Task definition

Lightning Fast Legal Overview

Exported on 11/01/2018

Table of Contents

1 Project attributes

Project name	Lightning Fast Legal Overview
Project case number	7699
Project task number	10
Project start date	01 Jan 2018
Project due date	31 Dec 2018
Project status	DONE

Stakeholders

See the Project role definition.

Name	Organization
Jens Bligaard	SEGES
Jonas Valhøj Kleffel Nielsen	SEGES
Søren Harrild Eriksen	SEGES
Per Hejndorf	SEGES
Christian Schou Oxvig	SEGES
Unknown User (nicr)	SEGES
Peter Fogh	SEGES

2 Task description

A lot of legal verdicts and laws that impact the daily operation of farms are constantly issued. It is difficult, both for the farmer and his agricultural advisor, to plan the optimal response to a given legal issue. In particular, it is difficult to assess if a given response is in line with related decisions. New data analysis tools, such as IBM's Watson, are able to process a large set of decision much faster than any human being with the purpose of finding patterns and systematize the decisions, such that one can clarify what the current practiced law is. The farmer or his advisor may, with such a system, be able to quickly decide on the optimal way to handle a legal issue, thereby saving time and money. As part of the project, legal decisions and verdicts for a given domain are given as input to Watson or a similar IT system with the purpose of assessing whether such a tool can find patterns for the current legal practice to ensure a good foundation for choosing the best legal response.

The focus is on the case of tax exemption when selling off a house, a farmhouse, or a summer house. In particular, we consider following overall goal:

- Identifying the 5 most relevant previous rulings on tax exemption for a given case of selling off a house, a farmhouse or a summer house (essentially creating a search engine). If appropriate, the user may be prompted to input keywords (say 5) in addition to the specific case.

3 Tasks prioritization

3.1 Main tasks for the data science team

- **Data collection**:
  - The data science team must collect a data set containing all publicly available rulings (text formatted in "Afgørelsesdatabasen[1]" and "SKAT-afgørelser[2]", see below) regarding tax cases to obtain a corpus of domain-specific text as large as possible - this includes cases with tax exemption when selling off a house, a farmhouse, or a summer house.
- **Analyses**:
  - Based on the collected data set, the data science team must define a representation (i.e. a data model) of the individual cases, which allow a systematic assessment of relations between cases.
  - Based on the systematic representation of the data set, the data science team must create a decision tool (e.g. semantic search tool using machine learning) that selects the 5 most similar legal cases to a given textual description of a new/old legal case of selling off a house, farmhouse, or summer house.
- **Evaluation**:
  - Based on the analysis results, the data science team must evaluate the proposed models and methods in order to assess their validity and performance.
  - Based on the analysis results, the data science team must prepare a set of 25 cases containing the input textual representation along with the resulting decision recommendations. This set of 25 cases is to be manually inspected by tax exemption domain experts in order to assess the performance of the decision tool.
- **Report**:
  - Based on the data collection, analysis results, and evaluation, the data science team must write an IMRAD report covering the results.

3.2 Prioritized additional tasks (if time permits)

1. We may expand our prediction models with data from the official legal guidance published by SKAT[3] and/or tax related messages published by SEGES[4].
2. We may include the PDF-based rulings from "Afgørelsesdatabasen" and "SKAT-afgørelser" in our prediction models. (Note: PDF-based rulings in the 'Afgørelsesdatabasen' from 'Skatterådet' can all be found as HTML-based rulings in the 'SKAT-afgørelse')

3.3 Not doing

- We do not investigate whether or not transfer learning using another danish (or another language) domain-specific corpus can improve the model.
- We do not train a machine learning system that predicts whether a given case of selling off a house, a farmhouse or a summer house is eligible for tax exemption (a Yes/No answer).

