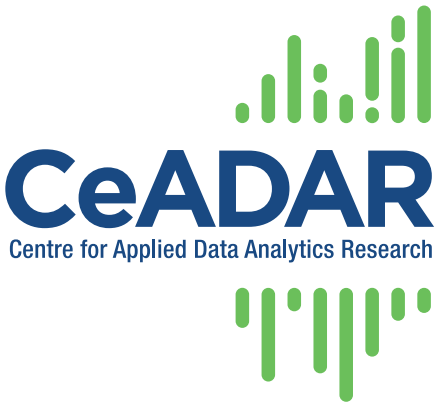

CeADAR – Centre for Applied Data Analytics Research

Enterprise Ireland Data Analytics Technology Centre

DocoPool – Harnessing Document Knowledge

Document Type:	Literature Review
Project Title:	CeADAR
DDN Theme:	1 – Intelligent Analytic Interfaces
Theme Leader:	Sarah Jane Delany (DIT)
DDN Sub-Theme:	Ease of Interaction
Authors:	Ingo R. Keck, Susan McKeever
Document Version:	1.0
Date of Delivery to ISG:	10th August, 2015
Number of pages:	13

CeADAR is a research partnerhip comprising University College Dublin, University College Cork, and Dublin Institute of Technology.

http://www.ceadar.ie

Contents

Executive Summary

In this review, we discuss the state-of-the-art of both research and development in Harnessing Document Knowledge (DocoPool):

- At present, commercially, documents are usually managed in Document Management Systems (DMS) or Enterprise Content Management (ECM) Systems. The market for DMS/ECM is highly diverse: a search on Capterra yielding 65 results of commercial solutions. Products like OnBase or ABBYY FlexiCapture offer document classification based on content and metadata;
- The workflow of DMS/ECM consists of document import, analysis, information extraction, management and archiving. DMS/ECM products offer import directly from a business workflow, manually or via scans of paper documents, with a view to achieving the "paperless office" and streamlined workflows. The purpose of DocoPool is to analyse specific document pools for knowledge gain. It will not focus on workflow or large-scale document management;
- DocoPool will require a document conversion from digital format to either a structured or unstructured text format. A number of digital document import solutions are reviewed;
- Ideally, the input documents will contain typed characters. Some companies rely heavily on paper originated documents stored as scanned documents. We will examine existing solutions to convert scanned documents using Optical Character Recognition (OCR), in case this functionality can be included within the timeframe.
- DocoPool will focus on applying text mining techniques to unstructured text, incorporating elements of document clustering, exploration and classification. The state of the art of these fields is investigated;
- We also discuss publicly available text documents that may be useful for the development of DocoPool;
- There are many commercial solutions that enable the capture and management of electronic documents. We conclude however that many commercial solutions fall down in their limited ability to analyse document data. The focus of DocoPool is to mine documents for knowledge, focussing on advanced document clustering, cluster visualisation and an exploration of similarities or differences across documents. In addition to supporting specific text mining use cases, DocoPool will serve a wider purpose of demonstrating the use of text mining techniques to CeADAR companies.

1 Introduction

Existing Document Management Systems are widely used in industry, supporting the capture, storage and mangement of electronic documents, in a variety of formats (such as PDF, Word). The aim of DocoPool is to support the extraction of knowledge from these document stores. DocoPool will support a number of specific text-mining use cases, allowing pools of documents to be searched, explored and analysed, with results presented in easily understood visualisations. This will support business users in gleaning knowledge from electronic documents in a rapid and automated way, reducing the need for manual effort.

2 Industry Benefit

Companies frequently need to find or analyse information that is held across a variety of documents and document formats. Such documents are difficult to parse and require substantial manual effort to examine or search. This project will focus on developing an easy-to-use tool that implements a set of text-mining based use-cases to derive knowledge from electronic documents. Documents will be converted into useful formats to support document exploration, search and additional analytics tasks. This tool would be valuable to any business that accumulates substantial numbers of electronic documents.

