ayfie

BASIC PRINCIPLES IN AYFIE'S INFORMATION RETRIEVAL ENGINE

Johannes Stiehler

Chief Technology Officer, ayfie Inc.

Version 1

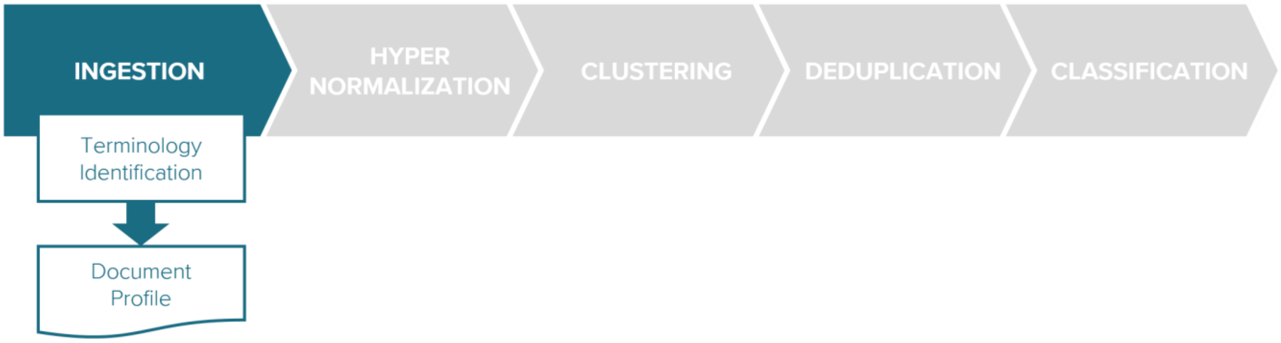
1 Document Profile

1.1 Ingestion

Almost all operations in ayfie are based on linguistically motivated document profiles.

Each document is initially ingested as a sequence of tokens (and is always searchable as such). During the **ingestion** process, all relevant terminology in that document is identified. Terminology can consist of **entity names**, **domain vocabulary** or **special expressions** such as acronyms, measurements etc. This terminology is often not represented by single terms but by multi-word expressions, e.g. “revenue sharing agreement” or “Jim McGhee”.

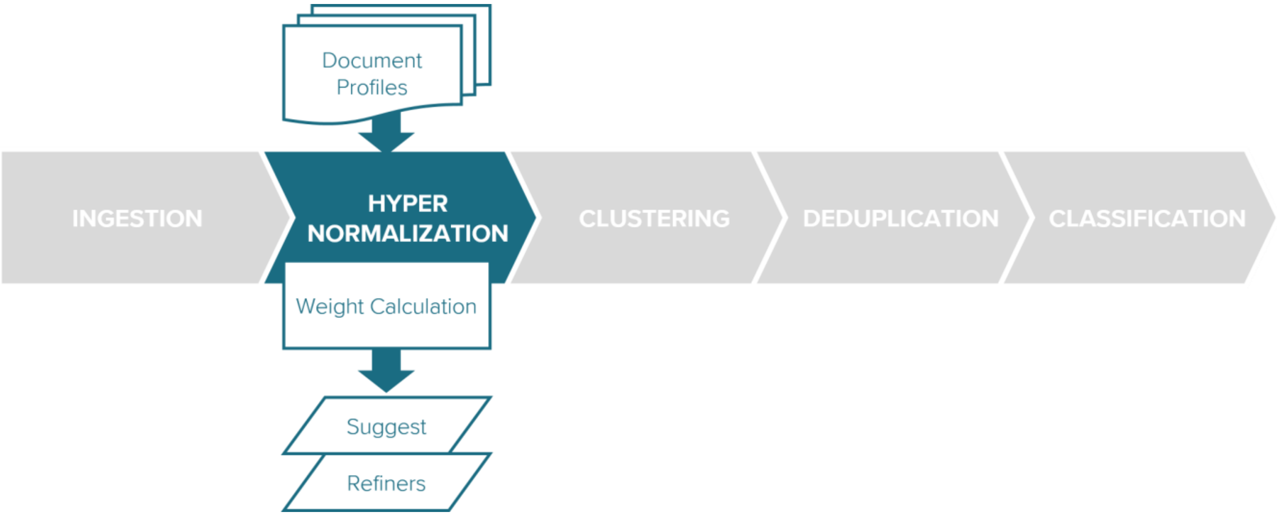
Ayfie groups these terms together based on spelling variants, inflectional forms etc. and enriches them with synonyms. The resulting term groups make up the **document profile**, i.e. a linguistic signature that describes the document content at an abstract level.



