Project abstracts

732A92

Automatic keyword extraction from patent documents

This project in text mining aims to develop an unsupervised algorithm for automatic extraction of keywords from patent data that could be used as input for ad-hoc searches as well as for standing queries. The project also aims to contribute to the text mining community by proposing an efficient graph-based ranking algorithm, HarmonicRank. The ranking algorithm is designed with a modular structure in mind, with the components being based on key attributes derived from graph information and frequency measures extracted from the patent portfolio.

House of cards topic modeling and characters analysis

As a big fan of TV series, I chose for my text mining project to do a script analysis of the show House of Cards. According to Wikipedia, House of Cards (HoC) is an American political thriller and drama series. The main character is Frank Underwood, a congressman, who is ready for anything to attain power, with the help of his wife, Claire Underwood. Through the five seasons, we will follow his ascent until the White House, at the expense of many people. The main themes of the series are politics, manipulation, media and love stories. After scraping and processing the data, two different tasks will be done as 'text mining work': topic models and character mentions.

The automation of the candidate choice for projects in a consulting firm

All consulting firms, without exception, have the need to know which skills their workers have to provide them with a suited project as fast as possible when the opportunity of a project arises. Thus, it is essential to have updated the skills of the employees in a CV to match the market. By doing that the first-mover advantage in the view of a project is reachable, which stands for a raise in the contracts and a higher utilization of your employee resources, translated in more inflows for the firm and

thus the possibility of growing and getting a larger portion of the niche of the desired market. For this reason, I thought I could try to automatize the process of evaluating emails concerning project opportunities for a group of consultants, trying to produce a retrieving algorithm that matches emails with projects to consultants with certain skills gathered on their CVs, giving out the 3 best candidates for a specific offer.

Named entity recognition in Swedish and English fiction

Before Gutenberg books were not something that was very common for an regular person to own by themselves but a thing for the privileged. Fast forward 600 years to 2018. In today's society books are a commodity that everyone can afford and an unimaginably size of books is available online, either free via libraries and open source, or provided by paid services. This project is written as a pre-study for the authors Master's Thesis with an undisclosed online audio book and e-book provider and is aimed towards finding a unsupervised method to match different editions and editions in different languages.

Evaluating Naive Bayes and multinomial logistic regression for tweet classification

Dagen för riksdagsvalet 2018 närmar sig och det vi ser framför oss är ett val som inte liknar något av alla de som hållits tidigare. Eftersom de olika partierna har markant skilda åsikter berörande alla ämnen som innefattar hur Sverige skall styras, är det intressant om det är möjligt att klassificera en för algoritmen osedd tweet till det parti som ursprungligen författade den. Det finns flertalet tillvägagångssätt för att gå till väga med detta, i detta arbete har algoritmerna Naive Bayes och multinomial logistisk regression valts för att klassificera tweets. Projektets frågeställning är således: Vilken algoritm presterar bäst när det gäller att klassificera en osedd mängd tweets till rätt parti, Naive Bayes eller multinomial logistisk regression?

Exploring text features in horse racing forecasts

In Hong Kong, the Hong Kong Jockey Club (HKJC) is a government-granted monopoly in operating horse racing events. Numerical data, along with some natural text description of each race, is published on their website. These published text description are usually very succinct, but it may contain information which is not found in other published numerical data, such as the writer's subjective observation about the horses health status. We attempt to make use of these text information in forecasting the racing results. We found that when added to a ridge regression model, these text information slightly improves the prediction result in a statistically significant sense.

The analysis of emotiveness of comments on statuses of UNICEF's Facebook page

The main objective of my project is to quantify the emotiveness of the comments, given on the statuses of the UNICEF Facebook page (UNICEF, Facebook page). In order to do so, three different psycholinguistic markers are calculated. Since emotiveness describes the emotional engagement of a writer at the time that he or she is writing the comment, I believe that this information could be useful to UNICEF. Emotional engagement is an important factor in the decision to support a non-profit organization and therefore it should not be ignored. First of all, this project investigates if there is a correlation between the average emotiveness of comments on a status and the number of likes that this specific status got. Furthermore, it intends to reveal whether or not the emotiveness of comments is significantly different for different statuses. This analysis is performed in the hope of discovering which (type of) statuses generate high emotiveness. Indeed, for non-profit organization it could be useful to know which statuses emotionally move their followers the most.

Generation of cover letters using LSTM networks

In this paper I attempt to generate cover letters using long short-term memory (LSTM) networks. Since I have a small corpus (around quarter of a million characters), I will attempt to answer the question: does pre-training a character-level predictive recurrent neural network on a larger corpus improve the predictive performance on a smaller corpus?

