

Text Mining: beyond the CAQDAS?

Davy Weissenbacher, Brian Rea,

Sophia Ananiadou

National Centre of Text Mining

{firstname.surname}@manchester.ac.uk

What a CAQDAS software do?

Source document

List of annotations

The screenshot displays the CAQDAS software interface. On the left, a list of documents is shown, including 'P 2: Revelat', 'Quotes', 'Codes', 'Horror %3', and 'Memos'. The main window shows the 'King James Version (Public Domain)' document. The text 'Revelation 9' is highlighted. Below it, the text '1. And the fifth angel and his star fall from heaven unto the earth: and his army was given the key of the bottomless pit.' is highlighted. To the right, a list of annotations is shown, including 'C: Fire (12-3) -> Q:2:1', 'Created by Admin (11/03/91)', 'Horror %4', 'Smoke', '1:10 <supports>', 'Earth', 'Fire', and 'Horror %5~'. Arrows indicate the relationship between the source document, the list of annotations, and the annotation describing the sequence.

King James Version (Public Domain)

Revelation 9

1. And the fifth angel and his star fall from heaven unto the earth: and his army was given the key of the bottomless pit.

2 And he opened the bottomless pit; and there arose a smoke out of the pit, as the smoke of a great furnace; and the sun and the air were darkened by reason of the smoke of the pit.

3 And there came out of the smoke locusts upon the earth: and unto them was given

Annotation describing the sequence

C: Fire (12-3) -> Q:2:1
Created by Admin (11/03/91)

