NORSI-TRANS
Closed joint-stock company

Analytics and visualization tools

Vitok-TEXT

˝Vitok-TEXT˝ is a special search system in the unstructured text data. The system is intended for accumulating, analysis and search in the unstructured text data, supporting a large number of source file formats and different means of data flow for processing. The system saves the accumulated information, including different attributes of source files, in the database optimized for quick search.

AREA OF APPLICATION

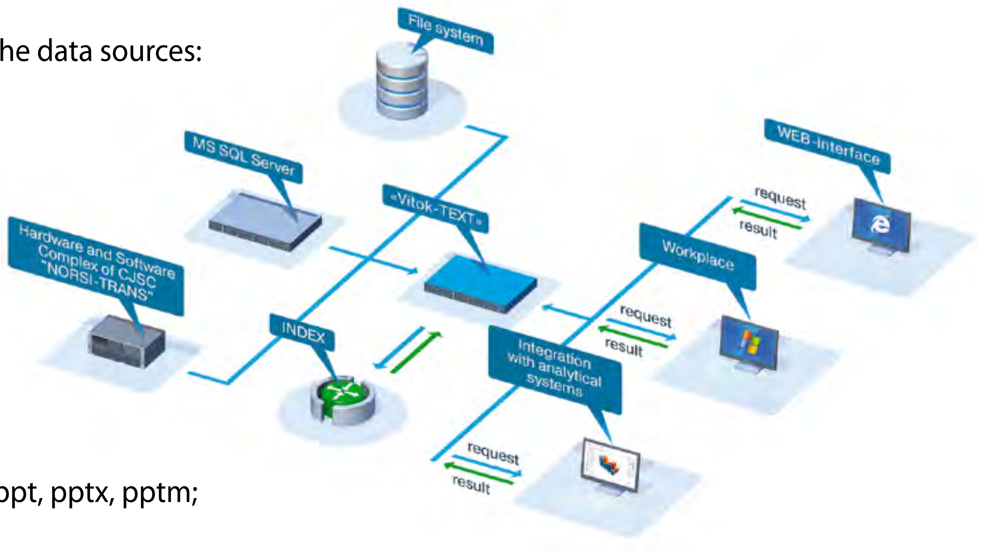
The system can be applied for the search in the unstructured text data, accumulated by different organizations.

DATA SOURCES

The following systems can be used as the data sources:

- file system (the folder in the disk);
- DBMS MS SQL Server.

Refillable sources have the function of monitoring new files, sending them for processing and keeping records of the processed ones.



FILE PROCESSING

The system extracts a text from a large amount of file formats:

- MS Office: doc, docx, xls, xlsx, xlsm, ppt, pptx, pptm;
- OpenOffice: odt, ods, sxw;
- the rest: txt, rtf, pdf, html, mht, xml, eml, wpd.

The system processes password unprotected archives (also self-extracting ones) of rar, zip, gzip, tar, tgz, bz2 formats. It is possible to develop specialized parsers of the structured files for extracting and saving the satellite information.

TEXT ENCODING

Automatic recognition of the processed text coding. A wide range of encodings is supported:

- Windows-125x family;
- ISO-8859-x family;
- UTF-x family;
- KOI-8-x family;
- IBM866.

LANGUAGE OF THE TEXT

Automatic recognition of the text language:

- CIS languages: Russian, Ukrainian, Belarusian, Kazakh and others;
- European languages: English, French, German, Italian and others;
- Arabian, Persian, Hebrew, Kurdish and others.

TEXT-PROCESSING

The texts are analyzed morphologically in order to recognize the word's initial form. The morphological support is realized for all recognizable languages. The correction of spelling errors in the source text and search query is possible for the Russian language.

OBJECTS SELECTION

The objects are selected during text processing. The types of selected objects form three basic classes:

- Template objects: telephone numbers, document numbers, vehicle identification numbers, etc. Conversion of different means of record of the same object is performed. (17.04.2014<->2014.04.17<->April 17, 2014).
- Word objects: surnames, names, patronymics, address objects, traffic centers, makes and models of cars, etc.
- Dates. Different types of methods of recording, full and not full (without year indication) dates.

Object dictionaries are refillable, including batch load from text files.

CATEGORIZATION

Text categorization is performed to define the thematic focus of the text on the basis of directory of words and word combinations. The directories are refillable with the possibility of creation of new headings.

CLASSIFICATION

Classification of texts is performed: definition of the thematic focus area of the text with the help of the classifier based on the use of training texts. Statistical analysis of the texts selected for training is performed to define the criteria significant for the specified subject. It is possible to integrate the function of training text selection to the operator's workplace.

DEFINITION OF THE TYPE OF THE DOCUMENT

The analysis of the text formatting is carried out (for file formats having this possibility) to define the type of the document in accordance with the set of templates. Document template editor allows to specify the arrangement of text blocks on the page, occurrence of certain words, some features of formatting.

SEARCH

- The query language is developed to form the search queries. It supports logical operators «and», «or», «not», the operator of the distance between words and the operator of morphology deactivation. Common words as well the objects can be used as query elements.
- The filters can be used for searching the attribute values. Examples of filters:
  - time range;
  - headings, subjects, type of the document;
  - additional attributes of the source text: telephone number.
- The query result includes the fragment of the found text including the occurrence of matching words and also saved text attributes. The view of the full text of the document with multi-colored highlight of the words and objects with the possibility of navigation between them is available.

THE USER INTERFACE

The system presents different means of realization of the user interface: web-interface, different variants of applications for installation at operators' workplaces, program interface.

POSSIBILITIES OF INTEGRATION

Possibility to integrate the systems of textual analysis with other products of the company on the base of:

- data input for processing;
- using the results of search queries to solve the analytical tasks.

NORSI-TRANS
Closed joint-stock company

CJSC «NORSI-TRANS»
12/15 Bolshaya Novodmitrovskaya, 127015, Moscow,
Russian Federation
Tel.: +7 (495) 748 74 83
info@norsi-trans.com, **www.norsi-trans.com**