Classifying patents into patent portfolios

A patent is for most people a paper that gives the owner or creator exclusive rights to the invention. The main purpose with patents is to promote innovation but it can also be extended to operate as a strong measurement for market leadership and ensure the freedom to operate in a country. If you treat patents as a collection of related information, you would be amazed of what could be extracted. In practice, technology intensive companies store patents in portfolios where the portfolio could represent a topic of interest for the company. For this project we are going to focus on automatically distributing patents to relevant portfolios based on its contents. This leads to the project research questions: How well can we classify patents into correct portfolios? Which model is the best performing one and should be used for future studies?

Authorship attribution of tweets

The aim of this project is to build a supervised classifier which is able to distinguish between tweets written by different authors based on quantitative measures. Term frequency (TF) and term frequency-inverse document frequency (TF-IDF) representations of tweets are used as quantitative measures. As a classifier a logistic regression is used. The classifier will be applied to a set of the 1,000 most recent published tweets by @realDonaldTrump. Finally, an exploratory analysis is carried out to make differences between the groups evident.

A classifier for emergency calls

This project belongs to the course Text Mining. The aim is to independently apply the methods covered in the course to self-defined problems. The problem that I intend to address is using information collected from 137,000 calls to the Regional Emergency Services (1-1-2) from Canary Islands (Spain) in order to predict the Type of Incident that occurred, Emergency Resource Used and the Assessment of the Incident. I chose this project due to the fact that I had worked with this data before during my Bachelor's degree final project. However, I only used this text variable to build a word cloud (which was visually interesting but did not bring too much useful information to the project) since I focused on the rest of the quantitative features available. Ever since, I have had the interest in exploring further this data and find interesting insights, but I lacked the knowledge. Now, after taking the Text Mining course I have learnt the way to address this problem in the best possible manner.

What makes a successful Facebook?

When it comes to marketing to today's young consumers, brands are desperately trying to figure out what catches the younger audiences attention. The thing is that we are witnessing the rise of a new generation, designated by millennials or native generation because they don't know what it feels like to live disconnected. Due to the emergence of this new generation, together with the fast growth of social media networks, the Influencer marketing has become extremely popular for targeting younger demographics. NN is a social media marketing agency that is aiming to disrupt the marketing market by using nano-influencers to increase post's engagement rates. To achieve its goal, besides using a big pool of nano-influencers, NN collect's all the information resulted from the campaign to identify the characteristics that make a successful post to improve and recommend for future campaigns. Therefore, this project aims to support the early development of NN's data analysis methods. The goal of the current project will be to identify key characteristics of a successful post on

social media, as well as learn if there are specific words that create more engagement than usual.

Visualisation of tweet texts and relevant tweeters

Twitter has approximately 330 million monthly active users according to their 2017 year-end report (Twitter Marketing Department 2018). Nearly one quarter of Americans use the social network (Pew Research Center 2018). This provides a vast amount of data ripe for text mining. The goal of this project is to find a method to graphically display similarity between a body of tweets and those of the top 50 tweeters (by number of followers). This would allow a twitter user to input their tweets and see who they are most similar to using a variety of Natural Language Processing techniques. In the search for appropriate methods, a novel and interactive way to visualize topic groupings was also found, and has been included.

Hotel review analysis

Nowadays, online hotel booking is becoming a popular choice when people are traveling around different countries and regions. After the traveling, usually people would post their reviews on the reservation website to express their experiences about the hotel or the city. It would be beneficial for the hotel owners or local tourism offices to read those comments for improvements in the future. But with the development, the data size is becoming too large so it's hard for humans to read every single review manually. Hence there might be some automatic text mining approach would be more suitable to analyze these reviews to get some conclusions from humans' aspects. In this project, topic modeling is applied to investigate visitors' reviews about the hotels when they are traveling to different continent (taking Asia and Europe as examples here for a comparison). During this process, held-out log marginal likelihood is introduced as evaluating method for selecting the optimal topic numbers. Besides, those data are divided into positive and negative contents for applying topic modeling separately for some more detailed comparisons.

TDDE16

Classifying movie genres by analyzing text reviews

This paper proposes a method for classifying movie genres by only looking at text reviews. The data used are from Large Movie Review Dataset v1.0 and IMDb. This paper compared a K-nearest neighbors (KNN) model and a multilayer perceptron

(MLP) that uses tf-idf as input features. The paper also discusses different evaluation metrics used when doing multi-label classification. For the data used in this research, the KNN model performed the best with an accuracy of 55.4% and a Hamming loss of 0.047.

