Dynamic Document Representation Using Auto Summarization

Shreya Desai[1], Shraddha Ravishankar[2], Rohit Dsouza[3], Abhijit Joshi[4]
[1, 2, 3, 4] Dept of Information Technology
[1, 2, 3, 4] D. J. Sanghvi College Of Engineering, Mumbai, India

Abstract- *With the diverse functionality of the internet and the immense wealth of information available on the web, our dependence on it as a primary source of knowledge has grown over the years. However, the overwhelming number of indexed sources can often leave us with incomplete information on the topic at hand. Further, the task of sifting through numerous pages of data can be time consuming and inefficient. In order to address this issue, we designed an auto-summarization system, which provides a concise document covering all the important points, relevant to the topic of research. This is achieved by intelligently scanning the web pages, while ignoring data that may not be a part of the core topic such as advertisements and banners. The system acquires the crucial data from the web pages and avoids any sentences those are trivial to a summary, as well as sentences that are repeated, in order to provide a gist of the entire topic. Further, it uses this summarised data to prepare power point slides, which are useful for the purpose of presentation. Consequently, the auto-summarisation system tries to learn from past history to make it more effective by increasing the accuracy of summarization process.*

Keywords- voice controlled, auto-summarisation, speech to text

I. INTRODUCTION

The field of Artificial Intelligence (AI) has not reached its full potential. AI attempts to create a system, which is known to be capable of exhibiting intelligent behaviour. In early days, the general view about AI was just to capture human like behavior only. The main view in AI became more and more that of a central information processor. But this view is off track as per Brooks's law, which shows the fundamental flaw with this view. It states that, AI can be used to build complete world models internally and then manipulate them through controlled situations [1].

Another big chunk of classical AI has been the parallel approach. Computers are fast serial computation machines. In nature, massive parallel slow computations are used. In AI, neural networks tried to incorporate this behaviour. But artificial networks learn slowly and the learning rate is tuned by hand, unlike its natural counterparts. Researchers have asked then, what is intelligence? The best known definition is the one from Alan Turing. Roughly this definition states that, when an observer cannot make a distinction between the behaviour of a computer and the behaviour of a human or other animal, then one must conclude that computer has the same level of intelligence as the animal in question [2].

Over the years, AI has rarely been used as a tool in the work environment. The implementation of AI in our daily lives has been restricted to mobile based voice assistants, entitled with tasks such as setting alarms or setting reminders. Our research work aims at introducing AI in to our daily workspace, not only as a voice assisted search engine but also as a tool to collaborate with, through the traditional stages of a project. Assisting the user through the phases of initiation, planning and execution, the software will intuitively guide the user and also enable a more efficient use of time by performing seemingly mundane tasks such as creating word documents or presentations that might be critical to the project. For example, a user may be faced with the task of researching an unknown topic, important to a module of the overall project. In this case, our software, once commanded, scans the web for the most reliable sources and picking only that information, which may be considered relevant and avoiding common mistakes such as conceptual repetition, creates a comprehensive document, which can easily be used for reference. Additionally adapting to the users patterns, it serves to provide a more personalized experience by learning with increased use. The ultimate goal of the research work envisions a world where human and artificial intelligence, co-exist independently and continue to grow simultaneously with time.

The rest of the paper is organized as follows. Section II covers the existing work in this regard along with the survey conducted by us. Section III describes our approach at providing a solution to this problem along with its potential benefits. Section IV presents the implementation details of our system. Section V presents the results obtained and their analysis. The paper ends with the conclusion and future scope.

II. LITERATURE REVIEW

Before reviewing the existing work done in this area, let us see the opinion of people about the research work undertaken by us.

- *Survey*

A survey was conducted using Google forms, over a period of three months. For that, a sample questionnaire was prepared comprising of following set of questions.

