Project abstracts

732A92

Topic Modelling of Movie Descriptions

As a big movie fan, I chose for my text mining project to analyse movie descriptions from Wikipedia. Since the start of the film production in the 1890's, movies has been categorized into and labeled with so-called movie genres. The division into movie genres are used daily to facilitate for movie viewers. for example it's widely used in search engines in movie-streaming services. Streaming services such as Netflix and Viaplay allow the user to influence what type of movies to display in the meny by selecting a certain genre. The aim of this project is to divide movie descriptions into natural groups and explore if the groups are consistent with the movie genres. This will be made my using the method Topic modelling with the model Latent Dirichlet Allocation. The results showed that there is a relation between group belonging for a movie description and the movie genre for some genres.

"This is not Italian!" The Science of Fake Italian Recipes

As proud Italian, the main aim of this project is to scientifically prove that some of the most famous recipes known as Italian are not actually authentic Italian. The idea behind this study is that national cuisines are made of specific combinations of ingredients which give special flavors to traditional dishes and clearly distinguish them from others. For historical reasons linked to the massive immigration of Italian people to America in the early '90s, the main fake Italian recipes found online and well known in the rest of the world (except Italy), come from USA. For this reason the dataset used here was built from labeled American and Italian recipes, scraped from well known cooking websites. The text mining task is a clustering problem based on text data of only ingredients. It will detect whether a recipe is or is not real Italian. Different text classifiers will be used: Multinomial Naïve Bayes (MNB), Linear Support Vector Machine (SVM), Decision tree, Random Forest and Neural Network. After training and validating, the classifiers will be tested on some of the most famous

variations of Italian dishes, the reasons of wrong/right classification will be explored and some more analysis about differences between American and Italian cuisine will be reported.

Predicting Votes of Short Stories

Writing has always been an artistic expression of an author, trying to put their feelings and thoughts into a story. With the purpose of entertaining the reader, the writer tries to create an atmosphere where the reader can get involved and absorbed by the story. However, what if we could predict how entertaining or well received a story would be, just by the words the writer used, or maybe the style of writing. Is there a more popular topic than others? In this study, we will test different algorithms such as Logistic Regression, Recurrent Neural Networks and Latent Dirichlet Allocation, to try to detect which are the characteristics of a popular story.

The Influence of Stop Words on Sentiment Classification of Movie Reviews

In this project models are examined for the classification into *positiv*e and *negative* of around 50000 movie reviews from the Internet Movie Database (IMDb). For this purpose, this paper investigates the influence of stop words on the classification performance for a Naive Bayes model, in contrast to a Long Short-Term Memory model (LSTM). Additionally, the impact of term frequency-inverse document frequency (tf-idf) as input for a Naive Bayes classifier is analysed. Similarly, a variation of the learning rate for the Adam optimiser in the LSTM is conducted. Overall, models excluding stop words perform slightly better. It can be seen that the use of tf-idf can improve the performance of a Naive Bayes classification. Nevertheless, a LSTM accomplishes superior results over a Naive Bayes. In particular, a LSTM with a learning rate of 0.001 and input data without stop words performs best with an accuracy of 88.18%.

Evaluating Naive Bayes and Random Forests Classifier Based on the Topic Model Outcomes of PubMed Abstract

PubMed is the most widely used database for biomedical and life sciences literature from MEDLINE. It is free and includes more than 29 million citations which may also link to the full-text content from PubMedCentral and publisher websites. Although millions of documents' abstracts are accessible by searching using different keywords, it is time-consuming to get the information behind those abstracts. In this study, we use the topic modeling method to find the groups of words of the documents

from PubMed search engine with the keywords of interest to help us save the time and effort to get the information behind the documents. Based on the weights of topic distribution for each document, Naive Bayes algorism and Random Forests algorism are used for the classification of the documents. Here, we found that the latent Dirichlet allocation (LDA) model with 16 topics showed the highest coherence score and the random forests classifier has a better performance when used to classify the documents based on the weights from topic modeling.