Sentiment analysis on Reddit comments

This study aims to explore the potential of using Reddit comments to express general sentiment over specific terms. How easily the correct sentiment of user comments may be extracted is also an area of interest. This is to be performed by creating a Python program, using NLTK and VADER to parse, process, and analyze Reddit comment data.

Semantic retrieval of TV-series episodes

Watching TV-series is a lot of fun for most of us. Sometimes, we like an episode so much that, later on, we go looking for it to watch it yet another time. When there is an overarching story encompassing the whole series, it is usually not hard to locate the episode, knowing what happened before and after it. But as it turns out, there are a lot of series that have almost completely disconnected plot lines from one episode to the next. In this case, it may be difficult to find what we are looking for. Many turn to Google, but since it is but a general purpose search engine, it may or may not succeed, or offer results that are entirely incorrect. A specialized search engine could thus be of interest, especially as an additional tool for websites that provide information about various series, but lack a way to find specific episodes if not by name, number, or by patiently skimming through each description, or as a standalone application. The purpose of this project is therefore to implement and test a specialized system to retrieve the most relevant episode based on an user query. Two different models are compared, the vector model –; a simple algorithm in information retrieval –; and a model based on dense word vectors, and discussed.

Spooky author identification

In this project an investigation and study of the problem of author classification on the bases of sentences was conducted. This problem is originally taken from a Kaggle competition. The problem of document/text classification is a problem that has gained a lot of interest due to the increasing amount of available data. Therefore the question that this report tries to answer is, is it possible to classify which author has written a specific sentence? In this report results from different classifiers using

hand-crafted features are compared with results obtained from using a convolutional neural network (CNN).

News article tagging using tf-idf for multi-label classification

Teknomedia is part of the Norrköpings Tidningars Media (NTM) corporation which is a company working in the Swedish news paper industry located in Norrköping, Östergötland. One of their areas of research is the development of system capable of suggesting relevant tags for news articles. Tags are labels that describe the content of the text. To determine if a tag is relevant is not a trivial matter as relevance in this sense infers some level of ambiguity where a tag can be seen as relevant by one observer and irrelevant by another. Simplistic methods for evaluation with precision, recall and F1-measures might therefore not give a fair representation of system performance. In this project we address the following questions: What could be a suitable evaluation method for a information retrieval system with a ambiguous gold standard? How can multi-label classification be performed with a undefined amount classes?

Sentiment classification of movie reviews

An important aspect of text mining is the automated extraction of the meaning from a text. This automation allows one to quickly quantify and summarize opinions from large amounts of text, which could be very useful when wanting to compare different products based on their reviews, or studying the public opinion on a certain subject. The task of extracting opinions from the text is called sentiment analysis. In this project sentiment analysis has been conducted on movie reviews, where the reviews were classified as either positive or negative. In doing this classification, different techniques and preprocessing steps have been utilized. The choice of project task is motivated by several factors. First of all, it captures the essence of text mining –; the non-trivial extraction of the meaning from a piece of text. Second, there has been some promising research done on the subject which shows that it is possible to achieve decent results. Third, it allows comparison of different techniques covered in the course, which both renders practical experience with them as well as knowledge in regards to how suitable they are in this sort of task.

Identifying 'explicit' song lyrics

In this project the aim is to investigate song lyrics containing explicit content and if it is possible to classify a lyric as 'explicit' or 'non-explicit'. Instead of manually classifying a lyric as explicit or not this implementation will classify the lyric based

on the words in the lyric using NLP techniques. It would be possible to only look if the lyric contains specific words that are inappropriate. Although, this could miss lyrics that obviously imply inappropriate content without directly using inappropriate words. This project is chosen based on the increased popularity of music streaming services and their use of lyrics in their applications. The author's big interest of music in combination with previous work at the music streaming company Spotify is also a factor when choosing the project. Since earlier experience in using Spotify's API exists, it also appeared possible to perform this project.

A voluntary filter bubble

This work will investigate the possibility to ignore hateful and mean comments on the social media platform Twitter. The intended effect of the work is to be a first step in creating a healthier type of content on online social medias by filtering out hateful or negative sentiment directed towards you. However, the intent and wish to do good in this work proved to be complex for the scope of this project and had to be delimited into only how effectively it can be done, and not if it will have a positive impact on mental health.