1.2 Normalization

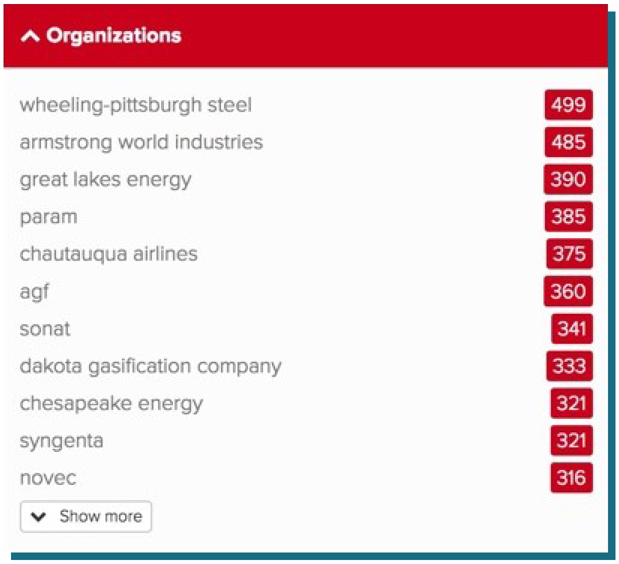
After ingestion, **hyper-normalization** is applied across all documents, i.e. all document profiles are put in statistical relation to each other. Through this process, additional normalization to the term groups is carried out and each term group is assigned a weight based on its importance for the document at hand.

Thus, each document is assigned a **weighted** document profile which is used in further computations. Additionally, the variants corresponding to the term groups also serve for improving recall during free-text searches.

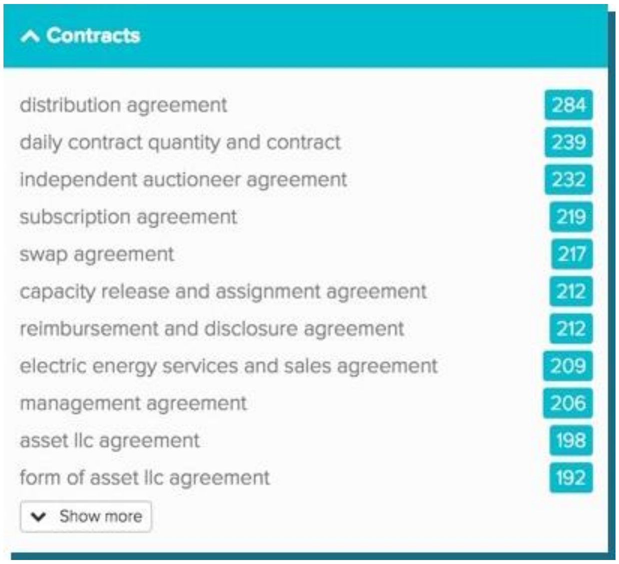


1.3 Typed Search

Terms in the document profile retain their **types** (ie. person name, organizations etc.) so they can later also be used for more precise retrieval and aggregations across the full corpus or the result set for a query:



^ Organizations	
wheeling-pittsburgh steel	499
armstrong world industries	485
great lakes energy	390
param	385
chautauqua airlines	375
agf	360
sonat	341
dakota gasification company	333
chesapeake energy	321
syngenta	321
novec	316



^ Contracts	
distribution agreement	284
daily contract quantity and contract	239
independent auctioneer agreement	232
subscription agreement	219
swap agreement	217
capacity release and assignment agreement	212
reimbursement and disclosure agreement	212
electric energy services and sales agreement	209
management agreement	206
asset llc agreement	198
form of asset llc agreement	192