Wine Classification Using Convolutional Neural Networks

This project aims to build a wine recommender which can, based on a text description, suggest a grape and a country to help you selecting a suitable wine. To reduce the number of classes (and the number of parameters to be estimated in a model), the countries have been merged into two classes "Old world" and "New world", which is a common division of wine. "Old word" corresponds to the traditional wine-making countries in Europe and Middle East and "New world" corresponds to countries such as USA, Chile, South Africa and Australia. The method used in this project is a CNN with pre-trained word vectors. Different values on the hyperparameter "Region size" have been tested and a value of 4 yields the highest F_1-score.

Investigating if the Structure of a Comment Significantly Contribute to Identifying Sarcasm or Non-seriousness in Reddit Comments

The majority of internet-users has experienced sarcasm or non-seriousness in texts on the internet. The way which a user experience sarcasm or non-seriousness can be very personal and therefore make it hard to identify. Comments from the internet forum www.reddit.com is analyzed to investigate if the structure of the comments contributes significantly to the identification of sarcasm or non-seriousness. Two models is proposed for this binary classification task. A Naive Bayes and a Long Short-Term Memory (LSTM) network, where two versions of each model are tested. The first version of the Naive Bayes model contains stop words while the second version is without stop words. For the LSTM network, one version contains an Unidirectional layer and the other version contains a Bidirectional layer. Data which is used in this project is a part of the Self-Annotated Reddit Corpus (SARC) and contains 1.01M comments where 505 413 comments are labeled as sarcastic or non-serious. Infrequent words in data is replaced with an unknown-token and comments with more than 40 words are removed. In terms of accuracy, the Bidirectional LSTM has the highest performance, but the general difference between the models are not sufficiently different. It can, from the results and discussion, be concluded that the

models may be to simple for this type of classification problem and it is necessary to perform further tuning and development of the models.

Similarity of Books in Philosophy and Religion

The aim of this project is to study the similarity in writing styles and themes between famous/important publications in the realm of religion and philosophy and the main topics of these books. We study this by transforming the books into numerical vectors by different methods and then clustering these vectors according to their similarity with a hierarchical clustering algorithm. We were successful in showing that it is possible to identify relevant topics from these books and that clustering brings new and interesting insights about our data.

Sex, Drugs and Swedish Restaurant Reviews

According to previous research, studies have shown that the language being used in reviews to describe a restaurant visit is different depending on how exclusive the restaurant is. Dishes from an upper end restaurant tend to use metaphors to sex and sensual pleasure such as, an ”orgasmic pastry” or a ”very naughty deep-fried pork belly”. In comparison, lower end restaurants is described with metaphors to drugs and addiction, such as ”the wings are addicting” or ”I’m craving for pizza pretty badly right now”. In this report, I intend to investigate whether similar metaphoric patterns is being used in Swedish restaurant reviews by using Latent Dirichlet Allocation. Also, this report will evaluate how well the sentiment of a review can be predicted with various classification models.

Text Review Based Classification of Single Malt Scotch Whiskies

Text review based classification models of Single Malt Scotch whiskies are trained to fit numeric score based dataset. This project aims to find a possible way of relating numeric score based reviews and text based reviews. Term-frequency based models with different n-gram settings are trained and tested with multiclass classification method given authorized datasets scrapped from websites. Possible improvements and methods of relating two different types of reviews are suggested by interpreting and discussing results.

Developing a System for the Automated Classification of Journalistic Content

This paper implements a machine-learning approach for the automated categorization of German news articles for the purpose of digitizing archived material. Based on a

dataset of approximately 12,000 documents and pre-trained fastText word embeddings, a shallow convolutional neural network is built to classify articles into ten different categories: Sports, Politics, Business, Miscellaneous, Entertainment, Motoring, Science, Technology, Consumer advice and Life advice. With an accuracy of 0.89 and a F1 macro average of 0.84, the convolutional neural network significantly outperforms naive baseline models. Still, model performance varies substantially among classes. While Motoring and Sports can be identified almost perfectly, the more generic classes such as Miscellaneous and Life advice are being misclassified more frequently. Suggestions on how to improve model performance are given, before a reference implementation of a possible end-user application using a graphical user interface is being presented.

