UK DATA ARCHIVE KEYWORD INDEXING WITH A SKOS VERSION OF HASSET THESAURUS

SKOS-HASSET

Funded by JISC

Led by UK • DATA ARCHIVE

Mahmoud El-Haj

m.el-haj@lancaster.ac.uk

PURPOSE AND MOTIVATION

Apply automatic indexing tool, KEA, to some of the UK Data Archive's document collection using HASSET thesaurus with aims to:

- see whether KEA could potentially be used to aid metadata creation.
- develop recommendation for the future use of automatic indexing with an existing thesaurus.

DATA COLLECTION

Corpus Name	Whole Corpus		Training Corpus	
	# Files	Size MB	# Files	Size MB
Nesstar bank of variables/questions	26,634	5.70	21,307	4.56
Survey Question Bank (SQB)	1,353	**88.00**	1,082	**70.00**
ESDS partial data catalogue records	5,610	14.50	4,488	11.60
Case Studies / Support guides	243	4.10	194	3.28

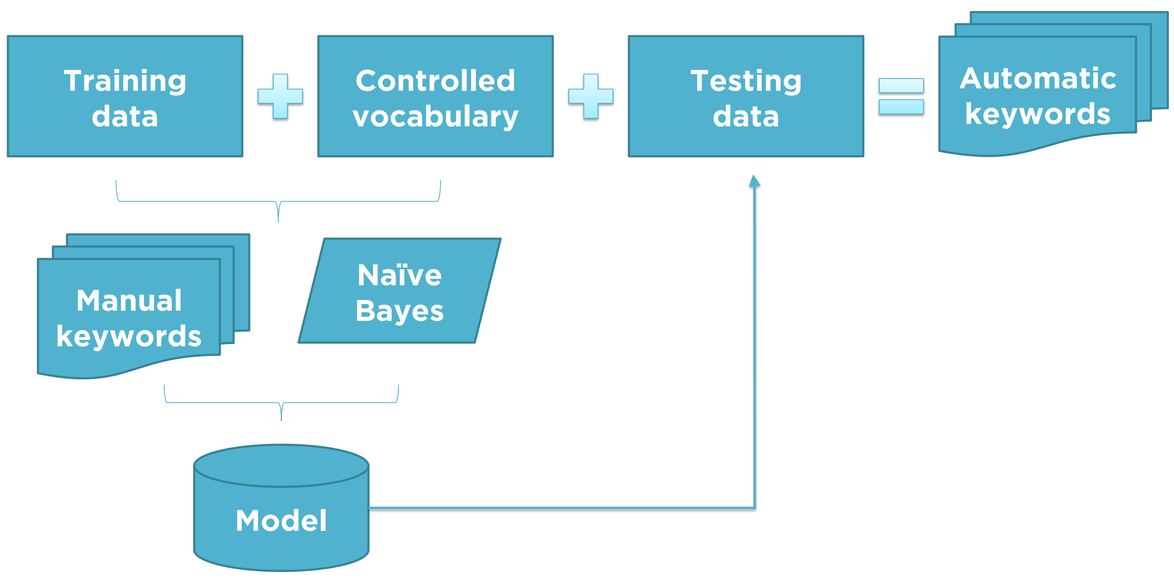
KEA (KEYWORD EXTRACTION ALGORITHM)

- an algorithm for extracting keywords from text documents
- calculates feature values for each candidate (TF.IDF, First Occurrence, Length)
- uses a machine-learning algorithm to predict which candidates are good keywords.

INDEXING PROCESS

1. Get PDFs
   - Extract Metadata (Manual-Keywords)
   - Convert PDFs to Text
2. Prepare In/Out (.txt/.key)
   - Apply KEA
   - Extract Auto Keywords
3. Automatic Evaluation
   - Manual Evaluation (Experts)

KEYWORDS EXTRACTION



EVALUATION

Human Evaluation:

Manually compare auto-keywords with manually assigned keywords.

- strictly relevant: 'exact match' of a manual keyword, or 'extremely suitable'.
- broadly relevant: 'exact match' of a manual keyword, or 'extremely suitable' or 'partially suitable' by the evaluator.

Automatic Evaluation:

$$Precision = \frac{Relevant\_Keywords\_Retrieved\_by\_Auto-indexer}{All\_Keywords\_by\_Auto-indexer}$$

$$Recall = \frac{Relevant\_Keywords\_Retrieved\_by\_Auto-indexer}{All\_Relevant\_Keywords}$$

$$F1-Score = 2*\left(\frac{Precision*Recall}{Precision+Recall}\right)$$

RESULTS

Individually, the best performance overall was seen in the SQB corpus. Catalogue records' low F1 score was to be expected, given that KEA had relatively little text to index from, compared to the manual indexers who indexed from the full catalogue records.

Corpus Name	Auto F1	Strict F1	Broad F1
Nesstar bank of variables/questions	0.12	0.14	0.34
Survey Question Bank (SQB)	0.14	0.33	**0.43**
ESDS partial data catalogue records	0.11	0.19	0.21
Case Studies / Support guides	0.06	0.27	0.36

As expected there was relatively little overlap between KEA keywords and manual keywords :

- on average KEA extracted 18.60 keywords per document across the four corpora.
- only 2.33 were exact matches with the manual keywords.
- a high percentage of KEA keywords were considered relevant/suitable even if they were not exact matches - 33 per cent for the SQB corpus, with an average of 25 per cent across all four corpora.

This suggests that KEA could be a very useful tool for indexers. The average number of manual keywords varies from 1.63 for Nesstar to 62.86 for catalogue records.

CONCLUSIONS AND RECOMMENDATIONS

- KEA is a useful tool for indexers of full text social science materials
- KEA would work best as a suggester of new terms, with moderation from a human indexer
- KEA could also be used as a quality assurance tool, to ensure that terms are not overlooked
- more work is needed to investigate KEA further and to see how it could be incorporated technically, and in terms of process, into ingest systems.