.com*

and the sentence selection algorithm. In this work, we try to improve the document's representation and proposing a new sentence selection algorithm.

The first contribution in this study is to use CA-LSA approach for Arabic text summarization that can capture the latent semantic structure of a document to generate an intelligible summary with a good coverage [4].

The second contribution is using Deep Learning approach with DRBM in extractive text summarization to enhance the text summarization and give a robust and accurate summary.

This paper is the first study to compare between two approaches, CA- LSA and DRBM in Automatic Arabic text summarization and this study applying Artificial Intelligent (AI) services and deep machine learning. The goal of this paper is to extract a summarization which we called the storylines to produce short important sentences from a long Arabic story. The advantage of storyline generation is it can speed up Information Retrieval (IR), and save time in reading a long Arabic story. A manual evaluation using human judges applied too. The CA-LSA and DRBM representation models in text mining (TM) fields including machine learning in text clustering algorithm (CA) [6,7] document classification [8,9] And Semantic Analysis (SA) [10], Deep learning in extractive text summarization.

State of the art

The Arabic language is a semantic language which consists of twenty-eight characters. Arabic writing start from right to left and the word in this language is connected, so one letter consists of three or more different shapes. While much research has been done for the English language, little has been done for the Arabic language. This paper presents two approaches in extractive text summarization one with neural networks NN which is CA-LSA, the other approach is DRBM with deep neural networks DNN to enhance the performance of text summarization. To enhance the text summarization we used deep learning approach which combines extracting various features from text with the power of Deep RBM method [4]. The framework in this paper consists of two main parts in extractive text summarization. The first is CA-LSA and the second one is the DRBM method. The frameworks will be described in section 4. The rest of paper is organized as follows. Section 2: presents the literature review in extractive text summarization and document clustering. Section 3: describes the Arabic text summarization model methodologies and frameworks based on CA-LSA and deep learning using DRBM methods. Section 4: present the proposed approaches in detail (Dataset, text pre-processing, tokenization, CA-LSA, DRBM post-processing), Section 5: Implementation and results. Section 6: Conclusion and future work and Appendix to show an example of Arabic text summarization. In this paper, we present a manual evaluation using human judgment. The evaluation results show that ML using CA-LSA success to generate a summarization but deep learning using DRBM success to generate robust, accurate summary with better results.

2. Literature Review:

Document summarization is one of the most difficult but promising application in natural language processing (NLP). There are two types of input document summarization: single document and multi-document summarization. There are also two types of output document summarization, extractive techniques and abstractive technique:

1. Extractive techniques perform text summarization by selecting the most important sentences from the original text and combine them into a new shorter version according to some criteria.
2. Abstractive text summarization it attempts to build an internal semantic representation of the original text and then create a summary closer a human-generated one [11].

Text summarization has increased the attention of the researchers since 1950. H. P. Luhn was the first to make an automatic text summarization system. It was grounded on the frequency of terms. In 1997, Inderjeet Mani and Eric Bloedorn developed a graph-based method in which the text was represented in the form of graphs [12]. In 2005, Evans and Klavans proposed a new method for summarizing the clusters of documents on the same events based on text similarity [12]. The Embra system for DUC 2005 uses LSA to build a very large semantic space to derive a more robust representation of sentences English and Arabic language documents were used [13]. In 2015, S. A. Babar and P. D. Patil proposed an approach to improve the performance of text summarization using Singular value Decomposition SVD and Fuzzy inference system [14]. In 2016 a method for text summarization using Restricted Boltzmann Machine proposed [15]. In 2017 an approach for auto text summarization was developed by Chopade and Narvekar using deep neural networks and fuzzy logic which provide significant –increase in the accuracy of Summary [16].

3. Methodology

3.1 Methodology using CA based on LSA

In our study, we proposed a CA using an adaptive latent semantic analysis LSA model-based Arabic document clustering to generate a summarized text called storylines from a single Arabic document. In the enhanced CA-LSA based text summarization framework we provide summaries with reasonable quality. As shown in figure (1) the CA-LSA framework consists of three main stages: first, Appling Pre-processing with NLP (tokenization, stop words removing, punctuations removing and stemming). The pre-processing steps aim to clean the text from unwanted data like stop words, punctuations that are not related to the concept and considered as noisy data. Although in our approach we use the stemmed words to extract roots from words to give better performance in Arabic document clustering. After pre-processing, we used the TF-IDF method in the second stage to develop matrix which presents the main concepts and its weights in the document for sentence selection algorithm. Third, we use SVD then LSA to generate the storyline which is the summary of an Arabic doc-

ument. The following sections explain these stages in more details.

