Effective Search Term Construction: Practical Advice and Strategies from a Linguist

Attorneys today use various search strategies to reduce collected document populations prior to review or production. Whether negotiating the search criteria in advance with opposing counsel or employing these strategies unilaterally, the goal is to reduce the population as much as possible, while fulfilling the obligation to produce what is required. Too often that goal is not fully met, leaving the legal team with too much to review and produce, or putting at risk the defensibility of their production.

This one-hour course, presented by Dan Brassil, Principal Consultant at H5, confronts these challenges head on. It shows how one can apply basic principles of linguistics and information retrieval science to substantially increase the accuracy of the keyword cull, while reducing risks and costs.

Topics include:

- Classifying search-based culling strategies and understanding their strengths and weaknesses
- Search syntax and common pitfalls
- Using linguistic and information retrieval principals to reduce false positives
- Using linguistic and information retrieval principals to ensure coverage of the key issues

The following is a detailed outline of the specific topics addressed during this session:

Keyword Culling Strategies

- Single word
- Phrase
- Boolean AND/OR
- Proximity
  - Bidirectional
  - Unidirectional
- Strengths and weaknesses
  - Simplicity vs complexity
  - Broad vs narrow coverage
  - False positives vs false negatives

Search Syntax

- Standard operators
  - Boolean and proximity
  - Wildcards
- Common Pitfalls

Dan Brassil
Principal Consultant, H5

Dan Brassil provides expertise in information retrieval, linguistics and solutions design. Since joining H5 in 2005, he has served in key leadership roles on H5 engagements, consulting with clients to identify strategies for capturing relevant subject matter from large document populations and overseeing the implementation of these strategies, ensuring they meet client needs and objectives.

He previously served as an associate director in H5's Professional Services group and was head of research, modelling and analysis. Prior to joining H5, Mr. Brassil taught linguistics at the University of California, San Diego and his work has been published in linguistics and information retrieval journals such as "Artificial Intelligence and the Law" and "Proceedings of the IEEE International Conference on Systems, Man and Cybernetics."

Mr. Brassil received his B.A. with honors from the University of California, Santa Cruz and his M.A. in linguistics from the University of California, San Diego.

- Stop words
- Ignore class characters
- Numbers
- "Short" searches (2-3 characters)
  - with wildcards

Addressing false positives

- Polysemy
  - Multiple meanings
  - Common names and acronyms
- Word frequency
  - Context-free vs context-dependent
- Reducing false positives via conceptual/contextual anchors
  - Selecting appropriate anchors
  - Anchors and operators
    - Co-occurrence vs topicality vs grammatical relationships

Addressing false negatives

- Words vs concepts
  - Mapping relevant concepts and searches
- Linguistic variability
  - Morphological variability
    - Inflection (e.g. number and tense)
    - Derivation (e.g. verb to noun (research→researcher), noun to verb (category→categorize), etc.)
    - Word-stems and wildcards
  - Lexical variability
    - Synonymy (similar meaning)
    - Meronymy (parts and wholes)
    - Hypernymy/hyponymy (generic descriptor of a class/specific member of a class)
    - Word-types and anchors
  - Syntactic variability
    - Subjects, verbs, objects and narrative flexibility
    - Bidirectional vs unidirectional operators