Extractive summary from Ground Lease Documents

Tom Verburg
University of Amsterdam
10769633

ABSTRACT

Due to a change in the ground lease plan options in the municipality of Amsterdam, the municipality is now facing thousands of ground lease documents which need to be processed. This research will focus on using a variety of machine learning techniques to perform extractive summarization on these ground lease documents, to automatize one element in the document processing chain.

1 PERSONAL DETAILS

My email mailto:tom_verburg@hotmail.nl
External supervisors email mailto:j.meurkes@amsterdam.nl
Internal supervisors email mailto:maartenmarx@uva.nl
The wiki on my github account https://github.com/tcjverburg/Master-Thesis

2 INTRODUCTION

Ground lease is an agreement which allows a tenant to live on a certain piece of property during a specified lease period, but after the lease the land and all the structures and improvements to the property on that land still belong to the original owner of the land. In the municipality of Amsterdam, it is very common to have a ground lease with the municipality and pay lease costs for living on a particular piece of land. Since 1896 [1], the municipality has not sold any more land and currently owns about 80% of all the land within its borders. The reason for this decision was because of the growing scarcity of available land within the city, and resulted in the need of the ground being governed efficiently. In turn, people lease this ground from the municipality, and this payment of the lease is called a canon. The price of the canon is dependent on a variety of variables and which were recalculated every year based on the real estate value and the interest rate of that moment and is referred to as the continuous system. Seeing as the canon is paid every 50 years, this meant that the price was recalculated every 50 years as well. However, the leasehold system has changed regarding to the ground lease and the determination of the price of the canon in Amsterdam. From the beginning of 2017 onwards, it is possible for leaseholders to change from a continuous leasehold plan to an everlasting leasehold plan.

An everlasting leasehold plan is different because it allows for how the leaseholder wishes to lease the property. Firstly, the leaseholder can choose to set the current canon price to be the price they will have to pay from that moment forth. This means that the calculated price at that moment in time for the canon is set and will not be recalculated when the next canon is due, regardless of whether the value of the lease increases or decreases. The second option is that the leasehold is bought off forever, and never has to be paid again. In both these situations it is important to note that the municipality will still remain the legal owner of the land.

Due to this significant change, it has resulted in a large amount of applications for a change in leasehold plan and result in countless documents applying for different leasehold plans in different scenarios. At the time of writing this proposal, the process of handling these documents and extracting the relevant information is done manually. This is a tedious job and employees apply crude methods such as using CTRL+F to search for signal words. These can hint for the information they search to be in the vicinity in that specific pdf ground lease document. This results in a growing backlog of 14 000 applications (consisting of 180 000 documents), and the focus of this research is to see to what extent the task of extracting information (mainly location and surface area) can be automized.

3 RESEARCH QUESTION

In this thesis, the following question is posed:

To what extent is it possible to apply extractive summarization techniques to extract the information regarding property location and surface area from the ground lease applications spread over various documents?

3.1 Subquestions

(1) What are the specific signal words which suggest important information regarding the ground lease application?
(2) What features, besides the signal words, can be extracted and used to identify important elements of a ground lease document?
(3) How can information residing in various documents be combined to summarize the essential information within these documents?

4 DATA

The exact data which will be available for this research is unknown at the moment of writing this proposal, but a subset of 300+ pdf documents with annotations has been shared, as well as a corresponding excel file which shows what information was extracted and relevant for processing the application. Furthermore, the signal words and important information is highlighted in the document itself, as well as noted in the excel file indicating on which page the information is present.

These documents represent one or more objects relating to specific ground lease agreements. Seeing as the data itself is entered manually, there are many inconsistencies in the exact spelling and semantics, which makes it very difficult to assign any numeric values to the preliminary data analysis. After cleaning up these inconsistencies, the actual shape and form of the data will become more evident.

"Property type" is one of the few columns which is rich in data and this seems to be the most important piece of data the municipality is able to extract from the ground lease documents. The surface

area of the property is also of importance, but at first glance of the preliminary data analysis it seems that a significant part of the documents do not contain information regarding the surface area. However, when it is, the page and location on that page is noted as well. This means that the amount of relevant data that can be extracted can differ greatly between documents, seeing as the data is often not mentioned in the document at all.

5 PREVIOUS RESEARCH

For this research, the main focus will be on supervised techniques which will enable the extraction of relevant information from individual documents. There is are various approaches concerning the categorization and classification of texts [7] and these differ from a regression [13], naive bayes [9], k-nearest neighbours [5] or decision trees [6].

The approach is to implement an extractive summarization model which will make it easier for the municipality to find relevant information in documents of varying lengths and content. Therefore, the focus will lie more on the extraction of relevant paragraphs and/or sentences instead of classifying the document itself, even though such methods could be adapted for the goal of this research(seeing as individual sentences can either be classified as either relevant or not for specific approaches).

The research of [3] suggests that there are two methods in summarizing a document or text. Firstly, there is the abstractive summarization approach, which is an understanding of the concepts and express those concepts in clear natural language. The extractive approach consists of selecting the important sentences, paragraphs etc. from a document and concatenating these into a shorter form. This research will mainly focus on the latter, seeing as it enables the experts of the municipality to make their own inferences. Babar et al. [3] suggest a variety of different approaches possible for the summarization of text. This research will focus on a machine learning approach, and what different approaches which are considered are discussed in the following section.