1. How often do you use a voice based assistant on a mobile device or personal computer?
2. What is your primary purpose for using such an application?
3. Do you feel that every function of the application provides satisfactory results?
4. Do you think a voice based assistant could possibly be a useful tool in project development?
5. For a work based environment, which of the following features would you wish to include?
6. If implemented effectively, would you like to use such an assistant for projects?

This survey was conducted amongst a sample population of 167 people, which includes both male and female. The targeted audience spanned across an age group of 18-45. Inferences drawn from the survey are as follows.

- The survey revealed that a majority of people do not make use of voice based assistants more than once a day, with some not even being inclined towards the need to use such software.
- Frequent users state their primary reason for doing so is restricted to a few mundane tasks such as setting alarms, taking notes or simply planning their day through reminders. This indicates that the full potential of a powerful tool such as AI has not yet been completely utilized.
- Moreover, a vast array of people have also agreed upon the fact that the few applications that do provide useful voice based functionality have failed to provide efficient results, showing an overall lack of robustness and poor build quality.
- However, a large population suggested that a voice based assistant could be capable of being a fully recognized project assistant, if the right features are incorporated.
- The demographic has also expressed a keen interest in using such an application, if built successfully, thus providing the undertaking of this project with necessary impetus.

B. Existing Systems

Currently there exist systems such as Siri by Apple, Cortana by Windows and Google voice, which utilises one or more features mentioned as a part of our system. But they suffer from the following drawbacks.

1. Inefficiency of Auto-Summarisation:

Many of the tools available can be used exclusively online and thus require users to be constantly connected to the internet. The algorithms focus primarily on the use of keywords and provide higher weightage to the more obscure word without considering the context of the sentence in which it is used.

2. Limited use of voice based application:

Existing applications have fixed responses for a set of questions, which have been hard coded previously. Therefore, a response for anything beyond the domain of the application provides unsatisfactory results.

3. Minimum collaboration:

The response generated for any particular query redirects the user to a search engine result page. This provides the user with a myriad of repetitive and sometimes irrelevant data that is difficult to channelize. Thus, collaboration between the user and the query responder is similar to the one between a user and any internet based search engine.

4. Domain specific machine learning:

Current applications for machine learning are limited to particular field in which they are developed. Thus, an application cannot be used for a varied audience, which limits it's use subsequently the scalability.

5. Adaptability:

Current systems rely on regular updates from the developers in order to augment the existing knowledge base. There is no scope for user-dependent learning and adaptation, making these applications rigid. One cannot personalize such an application and hinders the users' interaction.

Keeping in mind all these issues in existing systems and the positive outcome of the survey, it was decided to develop an auto-summarisation system, which dynamically generates a document as per the user's need. Now let us see the details of our approach in next section.

III. OUR APPROACH

We have developed a web based application that uses voice recognition to interact with the user and assist the user in a work based environment. The application serves as a voice based assistant, recognizing commands as well as simple speech that is processed using Natural Language Processing (NLP) techniques. Besides simple commands and operations, it also utilises web based content to perform a variety of tasks, encountered by an array of people ranging from students to executives. Students can use this software as a tool for research, saving hours of searching the web by using Web-To-Doc feature, which extracts web based information to provide a clean, well designed Word document. Additionally the Doc-To-Ppt feature can readily convert the Word document into a PowerPoint presentation not just by converting paragraphs into bullets but by only choosing those points, which manage to capture the purpose of the document in a lucid, concise format. At the same time, executives and project managers can use the functionality provided by our software to systematically plan out the path to approaching a problem definition. The system recognizes sentences and converts them into flowcharts, allow users to track their thoughts and manage them along the way, which makes sure that no ideas are left out.

With our approach, it is possible to integrate seamlessly all these features into a single environment making each of these independent tasks conveniently blend into each other, making for a better and smoother user experience.
In the E-Learning scenario, interaction is a very important aspect. So, rather than using a conventional keyboard and mouse, we have chosen the most natural choice of interaction, i.e., language. This makes the application more user-friendly than ever, adding layers of accessibility and intuitiveness, ensuring that it can be used by everyone without ever feeling the need to refer to any manual or tutorial.

