Applied informatics

Information validity analysis and formation of citations.

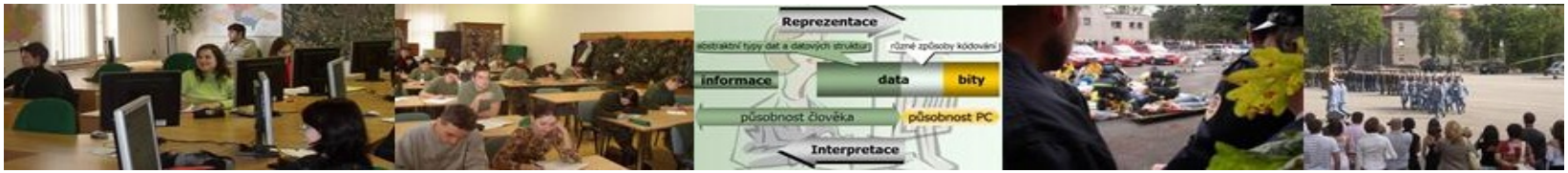
ZEMÁNEK, Z. - PLUSKAL, D. - SMETANA, B.

esf european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
UNIVERSITY OF DEFENCE
INVESTMENTS IN EDUCATION DEVELOPMENT

Information validity analysis and formation of citations.

1. **Analysis of the validity of information in selected in SW**
2. **Analysis of the information from the Data Mining**
3. **Adding citations for the assignment**
4. **Assignments**

Aims of the exercise



1. Provide students with basic information about the analysis of information about the selected software.
2. Introduce and explain the basics of creating citation with software support.
3. Introduce process of creating citation and analysis of information for the assignment.

Information validity

- When obtaining data from different sources, as well as for statistical assessment of technological processes (such as compliance with the prescribed standards), it is important to examine the validity, that is, the validity of the results obtained due to the fact.
- Qualitative or quantitative independent validation is particularly important where the phenomenon under review can not be completely separated from other influences and where the interpretation of results difficult.

Frequency analysis of words

- It is a basic method of extraction of unstructured texts.
- As a universal method it has found its application in cryptanalysis.
- Can analyze texts and can be expressed graphically.
- For text mining analysis of the importance of keywords (keywords) analysis.
- It is associated with a specific language – A. C. Doyle-Dancing Men.

Keywords density

- The importance of occurrence (keyword density) keyword analysis is given by the frequency of words.
- In all natural languages the Zipf law holds - then the product of that order of frequency, and the frequency of words remains approximately constant for all words.
- With the exception of the least and most frequent words, this rule works very well. [2]

Keywords density

- Corollary of the Zipf law - the basis of language is made up of a relatively small number of recurring words.
- When selecting insignificant words with a frequency of <10 (typos and errors ...), we reduce the number of words even at 17%.
- The list of so-called stop words = mostly clutches and prepositions - they can be omitted. [2]

Keywords density

- Feature extraction - replacing a new set of symptoms - merge into one dimension of words that have the same meaning.
- Lemmatization - convert word to its basic shape suitable for English.
- Stemming - convert word to his tribe - suitable for English.
  - For example, the basic form of the word “taught” is the infinitive „teach“

[2]

Keywords density

- A major problem stemming and lemmatization is Ambiguity.
- The problem of homonyms - words in unison (words with two meanings.
  - Accept (to receive) and Except (excluding)
  - Acts (things done) and Ax (chopping tool)
  - Ad (advertisement) and Add (short for addition)
  - Affect (to influence) and Effect (result)
  - Aid (to assist) and Aide (an assistant)
- Synonym problem – trouble x complication x obstacle

Data Mining - Text Mining in fulltext

- Acquisition (mining) knowledge from the data is called a non-trivial process of acquiring implicit, previously unknown and potentially useful and valid (valid) information from the data.
- Data-mining methods for working with numbers, or with nominal, or ordinal variables, such as category names, etc.
- Text mining is working with unstructured text. It can therefore be defined as the process of extracting valuable information from text.

Principle of Text Mining

- Text Mining is a scientific discipline at the frontier of data mining, machine learning and computational linguistics.
- It evolves with the need for automated processing of large amounts of information in the form of free text.
- Traditional methods of data mining work only with structured data (metadata important for processing), and most of their information remains inaccessible. [1]

Use of Text Mining

- Text summary
- Select the most important passages (e.g. sentences), and sort them in an appropriate way (summary extraction).
- Or it is possible to analyze the text more deeply, and based on its semantic representation paraphrase its contents (summary abstraction).
- The analysis of sentiment (sentiment analysis)
- Based on the occurrence of emotionally colored words can infer the author's positive or negative attitude towards the subject. [1]

Use of Text Mining

- Determining the type of text
- Assign a category - sport, politics, crime;
- According keyword frequency;
- Clustering texts / documents into groups based on their mutual similarities.
- Each document is assigned to exactly one group.
- The created groups may or may not correspond to the expected categories (of numbers).

