KADAXIS

DATA SERVICES GUIDE

ABOUT KADAXIS.

Our mission is to help publishers, authors and book retailers improve the discoverability of their titles, by helping them to quickly and easily generate high value metadata from the text of books.

Kadaxis has also developed Author Checkpoint, an online tool that lets authors upload their books, extract metadata, then discover and analyze relevant keywords to help sell more books online.

Our team has worked with big five publishers, built a book recommendation engine, a book search engine, and processed tens of millions of book records from data feeds from the top publishers, Bowker, Baker and Taylor, and LibraryThing.

Kadaxis is based in New York City, the heart of the book publishing industry in the United States, and participates in Catalyst, the SoftLayer startup incubator.

This brochure provides an overview of Kadaxis Data Services, which help publishers and retailers fully leverage their digital assets, by maximizing discoverability opportunities through effective application of metadata.

KADAXIS DATA SERVICES.

Kadaxis data services can be integrated into your metadata workflow, in real-time through JSON APIs. A batch export facility is also available for popular file formats, such as CSV, XML or JSON. Our technology is built to scale, and can handle a single title to hundreds of thousands of titles.

Metadata Extraction and Augmentation

Kadaxis extracts metadata from the full text of a book, using the latest in data science technology. Our metadata extracts can be applied in many use cases to support search and discovery of your titles. Please refer to the last section of this brochure for a case study in extracting metadata from a popular terrorism thriller novel.

Topics

Topic data is extracted by analyzing clusters of words that occur frequently together in a book. Topics provide a more granular representation of a book's content than BISAC categories, and more meaningful data than keywords. Our system identifies the most relevant topics for a book, and scores each based on their prominence. Topics are useful for book product pages on a retail web site, as input for book-to-book comparisons or to power a book search engine.

BISAC Categories (Fiction only)

We classify books into BISAC categories, providing a standardized and consistent classification of categories across a catalog of books. Only fiction categories are analyzed at this stage.

Entities (Locations, People, Organizations)

Entities, such as locations, people (and characters), and organizations are identified within a book, along with a count of how frequently they occur. This metadata is useful as keywords, or as attributes for use in search and recommendation engines.

Significant Terms

Significant terms are often used by search engines when indexing content. Our system identifies terms that occur more frequently in a book, than would typically be observed. Significant terms are particularly useful as keywords for non-fiction titles, or as attributes for use in a book search engine.

Editorial Quality Assessment

This service analyzes a book, and compares word statistics (overused words, adverbs, initial pronouns, etc.) against professionally edited books. While it makes no assertion of the actual quality of a work, it does provide an indication of how a book compares.

Readability Assessment

Similar to the Editorial Quality Assessment, this service measures data about a book's readability, then compares it to James V. Smith Jr.'s ‘Ideal Writing Standard’. Scores that fall outside Smith's standard are highlighted.

Point of View

This simple service analyzes frequently occurring sentences to identify whether a book was written in the first or third person.

Author Gender Prediction

This service provides a score denoting how likely a book is to have been written by a male or female.

Language Detection

French, English, German, Italian and Spanish languages can be detected with very high speed and accuracy, by reading only a book's description.

Book Keyword Discovery and Analysis

Kadaxis helps automate the challenging, technical work required to implement an effective keyword strategy. Our keyword database has been vetted to include only keywords used by people searching for books. The following services expose our keyword data:

- Search for keywords by BISAC code (e.g. find all keywords that relate to books with a FIC043000 Romance code)
- Search for keywords by Amazon Browse Node
- Search for keywords by book Topic (see: Metadata Extraction and Augmentation)
- Show similar keywords to a keyword. This function shows keywords that are similar by search results returned, not simply the similarity of two terms.
- Keyword volume analysis

Comparative Title Generation

The best comparative titles for every book in your catalog are identified, by comparing the text of each book with every other book, using our unique system that decomposes a book's text to a statistically unique but much smaller size. This allows for high speed comparisons over a large catalog. (This capability is accessible via batch only).

Metadata Extraction Case Study

Each of the metadata extraction functions are illustrated below using real metadata derived from a popular terrorism thriller novel.

Topics

The top ten most prominent topics identified in the novel were:

Terrorism 5.55%
Spies 4.31%
The CIA 3.84%
Diseases 2.23%
Criminal Investigation 1.59%
Police 1.43%
City Life 1.42%
Afghanistan 1.34%
Boats 1.19%
Violence 1.19%

BISAC Categories

FIC031000 : FICTION / Thrillers
FIC006000 : FICTION / Espionage
FIC030000 : FICTION / Suspense

Entities

The top six most frequently occurring entities for each entity type, along with their counts, are listed:

Locations
Bodrum (106)
Turkey (69)
New York (61)

US (57)

Afghanistan (54)

Paris (35)

Organizations

FBI (70)

CIA (52)

Harvard (13)

State Department (8)

NSA (6)

Taliban (6)

People

Lucy (209)

Keith (183)

Fletcher (144)

Ben (99)

Tlass (82)

Cameron (80)

Significant Terms

The top fifteen significant terms were:
smallpox, vaccine, mosque, zoologist, virus, passport, muscleman, greenway, arabia, nanny, uffizi, imam, albanians, soviets, beretta

Editorial Quality Assessment

Typical Range / Outside Range

Word Count: 105969

Page Count: 423

Sentences Count: 10265

Vocab (# different words): 15149

Overused Word Count: 5.35% (5671)

Generic Word Count: 2.39% (2531)

Adverb Word Count: 6.8% (7201)

Adjective Word Count: 6.12% (6481)

Initial Pronoun Word Count: 1.98% (2093)

Readability Assessment

Typical Range / Outside Range

Flesch Kincaid Reading Ease: 78.11%

Flesch Kincaid Grade: 4.93

Passive Sentences: 4.99%

Characters Per Word: 4.57

Average Sentence Length: 10.32 words

Long Sentences: 22.98% (2359)

Syllable Count: 148118 (1.4 per word)

Point of View

This book was written in the first person.

Author Gender Prediction

Books similar to this are typically written by Male authors (99.99%)

Language Detection

This book was written in English (97.9%).

CONTACT US.

Email

info@kadaxis.com

Websites

kadaxis.com

authorcheckpoint.com

Social Media

Twitter

@kadaxis

Facebook

facebook.com/kadaxis

Location

139 Fulton St, Suite 703
New York, New York, 10038