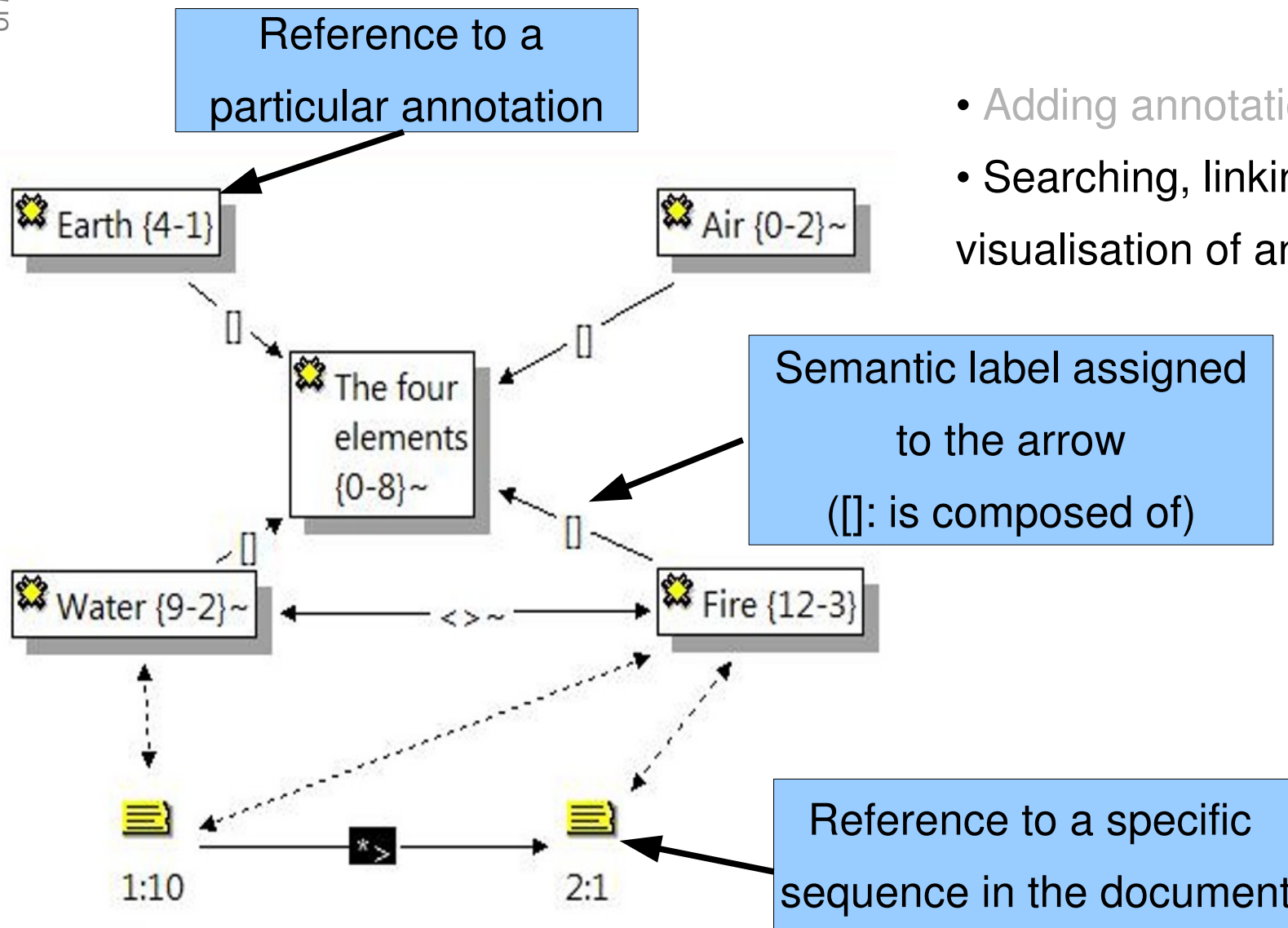
Horror %4
Smoke
1:10 <supports>

Earth
Fire
Horror %5~

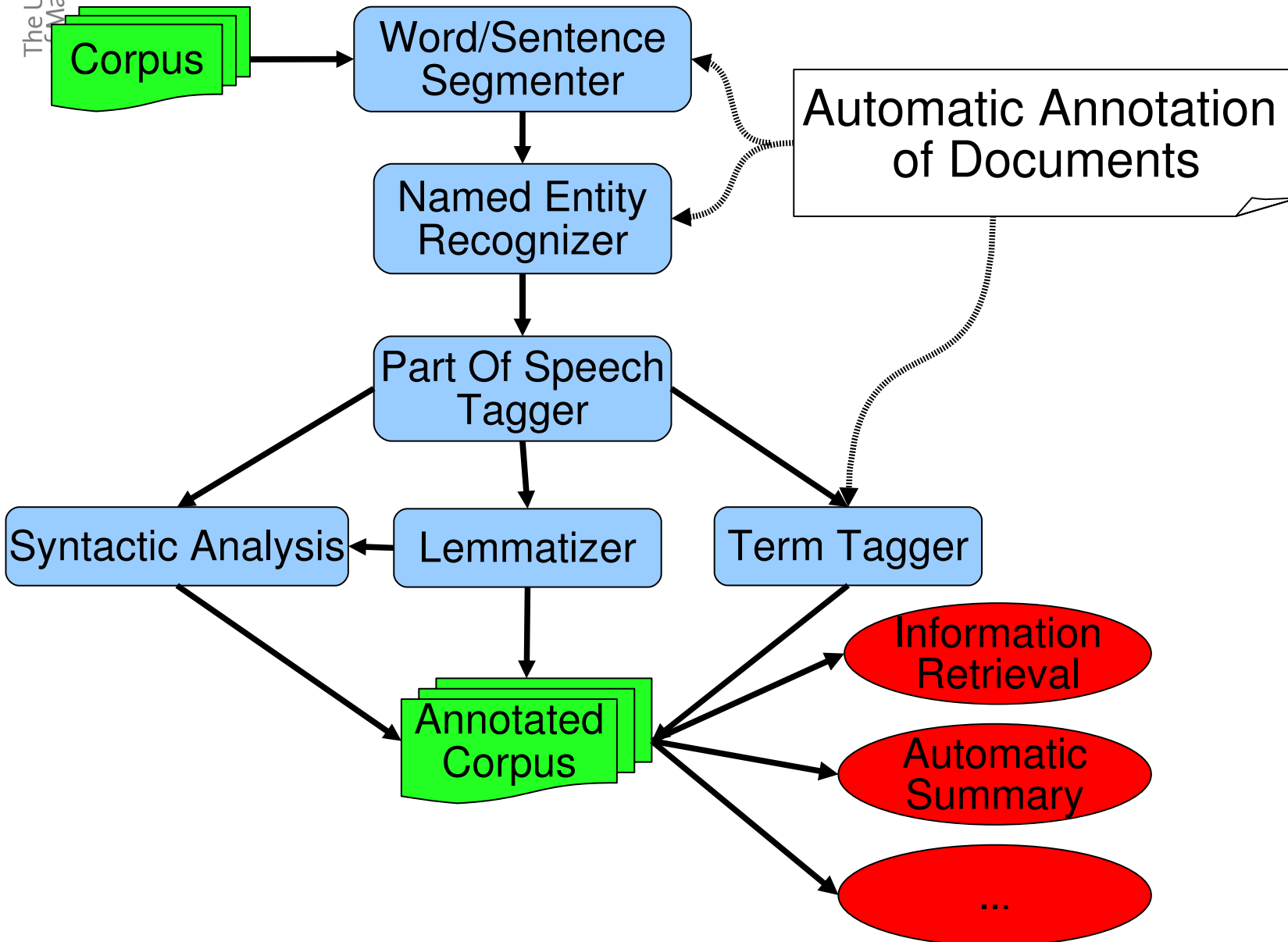
- Adding annotations
- Searching, linking and visualisation of annotations