Some particular "document heavy" scenarios include:

- Insurance companies that need to explore or analyse documents for a particular claim or set of claims;
- Legal practices that need to analyse documentation for particular cases;
- Advertising companies with reliance on unstructured advertising information and data held in documents;
- Recruitment companies with applicant and employer information.

The following is a list of potential use cases for DocoPool, based on dscussions with companies, to be further validated and refined during the next phase of the project:

- The ability to automatically cluster a pool of documents into a set of topics (topic modelling), allowing users to identify occurrences of rare and common words. Topic modelling is a key tool for the discovery of hidden structure in large collections of documents;
- The ability to perform Named Entity Recognition (NER) on a set of documents - whereby named entities such as names and addresses can be automatically extracted;
- The ability to analyse documents for multi-word phrases (n-grams), with visualisations, such as the mapping of occurrences of n-grams over time.

The above will be supported by easy to interpret visualisations. In addition, these functions assume the availability of documents in a suitable format for analysis. DocoPool will therefore support some level of document upload, as a precursor to enabling document analysis.

3 Technology Description

The demonstrator will allow the input of a number of related text files, likely to be in a number of different application formats, such as PDF, Word and scanned documents. The system will perform a user-directed search and /or analysis through these documents and extract

relevant information. There are several document search tools already freely and commercially available. The innovation in the DocoPool tool will be:

- The ability to perform higher-order analysis on document pools, beyond simple search results – such as similarity/differences/clustering on keywords and meta data;
- Effective visualisations to enable business users to easily understand the output of analyses;
- The application of text mining and analysis to specific company-driven use cases.

The core idea of DocoPool bears a similarity to existing enterprise document management systems. These systems usually define a workflow of document input either in digital or scanned format, while also allowing for document analysis, information extraction, document storage and archiving. We will discuss here the state of the art of these systems.

DocoPool will require a document conversion from either scanned or digital format to a structured or unstructured text format. Therefore we will look here at digital document import solutions.

DocoPool will incorporate elements of document clustering, exploration and classification. The state of the art of these fields will be discussed.

Finally, we will discuss publicly available text documents that may be useful for the development of DocoPool.

In the following sections we explore Document Management Systems (Section 3.1), Document Import (Section 3.2), Optical Character Recognition (Section 3.3), Document Clustering (Section 3.4) and Public Document Data Sources (Section 3.5).

3.1 Document Management Systems (DMS)

A wide variety of document archiving systems exist on the market to support the creation and management of document repositories, and paperless workflows. Here is a selection of the most commonly used systems (as, for example, reported by [12]) which often form the basis of more specialised solutions by other companies:

- Alfresco (`https://www.alfresco.com/solutions/document-management`) is a popular document managing system also offering a cloud service, citing KLM and NASA as customers. It offers full text search in documents, and search based on tags, but not automatic clustering of documents. All tags and the category of the document have to be added manually.
- Docstar (`http://www.docstar.com/`) is a DMS that optimises the business workflow. It allows the definition of document templates and can force the user to manually control documents if they contain all necessary information depending on the document type. It does not seem to have automatic tagging facilities or clustering of documents.
- Evernote (`http://www.evernote.com/`) is a *note centred* document managing system. Typical office files and paper scans can be attached to notes and are automatically converted to text, which can then be searched. Evernote does not cluster documents, all tagging has to be done by the user.