3.1.1 Text pre-processing

These steps are needed for transferring text from human language to machine-readable format for further processing. As shown in the next figure.

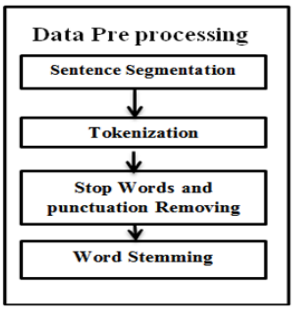

Fig 1: Data Preprocessing

i. Morphological Sentence Splitting and Words Tokenization

Tokenization in a single document is a way to split every paragraph into sentences and every sentence into words called (token). For this purpose, we use morphological decomposition based on punctuation. In this work, Arabic sentences split using punctuation marks that define the end of each sentence. A set of punctuation marks, including commas (,) semicolons (;), question marks (?), exclamation marks (!), colons (:), and periods (.), are selected to split the text into sentences. Words are split using the space (ex: "ذهب الحراس إلى القصر") after tokenization (ذهب / الحراس / إلى / القصر).

ii. Punctuations Removing

We remove the punctuations like these symbols in Arabic text [!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~] from text for dimensionality reduction.

iii. Stop Words Removing

Stop words removing and filtering any junk data in the document. Stop word are the most common words in a language like "the", "a", "on", "is", "all" in English or "هل" "ما" "إذا" in Arabic and we use the corpus for Arabic Stop words with Natural Language Toolkit (NLTK) and suite of libraries and programs for symbolic and statistical natural language processing in python. These words do not carry important meaning and are usually removed from texts.

iv. Words Stemming

A stemming algorithm is a process of linguistic normalisation, in which the variant forms of a word are reduced to a common form, for example, (حرس <--- حراسه - يحرس – الحراس - الحارس).The important of word stemming phase in document summarization is that it reduces the space of textual representation by using the root of the word instead of the whole word and this phase mainly which we consider to enhance our approach in Arabic document clustering because it achieves the dimensionality reduction.

3.1.2 TF-IDF

The TF-IDF is a short term for frequency-inverse document frequency which is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus .it determining the importance of a document and find the important concepts in the document. The TF-IDF converts the words in the text into vectors by calculating the TF-IDF for each term in the document and the equation is applied. In this matrix, each column represents a document and each row in the column represents a term frequency in that document.

Tft1, d3 = (1+ tft1, d3) (1)

Then Idf t1,d3 = 1+ (N / df t) (2)
Then we use a log to the base 10, to diminish the values of the results since we are dealing with a huge number of documents and terms [24]. Finally, the TF-IDF calculated by

w = (1+ tft1, d3)(1+ (N / df t) (3)

And this is applied for every term in the document. This step is essential to perform the Singular value decomposition SVD.

3.1.3 Singular Value Decomposition SVD

The purpose of the SVD procedure is to perform dimensionality reduction. The machine learning library in python tool contains an implementation of the singular value decomposition (SVD) that can handle enormous matrices. The singular value decomposition takes an m x n matrix and returns three matrices that approximately equal it when multiplied together
M (m x n) = U(m x k) S(k x k) VT(k x n) (4)

Where m, n, k are the number of the document, the number to terms and the number of concepts respectively [20]. In this paper, we use one document to extract its important terms to generate a summary. A key insight of LSA is that only a small number of concepts are important to represent the data.

3.1.4 Latent Semantic Analysis LSA

LSA is a fully automatic mathematical technique that transforms the original data into a different space so that two (or more) words about the same concept are grouped. LSA achieves this by Singular Value Decomposition (SVD) of the term-document matrix. The first step is to represent the input document as an m × n matrix (A). Each row in A matrix represents a term and each column represents a sentence A = [a1 j, a2j, . . .anj]. The cell value (aij) represents the importance of the word. Each cell contains the TF-IDF frequency with which the word of its row appears in the text denoted by its column. Next, LSA applies singular value decomposition (SVD) to the matrix. This is the mathematical generalization of which factor analysis is a special case [21].