How Well Can a Topic Model Built on Wikipedia Entries of States Model the Geographical and Developmental Similarities Between Those States?

Knowing the power and limitations of topic modeling is of high importance for many text mining technologies, as it has many applications both in dimension reduction and natural language processing. This report will present an example of Latent Dirichlet Allocation (LDA) topic modeling applied to descriptions of states taken from the English wikipedia and combined with k-means clustering. The wikipedia entries of states are always structured in a similar way and cover topics like the states' history, geography, politics, economy, demographics and culture. For this project, a LDA topic model with Gibbs sampling was trained on the wikipedia entries of 93 Asian and European states. The purpose was to find out whether there would be a correlation between the topics which occur in the descriptions of certain countries and their geographical and developmental similarities. Thus, the states that have been put in one cluster are compared in respect to geographical measures like their distance and to their human development indices (HDI). This project shows that topic modeling techniques are able to represent such similarities partly. However, several methods to improve the approach described in this report were found.

Topic Analysis of Myths and Fairy Tales

I believe that one of the most fascinating features of human civilization is its' ability to dream, observe the surroundings and create stories. The mythology of ancient cultures combined these three characteristics and formed the people's understanding of the world as well as shaped the core values of a society. In this project I use the probabilistic topic modelling method LDA and hierarchical clustering on a collection of books of mythology, legends and fairy tales. The aim is to compare the results of the analysis to the prior knowledge and assumptions about the cultures.

Evaluation Wine Based on Descriptions

Wine sensory examination and assessment is not easy, even with a wine specialist. However, the description printed on each bottle can bring us some helpful information. Based on that ideal, I decided to analyze the data about wine review (taken form Kaggle) and help the consumers make the decision on two different aspects. The first one is distinguish good and excellent wine. The second aspect is investigating the finance beneficial bottle (high value with an acceptable amount of money). Different natural language processing techniques are applied to process the text data. Then I compare Naive Bayes, support vector machine (SVM) Linear and SVM RBF kernels when building the best model for predicting. After the best kernel is selected, n-gram and Part-Of-Speech Tagger are used to improve the model. The final results are quite impressive with accuracy 78.69% for the first issue and 77.09% for the second one. Based on these models, some suggestions are provided for the consumers to choose satisfied bottles.

Discovering Topics in News Articles Published by American News Outlets June 2016 through May 2017

This study aims to find the topics of articles published June 2016 through May 2017 by six American news publications. It examines the topic distribution over time and if publications with different political alignments display different topic distributions. Using latent Dirichlet allocation and variational inference and estimation, several models were fitted to the data, each with different number of topics, and evaluated by how interpretable the topics of each model were. Some of the topics discovered were: the 2016 presidential campaign and election, sports, and business. The study shows that the distribution of topics, found by the LDA model, changes over time and that the differences in topic distribution for the six publications could not be explained by political alignment.

Predicting Ratings from Women's Clothing Reviews

Product reviews are a great source for customer feedback for online businesses. Usually, reviews come with an associated rating, but the reviewer could opt out of leaving a grade. For businesses accumulating a lot of textual data in the form of reviews, it would be too time consuming to manually go through everything to assess the sentiment behind the review. In this project, a collection of reviews written by customers of women's clothing items is analyzed by classifying the reviews as positive or negative. The classifiers used are multinomial naive Bayes and random forest, where the aim is to see how well the latter performs compared to the relatively simple naive Bayes model. TF-IDF vectorization is used to generate features from the raw document collection and an exhaustive grid search is used to learn the model parameters. When predicting previously unseen test data, both classifiers achieved similar results in terms of evaluation measures, with multinomial naive Bayes performing marginally better.