1 http://www.afgoerelsesdatabasen.dk
2 https://www.skat.dk/skat.aspx?oid=9464
3 http://skat.dk/skat.aspx?oid=124&ik_navn=footer
4 https://www.landbrugsinfo.dk/oekonomi/skat/skattemeddelelser/sider/startside.aspx

3.4 Risks and blockers

- **The textual descriptions of the rulings are very unstructured which may potentially make it impossible to use machine learning effectively. Specifically, the potential problems related to:**
  - **Different judges/secretaries may use different "textual representations" when describing the exact same ruling.**
  - **Some rulings are available in multiple source databases.**
  - **Some rulings are related in the sense that it is the same case considered by the different judging instances (LSR, Byretten, Landsret, etc.)**
  - **The rulings are not annotated with a clear indication of the subject of the rulings and the final verdict.**

3.5 Time estimation

As seen in the table below, we expect that we will use the 760 of the 857 allocated hours to solve the task defined in this document. Thus, providing the remaining 100 hours.

Time	Hours
Allocated time	857
Scoping time (already spent)	230
Estimated time needed for project management (meetings, task refinements, etc.) (to be spent)	130
Estimated time needed for solving the task (to be spent)	400
Remaining time	100

3.6 Description of end product

This is a proof-of-concept that is only to be used as a guidance for possible future uses of machine learning methods for law-related decision support.

3.7 Current state-of-the-art

The use of machine learning (ML) and artificial intelligence (AI) technologies, within the legal businesses, is currently very much on an experimental level in Denmark and elsewhere. However, the largest legal firms such as Bech-Bruun, Plesner, and Kromann-Reumert are actively investing in the area in order to maintain a competitive edge and reduce cost.

In November 2016 a working group within the professional organization “Danske Advokater” was established with the purpose of, among other goals, identifying the technological advances that are already in use or can be expected to come into use in the area of legal advice by the end of 2017. At this time there is no publicly available result of the deliberations of this working group [1].

The large legal firm, Bech-Bruun, is known to work with a platform from British company “Luminance”, which

markets itself as “the leading artificial intelligence platform for the legal profession”. Luminance aids in sorting, clustering and classifying legal documents [2]. The product description on the Luminance website seems to imply a focus on analyzing and comparing contracts [3]. No indication of the effort in setting up and running Luminance is stated.

ROSS is built on top of IBM Watson. ROSS is currently specialized in (US) bankruptcy law, IP law and labor & employment law. There does not seem to be anything currently in the pipeline for Danish language and Danish laws and practices [4].

There are a number of other start-up companies and solutions that are marketed to legal firms, often specialized in specific legal areas and/or legal functionality.

In conclusion, none of the existing solutions appear to specifically target the case of tax exemption when selling off a house, a farmhouse, or a summer house. Potentially, the "Luminance" system may be able to solve the problem, although further investigation is needed to decide if this is the case or not. However, this question is out of the scope of the present task.

3.7.1 Reference list

Number	Title	Author	Date	Link
1	Nyhed fra Danske Advokater: Hvad betyder AI og teknologien for advokatbranchen?	Ulrikke W. Krogbeck	November 11, 2016	https://www.danskeadvokater.dk/Nyheder.aspx?ID=17069&Action=1&NewsId=18292&currentPage=28&PID=28735
2	Bech-Bruun implementerer kunstig intelligens-teknologi i M&A	Christian Ejvin Andersen	August 30, 2017	http://www.bechbruun.com/da/Videncenter/Bech+Bruun+Nyheder/2017/BechBruun+implementerer+kunstig+intelligensteknologi+i+MA
3	Luminance | Platform	N/A	N/A	https://www.luminance.com/product/
4	ROSS Intelligence - Artificial Intelligence Meets Legal Research	N/A	N/A	https://rossintelligence.com/

3.8 Current related SEGES projects

None.

4 Tentative time schedule

This project is part of the time schedule of the Data Science team.

5 Data description

Publicly available rulings regarding tax exemption when selling off a house, farmhouse, or summer house are generally textual descriptions of the individual cases and the final rulings.