3.2 Methodology Using DRBM

The RBM framework consists of four main phases first: data pre-processing which made to reduce the document size by removing the unwanted data like stop words and punctuation marks then applying stemming algorithm to reduce the word size by finding its root and stem. Second: Feature vector extraction like sentence features, term weight, and sentence length. Third: Feature matrix generation. Finally: a deep learning algorithm applied using RBM method and enhanced feature matrix obtained from the deep learning phase which used in summary generation phase. Next figure shows the DRBM system Architecture.

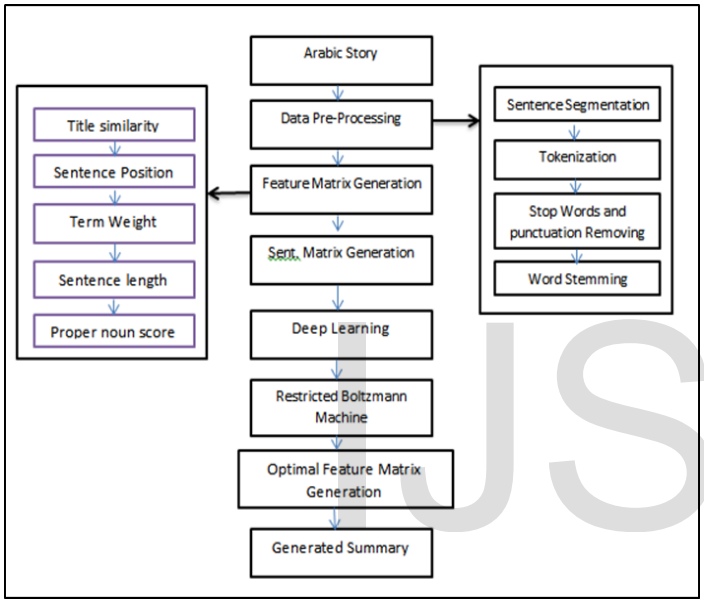

Fig 2: DRBM system Architecture.

3.2.1 Pre-processing

In this phase, we will repeat the same steps as done before. The data pre-processing will contain the text segmentation, the words tokenization, stop words and punctuation removing then stemming. This step is important to remove unwanted data from the text.

3.2.2 Feature Vector Extraction

The document prepared in the pre-processing phase is now structured into a matrix. A sentence matrix M of order n*v contains the features for every sentence of a matrix. Here, 'n' is the number of sentences in the document and 'v' is the number of features [12]. Five features are extracted of a sentence of text document namely: Title Similarity, Positional Feature, Term Weight, Sentence Length and Proper Noun Score.

i. Title Similarity

A sentence in the text is said to be important for the summary if it is similar to the title of the document. The similarity is calculated based on the common words occurring in the title of the text and sentences of the text. The title similarity is calculated using the sentence score which is defined as the ratio of the number of common words occurring between the title and the sentences in the text to the total number of words of the text3. The feature sentence of a sentence is said to be good if it has a maximum number of words common to the title.

ii. Sentence Position

The position of the sentence can determine the importance of the sentence for the summary. Usually, the sentences that appear in the starting and end of the text are of more importance. So, based on this the sentence score is calculated. The positional score in our case is calculated by considering the following conditions:

p2 = 1, if the sentence appears in the starting part of the text

p2 = 0, if the sentence appears in the middle part of the sentence

p2 = 1 if the sentence appears in the last part of the sentence Here, we have considered 20% of sentences from the start and the end of the text as important and marked their p2 value as1.

iii. Term Weight TF-ISF

Term weight means the term frequency and its importance. The term frequency gives the total number of times the term has occurred in the whole document which depicts the importance of the term in a document. The term frequency is represented by TF(f,d) where f is the frequency of the word and t is the text document. The inverse sentence frequency ISF tells whether the term is common or rare across the document. The total term weight TF-ISF is calculated by computing these two concepts.

iv. Sentence Length

The sentence length decides the importance of the sentence in summarization. The sentences that are too short do not give much information about the document. Whereas sentences that are too long will have unnecessary information about the document that will not be useful for summarization.

v. Proper Noun Score

In the process of summary generation, an important role is played by the Proper Nouns. It gives information regarding, to whom or to what the author is referring. Roles played by individuals or locations description will be a different number of times in a document.

3.2.3 Feature Matrix Generation

The above-calculated features' values are then stored in a matrix form where the columns represent the features and rows represent the sentences.

3.2.4 Deep Learning Algorithm

In this paper, we are using RBM for deep learning. The sentence matrix containing a set of feature vectors is given as input to the RBM phase as a visible layer [12].