- DocuXplorer (`http://www.docuxplorer.com`) is a document management system that allows the import and indexing of all kinds of text documents, includes optical character recognition (OCR) and allows full text search of archived documents.
- DocuWare is another popular DMS. An interesting function of DocuWare is the so called *intelligent indexing*, which can automatically extract meta data from documents by learning their position or by detecting keywords and applying this knowledge to similarly formatted newly imported documents[1]. At the same time it uses this information to automatically classify the imported document. DocuWare offers a complete workflow for this classification and gives visual feedback to the user of the information extraction and classification performance for each document at the import, so the user can manually train the system for documents that the system does not already know or has problems handling correctly.
- OnBase (`https://www.onbase.com/en/product`) is an Enterprise Content Management (ECM) solution that also includes document import, OCR of scanned documents, full text search, automatic document type classification based on document layout and semantic content (`https://www.onbase.com/en/product/onbase-anydoc/data-capture/classification`) and automatic information extraction of data with optional verification against existing data sets.
  OnBase has a use case for insurance claim handling: `https://www.onbase.com/en-GB/solutions/insurance/property-and-casualty/claims-processing`.
- Recall (`http://www.recall.co.uk`) offers tailored solutions based on OnBase.
- Alos Solutions offer a variety of products that cover document scanning and import, automatic document classification and information extraction based on ABBYY FlexiCapture, document management DOCUWARE, archiving and workflow management.
- Enadoc (`http://www.enadoc.com`) is a document imaging system that concentrates on data extraction from structured and semi-structured documents.

This list composes only some of the available products in the market. A search on Capterra for *document management system* with *document conversion, document indexing, full text search, OCR* yields 65 results of potentially interesting software solutions in this field. [2]

Several of the systems described incorporate workflow and process automation, which are outside the focus of DocoPool. We also note that while the systems described allow a user to conduct most typical document management tasks, they lack the ability to do more sophisticated analyses using text clustering, aided by visualisations.

3.2 Document Import

Most commercial document management systems provide import filters for several kinds of data formats [1]. Apart from that, for typical text processing document formats like *Microsoft Word, Excel* or *ODT*, an open source solution like LibreOffice (`http://libreoffice.org/`) or OpenOffice (`http://openoffice.org`) can be scripted from the command line to convert the content of the document in question into a pure text output. Using the scripting included in

[1] `https://www.youtube.com/watch?v=kR1oT7KxOek`
[2] `http://www.capterra.com/document-management-software`

both programs, an input filter could be programmed to extract information out of structured representations.[3]

AMI (`https://bitbucket.org/petermr/ami-core`) is able to work with PDF, XML, HTML data to extract text and structured information, even analysing scientific diagrams included in these files.

3.3 Optical Character Recognition Solutions

Company documents may originate as paper documents, and then stored electronically as scanned documents. Optical Character Recognition (OCR) is required to extract the text content from such documents. Many OCR systems exist in the market. However most of the solutions are not easily scriptable to be implemented in an automatic system. That mainly leaves three solutions:

- The open source project Tesseract (`https://github.com/tesseract-ocr`) is backed by Google and used for Google books. In comparisons it outperforms other open source libraries [9] and gives comparable results to the commercial alternative ABBYY FineReader Engine CLI.[11]
- ABBYY FineReader Engine 11 CLI (`http://www.ocr4linux.com/en:start`) is a commercial OCR solution that supports 190 languages and operates from the Linux command line. Compared to ABBYY FlexiCapture, it focuses on retaining the document layout and does not offer options for automatic document classification and information extraction [3].
- ABBYY FlexiCapture (`http://www.abbyy.com/flexicapture/`) is an intelligent system that analyses the content, the images and the layout of a scanned document (handwritten, printed or form) to automatically detect the type and class of a document and extract relevant information [2, 4]. It can learn to classify documents based on their layout, images (e.g. company logos at given positions) and their content. It is able to extract information from documents based on the position in the document or surrounding keywords. ABBYY FlexiCapture can detect information like delivery addresses in documents (see Figure 1) and also cites automated insurance claim handling as a use case. ABBYY FlexiCapture focuses on *document separation*, *classification* and *data extraction*.

These technologies are relevant since DocoPool may support the upload of scanned documents. As with the document import function, existing OCR solutions will be used within DocoPool where possible, if this feature is implemented.

3.4 Document Clustering

Document Clustering supports the analysis of text document contents, by grouping or clustering documents based on similarities in text across documents.

Lingo3G (`http://carrotsearch.com/lingo3g-overview`) is a commercial text mining software engine. It categorises documents based on their text and includes a front end to visualise and navigate the clusters for experienced users (Figure 2) as well as for non-expert users.

