Text extraction benchmark - test models definitions (PSMM - MS Word)
Text extraction benchmark - test models definitions (PSMM - Libre Office)
Model transformations
QvT language
12 model transformations defined
Text extraction benchmark - test models definitions (model queries)

• Ground truth and failure diagnostic data
• Model queries
• OCL language
• 20 model queries defined
• additional queries for text elements

Property : Font Color
Context : PIM (Document)
Query : self.pages.elements->selectByKind(TextContainer).fontColor
Test data generation
Test data generation
Test data generation
Test data generation
Generated dataset v1

- Size
  - 60 PDF document
    - 20 in version 1.4
    - 40 in version 1.5
  - 40 DOCX Transitional (Office Open XML)
  - 20 Open Document text Document

- Test case
  - document file
  - text snippets ground truth
  - metadata ground truth

- Published on Figshare together with benchmark results!

http://dx.doi.org/10.6084/m9.figshare.c.3683332
Software evaluation process
Software evaluation process

• Identifying a text snippet in the extracted text:
  • Based on the Levenshtein edit distance algorithm
    • operating on a word level (instead of character level)
  • Added modification which identifies portion of a text with the smallest edit distance
### Benchmark results - all formats (Apache Tika v1.13)

<table>
<thead>
<tr>
<th>nonjuristic</th>
<th>charlatan</th>
</tr>
</thead>
<tbody>
<tr>
<td>86</td>
<td>91</td>
</tr>
<tr>
<td>68</td>
<td>61</td>
</tr>
<tr>
<td>93</td>
<td>93</td>
</tr>
</tbody>
</table>

sociesthesia consentingly cyanem grimp ungagged skatoseine aquatical Plutarchial neuroapophysis cosmopolicy cinquefoil defrock caviya dote polymastie Halloplodaei Penelopef pouchd dishonorable teacherly preinvocation measurelessly noctambulism scribodeion Giils dorts clorigen circumbagious preconversion salramder rhysimeter Seltzer renmonstrance underbraced Ptarmica counterefficiency bonehead asklen Havakian consumativeness theorapiam decomposite superfulfilment nitrosulphate vulvouterine blaster phleborrhage zadruga terebraitunie Wilsonian relime corta craniographer

blossomtime consonantnessouchy floristically debarras purport overexpert necklet nototribe unwivingly clearance iceberg inequently Cylleius bosh arfoalo gel unacatholicaeness antifanatic swampable unstoicize hectogram oppressed somnumbule wouns nudibranchian Messines alchemistic unretained urateribolitasis outshake arish unvisoned Penolepias Bahasian claim oxalic demagnetize droplet cometwise nonglucose Pisian overegerness toule unprojeted kaiwi goffle pendeloque cheddite baggie hydraulican disrobenent clopghed gesticulacious anteresurrection fretty

<table>
<thead>
<tr>
<th>anteresurrection</th>
<th>fretty</th>
</tr>
</thead>
<tbody>
<tr>
<td>83</td>
<td>82</td>
</tr>
<tr>
<td>24</td>
<td>48</td>
</tr>
<tr>
<td>94</td>
<td>59</td>
</tr>
<tr>
<td>24</td>
<td>52</td>
</tr>
<tr>
<td>14</td>
<td>42</td>
</tr>
</tbody>
</table>

Buildmon canalboat elecampane geogenesis subepithelial physicalism hotelfeepkeeper argentine premixture asthenopia lambiy speculor hypothermic remagnetization gother phanerogam cytoproct bangy hobbless tickled Prosopism sweat shortschat unbordered reminiscential rampageously voluta deciphemer airdale respeak uncentrality oversale farish postillation unsted Sildago ineptly elagage confinginess intactness isocolic Nazarean bacteriostropın _ emblem plicable acreak_ aspersor overbrightness

sociesthesia consentingly cyanem grimp ungagged skatoseine aquatical Plutarchial neuroapophysis cosmopolicy cinquefoil defrock caviya dote polymastie Halloplodaei Penelopef pouchd dishonorable teacherly preinvocation measurelessly noctambulism scribodeion Giils dorts clorigen circumbagious preconversion salramder rhysimeter Seltzer renmonstrance underbraced Ptarmica counterefficiency bonehead asklen Havakian consumativeness theorapiam decomposite superfulfilment nitrosulphate vulvouterine blaster phleborrhage zadruga terebraitunie Wilsonian relime corta craniographer
Benchmark results - all formats (Apache Tika v1.13)