What a CAQDAS software do?

- Adding annotation
- Searching, linking and visualisation of annotations



What is a Text Mining (TM) Software?



Are the CAQDAS and TM software competitors?

- CAQDAS and TM software are designed to add annotations but:
 - CAQDAS: human annotation (Hundreds of documents)
 - TM : automatic annotation (Millions of documents)
 - CAQDAS: Semantic and Pragmatic annotations
 - TM : Syntactic and Simple semantic annotations

How can TM techniques complement CAQDAS software?

- TM techniques enrich CAQDAS:
 - QDA Miner + Wordstat: stoplist for word frequency, lemmatizer, thesaurus for retrieving sequence to annotate, clustering of documents
 - Qualrus: machine learning techniques to propose sequences to annotate
- TM techniques are used to:
 - Extend the user queries
 - Focus the user attention on the pertinent sequences
- ➔ The ASSIST Project: evaluate the benefits of TM for frame analysis of Media

ASSIST project

- Aims to deliver a service for searching and qualitatively analysing social science documents
- NaCTeM is designing and evaluating an innovative search engine embedding text mining components
 - **Domain knowledge** facilitates expansion of user queries
 - Real Time **clustering** of search results
 - **Term extraction** for improved browsing capabilities
 - **Semantic Information enrichment** for targeting the main topics
- Final deliverable will include a web demonstrator for further integration into JISC e-Infrastructure
- NaCTeM local project website: <http://www.nactem.ac.uk/assist/>

Technical Characteristics

Multi-format documents

Conversion tools

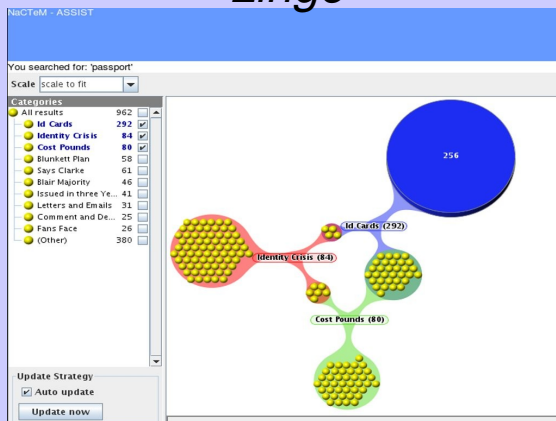
.PDF with *pdfbox*
.DOC with *POI*
.HTML with *Jtidy*
.XML

TM components

- Named Entity Recognizer
BaLIE
- Term Extractor
Termine
- Anaphora resolver
Bayaphora
- Lexical Chain extractor

Search Engine
Lucene

Search result clustering *Lingo*



Web Query Interface



NaCTeM - ASSIST Browser

ASSIST Browser

authors:Wintour AND date:2005 AND educat*

Search

Sample queries include:
10', 10" or 10 OR Scientist

Browse...

Curriculum subject and skills	Administration, governance and finance
School Sixth form Higher Adult learning...	Finance Buildings ICT Admissions Governance...
Learning/teaching needs and practice	Teachers, lecturers and support staff
Early years Primary Secondary Homework...	Teachers CPD Recruitment Teacher education...
Performance, assessment, standards	Home, community and society
Inspection Early years education Secondary...	Parents Equality and diversity Crime NEEETs...
Careers and work experience	Care, welfare and psychology
Career guidance Employment Apprenticeships...	Adoption Childrens Tendency...