Figure 1 shows the architecture of our system. As seen in Figure 1, the input to the system could either be through voice or text. The function of each module is detailed as follows.

Voice Recognizer: This module processes the voice based input from the user and works with the Text Parser to identify individual words and commands, linking them into atomic statements or sentences. First, the speech input acquired from the user is pre-processed and sent to a feature extractor. The feature extractor verifies the input and submits it to the aforementioned Text Parser, which accordingly generates the corresponding text [3].

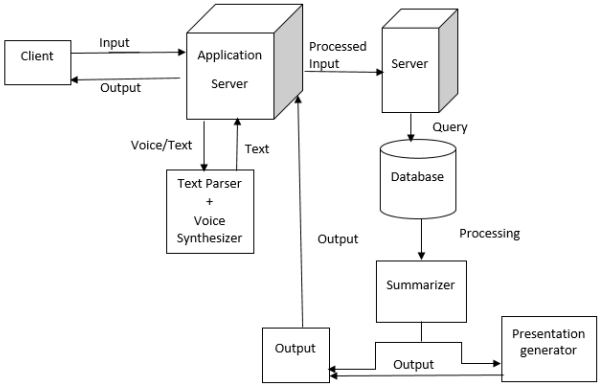


Figure 1: Architecture of the System

Text Parser: As mentioned earlier, this component works in collaboration with the Voice Recognizer, also accepting the direct text based input from the user. The Text Parser is directly responsible for comparing the keywords verified before, from those available in it's local database. Depending on whether a match is found, the voice based input is either accepted or rejected.

Auto-Summarization: The output generated by the text Parser is then submitted to the Auto-Summarization module. This module is responsible for extracting data from the web, removing redundant information, deciding the degree to which information needs to be retained and providing the user with the concise document.

The output received from the Text Parser is used as the primary query to fetch the required data from the web. In order to use the most relevant sources, we implemented Google's PageRank algorithm [4][5][6], which ranks web pages in the order of their relevance. Selecting only the most relevant web pages, the system extracts the required data using a Python library called BeautifulSoup [11]. This library allows us to select data only pertaining to the query, while ignoring unwanted data such as advertisements or banners. Once the data is obtained, it is processed using our Auto-Summarization algorithm to create a concise summary [7][8][9].

The algorithm works as follows –

1. Set the lower bound and upper bound as 0.2 and 0.8 respectively, in order to calculate the frequency of words. The above limits are selected in order to ensure that words with very low frequencies as well as high frequency English articles such as 'a', 'the', 'an', etc. are eliminated.
2. Using stop-words, common words and unimportant symbols like '\' are eliminated.
3. Relations between the word-tokenized sentences are then checked to verify continuity and their context

with the current query. This relevance is determined by calculating the score of each statement using the *dict()* function from the NLTK package in Python [10]. The scores are then compared to determine if the two sentences are related to each other.

4. If the statement is found to be relevant, the words marked to be eliminated above are restored and the statement is retained. This is carried out throughout the previously extracted data.
5. The selected sentences are finally verified against the upper and lower bounds to ensure that outliers are eliminated and the summary is generated.

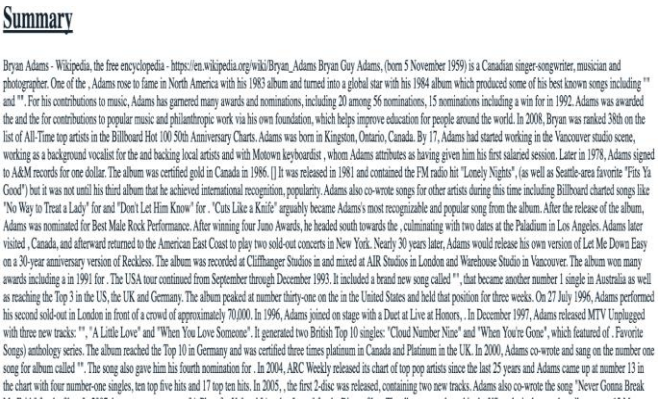
The output is then displayed to the user in a text-based format. The implementation details are discussed in the next section.