Use of Text Mining

Extraction of concepts; recognition of named entities (concept extraction, named entity recognition)

It's about identifying entities that are mentioned in the text (e.g., Article on Telecommunications, therefore, the terms "mobile operator" and "Vodafone" should be assigned to the same entity).

Determining the relationship between entities

Determining entity allows for an analysis of sentences (e.g. using frames - FrameNet) to determine their relationships (e.g. Expression of "Charles is married to Eva" can lead to relationship that Eva is the wife of Charles).

TextStat – frequency analysis

- Software TextStat - from text file statistics is generated:
  - number of lines
  - the number of words
  - number of characters
  - Number of sentences
  - number of spaces
  - number of uses tabs and some other information.
- The generated report can be saved to a file. = The program allows several other settings, such as entering uncountable characters, and the like. [3]

Citations

- Bibliographic references cited in scholarly works are important information for readers, reviewers, or other assessors work, as opponents ...
- The main reasons:
  - Prove their orientation in the topic.
  - Refer the reader to further literature.
- Respect the author's ethics and the copyright law. [4]

Citations

= Summary of the information about the cited publication, or parts thereof, that can be identify the work.

It is a quote of the source from which the information was taken, and often the place where the quote is selected, that page of the referenced document.

Any information of others, which we use in our own work, must be cited. [4]

Citation phases

1. Creating a bibliography citation = citation must clearly identify the work.
2. Create a list of citations = List of references.
3. Referencing the bibliographic citation used in the list of literature = Links to relevant publications in the list of references in the text to the point where the idea of the cited author's description of his methods is used. [4]

Citation norms

- ISO 690 Bibliographic references - Content, form structure ...
- = Diagrams and examples of citations of various kinds of documents, the order provides a method of placing the individual data in the quote
- ISO 690-2 bibliographic citations. Part 2: Electronic documents or parts thereof. [4]

ŠARMANOVÁ, Jana. *METODY ANALÝZY DAT - Učební text.* [online]. [cit. 2013-10-26] © 2012, Ostrava: VŠB-TU. 170 s.
ISBN 978-80-248-2565-6
Available at: http://www.person.vsb.cz/archivcd/FEI/MAD/

References

There are several methods how to write references for citations in the text. All methods have their pros and cons. Recommended methods of referencing are based on the requirements of the faculty or in agreement with the supervisor:

Harvard system - widely used especially in the USA, also known as "the method control and data"

The method of numerical references - the method used primarily in natural and technical sciences

Method notes - often combined with a list of sources at the end of the document

References

Harvard system - widely used in particular in the U.S. Also known as the "method and the data element"

For students who are not very good in working in a text editor. Harvard system that does not require setting up cross-references or footnotes. Moreover, it is widely used especially in the humanities. [4]

Examples: Citations in the text: The world, the reality is a concept too vague and broad (Šarmanová, 2012, p 7).

If the author's name is mentioned in the text, indicate the year in parentheses: Šarmanová (2001, p.31) believes that the results are inconclusive ...

References:

1. ŠARMANOVÁ, Jana. *METODY ANALÝZY DAT - Učební text.* [online]. [cit. 2013-10-26] © 2012, Ostrava: VŠB-TU. 170 s. ISBN 978-80-248-2565-6 Available at: http://www.person.vsb.cz/archivcd/FEI/MAD/

References

The method of numerical references - the method used primarily in Natural and technical sciences

Method of numerical references has the advantage that it is possible to quickly click through to the source, and it is not necessary to scroll to the end of the text.

This method requires the use of cross-references and check them carefully. It is therefore more demanding in terms of formatting, but the most widely used. [4]

Example:

The world, the reality is a concept too vague and broad. [1]

References:

1. ŠARMANOVÁ, Jana. *METODY ANALÝZY DAT - Učební text.* [online]. [cit. 2013-10-26] © 2012, Ostrava: VŠB-TU. 170 s. ISBN 978-80-248-2565-6 Available at: http://www.person.vsb.cz/archivcd/FEI/MAD/

