**Introduction to NLPIR**

       A set of specialized software for the processing and processing of original text sets provides a visual display of the middleware processing effects as well as a tool for the processing and processing of small-scale data. You can use the software to process your own data.

        NLPIR participle system was originally published in 2000 as ICTCLAS lexical analysis system. Since 2009, in order to make a great separation from previous work and promote NLPIR natural language processing and information retrieval sharing platform, the NLPIR word segmentation system has been adjusted to add Eleven functions.

        The NLPIR system supports multiple encoding (GBK encoding, UTF8 encoding, BIG5 encoding), various operating systems (all major operating systems such as Windows, Linux and FreeBSD), various development languages ​​and platforms including C / C ++ / C #, Java , Python, Hadoop, etc.).

new features

       Full text accurate search -JZSearch: support for multiple data types, multi-field, multi-language;

        New word discovery: tap new word list

        Segmentation tagging: Segmentation of the original corpus, automatic identification of names and other names of names and other unregistered words, new words and taglines tagging. And in the analysis process, import user-defined dictionary.

        Statistical Analysis and Terminology Translation: One-dimensional word frequency statistics, binary terms transfer probability statistics, and can be used for common terminology, will be automatically given the corresponding English explanation.

       Big data clustering and hot spot analysis -Cluster: Automatically analyze hot events and provide key feature descriptions of event topics.

       Big data classification filtering: screening out from the mass of documents to meet the needs of the sample.

       Automatically summarize -Summary: the ability to single out articles or multiple articles, automatically extract the essence of the content to facilitate users to quickly browse the text content.

       Keyword Extraction-KeyExtract: Can extract a number of articles or articles collection, representing the core ideas of the article vocabulary or phrases, can be used to refine reading, semantic query and quick match, etc.

       Document to weight -RedupRemover: can quickly and accurately determine whether the file collection or database records the same or similar content, and find out all the duplicate records.

       HTML Body Extraction - HTMLPaser: Automatically remove the navigation of the web page, excluding HTML tags on the page and navigation, advertising and other disruptive text, return valuable body content. For large-scale Internet information pretreatment and analysis.

      Coding automatic identification and conversion: automatically identify the encoding of the document content, and automatic conversion, the current support for Unicode / BIG5 / UTF-8 encoding automatically converted to Simplified GBK, Traditional Chinese BIG5 and traditional GBK complex conversion.

**Related technologies**

1. Real-time network information collection and text extraction

NLPIR big data search and mining demonstration platform According to Sina rss summary, the use of NLPIR accurate network capture system to capture Sina Sina latest news (each refresh will be re-crawled), NLPIR text extraction system will webpage navigation, advertising, etc. Content removal, the use of network text link density as the main parameter, the use of deep neural network model to achieve the text of the text of the automatic extraction. Here, you can also enter arbitrary articles manually by users.

2. Segmentation markup based on cascading hidden-horse model

NLPIR / ICTCLAS participle system, the use of cascading hidden horse model (algorithm details, please refer to: Zhang Huaping, Gao Kai, Huang Yan, Zhao Yanping, "Big Data Search and Excavation" Science Press .2014.5 ISBN: 978-7-03-040318-6 ), Segmentation accuracy close to 98.23%, with high accuracy, speed, adaptability and other advantages. It can really understand Chinese and use machine learning to solve ambiguity segmentation and POS tag ambiguity. Dr. Zhang has worked hard to create more than ten years, the kernel upgrade 10 times, the global users exceeded 300,000.

3. Based on the role of the label entity extraction

The NLPIR entity extraction system can intelligently identify thematic keywords of names, names, agencies, media, authors, and articles appearing in the text. The extracted words do not need to be pre-existing in the dictionary database and are a deep understanding of the laws of language And forecast. NLPIR entity extraction system based on character annotation algorithm automatically identifies the named entity (for details of the algorithm, please refer to: Zhang Huaping, Gao Kai, Huang Yan, Zhao Yanping "Big Data Search and Excavation" Science Press 2014.5ISBN: 978-7-03-040318-6 ), Based on which to build a variety of large data mining applications.

4. Based on the perfect double array TRIE tree word frequency statistics

The NLPIR's word frequency statistics algorithm is more efficient and uses our patented TRIE-tree algorithm (which has recently been further optimized), which is more than ten times faster than conventional algorithms, and the efficiency of the algorithm does not vary with the statistical results to be counted The dramatic increase in the number of exponential growth, generally sub-linear growth. It is recommended that you call NLPIR / ICTCLAS open word frequency statistics interface.