IV. IMPLEMENTATION

Figure 2 shows the snapshot where in the user can provide the input in the form of text or voice.

Figure 2: Input for the summary generation given through voice

Here, the user has given the input in the form of voice. The input has been identified by the system as 'Bryan Adams', as displayed in the snapshot. The interface provides the user with two options, 'Summarise', which generates the summary, and 'PPT', which designs the presentation. In this case, the user clicks on the 'Summarise' option, thus generating the summary shown in Figure 3.


Summary

Bryan Adams - Wikipedia, the free encyclopedia - https://en.wikipedia.org/wiki/Bryan_Adams Bryan Guy Adams, (born 5 November 1959) is a Canadian singer-songwriter, musician and photographer. One of the , Adams rose to fame in North America with his 1983 album and turned into a global star with his 1984 album which produced some of his best known songs including "" and "". For his contributions to music, Adams has garnered many awards and nominations, including 20 among 56 nominations, 15 nominations including a win for in 1992. Adams was awarded the and the for contributions to popular music and philanthropic work via his own foundation, which helps improve education for people around the world. In 2008, Bryan was ranked 38th on the list of All-Time top artists in the Billboard Hot 100 50th Anniversary Charts. Adams was born in Kingston, Ontario, Canada. By 17, Adams had started working in the Vancouver studio scene, working as a background vocalist for the and backing local artists and with Motown keyboardist , whom Adams attributes as having given him his first salaried session. Later in 1978, Adams signed to A&M records for one dollar. The album was certified gold in Canada in 1986. [] It was released in 1981 and contained the FM radio hit "Lonely Nights", (as well as Seattle-area favorite "Fits Ya Good") but it was not until his third album that he achieved international recognition, popularity. Adams also co-wrote songs for other artists during this time including Billboard charted songs like "No Way to Treat a Lady" for and "Don't Let Him Know" for . "Cuts Like a Knife" arguably became Adams's most recognizable and popular song from the album. After the release of the album, Adams was nominated for Best Male Rock Performance. After winning four Juno Awards, he headed south towards the , culminating with two dates at the Paladium in Los Angeles. Adams later visited , Canada, and afterward returned to the American East Coast to play two sold-out concerts in New York. Nearly 30 years later, Adams would release his own version of Let Me Down Easy on a 30-year anniversary version of Reckless. The album was recorded at Cliffhanger Studios in and mixed at AIR Studios in London and Warehouse Studio in Vancouver. The album won many awards including a in 1991 for . The USA tour continued from September through December 1993. It included a brand new song called "", that became another number 1 single in Australia as well as reaching the Top 3 in the US, the UK and Germany. The album peaked at number thirty-one on the in the United States and held that position for three weeks. On 27 July 1996, Adams performed his second sold-out in London in front of a crowd of approximately 70,000. In 1996, Adams joined on stage with a Duet at Live at Honors, . In December 1997, Adams released MTV Unplugged with three new tracks: "", "A Little Love" and "When You Love Someone". It generated two British Top 10 singles: "Cloud Number Nine" and "When You're Gone", which featured of . Favorite Songs) anthology series. The album reached the Top 10 in Germany and was certified three times platinum in Canada and Platinum in the UK. In 2000, Adams co-wrote and sang on the number one song for album called "". The song also gave him his fourth nomination for . In 2004, A&C Weekly released its chart of top pop artists since the last 25 years and Adams came up at number 13 in the chart with four number-one singles, ten top five hits and 17 top ten hits. In 2005, , the first 2-disc was released, containing two new tracks. Adams also co-wrote the song "Never Gonna Break

Figure 3: Output - Summary generated by the system

Figure 4 shows the snapshot of the intermediate text file generated by the system, when the user selects the 'PPT' option.