The NaCTeM ASSIST Browser is an beta version demonstrator for the ASSIST project funded by [JISC](#) to investigate the benefits of text mining in 2 case studies within the social science disciplines. This includes a review of the requirements gathering stage in order to advise future projects in this area and the development of high profile exemplars demonstrating how text mining solutions can solve, in part at least, major challenges facing e-Researchers across all domains.

- Automatic grouping of documents into useful categories
- Automatic creation of descriptive labels for categories based upon content
- Automatic identification of similar or related documents

User Query

Indexed Documents

Query interface

NaCTeM - ASSIST Browser

ASSIST Browser

Sample queries include:
'ID', 'ID*' or 'ID OR Scientist'

Browse...

- | | |
|--|---|
| Curriculum subject and skills | Administration, governance and finance |
| School Sixth form Higher Adult learning... | Finance Buildings ICT Admissions Governors... |
| Learning/teaching needs and practice | Teachers, lecturers and support staff |
| Early years Primary Secondary Homework... | Teachers CPD Recruitment Teacher education... |
| Performance, assessment, standards | Home, community and society |
| Inspection Early years education Secondary... | Parents Equality and diversity Crime NEETs... |
| Careers and work experience | Care, welfare and psychology |
| Career guidance Employment Apprenticeships... | Adoption Childcare Truancy... |

The NaCTeM ASSIST Browser is an beta version demonstrator for the [ASSIST](#) project funded by [JISC](#) to investigate the use of text mining in 2 case studies within the social science disciplines. This includes a review of the requirements gathering stage in order to advise future projects in this area and the development of high profile exemplars demonstrating how text mining solutions can solve, in part at least, major challenges facing e-Researchers across all domains.

- ♦ Automatic grouping of documents into *useful categories*
- ♦ Automatic creation of *descriptive labels* for categories based upon content
- ♦ Automatic identification of similar or *related documents*

Expanding the standard query interface

- ✓ Semantic operators to build complex queries
- ✓ Browsing documents through a domain taxonomy

Improving the rank of query results

- Resolution of Pronominal Anaphora relations to compute the real frequency of search words
(e.g. **The dog** eats **the cat**. **It** sleeps now)

Cluster results for: 'education' [Visualize](#)

	Query Results	Semantic Content
All Documents (684)		
Blair Seeks (101)		
ID Cards (70)		
Labour at MANCHESTER (64)		
Brown (50)		
Willfully (42)		
Election AFTERMATH (34)		
David Cameron (42)		
Blunkett Affair (31)		
Government (33)		
Just Week (25)		
New (28)		
Us' (22)		
Delay School Reform (20)		
Religious Hatred Bill (21)		
Britain (21)		
(Other) (275)		
	<p>TITLE: The State of Intellectual Property Education Worldwide AUTHORS: Lakhan, Shaheen E, Khurana, Meenakshi K FULL TEXT: http://cogprints.org/5640/1/IP_Education.pdf ABSTRACT SNIPPETS: , education in intellectual property is required and must be advocate We must make individuals ... property rights and provides a detailed overview of the state of IP education worldwide. The discussion ...</p>	
	<p>TITLE: Embracing ignorance in Higher Education AUTHORS: Soetendorp, Ruth FULL TEXT: http://eprints.bournemouth.ac.uk/3411/1/722.pdf ABSTRACT SNIPPETS: Ignorance receives a bad press which it doesn't deserve. Negative and unwarranted associations with stupidity and foolishness can make ignorance a quality from which to shy or for which to apologise particularly when education is on the agenda ...</p>	
	<p>TITLE: Exclusive Brethren: an Educational Dilemma. AUTHORS: Bigger, Stephen FULL TEXT: http://eprints.worc.ac.uk/241/1/EXCLUSIVES.pdf ABSTRACT SNIPPETS: An article from 1990 on the Exclusive Brethren and Education, particularly focusing on the ICT National Curriculum regulations that came out at that time, since Exclusive Brethren parents wished to withdraw their children from ICT on conscientious grounds. The paper follows their arguments. An update from 2007 has been added ...</p>	
	<p>TITLE: Citizenship education in the UK: devolution, diversity and divergence AUTHORS: Andrews, Rhys, Mycock, Andrew FULL TEXT: http://eprints.hud.ac.uk/563/1/MycockCitizenship.pdf ABSTRACT SNIPPETS: of education in the UK. But to what extent does citizenship education receive equal attention within the four UK Home Nations? And, what are the implications of</p>	