Open relation extraction enabling ontology involvement

Materials engineering is a domain that contains intensive knowledge during designs and manufactures. There are a number of research papers in materials engineering published to share new methods or discoveries. To extract the knowledge in such articles is meaningful to build or extend knowledge bases. Natural language processing techniques can support some practices and ideas for such extraction. Information extraction (IE) is a task to find structured information such as entities, relationships from unstructured data by analyzing human language with natural language processing (NLP) techniques. Relation extraction (RE) is one of the essential steps in IE which aims to extract semantics, namely relationships among entities. There are a number of methods to achieve RE according to different application scenarios with specific data sources and domain knowledge. There is a number of methods aiming at extracting relationships from text data such as supervised and unsupervised methods. This paper is going to present a framework that involves domain ontology in open relation extraction which is an unsupervised method.

Deriving personality type from written text

A personality type can tell a lot about a person. It can help when interacting with said person. It could also provide useful information about how this person prefer their work environment, working in teams or working alone. It might even tell some useful information about what attributes they prefer in other people. The personality type of a person might be reflected in how they communicate with others and how they choose to express themselves, even in text. If that is true then there should be textual patterns to analyse, to derive this personality type. This study is based on the Myers-Briggs Type Indicator (MBTI) personality test and uses texts written by people who has taken this test. A dataset of about 8600 people and 50 forum posts per person will be used to train two different text classifiers. One Multinomial Naive Bayes (MNB) and a Support Vector Machine (SVM) classifier, both as implemented in the scikit-learn library. These classifiers will be used to identify peoples personality types from texts they have written. If the prediction works well it should be possible to predict a persons personality type by reading their Facebook- or Twitter posts, or even emails. This can be integrated in pretty much any site as long as the users are prepared to provide information from for example their Facebook, Twitter and Google account.

Visualization of Topic Models for Research Papers

The purpose of this project work is to use different methods learned in the course Text Mining. Therefore, in this project the main Text Mining tools Information Retrieval, Natural Language Processing and Text Data Analysis are applied. In the remaining text, this study tries to answer the following research questions (RQ): How can one extract relevant information of a given text using unsupervised machine learning techniques? How can one visualize relevant information so it is intuitively understandable for humans? First of all, one must find the information that represents the text source in a comprehensive way. In the next step, one must visualize this information in a way that is intuitive for the human eye. This is different from just extracting and saving the results in a data structure which is accessible by a machine and, therefore, a big challenge. To achieve this goal, the project is divided in the following steps: In the first step, a search engine retrieves all papers for a given search term as the data source. In the next step, these papers are filtered for the most relevant. For the most relevant papers, the text is extracted and preprocessed. Based on the preprocessed text data, the topics and keywords are extracted which are then used to create edges, vertices and weights. In the last step, those are used to compute a network graph.

Spooky Author Classification

The goal of this project was to create a model which can predict which of three authors wrote a given sentence. The popular approach to these types of problems, and the one most closely examined here, is deep neural networks (DNN). The project idea was taken directly from a Kaggle competition called 'Spooky Author Identification'. The competition had a $25,000 reward for first place. Unfortunately, due to time limitations no solution was submitted before the deadline at December 15. Since at this point some interesting solutions had already been partially developed, the project continued as planned.

Predicting Myers-Briggs Type Indicator From Written Texts Using Supervised Learning

This work experiments with supervised learning classifiers to predict the Myers-Briggs type of a user based on written text. The models are tested with features extracted from the latent Dirichlet allocation model, the TF-IDF model and the bag-of-letters model. The Gradient Boosting Classifier is shown to achieve the best performance with a test accuracy of 30.3%. Improvements to the classifiers are proposed and the limitations of the data set are discussed.

Dota mining

From my personal experience of the competitive online gaming culture people tend to be expressive in what ever means of communication available to them. A win is much more enjoyed if you can rub it in the opponents face. A individual mistake can be mitigated if you put the blame on your teammates. I am interested in creating a classifier to predict the outcome of a match and how different features perform, by only looking only at the chat from games of Dota 2. Dota 2 is an online multiplayer game where ten players are divided into two teams with the objective to destroy the opposing team's base. To achieve this all five players within a team depend on each others performance. The teams can communicate with each other using the in-game chat.

Generating Linux kernel code

The Linux Kernel project suffers from a significant problem, thousands of people are collaborating to create one of the most popular contemporary Operating System Kernels. There are fears that software project of such magnitude may cause bugs which no single human is able to solve. In this paper we will examine how, by

training a Recurrent Neural Network to generate new kernel code, the productivity and efficiency of the Linux Project might increase as human involvement can be limited to oversight and managerial tasks.