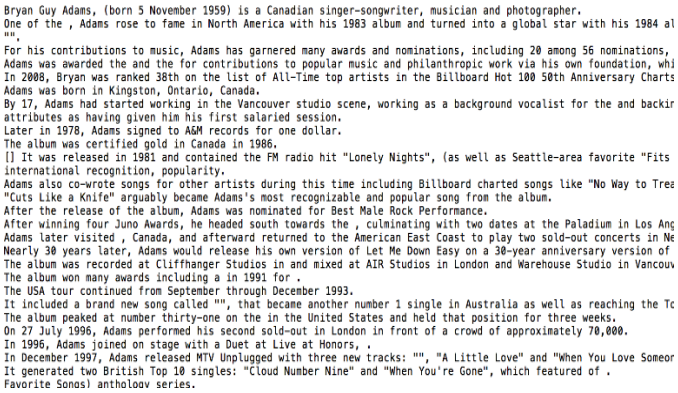

Bryan Guy Adams, (born 5 November 1959) is a Canadian singer-songwriter, musician and photographer.
One of the , Adams rose to fame in North America with his 1983 album and turned into a global star with his 1984 al
"",
For his contributions to music, Adams has garnered many awards and nominations, including 20 among 56 nominations,
Adams was awarded the and the for contributions to popular music and philanthropic work via his own foundation, whi
In 2008, Bryan was ranked 38th on the list of All-Time top artists in the Billboard Hot 100 50th Anniversary Charts
Adams was born in Kingston, Ontario, Canada.
By 17, Adams had started working in the Vancouver studio scene, working as a background vocalist for the and backir
attributes as having given him his first salaried session.
Later in 1978, Adams signed to A&M records for one dollar.
The album was certified gold in Canada in 1986.
[] It was released in 1981 and contained the FM radio hit "Lonely Nights", (as well as Seattle-area favorite "Fits
international recognition, popularity.
Adams also co-wrote songs for other artists during this time including Billboard charted songs like "No Way to Trea
"Cuts Like a Knife" arguably became Adams's most recognizable and popular song from the album.
After the release of the album, Adams was nominated for Best Male Rock Performance.
After winning four Juno Awards, he headed south towards the , culminating with two dates at the Paladium in Los Ang
Adams later visited , Canada, and afterward returned to the American East Coast to play two sold-out concerts in Ne
Nearly 30 years later, Adams would release his own version of Let Me Down Easy on a 30-year anniversary version of
The album was recorded at Cliffhanger Studios in and mixed at AIR Studios in London and Warehouse Studio in Vancouv
The album won many awards including a in 1991 for .
The USA tour continued from September through December 1993.
It included a brand new song called "", that became another number 1 single in Australia as well as reaching the To
The album peaked at number thirty-one on the in the United States and held that position for three weeks.
On 27 July 1996, Adams performed his second sold-out in London in front of a crowd of approximately 70,000.
In 1996, Adams joined on stage with a Duet at Live at Honors, .
In December 1997, Adams released MTV Unplugged with three new tracks: "", "A Little Love" and "When You Love Someor
It generated two British Top 10 singles: "Cloud Number Nine" and "When You're Gone", which featured of .
Favorite Songs) antholoov series.

Figure 4: Contents of the intermediate text file

The intermediate text file is further used by the system to generate the presentation. The output presentation generated by the system is shown in Figure 5.

