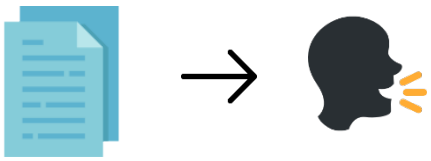
My first Machine Learning Adventure.

I loved *and tiny little bit hated* it. haha 😊

1. Frame the ML Problem:



The Idea

The idea is to create audio files and from PDF books. The idea results from the problem that I would like to listen to books while doing other physical work. If I could do that, I could use my time more efficiently and strike two birds with one stone.

In addition to that another Idea would be to summarize Text.

Inspiration

While I was collecting all the books that I wanted to transform, I realized how much I wanted to implement my idea and make it work, so that I would be able to listen to all these books. When I decided what idea I would be pursuing and as I continued to work on the project, more and more ideas came to me which I add to my project to build an entire ecosystem. A lot of these additional ideas would require Machine Learning as well and would make my ideas so much better awesome and better. For example being able to 'talk' to the book, saying things like, make me a summary, or to be able to skip a chapter or go back a sequence to listen to it again. Just in general being able to listening to the book with using voice commands.

My first idea (creating audio files) is a Machine Learning problem but one I was able to easily solve with the Google API.

2. Prepare the Data

To summarize text I I have to do 2 things to prepare my data. The first thing would be to get rid of all the extra data, such as page numbers, author biography, ISBNs, table of contents, publishing facts and so on.... Secondly I would have to collect enough data for the training and testing phase which hoe we learned could be about **80%/20%.**

Tagged data I important if I would want to train the a model to determine Meta Information about the text, like who wrote it which epoche it was written in, or what genre it is. Then I would tag the books by genre or epoch from which they come to train the model to determine such things when I give it a new book.

In the following I made a list of how and where and I collected my dataset.

- Data Format needs to be txt
- PDF files must be transformed into txt
- Data comes from free Data-Sets from Internet and Communities
- https://www.pdfdrive.com/ ← THIS is the library where I got my PDF books from
- https://cseweb.ucsd.edu/~jmcauley/datasets.html#goodreads
- http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm
- https://blog.cambridgespark.com/50-free-machine-learning-
- datasets-natural-language-processing-d88fb9c5c8da
- Aroud 0.5 million to 1 million available

3. Train the ML Model

- Even though I didn't completely solve the summarizing problem I know that I would have to use a supervised RNN - Recurrant neural network for my project. I was able to see the structure of the data preparation in the following tutorial:
- *"A recurrent neural network model is born with the capability to process long sequential data"* ← since I'm working with books this is the best solution
- the network takes time into account

4. Predict using the Model

The resulting audio books can be listened to in the car while cleaning or jut other physical work. The application could run on Mobiles, Smart Watches, MP3-Player etc. It would require a big data Set to train the model since books can

- The model probably doesn't need to be trained very often, because summarizing texts has often been done with machine learning and the community is very extended and this task has been done often.
- I don't think the model needs a lot of frequently retraining since language (text) changes relatively slowly. I would recommend every 5-10 years to

Conclusion & Learnings:

Conclusion, what works what doesn't work, what would I do if I want to work on in the future?

Conclusion:

I didn't have any ML background nor did I ever program in Python so I loved this opportunity to get to work with this. In the Github Repo I will be uploading the Python code that I've been working on even though It didn't work out that well.

What works:

- I can transmute PDF files into txt files with my own code locally on my Laptop
- I can turn text into speech with the google API
- I copied a code to summarize text from a tutorial, that worked halfway but not really well

What doesn't works:

- I can't really summarize text since I'm a new beginner with python
- Before I can create an audio book I have to prepare the text file and delete all unnecessary information
- The Google text-to-speech API doesn't sound that good (they have a better on, that I couls pay for)
- One of the biggest Problem was that the Google somehow didn't allow to let me use huge .txt files ?! I spent most of my time trying to solve this problem but still couldn't find a solution. The result is that I have to paste every single book into the code manually...

what would I do if I want to work on in the future?:

- I would love to open this project up to public and work on it with other programmers
- I also would love the implement all the additional ideas I mentioned in the beginning, because I think I could learn a lot from them, especially voice commands and machine learning
- If I had more time I would love to learn more about SSML where I can structure my text input

Learnings:

- Learning what google colab is and how I could link it to my drive.
- I learned what basic python is
- I understood the file system behind Google Colab
- I wrote my own PDF to txt transformer instead of using an online one
- I got more routine with node.js and the module/library system
- Learning that some links if they are old don't work in the code

- Learning that you have to add the exact pathnames and add own data to drive files
- Struggling big time with python but still learn something about it!
- A lot of problems you can just COPY AND PASTE
- A lot of problems I NEED MORE TIME TO BE ABLE TO SOLVE THEM

Fun Stuff:

Collecting the books was one of the nicest tasks. But I had to stop because after an hour it was consuming my time a little too much. In the end I was collecting Indian cooking books with more pictures then text, so this was when I knew I had to stop. 😊