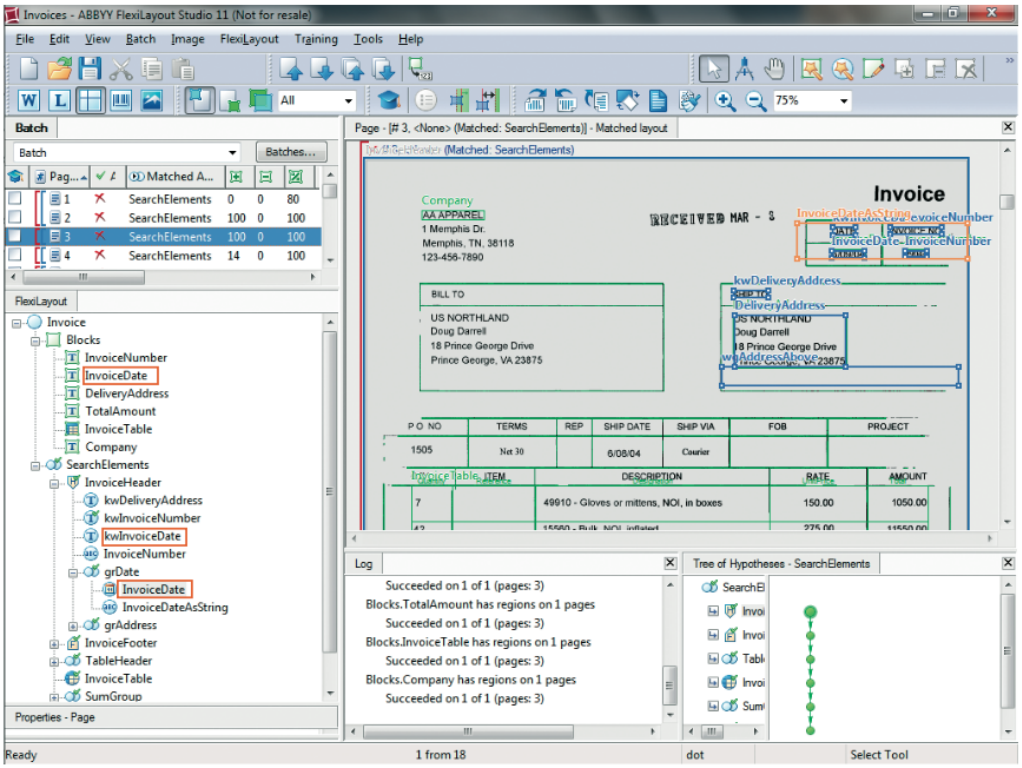

Figure 1: Document analysis and address detection on ABBYY FlexiCapture [5]