First Let S be a set of sentences

S=(s1,s2,….,sn)

where,

si = (f1,f2,……..f4), i<= n

Where, n is the number of sentences in the document. RBM contains two hidden layers and for them two set of bias value is selected namely H0 and H1: 1. H0={h1,h2,…,hn} 2. H1={h1,h2,…,hn}
These are sets of bias values which are randomly selected; the whole operation is performed with these two sets. Next, the same procedure will be applied to this obtained refined set to get the more refined sentence matrix set with H1 and which is given by:
s''=(s''1,s''2,….,s''n)
After obtaining the refined sentence matrix from the DRBM it is further tested on a particular randomly generated threshold value for each feature we have calculated.

3.2.4 Enhanced Feature Matrix and Summary Generation

An Enhanced feature matrix is obtained from the deep learning phase which is now used for the further summary generation phase. Now the summary can be generated

4. Arabic Text Summarization System Implementations

4.1 Implementation by Using CA-LSA

The Arabic test collection for the proposed approaches involves two datasets which are an Arabic document from Arabic educational textbook and the second from Arabic Story. We used Python tool for data pre-processing, with NLTK library. In this study, two summaries have been generated for each document the first one generated by LSA approach.

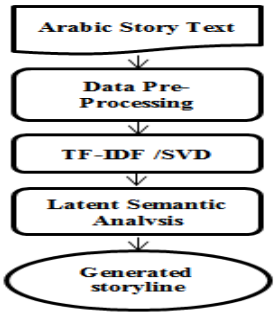

Fig 3: CA-LSA System Architecture

4.2 Implementation Using Deep Learning based on DRBM

In the implementation phase, we worked in Python using NLTK with Anaconda which is a free and open-source distribution of the Python programming languages for scientific computing [23]. The implementation starts with reading the Arabic text file which consists of a long story. Then prepare the data with the pre-processing step. Then we generated the feature matrix by calculated features values which then stored in a matrix form where the columns represent the features and rows represent the sentences. The sentence matrix having a set of feature vectors given to the DRBM as an input. The Deep RBM extracts the important features and generates the important storylines which are the summarized text.

5. Experiments and System Evaluation Results:

In this study, we applied two exesting approaches the CA-LSA and DRBM in a single Arabic document for extractive text summarization or storyline generated. The datasets used to create single document summaries was extracted from Arabic Education and storybooks. Moreover, the manual evaluation used which requires a group of human judges to read the whole original Arabic text and then read the generated summaries. And give their judgement for extractive summaries of those articles according to its quality. Each participant was given a document with two summaries: one generated using CA-LSA, and the second using Deep Restricted Boltzmann Machine DRBM. The participants evaluate each summary with (National Institute of Standards and Technology) NIST [22] which have driven several criteria for the judges to evaluate the summary performance like: grammatically, content and readability measures, non-redundancy, reflection clarity, focus, strong structure, robust and coherence.

TABLE 1: EVALUATION SCORE INTERPRETATION

Evaluation	Score	Interpretation
V. Poor	1	The summary is not related to the document at all.
Poor	2	The core meaning of the document is missing
Fair	3	The user is somehow satisfied with the summary, but he/she expects more.
Good	4	The summary is readable and it carries the main idea of the document.
V. Good	5	The summary is very readable and focuses more on the core meaning of the Document. The user is totally satisfied with the summary.

The manual evaluation Score defined interpretation as shown in table (1), (V. Poor 1), The summary is not related to the document at all (Poor 2), The core meaning of the document is missing (Fair 3), The user is somehow satisfied with the summary, but the user expects more (Good 4), The summary is readable and it carries the main idea of the document (V. Good 5), The summary is very readable and focuses more on the core meaning of the document. The user is totally satisfied with the summary. The judgments result obtained from the questionnaires for the methods that used LSA and RBM. The results shown in table 2 and it shows that the summarization with DRBM comparatively better than the summary generated by CA-LSA approache.

TABLE 2: CA-LSA AND DRBM EVALUATION SCORE

	Readability	Non-Redund.	Clarity	Strong structure	Robust
CA-LSA	2	0	1	1	0
DRBM	5	3	4	4	5

6. Conclusion:

In this paper, we investigate an automatic Extractive text summarization and compare between two different approaches to find which one give the better results in Arabic text summarization. The frameworks discussed for the single input Arabic text with Extractive text summarization. This research depends on document clustering using latent Semantic Analysis LSA which consuming Singular Value Decomposition SVD for sentence selection and filtering to select one sentence from each set of similar sentences. To enhance the summarization results we used deep learning Approach with Restricted Boltzmann Machine RBM method. By the experiment the Deep Neural Networks with RBM give better results than the Neural Networks using LSA. We evaluate the outputs quality from text summarization using manual evaluation by human judgments. The evaluation score from 1 to 5 where 1 is poor and 5 is very good as shown in figure (3), the evaluation was very good using text summarization using DRBM than CA-LSA.