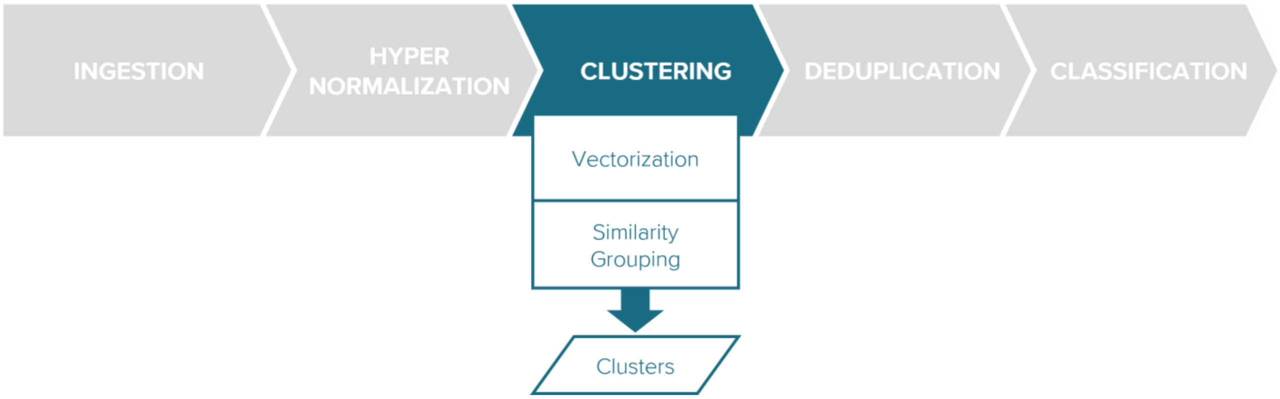
2 Clustering

Clustering essentially maps all documents into a **vector space** using the extracted concepts and their weights as vectors. Documents that are near in that vector space are grouped together into clusters.

Because k-means clustering and other common implementations exhibit several problematic properties, we apply a proprietary clustering algorithm to the document collections.

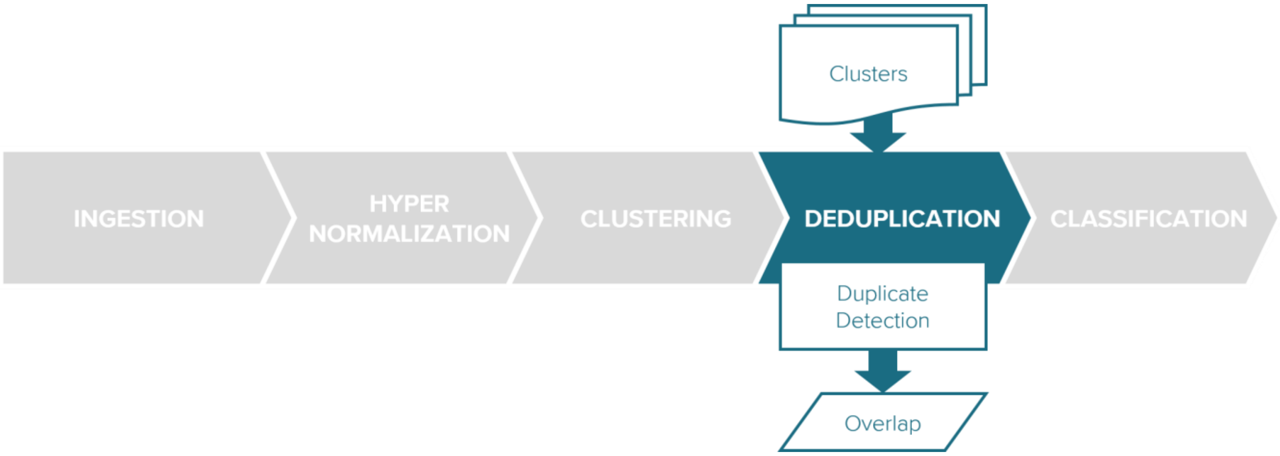
Cluster labelling is again based on the salient concepts in each cluster.

Since automated unsupervised clustering rarely matches the user's intuition about how the documents should be grouped together, other mechanisms for partitioning the document set such as dynamic drill-down and supervised classification should usually be preferred.



3 Deduplication and Near-Duplicates

Deduplication and near-duplicate detection is based on the outcome of the automated clustering process. Documents that are very close to each other in the concept space will be considered candidates for more specific duplicate analysis. Thus, only a fraction of the documents has to be analyzed.



Because of the duplicate detection process, overlap scores are computed. Ayfie can be configured to consider everything above a certain overlap score as a duplicate. Formatting and minor syntactic variance do not influence the duplicate detection process.

4 Email-Threading

ayfie analyzes the communication topology across the full document set. It considers normalized subject lines, people and timestamps and finds threads of people who communicate over time about a certain topic. This creates the initial candidate space of threads.

ayfie then analyzes all the emails in each such thread and splits them into their “contained messages”, i.e. all messages that have been quoted in all the emails in one thread. For each such message, it creates a signature that ignores differences in formatting etc.