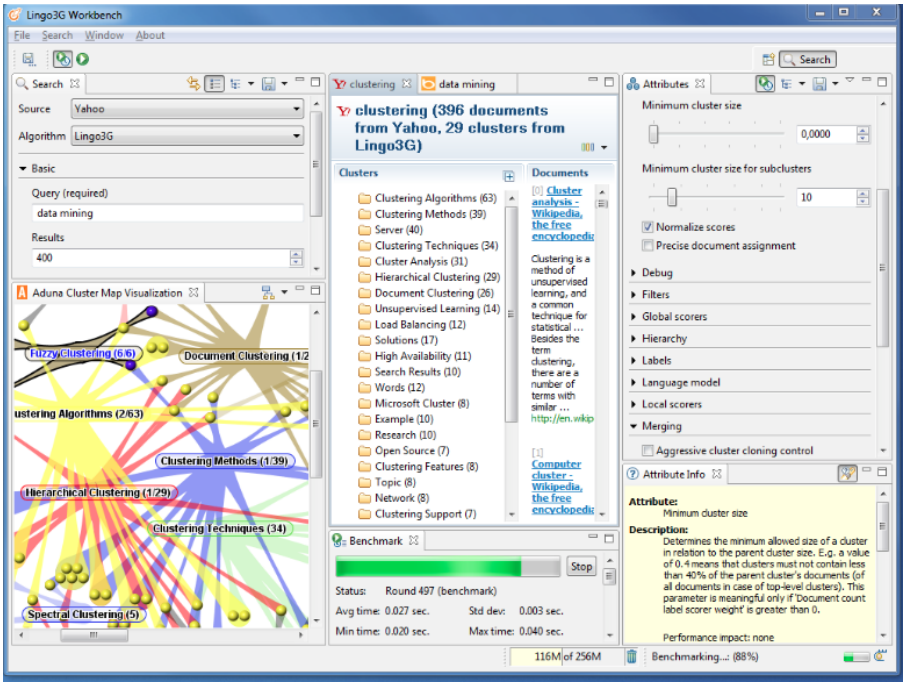

Figure 2: Lingo3G visualisation and benchmarking front end "Document Clustering Workbench" for the experienced user.

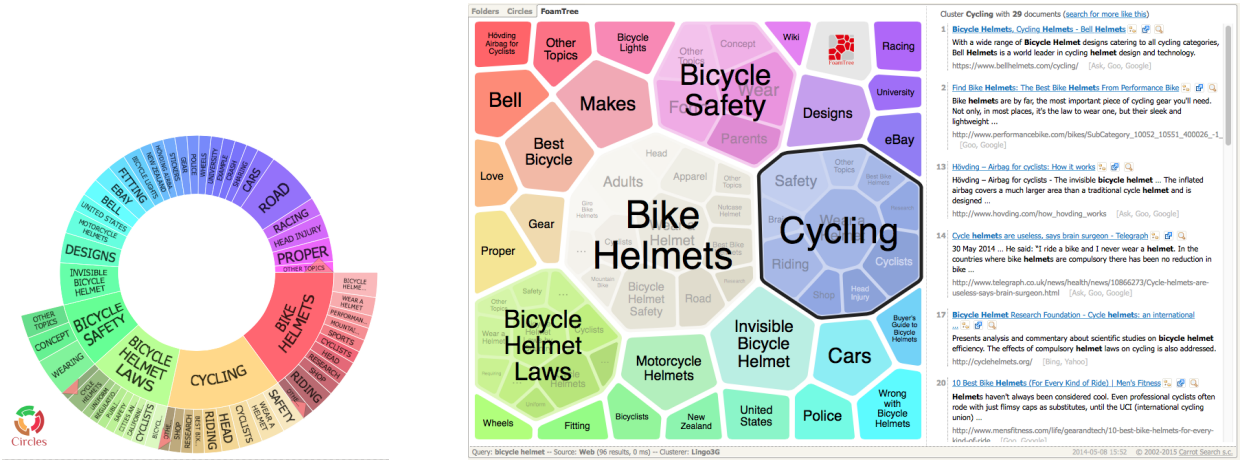

Figure 3: Lingo3G circle (left) and foam (right) visualisation of results for a web search for "bicycle helmet". The user can click on clusters and get an overview of the documents in the cluster.

Lingo3G also makes its capacities available as an API (java, C#), as a server (REST interface) or as a command line tool.

Lingo3G offers an online demo under `http://search.carrotsearch.com/carrot2-webapp/` that is focussed on non-expert users. Here Lingo3G provides a quick and easy to navigate overview of a set of documents after document classification. The user can move through hierarchies of document clusters either by topic or by navigating through a foam or circle visualisation (see Figure 3). They can also use word search to find interesting documents. Lingo3G offers an online demo under `http://search.carrotsearch.com/carrot2-webapp/`.

The underlying technology of Lingo3D is published as open source *Carrot*^{2} on `http://project.carrot2.org`. A similar front end to Lingo3G's webapp can be found on `http://search.carrot2.org/stable/search`. Figure 4 shows another example of a twitter data set that was clustered and visualised using *Carrot*^{2}.

