Extractive summarization from Ground Lease Documents: a supervised learning approach

Group Social

Tom Verburg 10769633

Problem

- Change in leasehold system in Amsterdam
- *Continuous* and *everlasting* system
- Citizens applying for transfer
- Enormous amount of ground lease documents to be processed
- Currently done manually
- How can a **Machine Learning** approach help in this process?

Internship

- 3 days at the office
- Data available in VAO
- Erfpacht department very helpful
- Assistance from Datalab
- Meeting planned with Kanoulas

Ground Lease Documents (data) #1

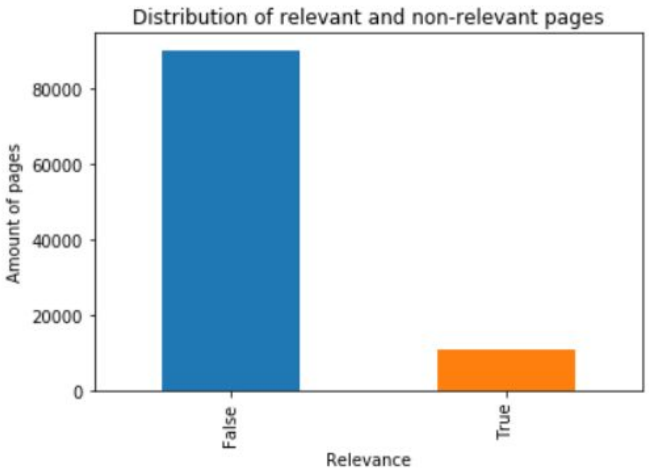
- 10 000 ground lease documents in PDF
- 5 (main) types of documents
- EDA of set of (processed) data currently implemented:

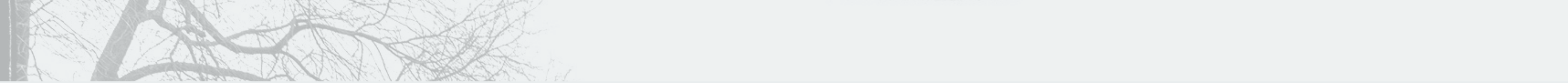
Amount of docs	Amount of pages	Amount of words	Mean # pages per doc	Mean # words per page	Range # pages	Range # words
3996	101118	21094705	26	209	2-233	0-1763

Ground Lease Documents (data) #2

Many inconsistencies

- Personal highlights
- Structure of documents
- No use of paragraphs
- Scanned or converted
- Handwritten documents





Research Question

How can extractive summarization techniques be applied to obtain information regarding the property from different types of ground lease documents?

Sub Questions

What are the specific signal words which suggest important information regarding the ground lease application?

What features, besides the signal words, can be extracted and used as features to identify which pages are ‘relevant’ and which are not?

Method

Pre-processing

- Extract the highlights
- Determine quality of individual documents (hand written, poorly scanned, corrupted)
- Lowering, tokenization, stop words, stemming

Supervised learning: Binary Classification

- Vectorization: Gensim doc2vec on page level
- Page relevancy is binary: highlight or no highlight present
- Logistic regression, naive bayes and SVM

Initial Results

- Some initial results on subset of **3000+** pages (200 documents)
- 3-fold cross validation

Logistic Regression:

- Average F1 score: **0.81802**
- Average Precision: **0.81434**
- Average Recall: **0.82414**

Challenges

- Quality of the ground lease documents
- Consistency of highlights
- Judicial documents are complex
- Difficulty aligning parsers
  - Tika better suited compared to PDFminer
  - Highlight extraction **only** with PDFminer

Plan for coming weeks

- Implement other classification methods **(eg. SVM)**
- Test other vectorization methods **(TF-IDF, LSA)**
- Feature engineering for structural features **(Key words, location in document)**
- Look on smaller scale **(Sentence level, split page in paragraphs)**
- **Writing, writing, writing..**