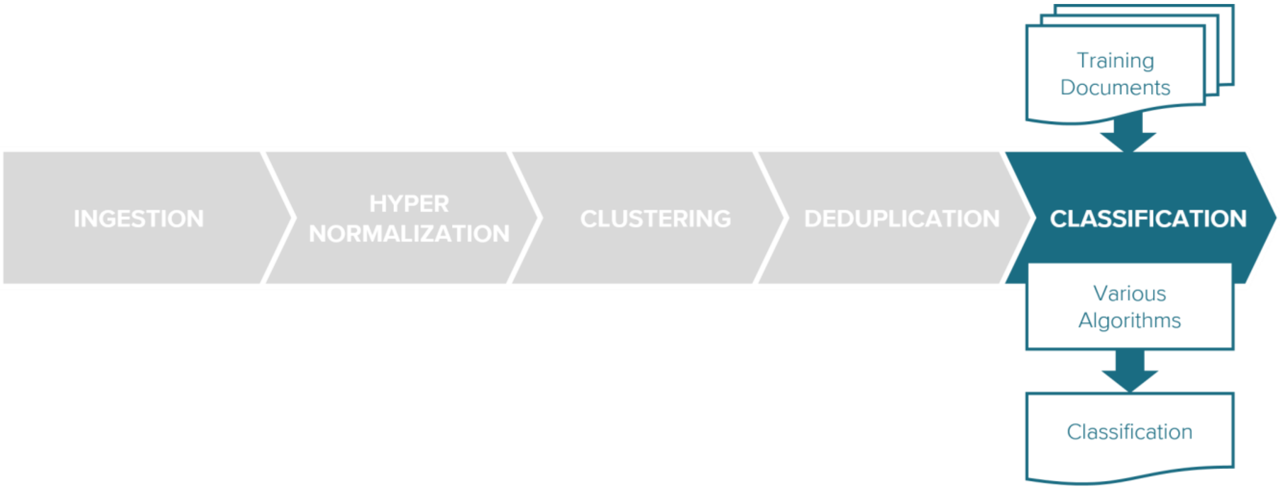
It then looks at the cross-product of all threads and their signatures and merges each two threads that share a certain amount of content signatures and recipients. The thresholds for this process are configurable.

After determining the threaded conversations, ayfie marks all emails in a thread that are "inclusive", i.e. that contain the maximum amount of the conversation in the thread as quoted messages. By only reading "inclusive" emails, a reviewer can be sure that he has seen all parts of the conversation.

This process can be imagined as ayfie looking at each full thread and “striking out” every mail that is not fully contained in another mail. What is left, are the inclusive mail(s).

5 Classification

Provided there is one or more pre-built taxonomies for the given domain with sufficient training examples for each category, ayfie can also apply high-precision and performant automated classification to the full result set.



Out of the box, two approaches are supported: Support Vector Machines with slightly higher precision, but only binary classification relevance (a document is either in a category or it is not) and Logistic Regression which assigns a probability score to each classification result. The latter allows for category-specific post-filtering on the classification results.

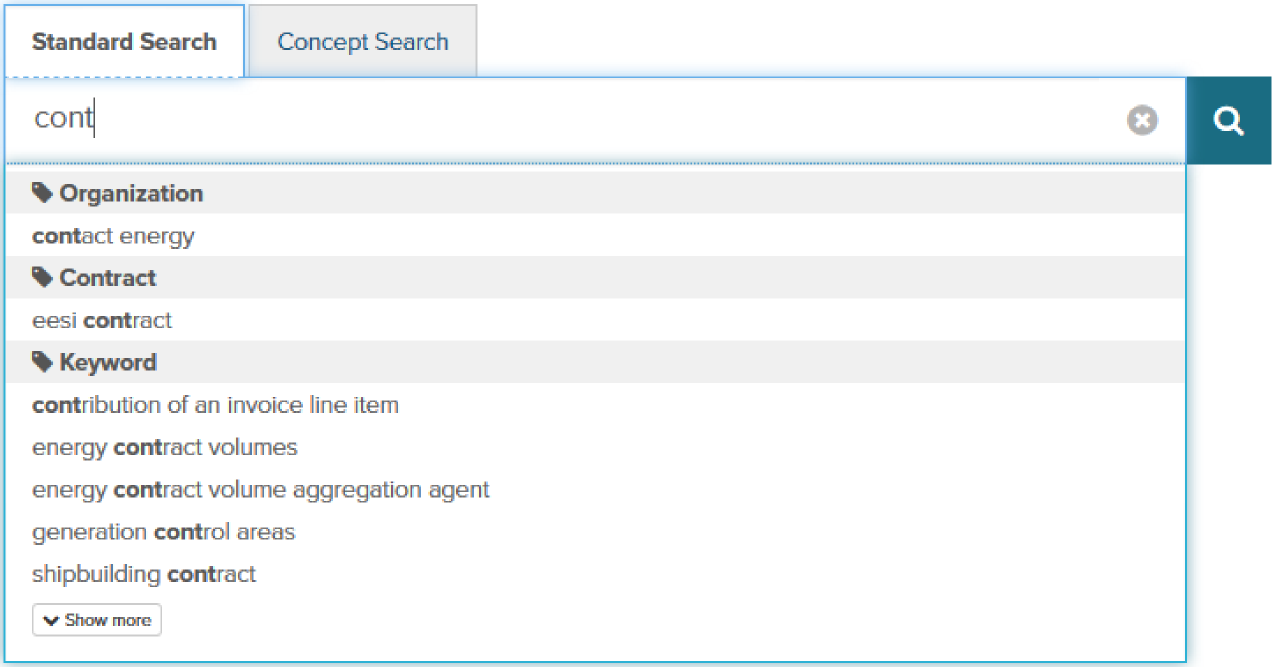
6 Suggest

Extracted entities, such as persons, organizations and locations and extracted terminology is automatically compiled into a suggest index.