In-Spire (`http://in-spire.pnnl.gov/index.stm`) is a commercial Microsoft Windows based tool for explorative document clustering and analysis. It mainly focusses on topic modelling and information extraction. In-Spire is able to extract information such as dates, names and places from documents and offers static visualisations or visualisations of changes over time in topics (see Figure 5).

Text document clustering is an active field of academic research, with the existence of tools such as Lingo3G and In-Spire demonstrating the scope for commercial growth in this sector. Both of these tools provide document clustering and visualisation capabilities. With DocoPool, we are aiming to build upon such State of the Art in document clustering by:

- Using document data as the core input, thereby allowing for drill down to underlying documents;

[3]OpenOffice can run macros on documents form the command line in the form of `ooffice -invisible macro:///Standard.Module1.SaveAsXHTML("/tmp/somefile.rtf")`

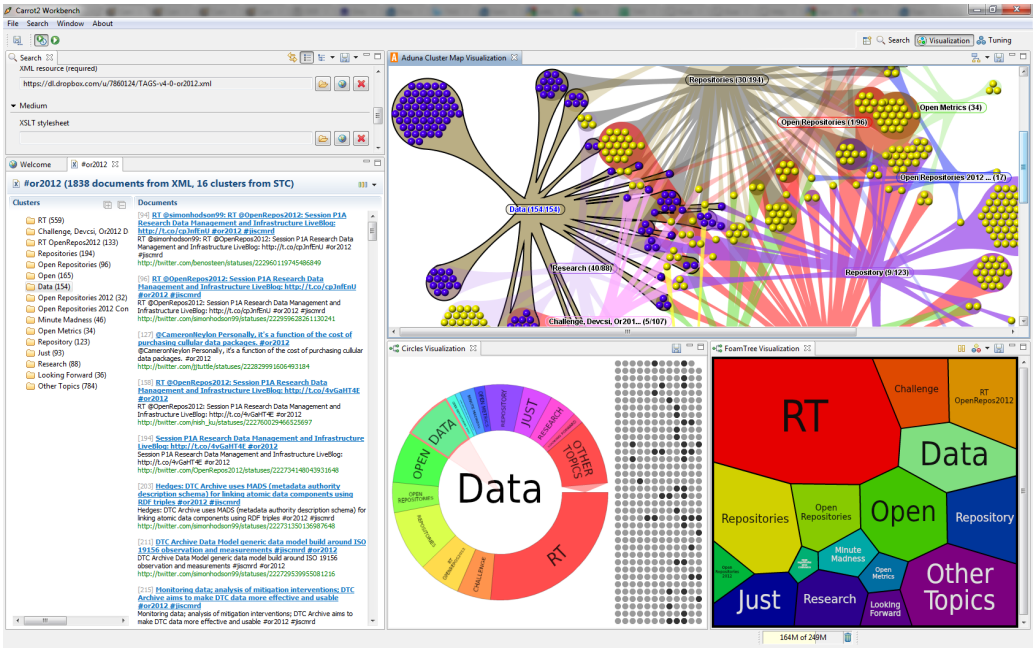

Figure 4: *Carrot*^{2} Text clustering visualisation of a twitter data set. Source: `https://mashe.hawksey.info/2012/07/quick-play-with-carrot2/`