✓ Clustering the query results in real time

Lingo algorithm merges instances of commonly occurring phrases, keeping the best candidate to describe each cluster

✓ A familiar presentation of query results including snippets

Search Result Interface

NaCTeM - ASSIST

Cluster results for: 'education' [Visualize](#)

All Documents (684)	Query Results	Semantic Content
Blair Seeks (101)	DEAR SUN -	
ID Cards (70)	AUTHORS: Unknown, DATE: May 27, 2005	
Labour at MANCHESTER (64)	TERMS: [MARK TAYLFORTH Kensington, ID card, tangible benefit, pension shortfall, health service, West London]	
Brown (50)	Reply: Letters and emails: Towards a surveillance society -	
Willfully (42)	AUTHORS: Robin Hull, DATE: April 26, 2006 Wednesday	
Election AFTERMATH (34)	TERMS: [legal compensation shakeup, false court action, Robin Hull London, compensation issue, dangerous issue, lingering sentence, ID card, human error, good faith, good evidence]	
David Cameron (42)	PUPILS FACE CLOCKING IN -	
Blunkett Affair (31)	AUTHORS: Unknown, DATE: April 19, 2004	
Government (33)	TERMS: [ID card scanning, human right issue, swipe card, St Thomas, High School, pilot project, Education chief, Marion Pagani, Glasgow Children]	
Just Week (25)	Learner numbers are a step towards ID cards -	
New (28)	AUTHORS: Unknown, DATE: February 15, 2008, Friday	
Us' (22)	TERMS: [national insurance number, income tax purpose, robert steel Salisbury, Government intent, unique number, birth certificate, HM Revenue, education purpose]	
Delay School Reform (20)	Learner numbers are a step towards ID cards -	
Religious Hatred Bill (21)	AUTHORS: Unknown, DATE: February 15, 2008, Friday	
Britain (21)	TERMS: [national insurance number, income tax purpose, robert steel Salisbury, Government intent, unique number, birth certificate, HM Revenue, education purpose]	
(Other) (275)	Whitehall poised for shuffle of top posts -	
	AUTHORS: David Hencke, Westminster correspondent, DATE: July 18, 2005	
	TERMS: [Sir John, permanent secretary, principal private secretary, Home Office, cabinet secretary, Gordon Brown, public service spending programme, Sir David, Mr Clarke, senior civil servant]	
	Labour reshuffle: The winners -	
	AUTHORS: Unknown, DATE: May 6, 2006 Saturday	
	TERMS: [John Reid Who Former communist bruiser, Alan Johnson Who Alan Johnson, Hazel Blears Who	

Document content is described using semantic information

✓ makes document analysis easier, faster and more efficient

NaCTeM - ASSIST

Cluster results for: 'education' [Visualize](#)