Whenever a user begins typing into the free-text search field, those expressions are presented in a dropdown list if they match the user's input approximatively.

Thus, the user gets a quick overview over what entities are actually present in the content and which specificity he can expect.

For instance, if a user types "cont", the different types of entities and keywords that appear anywhere in the document collection are displayed in the order of their conceptual weight.



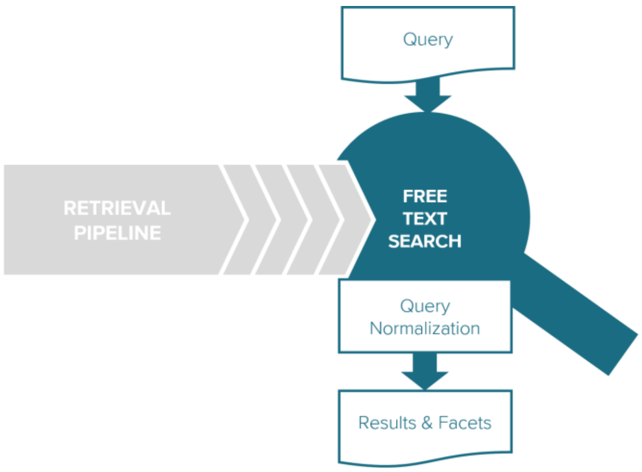
cont

7 Free-text Search

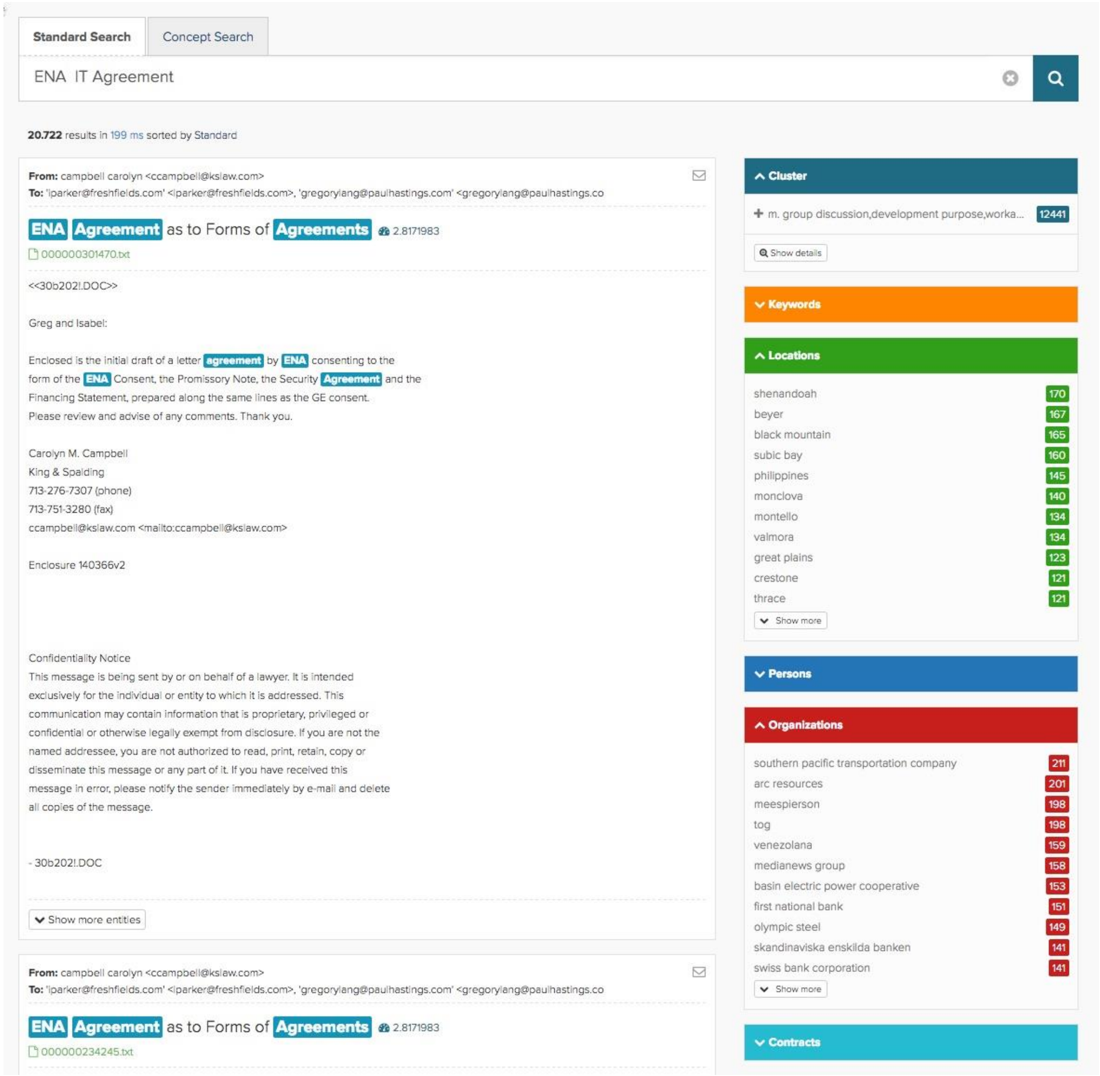
When using the free-text search form, the user can either select suggestions for the suggest dropdown or enter his own terms into the search field (which should rarely be necessary).