Detecting Sarcasm on the Microblogging Sphere

With an ever growing online presence in todays society many of social media posts are taken out of context or misunderstood. Online individuals often refrain from using "netiquette", i.e. good online manners. This leads to a series of problems when interpreting what you read online. My intention with this project is to demystify some of this miscommunication on Reddit by using Self-Annotated Reddit Corpus (SARC) dataset to train a neural network to detect sarcasm. Detection of online sarcasm could be important when reading online texts, microblogs etc. which aren't too obvious in their way of expressing sarcasm. This could also be used further when web-scraping and only want to retrieve "serious" content; content without veneer. Having a classifier for sarcasm will be useful when scraping online forums pertaining stocks for instance. Where one could incite or demotivate other buyers for their own gain. Being able to detect these posts would be most useful for my current development of an autonomous trading platform that is in part scraping these forums for stock sentiment.

Analysis and Text Mining of Data Scientist Jobs

The main objective of the project is to explore/inspect the data science job market. Online job portals/advertisements have become prevailing job searching tool and popular in all parts of the world. It is seen that 70-80% jobs in the World are posted online. I have used web scrapping to gather the summary/description related to data science related jobs posted in different cities of USA on indeed.com. The purpose is

to analyse the job market data by applying advanced text analytics by information retrieval and creating clustering models that can add significance to the data. To also achieve an insight about the skill set requirements of the job. It is also interesting to compare the job markets of east and west coast cities.

TDDE16

Classification of the Original Author-Party of a Political Document Using Logistical Regression, Stochastic Gradient Descent and Multinomial Naïve Bayes Classifiers

This paper outlines the procedure to mine web API:s, data cleaning, and general text mining pre-processing, to achieve a N-gram and TF-IDF model of the political documents collected. All this to be able to classify the original author-party of a political document using Logistical Regression, Stochastic Gradient Descent and Multinomial Naïve Bayes classifiers. The best classifier performed with an accuracy of 88.4%, reasons for that are found in the discussion.

Sentiment Classification on IMDB Reviews Using Convolutional Neural Networks and Bidirectional LSTMs

As a cinephile, I enjoy discovering new movies using the famous IMDB web application. One major factor when it comes to pick a movie, is the review section under each movie, which greatly influences me and of course many other cinephiles. This gave me the inspiration to build a basic sentiment analyzer whose task is to predict whether the movie review is good or bad. This report aims to take advantage of the recent advances in the field of deep learning and implement a Convolutional Neural Network architecture as well as a Recurrent Neural Network with bidirectional LSTMs for comparison purposes.

Prediction of Interacting Medicines from Summary of Product Characteristics for Healthcare Professionals

Doctors, pharmacists and nurses are vital parts of today's and tomorrow's healthcare. Their time is incredibly valuable and there are many areas where AI could facilitate and streamline daily work. This project examines how well a set of interacting medicines could be detected based on the medicine descriptions. The set of medicines used are examples from reality, where a doctor experienced that current systems sometimes forget to remind about their interaction. Dimensionality reduction techniques such

as principal component analysis and t-SNE were used together with unsupervised learning in order to get a better understanding of the TF-IDF representation of the data. With that knowledge in mind, it was examined how well three different binary classifiers could predict an interaction between two medicines. The decision three classifier managed to get the best results with a recall and accuracy just above 98% for the set of 550 medicines investigated.

Classifying Abstracts on DBpedia

Text classification, or document classification, involve methods of categorizing texts into one or more categories. In this report, multinomial naive Bayes and logistic regression is used to classify abstracts from DBpedia. DBpedia is a database containing structured data from the Wikipedia project. Abstracts on DBpedia are short descriptive texts explaining the subject in a concise manner. The dataset is created by compiling abstracts from 89271 DBpedia pages from 13 disjoint categories. The dataset is intentionally unbalanced, meaning that some classes are a lot more represented than others. The naive Bayes model is known to be biased towards classes with high representation and it is also shown in this report since not a single abstract was classified as one of the two least represented categories. The logistic regression model on the other hand performs better on all metrics. The dataset turns out to contain some errors due to some pages on DBpedia being falsely labelled.

