DT211 Year 4, Project Proposal Form

Student Name: Yassr Shaar	**Student Number:** C14328571

Project Title

Analysis and Summarization of Text & Audio using Natural Language Processing & Machine Learning

Summary (approx 200 words)

Summarization is to shorten a text document to make it much easier to consume and straight to the point. It identifies what an informative or important sentence is and creates a summary paragraph by extracting those sentences from the text document or any wall of text that is presented. When this is achieved with a software it needs to take into consideration many variables such as length, writing style and syntax.
With the use of Natural Language Processing we can extract the most informative sentences or paragraph and display it to the user, this can optimize the time it takes the user to do research on a topic or find answers within a wall of text.

The project will identify the different algorithms of Summarization and compare them to find what the best algorithm that can be used is. Trying to recreate this algorithm or improve it through machine learning for optimization and efficiency. Once that is achieved the project will perform the task of extracting an audio file from a video or on its own and transcribe it to text which will then utilize the algorithm for summarization and create a summary based on what was discuss in the video or audio file. This can then be taken a step further and using the timestamps of the summarized text the audio can be clipped to reflect the effects of summarization onto the audio itself.

DT211 Year 4, Project Proposal Form

Background (and References)
The idea for this project came while I was doing research on Natural Language processing. I found this topic to be very interesting as there are many approaches to it and the final product from it would be very beneficial.

Who would Benefit or Use this?
The program would create an application that would greatly benefit **Students** as they would consume online courses and video lectures almost daily. By creating a text summary and an Audio/Video summary of the lesson the application makes it much easier for students to consume the knowledge that is presented in the audio or video.

Lecturers who review video or audio research can save time by using the application as it would reduce the amount of content they would have to review by summarizing it.

Casual users who listen to podcasts or watch videos but would rather have a summary of their content with the most informative parts.

After considering the best options for programming I have decided to implement this project using Python. This would be the best option for this project as there are Natural language kits and tools that would make analysing the text documents more effective.
The idea for this project was to research summarization and create a web application that would demonstrate it with online articles but after meeting with my supervisor – Paul Doyle he suggested to me a way to increase the complexity of the project which was to implement the summarization onto audio or video files after extracting their text content and analysing it.

References and Related Links:
http://duc.nist.gov/ Document Understanding Conferences – Data Set
https://tac.nist.gov/ Text Analysis Conference – Data Set
http://rgcl.wlv.ac.uk/projects/CAST/ Computer-Aided Summarisation Tool – Data set

http://www.nltk.org/ Natural Language Tool Kit
http://www.pydev.org/ Python Development Tool Kit

https://dev.havenondemand.com/apis/recognizespeech
https://www.mashape.com/
} Audio to Text API

https://gist.github.com/alotaiba/1730160
https://goo.gl/uNyW6c
} Google Audio to Text API

I have also enrolled in this course to help educate myself on machine learning & NLP
https://www.udemy.com/from-0-1-machine-learning/

DT211 Year 4, Project Proposal Form

Proposed Approach

I will have three main stages in my proposed approach for this project.

1. **Research of Summarization Algorithms & Methods** –
   This is to learn all there is about summarization and the different technics used in achieving it. Also I will need to find out what the most used algorithms of summarization are and where they are used. How do current summarization tools detect what is regarded as an "informative sentence" and why would one algorithm be chosen above others.

2. **Testing and evaluation** –
   Testing of different types of summarization algorithms and comparing them. I will have to find a measure of summarization to score the algorithms on their performance in order to decide on the most efficient method.

3. **Implementation** –
   Once the research has been complete and I have replicated the best algorithm or improved it I will implement it into an application. This application will use the algorithm to summarize the text that has been extracted from an audio file and present it to the user.

The use of Machine learning and data sets will be very important for this project. As I mentioned in the references I have done some research on data sets I can use and will be exploring my options with this regard.

I will be encompassing the use of iterative and incremental development cycles, this will help divide my work and allow me to return to the design and implementation stages when new requirements emerge.

The techniques that can be researched for this project include using dictionaries and APIs. Tools such as Natural Language Tool Kit can help with this and give a better understanding of the way a short text is structured.

Deliverables

- Interim Report.
- Project dissertation.
- Help/User manual.
- A functional application for demonstration.

<u>Priority Features</u>

- Research of Summarization technics and algorithms
- Extract an audio file from a video and using an API transcribe the audio to text
- Summarize the text from the audio extracting the most informative sentences

<u>Secondary Features</u>

- Using timestamps from where the summarization of the texts occurred clip the audio to match the summarized text
- Match the clipped audio with the clipped video and merge them together creating a summarized video file.

DT211 Year 4, Project Proposal Form

Technical Requirements

- Natural Language dictionaries
- Machine learning
- Audio to Text API
- PyDev / Python IDE
- HTML, CSS, XML, Python, Django

Project Reviews – Please include reviews of two of LAST years projects from your programme.

Project 1
Title: Cultural Bias in Wikipedia

Student: Aadam Bari

Description (brief):
Identifying the cultural biases in Wikipedia by comparing a search result on Wikipedia through different languages. Using Natural Language Processing it is possible to identify the sentiment that is expressed in the article and so the web application would translate the results and confirm whether or not there is a bias on the topic. With the aid of visualization tools the application is able to show comparisons in an interactive format for the user to see just how much of a different or a bias there is on the same result in two different languages.

What is complex in this project
Using Natural language processing libraries to identify bias in a document.
Translating and comparing the translated document to the English document.
Visualization and analysis of differences between two Wikipedia pages of the same topic.

What technical architecture was used
http://www.nltk.org/ - Natural Language Tool Kit
http://www.chartjs.org/chart - Chart js - Produces Graphs

Natural Language Processing Libraries were used along with Chart Js for visualization of the analysis and comparison between the queries.
Iterative and incremental development cycles was used throughout the project.

Explain key strengths and weaknesses of this project, as you see it.
Iterative and incremental development cycles allowed for the developer to divide their work and allows them to return to the design and implementation stage if new requirements were to emerge.

The queries had to match in both languages if not than there wasn't anything to deal with that instead the user would be told that the comparison was not possible. This is a clear weakness in the project.

Dublin Institute of Technology

DT211 Year 4, Project Proposal Form

Project 2
Title: Detecting Bot Twitter Accounts using Machine Learning

Student: Emmet Hanratty

Description (brief):
The projects purpose is to detect whether or not a twitter account is a bot. These bots can be very harmful to the everyday user of twitter. This is performed using machine learning by training the program to detect bots. A database is used to store all of the Twitter accounts capable of being queried to use in the machine learning model. The text from these accounts is translated so that they are all in the same language and then it is analysed by the program using the machine learning knowledge that it has been trained on.

What is complex in this project:
Using machine learning and training the program to detect a bot by storing many samples of bots in a database that the program can refer to when analysing a twitter account.

What technical architecture was used
Django framework, also including front-end technologies such as JavaScript, Bootstrap, HTML, and CSS. IT uses a MySQL database to store and query a user's Twitter account and display the likelihood of that user being a bot.

Explain key strengths and weaknesses of this project, as you see it.
The machine learning database is a strength of this project as it allowed the program to query and transform data from twitter and be fitted into the machine learning models.

Proposal Sign off:

Lecturer Comments

plenty of challenges. Need to find a way to evaluate algorithms. video summarisation would be a good stretch goal.

Student Signature Eoghan Hanratty	**Date** 10/10/17
Lecturer Signature Paul Stynes	**Date** 10/10/17