Gedetailleerd overzicht van de use case voor de hackathon

Inleiding

Deze hackathon wordt georganiseerd door de Dienst van de Bestuursrechtscolleges (kortweg de DBRC). De Dienst van de Bestuursrechtscolleges is specifiek opgericht door de Vlaamse decreetgever om tot synergiën te komen tussen de verschillende participerende rechtscolleges om zo onder meer schaalvoordelen te bekomen, beheerskosten te verminderen en efficiënter te kunnen werken.

De DBRC ondersteunt en overkoepelt volgende drie onafhankelijke Vlaamse administratieve rechtscolleges:

1. het Handhavingscollege
2. de Raad voor Vergunningsbetwistingen
3. de Raad voor Verkiezingsbetwistingen.

In de toekomst zouden deze verantwoordelijkheden kunnen uitbreiden met andere rechtscolleges. De DBRC wil digitalisering omarmen en de procespartijen efficiënter gebruik laten maken van haar diensten.

Deze hackathon heeft als doel de DBRC een eerste zicht te geven hoe de DBRC-arresten beter kunnen ontsloten worden voor de buitenwereld. De DBRC wil onderzoeken of NLP hun huidige zoekfunctie kan verbeteren en de manuele tagging van arresten overbodig kan maken. De huidige zoekfunctie voldoet niet meer aan de wensen van de gebruikers en is daarom toe aan vernieuwing. Meer specifiek wil de DBRC een bepaalde passage van de arresten van haar rechtscolleges beter doorzoekbaar maken. Momenteel kunnen de arresten teruggevonden worden op de volgende site: https://www.dbrc.be/rechtspraak.

Het resultaat zal **iedereen** de mogelijkheid bieden om beter door alle arresten te zoeken. Deze zoekfunctie zal meer specifiek kunnen zoeken in de data en ook verbanden kunnen leggen tussen verschillende arresten. Voor interne gebruikers kan het de werkdruk verlagen door de automatisatie.

Hoe inschrijven en deelnemen?

Inschrijving gebeurt via het inschrijvingsformulier:
https://docs.google.com/forms/d/e/1FAIpQLSfN8WaQoX7jQ2RV3DJhmWSXUnBz1MPrX3Mz1x94AUvE3vQn7Q/viewform

Uitleg van de opdracht

Momenteel gebeurt er veel handmatig werk bij de DBRC om de arresten te ontsluiten (arresten worden handmatig gelabeld). Het automatiseren van deze taak zal de juristen ontlasten. Om dit probleem op te lossen wordt gedacht aan automatische tagging op basis van NLP zodat het opzoeken ook eenvoudiger wordt. Documenten die relevant zijn voor een bepaald thema worden immers bij elkaar gebracht en gebruiksvriendelijk getoond worden.

De labels zijn deels hiërarchisch gestructureerd, maar er is geen (volledige) standaardisatie. Het kan ook zijn dat gelijkaardige labels bestaan die eigenlijk hetzelfde zijn en gelijkaardige labels die verschillend zijn.

Het is de bedoeling om de bestaande labels toe te kennen aan de arresten, maar dan enkel voor de onderdelen "**beoordeling door de Raad**" uit het arrest.

Voor elke PDF gaat het relevante gedeelte van het document uit de tekst moeten gehaald worden. Aangezien het gedeelte telkens voorafgegaan wordt van de titel "Beoordeling door de raad", kan je hierop parsen, zoals in het stukje R code hieronder.

```
library(textreadr)
library(tm)
library(pdftools)
library(stringr)
library(NLP)
test_pdf  <- pdf_text("C:/Users/USER/Downloads/RVVB.A.1819.0002.pdf")
relevant_part <- sub( ".*Beoordeling door de Raad", " ",
as.String(test_pdf))
```

Hoe matchen?

De hackathon wordt georganiseerd om nieuwe en creatieve oplossingen voor dit probleem te vinden. Deelnemende bedrijven kunnen vrij kiezen welke tool ze gebruiken. Hieronder worden enkele voorbeelden gegeven hoe deze opdracht zou kunnen uitgevoerd worden. Het voorbeeld is louter een illustratie.
Deze methodiek hoeft dan ook niet gevolgd worden voor het uitvoeren van de opdracht.