<table>
<thead>
<tr>
<th>nonjuristic</th>
<th>charlatan</th>
</tr>
</thead>
<tbody>
<tr>
<td>86</td>
<td>91</td>
</tr>
<tr>
<td>68</td>
<td>61</td>
</tr>
<tr>
<td>93</td>
<td>93</td>
</tr>
</tbody>
</table>

societism consentiently cyanem grimp ungagged skatosine aquatical Plutarchial neuropophysics cosmopolicy cinquefoil defrock caviya dote polymasty Halloipodidae Penelopephon pouched dishonorable teacherly preinvocation measurelessly noctambulation scribbledon Gilis dorts cloragen circumbagambo preconversion salamander rhyngemony Seltzer remonstrance underbraced Pterancia counterefficiency bonehead askil Havian consumptiveness thephantom decomposite superfulfilment nitrutesulphate vulmouteine bloater phleborrashage zadruga tererebutiline Wilsonian relime corta craniographer

blossomtime consistimentioughly floristically debarrass purport overexpert necklet nototrib unwinningly clearance iceberg ineperctly Cyllenus bosh afoiog gel uncatasonic antifanatic swampable untozitise hectogram oppressed sommambule woons nudibranchian Messines alchemistic unretained urerolithiasis outshake arrish unvisioned Penelopeinae Bahanism claim oxalic demagnetize droplet cometwise nonglucone Pisian overagenerness toule unprojected kiwi goffle pendelico cheddite baggie hydraulician disroovement cloathed gesticulacious anteresurrection fretty

<table>
<thead>
<tr>
<th>anteresurrection</th>
<th>fretty</th>
</tr>
</thead>
<tbody>
<tr>
<td>83</td>
<td>82</td>
</tr>
<tr>
<td>24</td>
<td>48</td>
</tr>
<tr>
<td>94</td>
<td>59</td>
</tr>
<tr>
<td>24</td>
<td>52</td>
</tr>
<tr>
<td>14</td>
<td>42</td>
</tr>
</tbody>
</table>

Buildmon canailoat eclempame geogenesim subepithelial physicialism hotelkeeper argentine preminture asthenopia lamby specular hyperthermic remagnetization getter phareogram cytoptic bingy hobbyless tickled Prospisum sweat shortschat unbordered reminisental ramagously vuluta decipherment airedale respeak uncentrality oversale farish postillation untinged Solidago inepte elagate confindingness intactness isocolic Nazarese bacteriotrop emblem plicable acread asensor overbrightness
Conclusions

- Software benchmarks are needed in the digital preservation field
- Synthesised test suits are an important alternative to manually annotated test suits
  - less resources needed
  - easy to share
- MDE approach offers several important features
  - **Scalability**
  - **Configurability and extensibility**
  - **Representativeness**
    - feature distributions
    - using real world software environments
  - **Rich ground truth**
    - reusability of generated test suites
    - visual ground truth
Conclusions

- Text extraction tools
  - tools are not perfect
  - majority have problems with order and structure in the presence of specific features
    - trying to be too smart is not the best idea!
  - we have uncovered already known bugs
  - covered visual aspects
Conclusions

• Challenges
  • Shift in human expertise from manual effort intensive labelling to modelling
    • domain knowledge is important
  • How to generate erroneous test cases?

• Other scenarios?
  • Easy to extend to cases such as HTML and Latex (PDF)
  • What about binary file formats?
  • What about scenarios with multiple file formats
    • format identification
Conclusions

- Challenges
  - features are not independent
  - we are doing first order approximation

- How do we extend to second order?
- Smarter transformations?
- How many and which features would be needed to describe a document structure?