References

- Method of footnotes - often combined with a list of sources at the end of the document.
- Method offers the citations directly at the specific page, below the line, which is very convenient. Setting up the footnotes is also very simple. The bibliographical references refer to the serial numbers of the notes.
- The number must be distinguished from the actual text using the upper index.
- It is recommended to add the alphabetical list of bibliographic references at the end of the document, so you cite twice each document, which is of course extra work. [4]

---

1. ŠARMANOVÁ, Jana. *METODY ANALÝZY DAT - Učební text.* [online]. [cit. 2013-10-26] © 2012, Ostrava: VŠB-TU. 170 s. ISBN 978-80-248-2565-6 Available at: http://www.person.vsb.cz/archivcd/FEI/MAD/

Rules to remember

- The author name should always be in capital letters, in inverted form (= surname, first name, e.g. Hawaii, Petr)
- works without authors – take the name (alphabetical list) – Never use: Anonymous, the collective of authors, etc.
- name of the source document always in italics (the book's title, journal name, the name of Proceedings)
- each entry is separated
- for each page of the document
- Quotes must be complete, clear and uniform
- quote exclusively from primary documents [4]

The primary responsibility = author / authors

***The form:* FAMILY, First.**

- **Single author:**
  - HÁVA, Petr.
  - HÁVA, P.
- **Multiple authors:**
  - HÁVA, P., MAŠKOVÁ, P., POTŮČEK, M. et all.
  - HÁVA, P. - MAŠKOVÁ, P. - POTŮČEK, M. et all.
  - HÁVA, P. (ed.). [4]

Environment (http://www.citace.com)

- Option to generate your quotes according to ISO 690, and ISO 690-2.
- Accepting entries from other users or resources.
- Citation management (treatment, sort into folders, add your own notes, tables of contents, reviews).
- Accessing records via RSS feeds, sharing with other users.
- Export records to Word and RTF to HTML. [5]

For registered users

- **Administration**
- **My citations**
  - Generate citations
  - Import citations
- **My citations**
  - Newly generated citations
- Stored in folder
  - My citations
  - My RSS channels
- **My folders**
  - Uncategorized citations
  - User folders [5]

Inserting records

- Inserting records:
  http://www.citace.com/generator.php
  - Mandatory fields
  - Optional fields
  - Interactive fields
- Copy from other users:
  http://www.citace.com/hledat.php
- Copy from MU catalogue
  http://aleph.muni.cz/F/ [5]

Working with records

- Editing bibliographical references (edit)
  - Adding details (details)
  - notes
  - contents
  - review
- Tags
- Remove
- Move to Folder [5]

Working with folders

- Create a new folder
- Publication folder (the folder for access to other users)
- Export folder
- Creating an RSS feed
- Other Features
- Generate HTML code to insert into a web page
- Renaming
- Delete [5]

Assignment

- Select keywords to Data Mining.
- From the first three pages of search result choose the best matching document.
- Save as text.
- Analyze by TextStat.
- Document the occurrence of key words by a chart, and insert this into the assignment.
- Cite all documents including the graphics by using the Citace.com for use in the assignment.

Assignments

- Test your knowledge of text mining. Use the presentations and resources.
- Create citations according to the assignment in the presentation.
- Calculate the basic statistics of text mining. Document the results in a graph.

esf european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
UNIVERSITY OF DEFENCE
INVESTMENTS IN EDUCATION DEVELOPMENT

Resources:

1. ŠARMANOVÁ, Jana. *METODY ANALÝZY DAT - Učební text.* [online]. [cit. 2013-10-26] © 2012, Ostrava: VŠB-TU. 170 s. ISBN 978-80-248-2565-6 Dostupné z: http://www.person.vsb.cz/archivcd/FEI/MAD/
2. Semanticka-analyza-textu-3. *Http://fulltext.sblog.cz* [online]. 2008 [cit. 2013-10-27]. Dostupné z: http://fulltext.sblog.cz/2011/09/22/semanticka-analyza-textu-3/
3. TextStat. *Http://www.stahuj.centrum.cz* [online]. 2005 [cit. 2013-10-29]. Dostupné z: http://www.stahuj.centrum.cz/utility_a_ostatni/prace_se_soubory/porovnavani/textstat/?g[hledano]=textstat&g[oz]=3.0
4. Bibliografické citace. *Samba.fsv.cuni.cz* [online]. 2010 [cit. 2013-11-04]. Dostupné z: samba.fsv.cuni.cz/~tomandlo/JSM514/Bibliografické%20citace.ppt
5. Citace snadněji a rychleji. *Https://is.jabok.cz/www* [online]. 2010 [cit. 2013-11-05]. Dostupné z: https://is.jabok.cz/www/4106/495906/Citace_snadneji_a_rychleji.ppt