Mogelijke methodiek:

Als eerste stap zou één van deze pre-trained woord vector modellen gebruikt kunnen worden:

- https://fasttext.cc/docs/en/crawl-vectors.html
- http://vectors.nlpl.eu/repository/

Via deze modellen kunnen woorden en termen gecodeerd worden in een ‘betekenis vector’ (meaning vector). De bedoeling is om de teksten:

1. verschillende tag-woorden te geven.
2. met elkaar te matchen.

Het doel is dit zo goed mogelijk te verwezenlijken. Dit kan je onder meer doen door de teksten te ontdoen van frequente woorden. Hierna kan de gemiddelde word2Vec vector berekend worden van de overgebleven woorden (we raden aan om de tekst in stukken te knippen, aangezien verschillende paragrafen vermoedelijk andere thema’s kunnen bevatten). Een voorbeeld hiervan is: http://www.rpubs.com/mukul13/rword2vec (https://github.com/mukul13/rword2vec)

Deze gemiddelde vectoren kunnen dan vergeleken worden met de vectoren van de tag-woorden op hun similariteit. Deze mate van overeenkomst kan bijvoorbeeld bepaald worden door cosine similarity. Als de matching een bepaalde score (bvb. 0.85) overschrijdt, kan je aannemen dat de tekst en het tag-woord bij elkaar horen.

Het resultaat zal dus een reeks van triples of quadruples zijn:

Identifier	Type relation	Keyword	Matching probability
(text 1)	(has keyword)	Ontvankelijkheid middel	.78
(keyword) Raakt aan vergunbaarheid project	(is a subclass of)	Ontvankelijkheid middel	1.00
(text1)	(similarity)	(text2)	0.81

De bedoeling van deze hackathon is om een zo goed mogelijke mapping te hebben van de onderwerpen met de teksten.

Andere methoden die aanvullend kunnen werken zijn klassieke regex (al dan niet met fuzzy matching), bag-of-words, eventueel transfer learning van de bestaande word2vec modellen richting meer juridische taal, LDA of andere NLP technieken.

De aangeleverde data & evaluatie van de resultaten:

De DBRC heeft al verschillende databronnen voor het bereiken van betere ontsluiting van de rechtspraak. Meer specifiek zal de informatie uit deze documenten gedeeld worden:

- 1000 tal **PDF's van de Arresten**, die je ook publiek hier kan terug vinden : https://www.dbrc.be/rechtspraak
- Eén **woordenlijst digitalisering**: hoofd-labels met daaronder meerdere sub-labels.
- Overzicht **EVR - documenten:** Deze bevatten per dossier het type arrest en handmatig toegevoegde tag-woorden

Een deel van de teksten die handmatige getagd zijn (EVR-documenten) zullen gebruik worden voor de **evaluatie van de resultaten**. Er zal gekeken worden naar de **accuraatheid**, **recall** en **precision**. (https://en.wikipedia.org/wiki/Precision_and_recall).

Voor de woordenlijst en de EVR-documenten geven we ook een iets handiger formaat mee (alsook de code in R waarmee deze parsing is gebeurd), zodat je niet meer zelf alle data-extractie moet doen.

Wat matchen?

1) Een matching dient te gebeuren met de woordenlijst digitalisering
2) Een matching dient te gebeuren met de woordenlijst EVR
3) Een matching tussen arresten onderling

Bijkomend kan er een clustering gedaan worden op basis van de EVR woordenlijst (gelijkaardige woorden of zinnen met eenzelfde betekenis kunnen gegroepeerd worden).