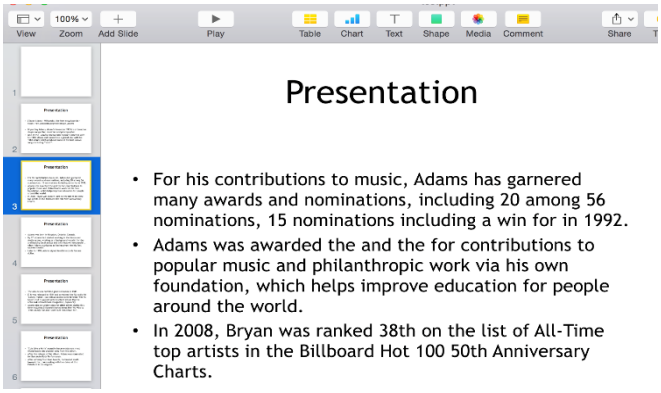


Figure 5: Output - Presentation generated by the system

V. RESULTS ANALYSIS

In order to measure the accuracy of our system, we carried out a comparison between the output generated by our

model and a few available examples. Figure 6, 7, 8 and 9 shows the presentation slides used for comparison.

On conducting an analysis of the summary provided by our system, we can conclude that the summary provide a well-defined overview of the query. To confirm this, we compared it with a summary reviewed by 1,203 people and widely deemed self-sufficient. Our findings suggest that the summary generated by the system substantially covers the key points of the topic provided as the query. Throughout our analysis the system consistently achieved an accuracy of 65% to 75%. Our survey, conducted to test the quality of the output, amongst 48 women and men, revealed that approximately 94% (45) of the survey demographic were more than satisfied with the output. Alternately, the remaining 3 people suggested that while all the major points were well specified a few more points could be included in addition. From the comparison of a sample presentation widely accepted by users and the presentation generated by the system, we can infer that the presentation generated by the system meets the requirements of the user. The points are independent, well-distributed and help to define the topic in a concise manner. Each point reveals more information about the topic and the system displays very little redundancy. Once again, the survey demographic provided substantial evidence that the system prevailed with high accuracy. Fairing even better, almost 96% (46) were satisfied with the output, without feeling the need to edit any of the slides in the presentation.