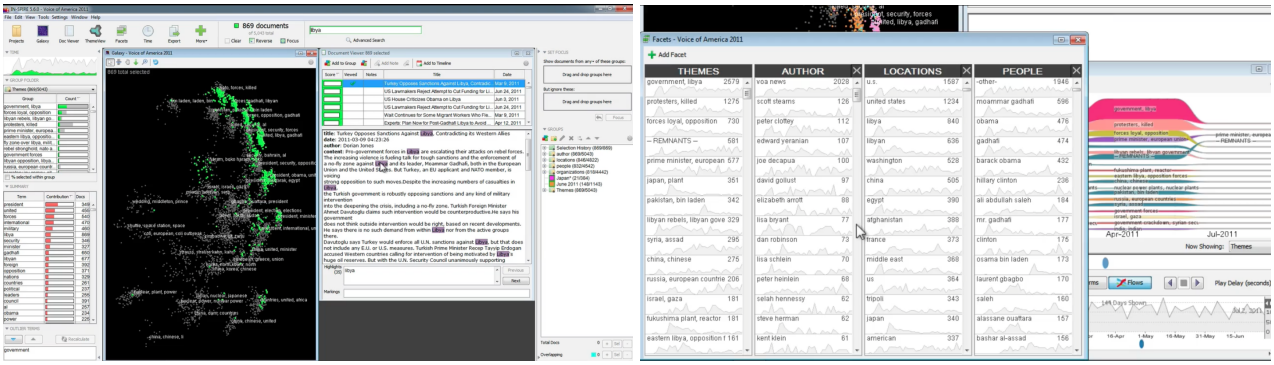

Figure 5: *In-Spire* tool for explorative document clustering. On the left is the so called "Galaxy" visualisation of describing topics in the document data set, on the right the temporal evaluation of topics in it.

- Including effective simple visualisations, aimed at business users, in an easy-to-use tool (we will avoid over-complex visualisations that can occur in existing tools);
- Aiming to advance topic modelling approaches in existing tools by using improved algorithms, such as the Non-Negative Matrix Factorization (NMF) algorithm. Existing tools typically rely on probabilistic techniques, such as the Latent Dirichlet allocation (LDA) algorithm.

See [10] for a recent overview of unsupervised clustering and [6] for a general introduction.

3.5 Public Document Data Sources

Whilst companies will be encouraged to supply sample data for DocoPool, it may be useful to also have access to publicly available text documents for the development of DocoPool.

This list contains text categorisation benchmark datasets:

- The Layout Analysis Dataset `http://dataset.primaresearch.org` contains scans of documents with a wide variety of layouts.[8]
- Based on Reuters Corpus Volume I (RCV1), this is an archive of over 800,000 manually categorized newswire stories recently made available by Reuters, Ltd. for research purposes: `http://www.jmlr.org/papers/volume5/lewis04a/lewis04a.pdf`
- The Ling-Spam data set, email data containing spam messages, based on [7]: `http://csmining.org/index.php/ling-spam-datasets.html`
- Web document clustering data set: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.11.6691&rep=rep1&type=pdf`
- classic3 and classic4 data set from Cornell University: `http://www.dataminingresearch.com/index.php/2010/09/classic3-classic4-datasets/`
- The Large Scale Hierarchical Text Classification challenge, a wikipedia page dataset: `https://www.kaggle.com/c/lshtc`
- The 20 newsgroups dataset: `http://mlcomp.org/datasets/379`

The following is a list of directories of further text data sets for machine learning:

- University of Salford: `http://www.primaresearch.org/datasets`
- IAPRTC11 Data set list: `http://www.iapr-tc11.org/mediawiki/index.php/Datasets_List`
- Corpora for Text Categorization: `http://disi.unitn.it/moschitti/corpora.htm`
- Technion Repository of Text Categorization Datasets: `http://techtc.cs.technion.ac.il`
- The UCI Machine Learning repository `http://archive.ics.uci.edu/ml/datasets.html`

4 Conclusions

Our conclusions are separated into a summary of the limitations on the existing State of the Art technologies, and our overall general conclusion for the DocoPool project.

4.1 Limitations on the State of the Art

Document Management Systems serve as a repository and capture point for electronic documents, enabling automated workflows, paperless environments and the creation of document repositories. The DocoPool project is based upon *exploiting* such document repositories for knowledge, but will not focus on the storage or repository aspects.

The ability to load documents, using document import and OCR technologies is another core feature of document management systems. These are relevant to DocoPool in that DocoPool needs access to document data, and will therefore need to provide some level of document upload. Whilst DocoPool will aim to make the upload/import functions simple, clean and easy to use, DocoPool is not aiming to advance the technologies used for this, and will where appropriate use existing open-source libraries for import. OCR adds less value for DocoPool as it is applicable for scanned documents only - and is already available via open-source tools, so will only be included if time allows.