Furthermore, the official legal guidance defining the current practice is available:

- C.H.2.1.15 "Reglerne for hvornår fortjenesten ved salget af en ejerbolig er skattefri": http://www.skat.dk/skat.aspx?oID=1948720&chk=214580
- C.H.2.1.16 "Afståelse af landbrugsejendomme med videre med stuehus og af skovbrugsejendomme og blandet benyttede ejendomme med ejerboliger": http://www.skat.dk/skat.aspx?oID=1948733&chk=214580

5.1 Data sources

Source name	Source format	Access	Storage and usage restrictions	Additional information
Afgørelsesdatabasen	Plain text (HTML) or PDF (both are available).	Scrape http://www.afgoerelsesdatabasen.dk/	We have obtained permission to scrape the website. However, the scraping has to be done outside of normal work hours, so it does not disturb the users.	Contains about 16000 rulings dating back to July 1st, 2008. Only about 300-400 (30-40 each year for a period of 10 years) of these are assessed to relate to tax exemption when selling off a house, a farmhouse, or a summer house. About 1000 rulings are only available in PDF format. The site contains all rulings from both 'Landsskatteretten' and 'Skatterådet'. However, note that rulings are not published when they are ruled. Thus, old rulings can be published in 2018 even though they were ruled in 2013.

Source name	Source format	Access	Storage and usage restrictions	Additional information
SKAT-afgørelser	Plain text (HTML).	Scrape http://www.skat.dk/skat.aspx?oid=9464&[5]	We have obtained permission to scrape the website. However, the scraping has to be done with 5-second delay per request.	Contains about 12000 rulings dating back to 2000. Only about 340 (20 each year for 17 years) of these are assessed to relate to tax exemption when selling off a house, a farmhouse, or a summer house. The site contains all rulings from both 'Højesteret', 'Landsretterne', 'Byretterne' and 'Skatterådet'. However, it only contains some of the rulings from 'Landsskatteretten'.

We conclude, based on the additional information: we will only scrape all rulings 'Afgørelsesdatabasen' and 'SKAT-afgørelser', even though there are overlapping instances of the same rules on both sites.

We will only scrape rulings published by 2017-12-31 and back in time.

5.2 Data attribute information of Afgørelsesdatabasen

The web address http://www.afgoerelsesdatabasen.dk/SearchResult.aspx?sb=-date&d2=31-12-2017 shows a list of all rulings published by 2017-12-31 and back in time (both 'Skatterådets afgørelse' and 'Landsskatterettens afgørelse') sorted by date descending. This list is divided into pages, where each page can be selected by the arrow indicators/buttons at the top and bottom of each page.

We can utilize this "verdict_list" to get the following data attributes. The verdict_list provides us with the link to each individual verdict/ruling page, which we can crawl to get the specific text regarding the verdict.

Note that the web address can also be parameterized with 'fn=monthyear%3a201701' such that the list of verdicts only contains items from a specific month in a year, e.g. http://www.afgoerelsesdatabasen.dk/SearchResult.aspx?sb=-date&fn=monthyear%3a201701 only returns verdicts from January in 2017.

5 http://www.skat.dk/skat.aspx?oid=9464&lang=da

Data source	Data attributes name	Data type	Number of records	Number of null	Validated	Cleaned	xpath expression	Additional information
verdict_list	title	text	16.120 (per 01-03-2018)	TBD	☐	☐	//div[@class="itemOverview"]/h4/a/text()	The title is assumed to often contain the string 'Journalnr. <verdict_number>' where verdict_number contain the string format xx-xxxx, e.g. '08-00024'.However some titles contain multipule verdict numbers like 'Journalnr. 14-0381804 og 14-0381808' This verdict_number can be extracted to become an identifier.
verdict_list	link	URL	- ll-	- ll -	- ll -	- ll -	//div[@class="itemOverview"]/h4/a/@href	The link to the individual verdict page, containing the verdict_data.
verdict_list	overview_p1	HTML	- ll -	- ll -	- ll -	- ll -	//div[@class="itemOverview"]/p[1]	Suspicious <p> tag between the title and date attributes, which do not contain any text in the verdict_list pages we have sampled so far, but could contain information. Thus, it should be saved.
verdict_list	date	HTML	- ll -	- ll -	- ll -	- ll -	//div[@class="itemOverview"]/p[2]	The data type is text but can be transformed to datetime.
verdict_list	judicial_authority	HTML	- ll -	- ll -	- ll -	- ll -	//*[@class="itemDetails"]	Contains text to classify a verdict as from 'Skatterådets afgørelse' or 'Landsskatterettens kendelse', i.e. the judicial_authority.