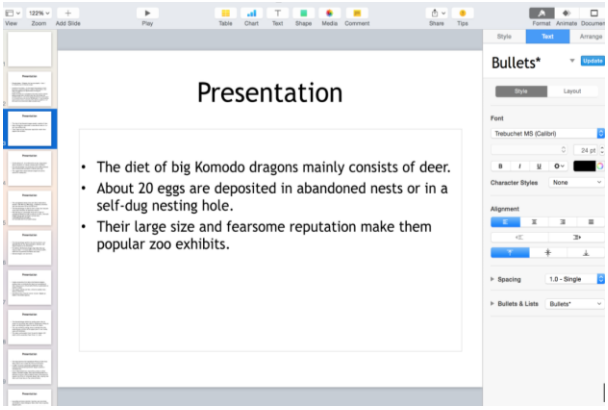


Figure 6: Output - Presentation generated by system for komodo dragons

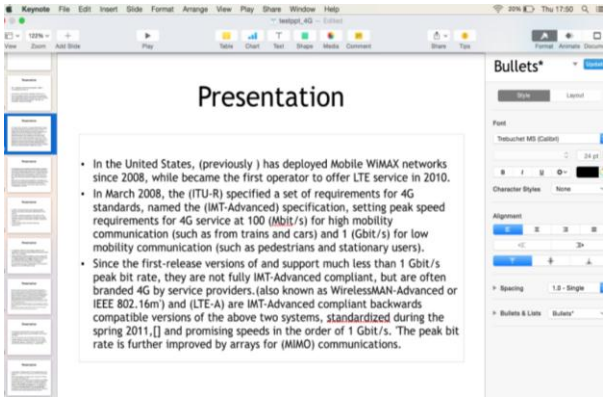


Figure 7: Output - Presentation generated by system for 4G

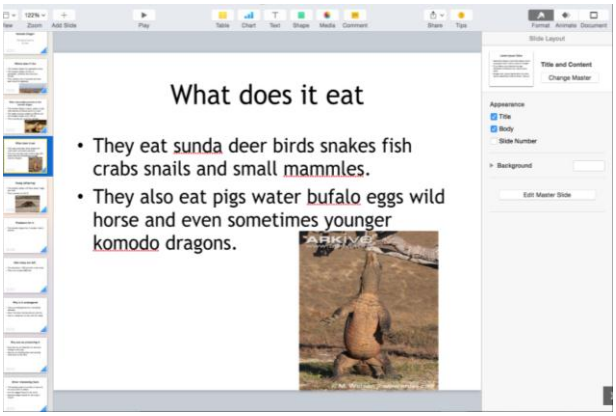


Figure 8: Presentation for komodo dragons

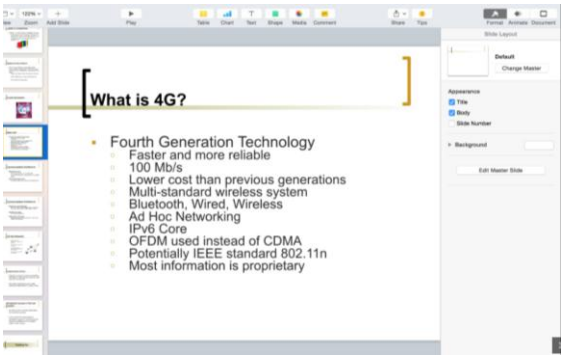


Figure 9: Presentaion for 4g

VI. CONCLUSION AND FUTURE SCOPE

With our system, we expect to bring the functionality of auto-summarisation and it's applications into the study/work environment. The versatile mode of operation of the system and feature set allow a smooth flow of information from the web to a ready-to-go document set. With the capability to improve in performance and it's platform independence, the system can run on any machine enabled with an internet connection and persist over time. Moreover, the voice based functionality makes this system apt for multi-tasking as well as the physically disabled. Thus, we can

conclude that this system can benefit a wide audience and play a pivotal role in saving time on a daily basis.

After analysing the output of our system, we can infer that the accuracy of the system has scope for improvement in the future. However, this cannot be achieved by a static algorithm. On integrating the algorithm with a neural network, the quality of the summary and accuracy of the system could be pushed even further over time. This would only be possible after prolonged use of the system and based on the feedback received from the user. Moreover, while the current system uses the algorithm for summarising web documents and generating presentations, the application of the algorithm can be expanded to cover a wider array of documents. This could prove to be helpful especially for summarising large pieces of information such as news articles or entire books, in order to enable faster consumption of the source material.

REFERENCES

[1] Ronald Chrisley, Sander Begeer: Artificial Intelligence: Critical concepts, 2000.
[2] Conventional AI : http://www.ai.rug.nl/~alberth/afstudeer/scriptie/scriptie/node5.html
[3] Text Prompted Remote Speaker Authentication: http://www.slideshare.net/gtiwari/text-prompted-remote-speaker-authentication-joint-speech-and-speaker-recognitionverification-system-midterm-project-presentation
[4] http://www.sirgroane.net/google-page-rank/
[5] Web Speech API, Google Developers: https://developers.google.com/web/updates/2013/01/Voice-Driven-Web-Apps-Introduction-to-the-Web-Speech-API?hl=en
[6] Page Rank Algorithm in Python : http://dpk.io/pagerank
[7] Automated text summarization in summarist: http://research.microsoft.com/en-us/um/people/cyl/download/papers/mit-book-paper-final-cyl.pdf
[8] Summarization handbook by Edward Hovy: http://www.isi.edu/natural-language/people/hovy/papers/05Handbook-Summ-hovy.pdf
[9] Evolutionary Algorithm for Extractive Text Summarization: http://www.scirp.org/journal/PaperInformation.aspx?paperID=902
[10] Nltk documentation: http://www.nltk.org/api/nltk.metrics.html
[11] BeautifulSoup documentation: https://www.crummy.com/software/BeautifulSoup/bs4/doc/