6 METHOD

In this section the method is proposed with the knowledge and data available at this time, as well as which machine learning methods will be applied and software implemented.

6.0.1 Data cleaning and processing. The first order of business is to aggregate the data available and remove the inconsistencies and synonyms in labelling. These range from simple spelling mistakes, to different notations for the same label. To be able to do this, the excel file must be loaded into python and using NLP (Natural language processing), the data will be cleaned up. When having cleaned up the data from the excel file, the text from the PDF documents needs to be extracted and processed. This will be done with a python library such as PDFminer[2].

However, the quality of the processed text needs to be evaluated, as well as how to extract the highlighted text present in the document which signals where the information is present. This is an essential part of the process, seeing as it will enable the possibility of labelling which sentences are relevant for the reader of the document of the text, and which are not. Due to often only having a part of the sentence highlighted, it has to be evaluated if this needs to be taken into account.

To have an algorithm which can do this efficiently and accurately will consist of a large part of the challenge posed in this research.

6.0.2 Feature extraction. Having extracted the labels for the sentences, this makes it possible to implement supervised learning methods. However, for this features need to be extracted to feed to the algorithm. One option is to implement a continuous-bag-of-words model which enables the document to be expressed in a vector, and makes it possible to represent (parts of) a document. Other options to extract the content based features from subparts of the text. These include the length, TF-IDF values and other ranking scores.

6.0.3 Algorithms. Dependable on what features can efficiently be extracted and used, there are various approaches which can be implemented. [12] implements convolutional neural networks to generate abstractive summaries with a topic modelling approach whilst Nallapati et al.[10][11] implements sequence-to-sequence recurrent neural network using key words and capturing the hierarchy of sentence-to-word for both extractive and abstractive summaries. However, this research will mainly focus on the extractive summaries, seeing as the context of a ground lease document may be essential and therefore it makes sense that it has to be quoted. Therefore, the approach concerning a RNN [10] which classifies each sentence as relevant or not, would be a suitable option for this research. Other methods which implement extractive summaries, such as Dong et al.[4] using reinforcement learning, could be applicable as well.

6.0.4 Evaluation. Evaluation can be done through an accuracy score (seeing as there is labelled data), as well as the results being evaluated by a human evaluator. However, the option of a human evaluator option still needs to be discussed with the external supervisor.

Another third option could be the ROUGE package [8] which is an automatic evaluation of summaries package which is available. However, ROUGE may not be sufficient as the information which needs to be extracted is very specific domain. The ROUGE evaluation might be too general for the scope of this research.

7 RISK ASSESSMENT

If the data or internship would unexpectedly not continue, the possibility is always to find another corpus and perform a unsupervised abstractive or extractive approach, and use the ROUGE package for evaluation.

Another problem could come forward regarding the extraction of the highlighted text from the pdf documents. Regular pdf extractors do not supply the option to differentiate between different background colors of text in the pdf. Seeing as this is one of the cornerstones of this research, it is important that the highlighted text is extracted as soon as possible.

In case this does not work or proves to be overly complicated, an alternative could always be to use the text in the excel file which quotes parts of the sentence of relevance for either the location or the surface area of the object described in the ground lease document.

8 PROJECT PLAN

Week	Plan
Week 1	Data is explored as well as the text and highlights extracted from the PDF's. Start with literature to set definitive approach regarding to method.
Week 2	Relevant features are extracted for supervised learning. Introduction is written and finished. Start with literature section.
Week 3	Initial model is done and first results and method. See whether this works or another method possibly needs to be explored.
Week 4	Create own model(s) and perform initial testing with various sets of features. Finish literature section.
Week 5	Try different models and see whether I can increase prediction accuracy, as well as see what the decisive features seem to be.
Week 6	Analyse intermediate results and finetune method and start writing results.
Week 7	Finish results and start writing discussion + perform research about other potential final tests.
Week 8	Finalize discussion, conclusion and implement final tests.
Week 9	Hand in first draft for evaluation and give it another thorough read myself.
Week 10	Read thesis and evaluate potential shortcoming and implement final additions.
Week 11	Hand in thesis and start on defence.
Week 12	Prepare defence.

REFERENCES

[1] Brochure erpacht in amsterdam: de meest gestelde vragen over erfpacht. https://www.erfpachtinamsterdam.nl/archief/Gemeente-2008-01-01-Brochure-Erfpacht-in-Amsterdam-de-meest-gestelde-vragen-over-erfpacht.pdf, 2019.

[2] Pdf miner library. https://github.com/pdfminer/pdfminer.six, 2019.

[3] SA Babar and Pallavi D Patil. Improving performance of text summarization. *Procedia Computer Science*, 46:354–363, 2015.

[4] Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. Banditsum: Extractive summarization as a contextual bandit. *arXiv preprint arXiv:1809.09672*, 2018.

[5] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. 1998.

[6] Tin Kam Ho, Jonathan J. Hull, and Sargur N. Srihari. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1):66–75, 1994.

[7] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, and Khairullah Khan. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1):4–20, 2010.

[8] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

[9] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

[10] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[11] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.

[12] Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*, 2018.

[13] Yiming Yang, Xin Liu, et al. A re-examination of text categorization methods. In *Sigir*, volume 99, page 99, 1999.