All terms provided by the user are then searched in the document collection, considering their variants such as inflectional forms and synonyms to achieve the best possible recall and ranked per their salience, thus improving the relevance for the user.

Alongside the actual results for the search, several navigational elements (facets) are returned that help the user drill down into the result set and thereby zooming in on the documents that best match his actual search intent.



This makes the process of finding very relevant and representative seed documents much easier and quicker and at the same time gives the user a perfect overview over the salient concepts present in the content.



ENA IT Agreement

20.722 results in 199 ms sorted by Standard

From: campbell carolyn <ccampbell@kslaw.com>
To: 'iparker@freshfields.com' <iparker@freshfields.com>, 'gregorylang@paulhastings.com' <gregorylang@paulhastings.co

ENA Agreement as to Forms of Agreements 2.8171983
000000301470.txt

<<30b202!.DOC>>

Greg and Isabel:

Enclosed is the initial draft of a letter agreement by ENA consenting to the
form of the ENA Consent, the Promissory Note, the Security Agreement and the
Financing Statement, prepared along the same lines as the GE consent.
Please review and advise of any comments. Thank you.

Carolyn M. Campbell
King & Spalding
713-276-7307 (phone)
713-751-3280 (fax)
ccampbell@kslaw.com <mailto:ccampbell@kslaw.com>

Enclosure 140366v2

Confidentiality Notice
This message is being sent by or on behalf of a lawyer. It is intended exclusively for the individual or entity to which it is addressed. This communication may contain information that is proprietary, privileged or confidential or otherwise legally exempt from disclosure. If you are not the named addressee, you are not authorized to read, print, retain, copy or disseminate this message or any part of it. If you have received this message in error, please notify the sender immediately by e-mail and delete all copies of the message.

- 30b202!.DOC

From: campbell carolyn <ccampbell@kslaw.com>
To: 'iparker@freshfields.com' <iparker@freshfields.com>, 'gregorylang@paulhastings.com' <gregorylang@paulhastings.co

ENA Agreement as to Forms of Agreements 2.8171983
000000234245.txt

Cluster

+ m. group discussion,development purpose,worka...	12441

Keywords

Locations

shenandoah	170
beyer	167
black mountain	165
subic bay	160
philippines	145
monclova	140
montello	134
valmora	134
great plains	123
crestone	121
thrace	121

Persons

Organizations

southern pacific transportation company	211
arc resources	201
meespierson	198
tog	198
venezolana	159
medianews group	158
basin electric power cooperative	153
first national bank	151
olympic steel	149
skandinaviska enskilda banken	141
swiss bank corporation	141

Contracts

8 Concept Search

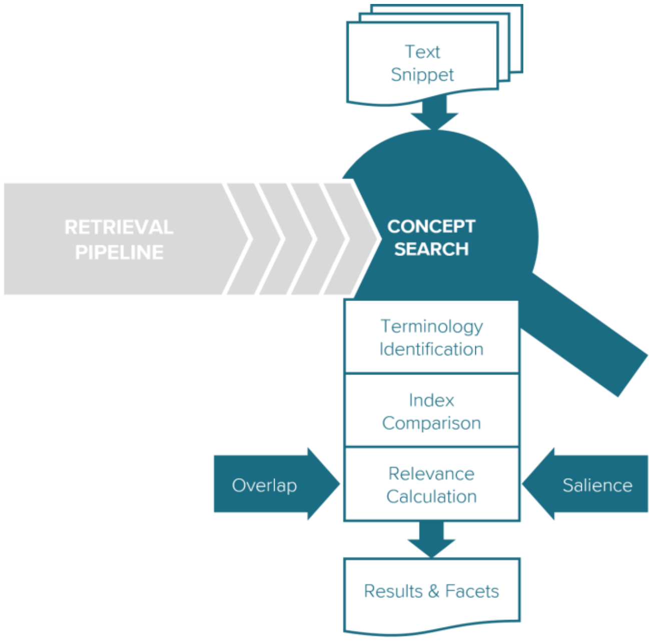
During a concept search, a relatively small text snippet is sent to the engine to retrieve all documents that have conceptual overlap with the given sample.