All Documents (684)	Query Results	Semantic Content
Blair Seeks (101)	DEAR SUN - AUTHORS: Unknown, DATE: May 27, 2005 TERMS: [MARK TAYLFORTH Kensington, ID card, tangible benefit, pension shortfall, health service, West London]	
ID Cards (70)		
Labour at MANCHESTER (64)		
Brown (50)	Reply: Letters and emails: Towards a surveillance society - AUTHORS: Robin Hull, DATE: April 26, 2006 Wednesday TERMS: [legal compensation shakeup, false court action, Robin Hull London, compensation issue, dangerous issue, lingering sentence, ID card, human error, good faith, good evidence]	
Willfully (42)		
Election AFTERMATH (34)		
David Cameron (42)		
Blunkett Affair (31)	PUPILS FACE CLOCKING IN - AUTHORS: Unknown, DATE: April 19, 2004 TERMS: [ID card scanning, human right issue, swipe card, St Thomas, High School, pilot project, Education chief, Marion Pagani, Glasgow Children]	
Government (33)		
Just Week (25)		
New (28)	Learner numbers are a step towards ID cards - AUTHORS: Unknown, DATE: February 15, 2008, Friday TERMS: [national insurance number, income tax purpose, robert steel Salisbury, Government intent, unique number, birth certificate, HM Revenue, education purpose]	
Us' (22)		
Delay School Reform (20)		
Religious Hatred Bill (21)		
Britain (21)		
(Other) (275)		
	Learner numbers are a step towards ID cards - AUTHORS: Unknown, DATE: February 15, 2008, Friday TERMS: [national insurance number, income tax purpose, robert steel Salisbury, Government intent, unique number, birth certificate, HM Revenue, education purpose]	
	Whitehall poised for shuffle of top posts - AUTHORS: David Hencke, Westminster correspondent, DATE: July 18, 2005 TERMS: [Sir John, permanent secretary, principal private secretary, Home Office, cabinet secretary, Gordon Brown, public service spending programme, Sir David, Mr Clarke, senior civil servant]	
	Labour reshuffle: The winners - AUTHORS: Unknown, DATE: May 6, 2006 Saturday TERMS: [John Reid Who Former communist bruiser, Alan Johnson Who Alan Johnson, Hazel Blears Who	

Document content is described using semantic information

- **Metadata**: informing the origin of documents
- **Terms**: most significant multi-words phrases in the document
- **Named Entities**: main discourse objects belonging to predefined categories
- **Lexical chains**: gathering terms to build up concept representations

- Examination of cluster memberships via a friendly visualisation interface
- Graphical representation of the intersection between the clusters provides immediate visualization of cluster relations
- ✓ Information regarding membership of particular cluster

Document: LNDocument1135.xml

Interviews to cut passport fraud

Fingerprinting and eye scans may also be brought in to tighten security.

Cath Urquhart reports No, it's not Dubai, it's Portsmouth.

This is the Spinnaker Tower, due to open this month, which at 170m (547ft) is higher than the London Eye or Blackpool Tower.

The tower, which has three viewing decks looking towards the Isle of Wight, is part of the Harbour Renaissance of Portsmouth Project. The **new passport system**, it is claimed, "will be simple for people who really are who they claim to be". THE INTRODUCTION of "e-passports", starting **early next year**, will see passport prices rise steeply, and new applicants being called in for one-to-one interviews to obtain the travel document.

E-passports, or **biometric passports**, are to be phased in from February.

They will bear a chip containing biometric data -initially, a facial scan taken from a photograph, although a fingerprint scan is likely to be included from 2008.

Britons who need to apply for their **first adult passport**, or whose passport is lost or stolen, will have to attend a face-to-face interview before they will be granted a **biometric passport**.

A network of 70 **new passport offices** will be created across the country, to supplement the existing seven offices, where the interviews, likely to start from October 2006, will take place.

This autumn the **UK Passport Service** (UKPS) is likely to announce a **huge price rise** to cover the cost of **biometric passports**.

Figures are not yet available, but the projected **unit cost** of the passport in 2006-07, according

Related Documents

A safer, more convenient passport. Now would you like chips with that?

A face-to-face interview to get your first passport

Face-to-face grilling for 600,000 first-time passport applicants

National: Passport price to rise for third time in less than two years: Increase to fund consular service, says Foreign Office

Bill is underwriting cost of ID cards, say opponents

Fingerprints plan for new passports

Related Topics

LOW MED HIGH

[biometric passport](#)
[ID card](#)

[human right group](#)
[adult passport](#)
[unit cost](#)

[early next year](#)
[first adult passport](#)
[full adult passport](#)
[huge price rise](#)

[ID card centre](#)
[ID card legislation](#)
[interesting moral dilemma](#)
[new passport office](#)
[new passport system](#)
[passport price increase](#)
[UK Passport Service](#)

- ✓ Identification of conceptually similar documents using the most commonly occurring terms and words in the source document
- ✓ Highlighting selected semantic information within the document
- ✓ Selecting terms according to their importance and using them to browse documents

Conclusion

- Both applications designed for annotating documents but TM software complements the CAQDAS software
- TM techniques help the fastidious annotation stage of the qualitative analysis
- Presentation of the ASSIST project for evaluating the benefits of a tool based on TM for frame analysis of Media