Hate Speech Identification in Tweets, Comparing Convolutional Neural Networks, TF-IDF + Logistic Regression and XGBoost

This article covers the topic of tweet classification, with the intent of classifying different tweets into either being offensive, using offensive language or not being offensive at all. For this, several methods to compare the results are attempted. The tweets are preprocessed and converted to vector representations using Word2Vec or TF-IDF combined with n-grams. Secondly, a convolutional neural network is investigated using a grid-search approach for different values of several hyper-parameters. As a comparison, smaller investigations of the performance of logistic regression and XGBoost are also made. The results show that a TF-IDF approach, combined with logistic regression, achieve similar results as a CNN trained using Word2Vec-representation, with significantly less training time, while XGBoost achieves an almost impeccable precision specifically on hateful tweets, but quite poor results on other metrics compared to the other two methods.

Representing Quora Question. Comparing tf-idf and doc2vec Using Variable Amounts of Pre-processing

A good representation of the data makes the classes more distinct from one another and therefore the task of the classifier much easier. The representation can be impacted by both the pre-processing done and the how the representation is generated. In this project the classic tf-idf representation and the newer doc2vec with distributed memory representation was used to represent Quora questions. The representation was then used with a linear SVM classifier to predict if the question was sincere. Also, the impact of different pre-processing methods was evaluated on this data set. It was shown that the tf-idf representation had a reasonable performance while the doc2vec performed very poorly. When pre-processing the text three different versions where created, all of them had non-alpha numeric characters removed and all characters lower-cased. The first had no further processing while the second one had stop words removed and the third had stop words removed and was stemmed. For this data set the first version, with the least amount of pre-processing, was the best performing and the more the data was processed the worse the performance.

Sentiment Analysis of Movie Reviews

In this report, sentiment analysis is performed using movie review data coming from different sources. The aim is to evaluate a few selected machine learning methods for the purpose of sentiment analysis and explore if the models that are trained using one particular source will generalize well when presented with new data from another source. To do this, data was collected from both IMDB and Rotten Tomatoes and a few experiments were done using different combinations of training and testing data. These experiments reveals that simple linear models performs the task of sentiment analysis very well and that tree-based methods have some potential as well. Additionally, it is shown that a high level of generalization is possible using the selected methods, especially those that are trained using the Rotten Tomatoes data.

Multi-Paragraph Question Answering

In this work, we consider the problem of multi-paragraph question answering (mpQA). In this task, the machine is given a text consisting of several paragraphs as input and is required to answer any questions based on the given text. We investigate several approaches to extend a single paragraph QA (spQA) system for multi-paragraph QA. Specifically, we investigate i) a baseline approach where the raw output score of a spQA system is used as a confidence measure to select an answer across paragraphs,

ii) a pipelined system which uses TF-IDF to first select a single paragraph given the question, and then processes it using a spQA system, iii) a binary classification model where every word in the given text is classified into "start of the answer" or "not start of the answer", iv) a paragraph classification model which predicts whether a paragraph contains the answer to a given question and v) a model which combines paragraph classification and spQA outputs to give a probibility score for a word being the start of the answer. Experiments are performed on the Stanford Question Answering Dataset. The results indicate that the baseline approach, along with the paragraph classification approach provides the best performance.

Explainable Tweet Classification with a Convolutional Neural Network

This work explores the use of a small convolutional neural network (CNN) to produce explainable classifications of tweets made by US politicians. A dataset of approximately 175 thousand tweets made by 1159 US politicians has been constructed using the publicly available knowledge base Wikidata and the Twitter API. A CNN classifier has been trained to distinguish between the tweets made by democrats and republicans. By inspecting how the presence of different n-grams in a tweet contribute to the final classification, a short explanation can be made available with each prediction. The quality of the classifier, as well as the usefulness of the provided explanation have been analyzed.