Many document management systems provide document search and extraction functionality to support drilling documents for information. With the advancement of text mining, some systems now provide more advanced text classification and clustering functions. Of these, DocoPool will aim to improve upon existing state of the art technologies by:

- Providing visualisations that are easy to interpret by business users;
- Providing targeted use cases for CeADAR companies that can demonstrate the benefit of text-mining, such as Named Entity Recognition, n-gram viewing and easy-to-use topic modelling;
- Using a more advanced topic modelling algorithm to achieve improved results over existing tools;
- Bringing text-mining research capabilities "out of the lab" in an easy to use tool for CeADAR companies.

4.2 Conclusion

Many solutions exist for the storage and management of documents by businesses. More recently, the ability to mine such documents for knowledge is being enabled via advances in text mining. DocoPool will focus on demonstrating several useful text-mining use cases. The demonstrator will use advanced document clustering, the visualisation of these clusters and an exploration of similarities or differences in extracted keywords/metadata. More basic functionality of DMS such as document import, while necessary for DocoPool, will be included where possible, without re-inventing the wheel. OCR is less value-added and should only be included if time allows.

References

[1] List of libre office input filters. URL: `http://cgit.freedesktop.org/libreoffice/core/tree/filter/source/config/fragments/filters`.

[2] ABBYY. Classification differences in finereader engine & flexicapture engine. URL: `http://www.abbyy-developers.eu/en:tech:comparisons:classification_fre_fce` [cited 2015-08-14].

[3] ABBYY. Fulltext ocr sdk vs. data capture sdk. URL: http://www.abbyy-developers.eu/en:tech:comparisons:fre_fce [cited 2015-08-14].

[4] ABBYY. Sample auto document classification engine 11. URL: http://www.abbyy-developers.eu/en:tech:samples:classification [cited 2015-08-14].

[5] ABBYY. Whitepaper: Insights into abbyy® flexicapture. URL: http://www.abbyy.com/media/4260/abbyy-flexicapture-technology-brochure.pdf.

[6] CharuC. Aggarwal and ChengXiang Zhai. A survey of text clustering algorithms. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 77–128. Springer US, 2012. URL: http://dx.doi.org/10.1007/978-1-4614-3223-4_4, doi:10.1007/978-1-4614-3223-4_4.

[7] Androutsopoulos, J. Koutsias, K.V. Chandrinos, George Paliouras, and C.D. Spyropoulos. An evaluation of naive bayesian anti-spam filtering. In G. Potamias, V. Moustakis, and M. van Someren, editors, *Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000), Barcelona, Spain*, pages 9–17, 2000.

[8] A. Antonacopoulos, D. Bridson, C. Papadopoulos, and S. Pletschacher. A realistic dataset for performance evaluation of document layout analysis. In *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on*, pages 296–300, July 2009. doi:10.1109/ICDAR.2009.271.

[9] MS Dhiman and P Dr AJ Singh. Tesseract vs gocr a comparative study. *International Journal of Recent Technology and Engineering*, 2(4):80, 2013.

[10] Justin Grimmer and Gary King. General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, 108(7):2643–2650, 2011. URL: http://www.pnas.org/content/108/7/2643.abstract, arXiv:http://www.pnas.org/content/108/7/2643.full.pdf, doi:10.1073/pnas.1018067108.

[11] Marcin Heliński ski, Miłosz Kmieciak, and Tomasz Parkoła. Report on the comparison of tesseract and abbyy finereader ocr engines. Technical report, Poznań Supercomputing and Networking Center, Poland, 2012. URL: http://www.digitisation.eu/download/IMPACT_D-EXT2_Pilot_report_PSNC.pdf.

[12] technavio. Global document management systems (dms) market 2014-2018. URL: http://www.technavio.com/report/global-document-management-systems-dms-market-2014-2018 [cited 2015-08-15].