5. Text Categorization Based on Deep Machine Learning

NLPIR uses a deep neural network to conduct a comprehensive training of the classification system, the current training category is only the manufacturer's political, economic, military and so on. Our built-in algorithm supports category-specific training, which has a high accuracy of classification of regular texts and an F-score of nearly 86% for comprehensive open tests. NLPIR depth text classification, can be used for news classification, resume classification, mail classification, office document classification, regional classification and many other applications. In addition, you can implement text filtering, quickly identify and filter out information that meets special requirements from a large amount of text, and can be applied to areas such as brand report monitoring, spam blocking, and sensitive information review.

6. Text Sentiment Analysis Based on Depth Neural Network

NLPIR sentiment analysis offers two modes: full-text sentiment (left) and affective sentiment (right) for a given object. Emotional analysis mainly adopts two techniques: 1.The automatic recognition of emotion words and the automatic calculation of the weight make use of the co-occurrence relationship and adopt Bootstrapping strategy to generate new emotion words and weights iteratively and iteratively; 2. The deep neural network of emotion discrimination : Based on the depth of the neural network to expand the calculation of emotional words, integrated into the final result.

7. Key word extraction based on context conditional entropy

The NLPIR keyword extraction can extract a number of words or phrases that represent the semantic content of the article based on a comprehensive grasp of the central idea of ​​the article. Relevant results can be used for refined reading, semantic query and quick matching. NLPIR mainly uses the cross-information entropy to calculate the contextual entropy of each candidate word. The documents to be processed are not restricted by the industrial field, and the newly appeared new words can be identified. The output words can be weighted with the weights.

8. Word2vec semantic extension based on POS-CBOW

The POS-CBOW method synthesizes part-of-speech and word distribution features, uses word2vector to improve the model, trains 5GB news corpus, and automatically extracts the semantic association. If the training texts are adjusted to the corpus of professional fields, the model can also generate the ontological associations of the professional fields.

9. Transition-dependent syntactic analysis based on global structure prediction model

NLPIR puts forward the structured transfer-dependent syntactic analysis model using Yamada algorithm. Based on Yamada's algorithm, it adds global training and prediction, and optimizes the feature set. The accuracy of this model (85.5%) is close to the best transfer-dependent syntax (86.0%) and achieves the fastest analysis speed in all dependent syntactic models with over 85% accuracy.

10. Simplified and Traditional conversion

NLPIR simplified Chinese based on the thesaurus, control extraction extraction.

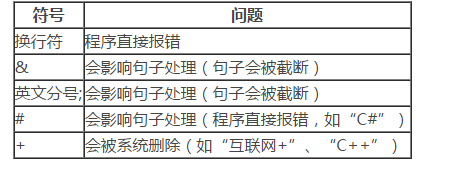
11. Auto-phonetic based on hidden horse model

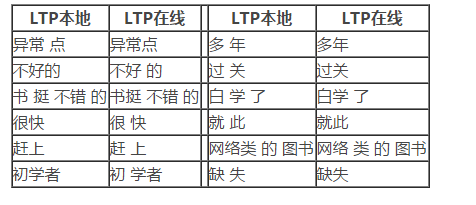
NLPIR can automatically annotate words based on thesaurus based on semantic understanding. 99% accuracy

12. Automatic summarization based on keyword extraction

Automatic text digest middleware can refine the text content, automatically extract the key sentences and key paragraphs from the long articles to form abstract content, which facilitates users to quickly browse the text content and improve work efficiency.

Automatic Abstract Middleware not only generates a summary of a coherent process for a document, but also removes multiple documents with the same subject redundancy and generates a concise summary; users are free to set the length, percentage, etc. of the summary Parameters; processing speed of 20 articles per second.





1.LTP automatic clause function, NLPIR no clause function;

LTP clause is based on the Chinese punctuation in the period, question mark, exclamation point, semicolon, ellipsis.

There are two kinds of NLPIR participle, the result of the participle is exactly the same.

(1) the use of segmentation function NLPIR\_ParagraphProcess;

(2) using NLPIR\_FileProcess word segmentation of the entire text.

The first (2) bug sometimes appears, the participle goes to the text somewhere will not go on.

3.LTP online API and local two:

(1) The online API is limited by URI construction rules, can not submit some special characters (such as #, &, +, newline, English semicolon, etc.), and there are a few unknown combination of bugs (such as and in actually not at the same time submit);

4. User Dictionary:

(1) NLPIR participant will give priority to user dictionary;

(2) LTP User Dictionary: The official added that "LTP segmentation module does not adopt the strategy of dictionary matching. External dictionaries are added into machine learning algorithms in a characteristic way, and there is no guarantee that all the words are segmented according to the way in the dictionary."

So some LTP words into the user dictionary is invalid, such as "C + +", "C #", "Internet +".