Investigating Sentence Vector Representations for Extractive Text Summarization Using TextRank

In the area of machine learning and Natural Language Processing highly data-driven models have shown great promise. These models crave large resources in the form of run-time, hardware and data-sets. This paper investigates the difference between a data-driven word2vec sentence vector representation, and a naive bag-of-words approach, for extractive text summarization using the graph-based algorithm TextRank. The research shows a slight improvement when using the word2vec based vector representation, measured using a BLEU and ROUGE based F-score.

Monolingual Written Natural Language Identification with Naïve Bayes and LSTMs

his project aims to correctly classify the language of written text paragraphs. Given a paragraph from the WiLi dataset, one out of 235 languages can be classified. Chosen model solutions to the classification problem is a Naïve Bayes model and an LSTM. They reach a classification accuracy of 93.87% and 63.87% respectively. The Naïve

Bayes model performs a lot better than the author of the WiLi dataset's neural network model, which achieves an accuracy of 88%.

Sentiment Analysis on Movie Reviews

In this project, random forest is investigated to see if it is fit to be used for sentiment analysis. IMDb reviews are used to train a model which should be able to classify reviews as positive or negative. TF-IDF features are extracted from a document set of 75000 reviews. Random grid search is used to optimize the hyperparameters of the model. Final model achives 86,5% accuracy on a class balanced test set.

Classifying Semantic Relations in a Cheaply Constructed Data Set Using Recurrent Neural Networks

This project aims to explore the effectiveness of state-of-the-art machine learning models for semantic relation classification when applied to a minimally preprocessed data set. A hierarchical RNN-based model trained on the very popular SemEval-2010 data set was selected for this study. One part of the project was creating a data set conforming with the SemEval-2010 data format. This entailed extracting sentences from webpages, performing part-of-speech tagging and named entity detection, and finally selecting two nominals among nouns and named entities to be marked up in accordance with the SemEval-2010 data. Since having simple and cheap preprocessing is an integral part of the study, the entities were selected at random. The data set that was finally used to evaluate the model performance consisted of sentences taken from the English Wikipedia page for Cars. The selected RNN model was then used to classify the semantic relations between the marked nominals over this constructed data set. When evaluating the results, the overall performance of the model was poor. Performance was very good with regards to recall, but very poor when it came to precision. This report ends with some discussion surrounding the results together with conclusions and suggestions for future work.

Scale-based Sentiment Analysis on Kindle Reviews

This paper aims to explore different pre-processing techniques in the domain of text categorization. We attempt to predict the corresponding number of stars given for each Kindle eBook review, which are specified on a scale from 1-5 stars.

Predicting IMDB Score of Movies

A movie on IMDB gets a score from people voting on how good it was. What if it was possible to predict this score by only using the information on the site IMDB. This could then be done before the movies are released and could then lead to a way of knowing before you see a movie if it is worth to see. Using the description, year and categories I tried to predict the score of a movie. Trying with models like a tri-gram Naive Bayes model no connection were found between the score and the rest of the information.

Automatic Labeling of GitHub Issues

This project explores the possibility of automatically classifying GitHub issues by their labels in two projects, Angular and Symfony, where the former had multiple classes and the latter had only binary classes. The documents were vectorized using tf-idf, and then classified using the following methods: linear SVMs, random forests, logistic regression, and multinomial naive Bayes. Two validation datasets were also collected to validate the chosen methods. The results are that the Angular dataset is difficult to accurately classify using these methods, while the methods have more success on the Symfony issues and the two validation datasets.

LSTM Neural Networks for Generating and Classifying Song Lyrics

This study explores the possibility of generating song lyrics by using recurrent neural networks. Generating meaningful text is a hard task for computers, but song lyrics are often more simple in structure and meaning which might make generating them a suitable task for neural networks. Another neural net is trained on predicting billboard score of the generated song lyrics. The predicted score is then fed back into the generation process for it to optimize a generation parameter. Some success with generating text is had, while the predicting neural network is not sufficient enough to provide meaningful insight.

