Features for text comparison

Marek Krótkiewicz
Krystian Wojtkiewicz

Published article info

Presentation outline

1. Basic terms
2. Parameters
3. Features
4. Application
5. Future development

Aim of project

Determining of text similarity in the aspect of wide texts comparison. The article describe features used to determine text comparison.

The method:

- works on morphological similarity of texts it doesn’t work on semantic similarity,
- is set to count exactly the same elements in two texts,
- gives number of features usable in screen-examination,
- It works on an assumption that similarity relation is not symmetrical,
- works in the field of comparison many texts in short period of time,
- is prepared to compare any two units that have specific morphology.

Basic terms

Sign – one of alphanumerical chars – {a-z, A-Z, 0-9},

Word – sequence of signs without *empty sign,*

Phrase – sequence of words occurring in uninterrupted flow,

Sentence – phrase ending with a sign of the end of sentence,

Text – finished set of sentences,

Similarity rate – degree of repeated words, phrases or sentences in two texts.

Basic terms

I level units (words)

a1 a2 a3 a4 a5. a6 a7 a8 a9 …

II level units (phrases)

{a1, a2}, {a1, a2, a3}, …

{a2, a3}, {a2, a3, a4}, …

…

III level units (sentences)

{a1, a2, a3, a4, a5}

…

Parameters

Parameters of features:

- minimal number of repeated words **(NRW)**
- minimal number of words in sentences **(NWS)**
- minimal participation of number of repeated words **(PNRW)**
- minimal length of phrase **(LPh)**
- minimal length of analyzed sentence **(MINLS)**

Features

There are three groups:

- on the level of ***words***
- on the level of ***phrases***
- on the level of ***sentences***

Each group consist of number of features build in similar way

- some of them deal with number of repeated units (words, phrases, sentences)
- some of them deal with number of repeated units **normalized** on the base of total number of appropriate units in the source text

Possible applications

Plagiarism

- Prepare the text
- Find basic units
- Features selection
- Similarity measure definition
- Comparison algorithm
- Decision making and appropriate action

Possible applications

DNA comparison

1. Prepare DNA sequence
2. Find basic units
3. Features selection
4. (Similarity measure definition)
5. Comparison algorithm
6. (Decision making)

Future development

New features development:

- Text is prepared in the most simple way, independent from:
  - Flexion
  - gender
  - multiplicity
  - tenses
  - etc.
- Grammar structures similarity (sequence of grammar units)
- At-hock ontological structure similarity (concepts connections structures)
- Semantic similarity (impossible at this time)

Thank you !!!

Krystian Wojtkiewicz - krystian.wojtkiewicz@kieg.science
Marek Krótkiewicz - marek.krotkiewicz@kieg.science