REFERENCES

[1] Binwahlan, M.S.; Salim, N.; Suanmali, L.: "Fuzzy swarm diversity hybrid model for text summarization", Inf. Process. Manag. 46(5), 571–588 (2010).

[2] Qumsiyeh, R.; Ng, Y.-K.: "Enhancing web search by using query-based clusters and multi-document summaries". Knowl. Inf.Syst. 47(2), 355–380 (2016).

[3] Sarkar, D.: "Text Summarization. In: Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from your Data", pp. 217–263. Apress, Berkeley, CA (2016)

[4] Al-Sabahi , Z. Zhang1,J. Long1 , K. Alwesabi: "An Enhanced Latent Semantic Analysis Approach for Arabic Document Summarization", Arabian Journal for Science and Engineering,(2018)

[5] Al Qassem, L.M.; Wang, D.; Al Mahmoud, Z.; Barada, H.; Al- Rubaie, A.; Almoosa, N.I.: "Automatic Arabic summarization: a survey of methodologies and systems". Proc. Comput. Sci. 117, 10–18 (2017).

[6] Hammo,B.H.: "A hybrid Arabic text summarization technique based on text structure and topic identification".(2011)

[7] Ferreira, R.; de Souza Cabral, L.; Lins, R.D.; Pereirae Silva, G.; Freitas, F.Cavalcanti, G.D.C.; Lima, R.; Simske, S.J.; Favaro, L.: "Assessing sentence scoring techniques for extractive text summarization". Expert Syst. Appl. 40(14), 5755–5764 (2013).

[8] Triantafillou, E.; Kiros, J.R.; Urtasun, R.; Zemel, R.: "Towards generalizable sentence embedding. In: Proceedings of the 1st Workshop on Representation Learning for NLP", Berlin, Germany, (2016)

[9] A. Sana, M. Saleh and A. Osama, "Semantic Sentiment Analysis of Arabic Text", International Journal of Advanced Computer science and Applications (IJACSA) , Vol .8, No .2 (2017).

[10] A. Hoto, S. Staab and G.Stumme, "WordNet Improves Text Document Clustering", (2003).

[11] Z. Elberichi, A. Rahmon and M.A., "Using WordNet for text Categorization", International Arab Journal of Information Technology, Vol .5, (2008).

[12] S.A. Younif, V.W.Smawai, I. Elkbani and R Zantout,"The Effect of Combining Different Semantic Relations on Arabic Text Classification", Word of Computer Science and Information Technology (WCSIT), Vol.5 , No.6 ,(2015).

[13] G. Gutam and D. Yaduv, "Sentiment Analysis of twitter Data Using Machine Learning approaches and Semantic Analysis", Seventh International Conference IC3, (2014).

[14] Wu, Z.; Lei, L.; Li, G.; Huang, H.; Zheng, C.; Chen, E.; Xu, G.: "A topic modeling based approach to novel document automatic summarization". Expert Syst. Appl. 84, 12–23 (2017).

[15] S. Sayed Priti S. Sajja, "Literature Review on Extractive Text Summarization Approaches, International jurnal of computer application", (volume 156) (2016).

[16] Hachey, B., Murray, G., and Reitter, D. (2005). The Embra system at DUC 2005: "Query oriented multi-document summarization with a very large latent semantic space". In Proceedings of the Document Understanding Conference (DUC) 2005.

[20] Al-Saleh,A.B.; Menai, .E.B.: "Automatic Arabic text summarization: a survey". Artif. Intell. Rev. 45(2), 203–234 (2016).

[21] Nadera, B.: "The Arabic natural language processing: introduction and challenges". Int. J. Engl. Lang. Transl. Stud. 2, 106–112 (2014)

[22] Froud, H.; Lachkar, A.; Ouatik, S.A.: "Arabic text summarization based on latent semantic analysis to enhance Arabic documents clustering".arXiv preprint arXiv:1302.1612 (2013)

[23] https://en.wikipedia.org/wiki/Anaconda_(Python_distribution)