Ayfie executes this search by applying the same document profile building process described in Section 1 to the sample document and doing a weighted disjunction search across all indexed document profiles.

Thus, the same entities and vocabulary are identified in the sample and they are weighted based on their importance in the full document collection. The results are sorted by relevancy and cut off at the level of “minimum overlap”.

The relevance order is computed based on the location where the term group overlap occurs – for instance document title or subject is better than document body or email text – and the importance of the individual term group. Thus, documents that contain few very salient term group (e.g. “revenue-share agreement”) will be ranked before documents containing many very general term groups (e.g. “will” and “Enron”). Please keep in mind that the examples only list one term from the term group. In reality, these are made up of multiple terms.

As opposed to Euclidean distance computations inside a term space, this allows for a very tune-able relevancy experience. If something more similar to Euclidean distance is required, most of the advanced ranking features can be switched off.



When doing concept searches, two different relevance measurements are combined to compute the ranking and the cut-off:

- Concept Salience: Concepts that have a higher average importance in the document collection, are preferred to concepts that are very general.
- Concept Overlap: The more of the concepts in the sample match a document, the better a match it is considered.

These two measurements can be combined to arrive at the desired ranking and cut-off.

An example of such a ranking specification could be:

Match at least 60% of the top 30% concepts in the sample (based on salience) OR match at least 40% of the top 60% concepts in the sample OR match at least 50% of all concepts in the sample.

Additionally, it can be specified how broadly to group the terms into term groups. For instance: group only syntactical variants, group strict synonyms or group hypernyms and hyponyms together.

9 Frequently Asked Questions

9.1 What kind of hardware requirement will be necessary to store a list of synonyms?

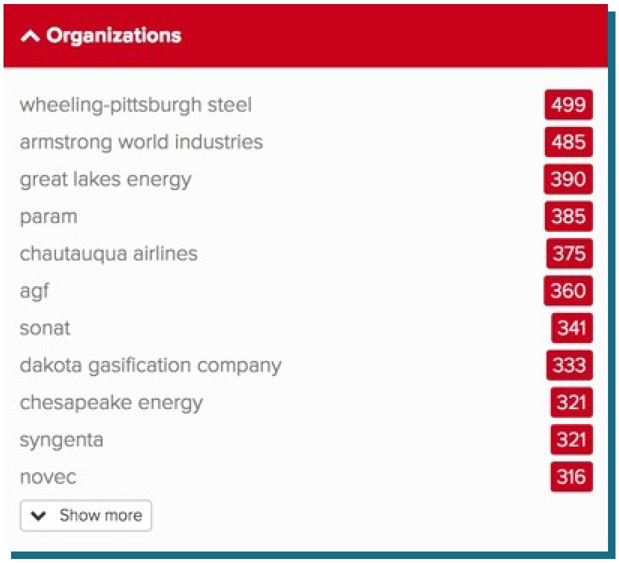
If large amounts of synonyms were to be used, the hardware requirement would still be negligible. However, the processing time (the time for initial ingestion) would increase by 10-20%. Query performance is not affected.

9.2 Does the tool require an active internet connection to access the library of synonyms and if not how often are updates required to keep up to date with changing acronyms, abbreviations and language in general?

No, it does not require an active internet connection. Linguistic resources are updated during the normal software maintenance cycle. We are currently working with quarterly feature updates and intermediate patch releases for fixing important bugs.

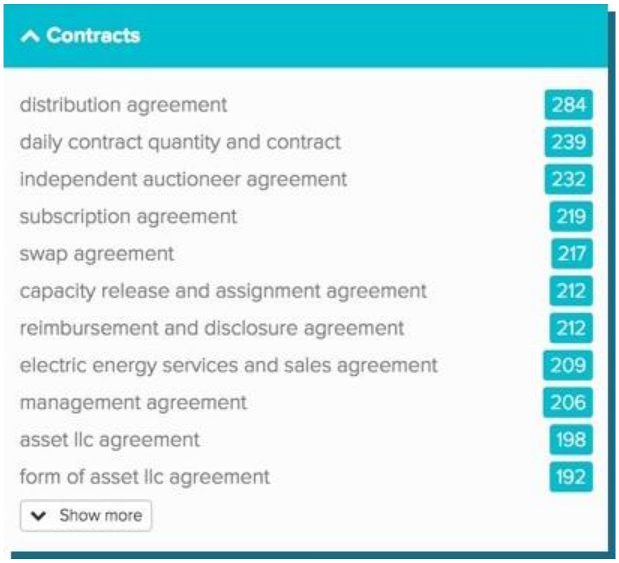
9.3 Does Ayfie allow a comparison of topics not based on the sample but compared to ingested data?