Data source	Data attributes name	Data type	Number of records	Number of null	Validated	Cleaned	xpath expression	Additional information
verdict_data	verdict_HTML	HTML	- l l -	- l l -	- l l -	- l l -	// div[@id="fullviewHtmlContent"]	The verdict text as one large HTML tag with nested tags. We assume this HTML verdict text is available for all 'Landsskatterettens kendelse' our sample of verdicts from different years (2007-2018) all had this HTML content. Based on the sampled verdicts from 'Skatterådets afgørelse', from 2008-2017, the verdicts from 2008 to some time in 2013 also contain this only-HTML content. From 2013 the verdicts from 'Skatterådets afgørelse' begin to be shown as a PDF in a <iframe> tag, where the web address to the PDF can be selected using this xpath expression: '// div[@id="fullviewHtmlContent"]/p/iframe/@src'

5.3 Data attribute information of SKAT-afgørelser

The structure of the SKAT-afgørelser is like Afgørelsesdatabasen, as it also has a "verdict_list" with a link to each individual verdict/ruling page. However, verdicts from different years are split into different lists, accessed via the fans in the top of the list. Thus, we present the link for each year-list, as their difference is the 'oid' URL-parameter (Number of verdicts is record the 01 Mar 2018):

Year	URL	Number of verdicts
2018	http://skat.dk/skat.aspx?oid=9464	92
2017	http://skat.dk/skat.aspx?oid=2924	745
2016	http://skat.dk/skat.aspx?oid=2393	633
2015	http://skat.dk/skat.aspx?oid=535	808
2014	http://skat.dk/skat.aspx?oid=79882	876
2013	http://skat.dk/skat.aspx?oid=69744	919
2012	http://skat.dk/skat.aspx?oid=17412	756
2011	http://skat.dk/skat.aspx?oid=4814	864
2010	http://skat.dk/skat.aspx?oid=1868376	866
2009	http://skat.dk/skat.aspx?oid=1796563	827
2008	http://skat.dk/skat.aspx?oid=1720451	1054
2007	http://skat.dk/skat.aspx?oid=1567793	950
2006	http://skat.dk/skat.aspx?oid=364282	806
2005	http://skat.dk/skat.aspx?oid=219700	550
2004	http://skat.dk/skat.aspx?oid=171725	524

Year	URL	Number of verdicts
2003	http://skat.dk/skat.aspx?oid=160169	591
2002	http://skat.dk/skat.aspx?oid=157410	682
2001	http://skat.dk/skat.aspx?oid=157408	663
2000	http://skat.dk/skat.aspx?oid=157406	28

Thus resulting in 12.397 verdicts published by 2017-12-31 and back in time.

These lists all contain multiple types of verdicts published by SKAT, i.e. verdicts from: 'Højesteret', 'Vestre/Østre Landsret', 'Byretterne', 'Landsskatteretten', or 'Skatterådet'. Note this website contains all verdicts from these judicial authorities, except for the 'Landsskatteretten', where only a subset is found.

Data source	Data attributes name	Data type	Number of records	Number of null	Validated	Cleaned	xpath expression	Additional information
verdict_list	list_SKM_number	text	12.397	TBD	☐	☐	//tr[contains(@class, "TableRow")]//td[2]//text()	The SKM number contains the string 'SKM<year>.<sequential_number_for_verdicts that_year>.<judicial_authority>' or the value 'Ikke nummereret'. The SKM number can be extracted to become an identifier.
verdict_list	list_title	text	- ll-	- ll -	- ll -	- ll -	//tr[contains(@class, "TableRow")]//td[3]//text()	The title of the verdict.
verdict_list	list_link	URL	- ll-	- ll -	- ll -	- ll -	//tr[contains(@class, "TableRow")]//td[3]/a/@href	The link to the individual verdict page, containing the verdict_data.