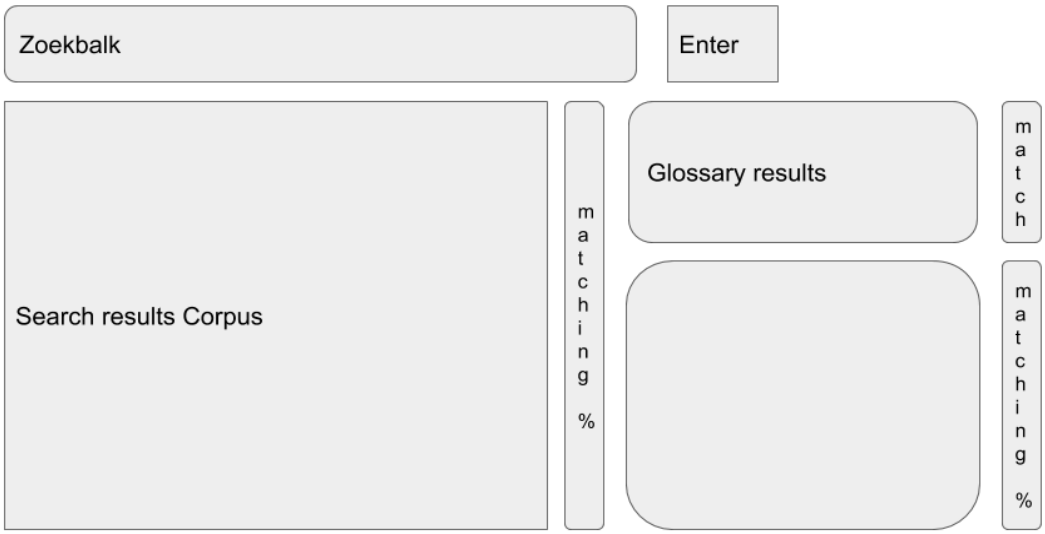
En daarnaast kan het interessant zijn om tijdens de hackathon (naast de bovengenoemde matching) ook functionaliteit te ontwikkelen om nieuwe onderwerpen te zoeken in de arresten. Dit gedeelte is optioneel.

Front-end en visualisatie

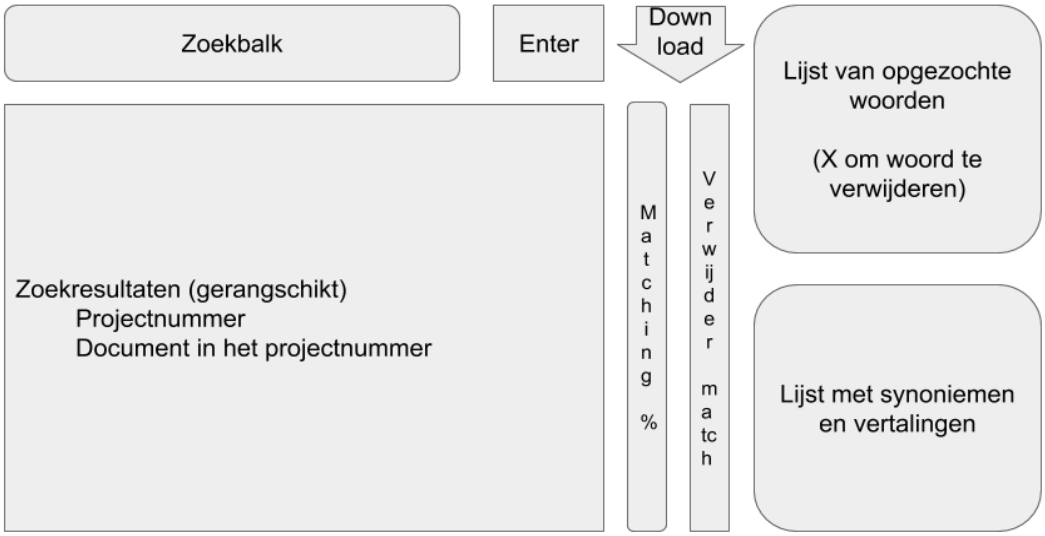
Naast de back-end engine, willen we ook een demo zien van de front-end.
Vanuit de verzamelde feedback van de gebruikers, is er alvast een bestaand voorbeeld dat handig bevonden wordt: http://juridict.raadvst-consetat.be/index.php?lang=nl.
Een gelijkaardige structuur van opzoeken, lijkt daarom ook een nuttige manier van werken.