Ayfie can use samples to find related documents, but it can also uncover topics from the ingested data. This can be exposed in the form of a table of contents into the document space on the right-hand side. There are many ways this could be used with or without aligning with a pre-built taxonomy.



^ Organizations

wheeling-pittsburgh steel	499
armstrong world industries	485
great lakes energy	390
param	385
chautauqua airlines	375
agf	360
sonat	341
dakota gasification company	333
chesapeake energy	321
syngenta	321
novec	316



^ Contracts

distribution agreement	284
daily contract quantity and contract	239
independent auctioneer agreement	232
subscription agreement	219
swap agreement	217
capacity release and assignment agreement	212
reimbursement and disclosure agreement	212
electric energy services and sales agreement	209
management agreement	206
asset llc agreement	198
form of asset llc agreement	192

ayfie also comes with a built-in logistic regression classifier that can ingest a taxonomy and samples and classify the ingested corpus into the same taxonomy. This will yield more precise results than using the samples to do concept searches.

9.4 What are the languages ayfie works in and what are the options for those languages that are not included?

Ayfie works in all major European languages, among them English, German, Spanish and French. We are continuously working on expanding language support even further. In case of an unsupported language, ayfie falls back to a more basic matching algorithm that recognizes less variants.

9.5 How is this process different from LSI?

In theory, LSI applies a singular value decomposition to the matrix of documents and the terms therein. However, due to the high dimensionality of the initial matrix, a lot of prefiltering and reduction of the term space is usually applied to make it computationally feasible for larger document sets.

Thus, it usually misses a lot of obvious variants of words (often not even collapsing inflectional forms into one concept) and collapses a lot of terms that are not related. Additionally, it normally works on the space made up of single terms of all documents while a lot of the salient vocabulary comes in the form of multi-term expressions. LSI will therefore normally not consider "reimbursement and disclosure agreement" a dimension, but look at the individual words instead.

ayfie's term space is made up of the (multi-term) expressions that were found in certain linguistic contexts in the documents. Variant reduction and synonym expansion is applied to these entities rather than to individual terms.

The effect of LSI is largely dependent on term distribution in the document space. LSI might therefore work very well for one specific use case and fail opaquely in another setting. ayfie does not depend on statistical properties of the documents at all.

9.6 How can it complete the same functions as an LSI engine?

Where an LSI engine deals with the "latent semantics" of documents, ayfie is based on 30 years of research into the actual semantics of documents. By applying codified linguistic knowledge to the document set, it can detect term variants without even computing a document, term-matrix.

An additional advantage of this approach is that it is transparent to the user because it is possible to document why a certain document matched, reporting all the different processing stages that led to the match.

Additionally, the linguistic knowledge can continuously be extended to yield higher precision or recall depending on the requirements of the specific use case while LSI is basically untuneable.

9.7 What is the fuzzy logic or the logic that looks for misspellings based off of?

After extracting the salient vocabulary from the individual documents, a similarity matrix is computed across the full document set, considering misspellings based on Levensthein distance, known inflectional forms, stopwords and synonyms (optionally). Based on that matrix, all variants are folded into their most dominant representatives which are then used for all further computations.

9.8 How are inflection forms identified and calculated?

We have compiled large dictionaries, formalizing inflectional classes and the vocabulary for each language we support. Each of those dictionaries has millions of entries but only takes milliseconds to apply to a document thanks to a proprietary finite state representation of those dictionaries. Additionally, we have implemented heuristics for dealing with special phenomena in certain languages such as decomposition in German and Norwegian.

These resources also include synonym lists and taxonomies that can be used if broader matching is required.

9.9 How is the “weight” that is assigned to each term group calculated?

The weight corresponds to the average salience of that group in the content. For every group and document, a salience factor is calculated that describes the importance of that group for the current document. When doing searches, that factor average is used to weigh every term group occurring in the query.

Contact

For more information please do not hesitate to contact us at the addresses given below:

Rob Wescott · rob.wescott@ayfie.com
Chief Revenue Officer, ayfie Inc.
www.ayfie.com