Data source	Data attributes name	Data type	Number of records	Number of null	Validated	Cleaned	xpath expression	Additional information
verdict_list	list_published_date	text	-ll-	-ll-	-ll-	-ll-	//tr[contains(@class, "TableRow")]//td[4]//text()	The data where the verdict was published on the website. Note that the data type is text but can be transformed to datetime.
verdict_list	list_topic	text	-ll-	-ll-	-ll-	-ll-	//tr[contains(@class, "TableRow")]//td[5]//text()	The data type is text but can be transformed to datetime.
verdict_list	list_document_type	text	-ll-	-ll-	-ll-	-ll-	//tr[contains(@class, "TableRow")]//td[6]//text()	Contains text to classify a verdict as a : 'Afgørelse', 'Bindende forhåndsbesked', 'Bindende svar', 'Dom', 'Forlig', 'Kendelse', 'Kommentar', or 'SKAT-meddelelse'
verdict_data	verdict_HTML	HTML	-ll-	-ll-	-ll-	-ll-	//div[contains(@class, "MPfullText")]	The verdict text as one large HTML tag with nested tags and NO PDFs. A verdict on this website always starts with a list of structured data, specified in the following data attributes. Note that this list of data is not consistant between different verdicts. e.g. some have 1 'Henvisning' row and others have 2 'Henvisning' rows, thus their xpaths expressions by the index on '...//tables[*]' will differ. After the list of structured data, the verdict text itself is found as individual HTML text tags, like <p>, and <h1>.

Data source	Data attributes name	Data type	Number of records	Number of null	Validated	Cleaned	xpath expression	Additional information
verdict_data	verdict_tilte	text	-ll-	-ll-	-ll-	-ll-	//div[contains(@class, "MPfullText")]/h1/text()	The title of the verdict. Contains same value as the data attributes 'title'.
verdict_data	verdict_date	HTML	-ll-	-ll-	-ll-	-ll-	//div[contains(@class, "MPfullText")]//table[1]	'Dokumentets dato'.
verdict_data	verdict_published_date	HTML	-ll-	-ll-	-ll-	-ll-	//div[contains(@class, "MPfullText")]//table[2]	'Dato for udgivelse' same value as the data attributes 'published_date', however, with timestamp and other data format.
verdict_data	verdict_SKM_number	HTML	-ll-	-ll-	-ll-	-ll-	//div[contains(@class, "MPfullText")]//table[3]	'SKM-nummer' same value as the data attributes 'SKM_number'.
verdict_data	verdict_judicial_authority	HTML	-ll-	-ll-	-ll-	-ll-	//div[contains(@class, "MPfullText")]//table[4]	'Myndighed'.
verdict_data	verdict_case_number	HTML	-ll-	-ll-	-ll-	-ll-	//div[contains(@class, "MPfullText")]//table[5]	'Sagsnummer'.
verdict_data	verdict_document_type	HTML	-ll-	-ll-	-ll-	-ll-	//div[contains(@class, "MPfullText")]//table[6]	'Dokument type' same value as the data attributes 'document_type'.
verdict_data	verdict_topic_words	HTML	-ll-	-ll-	-ll-	-ll-	//div[contains(@class, "MPfullText")]//table[7]	'Emneord' NOT the same value as the data attributes 'topic'.

Data source	Data attributes name	Data type	Number of records	Number of null	Validated	Cleaned	xpath expression	Additional information
verdict_data	verdict_resume	HTML	-ll-	-ll-	-ll-	-ll-	//div[contains(@class, "MPfullText")]//table[8]	'Resumé' short text which summarising the verdict.
verdict_data	verdict_warrent	HTML	-ll-	-ll-	-ll-	-ll-	//div[contains(@class, "MPfullText")]//table[9]	'Hjemmel' link or text to the law which is the bases of the verdict. Initial investiages shows this attributes not alway exists for a verdict.
verdict_data	verdict_references	HTML	-ll-	-ll-	-ll-	-ll-	//div[contains(@class, "MPfullText")]//table[10]	'Reference(r)' link or text to the law related to the verdict.
verdict_data	verdict_guidance	HTML	-ll-	-ll-	-ll-	-ll-	//div[contains(@class, "MPfullText")]//table[11]	'Henvisning' link or text to source for guidance related to the verdict.
verdict_data	verdict_text	HTML	-ll-	-ll-	-ll-	-ll-	Concat elements fetch by the Xpath expressing: //div[contains(@class, "MPfullText")]//hr//following-sibling::*	The text of the verdict.

6 Additional information

This project is part of the PAF project: "Øget konkurrencekraft i landbruget gennem brug af kunstig intelligens." - Ansøgning til promilleafgiftsfonden for landbrug 2018.pdf