Identifying toxic comments

The social environment on the internet is not always the most inviting. In every forum, no matter how innocent it might seem, there is always at least someone posting inflammatory or derogatory comments. These type of comments are commonly referred to as 'toxic'. This has become more of a problem recently as information about foreign states affecting the opinions of the people in a different country by mass posting such comments has surfaced. One alleged instance of this is the fake news that were spread on social media leading up to the election in the USA 2016. These fake news posts are often toxic in nature in order to induce a response in the readers. In a study by Qiu et al. the authors claim that humans tend to become less rational in their efforts to distinguish good from bad when being overloaded by information. According to the paper it causes individuals to choose the most popular over the more qualitative content. This calls for a solution which can process information quickly and provide the users with a good indication to whether a post is inflammatory or not. This paper aims to address this issue by developing a system which can detect whether a comment is indeed toxic or not. The idea is to use text mining as a tool for solving this problem, more specifically clustering and text classification.

A comparison of multinomial and multivariate Bernoulli Naive Bayes classifier in heterogeneous data sets

Today, social media is an obvious part of many peoples life. The availability of the internet, and though it communication, enables people to interact anonymously with each other through a variety of social media platforms. The anonymity can often lead to animosity and foul language in posts and comment sections, and with the vastness of many platform it can be impossible to moderate the comments and posts of often anonymous users. Automatic offensive language detection is vital for these situations, and text classification systems are of great use. Text classification systems are often developed with a sole purpose in mind, and trained thereafter. The objective of this project is to create a text classification model and evaluate it on a different data set than the training set. The model will be trained to detect foul language on web based social media platforms. The popular Naive Bayes classifier will be implemented with some variations in algorithms, to then be tested and evaluated.

Classy: Feature evaluation using Naive Bayes for genre classification

What defines a genre? What are the key elements that characterizes rap, pop and electronic music? In this project a corpus was created, including 6 genres and 3000+ tracks. Naive Bayes was used to classify the genres. A feature analysis and evaluation was made to identify good features when classifying the lyrics.

Detecting personality types

Today, semantic analysis of Internet posts are done to detect patterns which could generate information about the persons who wrote the posts. The purpose of this project is to analyze if it is possible to detect patterns in a text and map these to a person's style of writing. The questions that will be answered in this report are the following: Is it possible to determine if a person's personality type is containing feeling or thinking, depending on their style of writing? Are the most common informative words positive or negative in a text written by these personality types? To answer these questions, two code implementations of Naive Bayes classifier was made to generate a baseline. The settings of the one with the highest accuracy and F1-score, were used when implemented types of classifiers to analyze the results. The Naive Bayes classifier with the highest and second highest accuracy and F1-score were used to analyze the most common informative words. This was made with an opinion lexicon.

Using movie scripts and text mining to predict movie genres

In this paper, the effectiveness of using text mining to predict genres for movie scripts is described. The model which proved most successful was a combination of a multi-layer perceptron and a decision tree; this combination scored 11.1% accuracy. Analyzing the result, the model was found to have potential for significant improvements.

The challenges in learning from recipes

Machine learning tasks need large amounts of data to learn effectively without overfitting. Unsupervised tasks like training word embeddings require massive corpora, such as the entire collection of Wikipedia articles, to be general enough. Supervised tasks like part-of-speech tagging might not need the same quantity of information, but rely heavily on the quality of hand-labeled corpora available. Even then, learned models might still consistently fail if the testing data is different enough. Regardless of the task, a good data set is key to a good result. This project explores what is

necessary to curate a large-enough data set of recipes collected from the internet. While the original goal was to create an algorithm to learn patterns in recipes in order to generate new ones, the final scope has been narrowed to a prediction task based on each recipe's category.

Clustering of song lyrics

In this report I will consider the case of clustering lyrics for various famous songs by different artists. The data set is taken from kaggle.com and consists of 57,650 song lyrics by 643 artists. The goal of the project is to gain some knowledge about how lyrics behave as a corpus, and will finish with an attempt to predict an artist, given the lyrics from a song. I chose this data set mainly because I like music, but I have not been very good at listening to lyrics. I would therefore like to investigate them further, and hopefully learn something.

Representing documents as topic vectors for text classification

In this report we compare how well a document can be classified when it is represented as a distribution over topics compared to the more traditional bag-of-words. The distribution over different topics are generated from a latent Dirichlet allocation model, and the documents are classified using the naive Bayes classifier, support vector machines and logistic regression. We show that on the given data set, the classification using topic distributions is not as accurate as using bag-of-words. However, the topic distribution provides a much more low-dimensional way to represent the text.

Can Wikipedia be a millionaire?

In this paper I describe the development and evaluation of four different algorithms, trained on raw textual data from Wikipedia articles, to answer multiple-choice questions. The algorithms were evaluated on 1000 questions from the game show Who Wants to be a Millionaire and compared to the result of a human. I found that the human beat all the algorithms, but that all of the algorithms were clearly better than a naive, randomized baseline system.