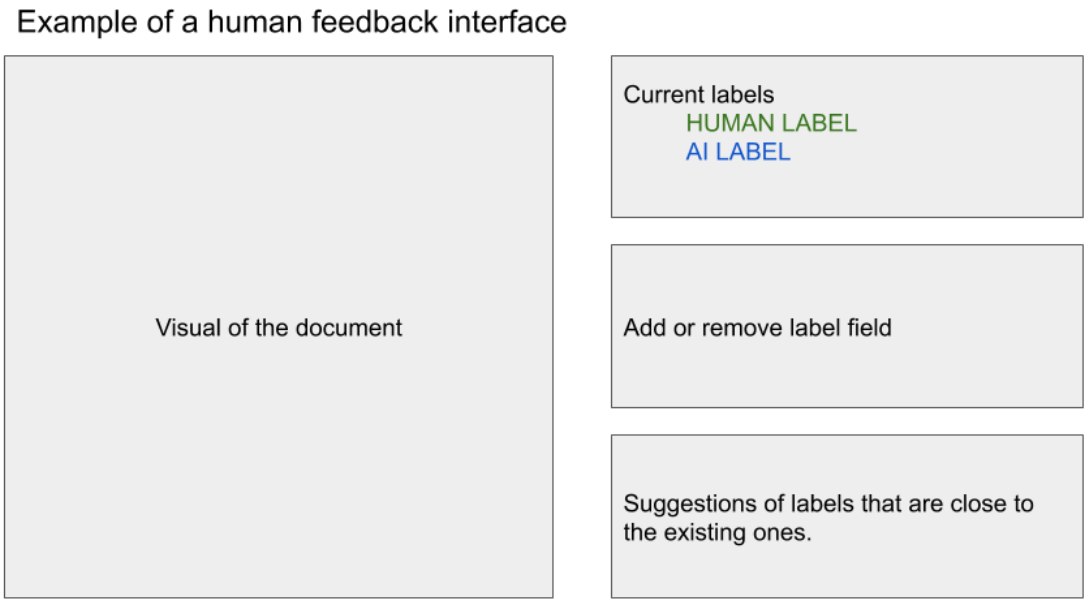
Daarnaast zijn er nog manieren van visualisatie zoals de voorbeeld schema's hieronder:

Voorbeeld gebruikersinterface



Zoekmachine: front-end voor de eindgebruiker





De exacte lay-out is minder belangrijk dan het aantonen hoe de zoekfunctie werkt en wat er getoond kan worden. Het is belangrijk dat de userinterface als webpagina kan gebouwd en gebruikt worden.

Oplevering resultaten

Volgende documenten moeten opgeleverd worden voor de finale evaluatie:

- Een databestand met de linken tussen:
  - Documenten en hun zoekwoorden
  - Documenten en hun onderlinge verwantschap
  - Zoekwoorden en hun hiërarchie.

- Een apart databestand in hetzelfde formaat als hierboven voor de ‘test-set’ : deze zal gebruikt worden om het model te evalueren. Dit bestand dient in .txt, .csv of .xlsx aangeleverd te worden.

- Een demonstratie van de front-end op vrijdag 13 maart

Alle bereikte resultaten worden op een open manier getoond en gedeeld met alle kandidaten aan de hackathon.

De databestanden mogen doorgestuurd worden naar christophe.cop@pwc.com

Indien er nog bijkomende vragen of onduidelijkheden zijn met betrekking tot de hackathon, kan u ons steeds bereiken op bovenstaand e-mailadres.

Tijdlijn

De Hackathon zal één week duren. Tijdens deze week zullen er **twee contactmomenten** zijn. Het **eerste moment** zal doorgaan op vrijdag 6 maart in de voormiddag. Dit is dan ook de start van de hackathon. Dit is de agenda voor deze voormiddag:

Planning vrijdag 6 maart (voormiddag)		
Uur	**Wat**	**Wie**
09u30:	Onthaal + koffie	PwC
10u00:	Welkom + inleiding	CM - PwC
10u10:	Woordje DBRC	DBRC
10u20:	Uitleg van de opdracht	PwC
11u00:	Q&A	
11u20:	Overhandigen data + Koffie + nabespreking	

Het **tweede contactmoment** wordt georganiseerd op vrijdag 13 maart in de voormiddag. Tijdens deze voormiddag worden de laatste loodjes gelegd en wordt de hackathon afgerond. De planning voor vrijdag is hieronder te vinden:

Planning vrijdag 13 maart (voormiddag)		
Uur	**Wat**	**Wie**
09u30:	Onthaal+ koffie	PwC
10u00:	Welkom + inleiding	MC - PwC

10u10:	Welkom + inleiding + deadline validatie datase	
10u30:	Presentatie van team 1	
11u00:	Presentatie van team 2	
10u50 - 11u10	[etc: 20’ per team : strikte timing!]	
10u10 + 20*n	Koffie & (indien validatie al gebeurd) : bekendmaking winnaar.	