Finding Document Clusters in Patent Databases

This report details the results of an effort in exploratory data analysis. In particular, the aim was to cluster documents in a patents database represented as document embedding vectors, by way of k-Means clustering. For that purpose, a publicly available sample of the European Patents Office (EPO) DOCDB format served as the source of data. Data pre-processing efforts included dropping unwanted XML tags, extracting the text of claims and further processing to filter stop-words and punctuation.

Document embedding vectors were created from the resulting corpus with Google's Doc2Vec. These were then mapped to 3D space using T-distributed Stochastic Neighbor Embedding (tSNE), normalized (L2) and fed to *k*-Means algorithm. The results obtained lead to the conclusion that intellectual property databases can be clustered and document relationships realized, based on a chosen feature or set of features. However, finding the optimal *k* and the validity of cluster allocations motivate further discussion this and other clustering techniques. For the dataset utilized, the optimal *k* was found to lie within a 55-58 range (inclusive), at a ceiling of 60 *k*-Means clustering iterations.

Easy Meta Embeddings and Ensemble Methods for Text Classification

Word embeddings are often used in conjunction with different models in natural language processing, however different pre-trained word embeddings may not always be well suited for a certain problem. This projects aims to investigate methods for combining word embeddings. Simple methods for creating meta embeddings and ensemble methods are shown to increase performance.

Automated Forum Moderation

It is not uncommon to hear of online discussions that derail into profanities, personal insults and racial slurs. This is a problem for online communities, which is traditionally solved by giving employees the tedious and time consuming task of manually moderating online discussions. However, with the recent advances in natural language processing and machine learning techniques, it is possible to automate this task. This text mining projects aims to evaluate the performance of a machine learning model called "Long Short-Term Memory Networks" using FastText word embeddings on text classification tasks, to see if it can be used to automatically detect content in need of regulation, automating the process of forum moderation.

Age Classification of English Literature

This project delves into age classification of English literature by trying to solve two related problems. Firstly, five different classification methods are used to try and predict whether a book was written before or after 1880: Cosine Similarity Weighing, Linear Regression, Logistic Regression, Support Vector Machine and Random Forest. Secondly, those five methods are used to classify according to which decade the books belong. As input to all five methods the term frequency–inverse document frequency (tfidf) is used. The project shows that out of the five aforementioned classification

methods, the Support Vector Machine and the Random Forest perform the best. The other three methods do not significantly exceed accuracies reached through random guessing. This report will make evident that data composition and the size of the used dataset plays an important role to why classification accuracy cannot be improved further.

Family Tree Extraction From J.R.R. Tolkien's World

There are numerous vast fictional worlds to deep dive into. One of them is J.R.R.Tolkien's epic universe, a world filled with complex characters, family trees, and relationships which occur over multiple books. To draw a family tree and visually see the whole picture therefor requires thousands of pages to be read by a fan or a computer. In this project, the goal is to extract family trees of characters from Tolkien's world automatically with natural language processing. To be able to validate the results the wikipedia-like web page for Lord of the Rings (www.lotr.wikia.com) is used as the text source and the infobox section of each character page is used as the ground truth. After training two neural models to recognize character names extraction of family trees were achieved with recall of 28% and precision of 0.15% at best.

Predicting Spam in SMS Text Messages

In this project a method for using text mining techniques to predict whether or not a SMS text message is spam or not is described and implemented. At first a extensive data exploration is conducted, resulting in two new features being engineered. These are then used to train two different models, the best of which is optimized using hyperparameter optimiziation. The final optimized Naive Bayes model achieves a accuracy score of 99% and $F_{1.5}$ score of 97%. This beats the benchmark models, which classifies everything as non-spam, 87% accuracy and a $F_{1.5}$ score of 48%.