Machine Learning Engineer Nanodegree

Capstone Proposal

Anne Line Stampe, November 15th, 2018

Text analysis of document-type text objects using NLP

Domain Background

Text interpretation is a new field of interest within the bank where I work; DNB. The bank has an ambition to digitalize the majority of analog and unstructured inputs, aiming to achieve a high degree of process automation. We still have several incoming channels which demands human handling of information, among these reading documents for extracting and processing information. We have several OCR initiatives for extracting text from pdf documents (scanned physical documents), one of which I have been responsible for. No NLP project is running for this purpose at the moment.

I have chosen NLP for text analysis and extraction as my Capstone proposal to investigate possible benefits of analyze text objects in general and, more specifically, to extract the essence of a larger text document. I will do the project using these english texts and, if successful, do an effort to try to apply the model on local DNB texts.

The project is planned to be performed using a corpus established from a set of ebooks, as the internal bank documents are not classified for external use and would have to be modified and anonymized, thus reducing the value for the analysis.

NLP is only briefly covered in the Udacity Nanodegree and the experience from lessons is limited, hence I will spend some time initially to on the subject, starting with the newly acquired book ‘Applied Text Analysis with Python’ from O’Reilly. Link to book

The project is also inspired by the GitHub project ‘Comparing Books’, Link to project

Problem Statement

DNBs process of handling several instances of incoming unstructured data, mainly documents, requires a time-consuming and skill-based human reading and interpretation of the full content. The proposed solution aim to establish a skillset and first model for text analysis, text comparison to other documents and a content summary. Re-applying this internally will hopefully provide sufficient information to process a bank transaction in a chosen process. In addition, data accumulated over time can form a basis for further analysis and increase the quality of a the solution.

For the project I see 3 distinct goals

- Analyse a text object and create descriptive data
- Compare with other texts and find ‘similar’ objects
- Create a text object ‘summary’ / topic.

Business value

- Reduce the need of human labour and skill for extracting core info
- Based on earlier project assessments this has a potential of replacing somewhere between 10 and 40% of manual effort in back-office processes.

Datasets and Inputs

Basis for this project is a robust corpus which serves the purpose of the analysis and functionality. I have chosen ebooks from the Gutenberg project to be my data set for building the corpus. They are all available, free of charge, on a .txt format.

I will load 40 full ebooks to form the corpus, and expect they will provide a wide vocabulary and a variety of language style. To be able to assess the results with a level of insight in the material I only choose books I have read myself.

The texts are not fully representative for the real-life problem, but they will at least be similar as they are full-sentence texts, not messages with an informal language. If time permits I will re-run the steps on internal DNB documents as mentioned above.

Solution Statement

The solution delivered from the project will consist of following items;

- A transformed and processed corpus established from the ebooks
- A pipeline for vectorisation and transformation
- A trained model based on the transformed corpus
- A coded interface for testing new text objects for similarity and topic extraction
- A report presenting the results, including an assessment based on insight

Benchmark Model and Evaluation Metrics

The nature of this projects implies that benchmarks must include a human assessment.

A set of metrics must be prepared even if not all results are directly calculated; eg a value from 1 to 5 for ‘similarity score’ and ‘topic precision’.

A correct benchmark would include a setup for comparing the time and effort a human spends to read text and describe content - to the coded solution. This is not planned.

The results will similarly need to be presented with more descriptive terms than for purely numerical experiments. Runtime logs will be useful for evaluating resource time and cost, although I assume these will be moderate with the planned corpus and case.

Project Design

The project structure will follow the recommended steps from the book mentioned above, although there will probably be a need for iterations within the steps :

- Building a corpus
    - Identifying the list of ebooks and loading them from the Gutenberg website
    - Using Codex for loading text
- Corpus preprocessing and wrangling
    - NLTK functions for
        - Tokenize, Tagging, Counting, Stemming and Removing stopwords
- Text vectorization and Transformation pipeline
    - Sklearn, Word2Vec or Gensim functions for
        - Vectorize, Transform, Pipeline build and Grid Search
- Classification for text analysis
    - Sklearn
        - Build and train a model, Perform simple classification
- Clustering for text similarity
    - Sklearn
        - KMeans-based clusters, Modeling text topics
- Visualisation and graph analysis
- Test and verify the code / models
- Discuss and document results
- Assessment of value in solution related to DNB use-case