**Kraków University of Science and Technology**

# LSA

## as an associative text retrieval tool

Agnieszka Figiel

Language and Technology 2007 Poznań

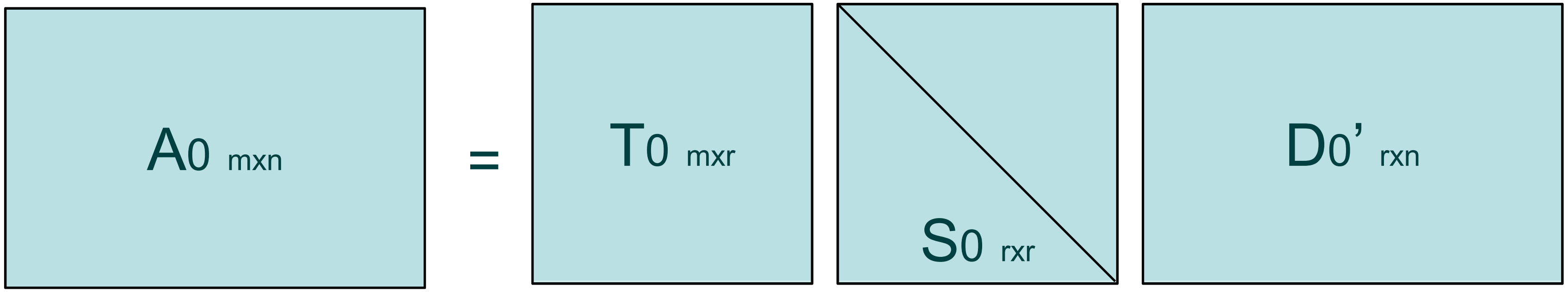
# Agenda

- ❑ LSA (LSI) fundamentals
- ❑ Some inherent problems in LSA IR
- ❑ Proposition: deviance indicator
- ❑ Experiments
- ❑ Conclusions

# LSA (LSI) fundamentals

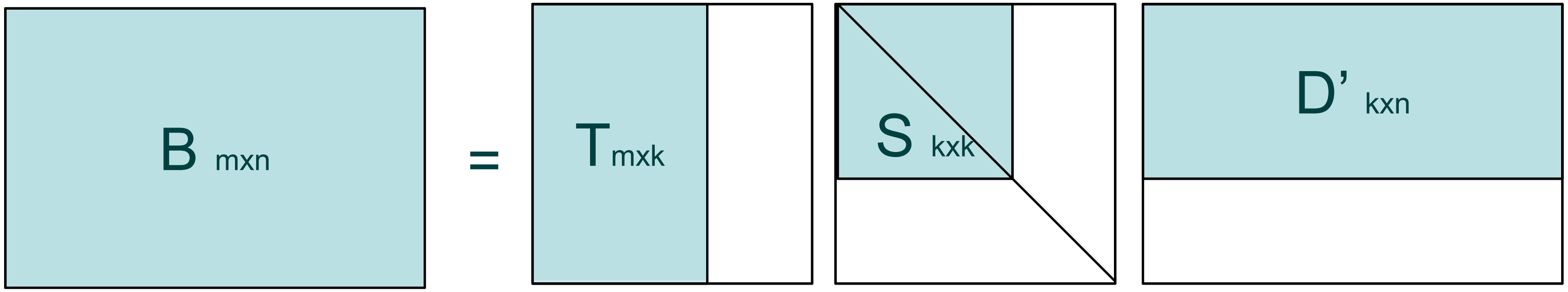
- variable term usage deteriorates retrieval
- synonymy, polysemy
- discovering the latent semantic structure in text
- „deriving a set of uncorelated indexing variables”
  *(Deervester et al. „Indexing by Latent Semantic Analysis” 1991)*
- projecting the term – document space on a low dimensional space
- dimension reduction should eliminate noise information
- comparisons in the new space should be more tolerant regarding variable term usage

# Singular Value Decomposition



- m – number of terms, n – number of documents
- r – rank <=min(m,n)
- $T_0$ and $D_0$' hold singular vectors
- $S_0$ holds the singular values on the diagonal
- singular values are ordered by size

# Dimension reduction



- singular values can be interpreted as degree of data variation
- keep only k largest singular values
- B is an approximation of the original matrix of rank k
- k should be big enough to reflect the data structure, but smaller than that to reduce the semantic noise

# Sample term associations

3000 short news items
Associations for „rynek” (market)

- NFI (national invest fund)
- MIDWIG (index name)
- kurs (rate)
- WIRR (index name)
- wartościowy (~securities)
- ustanawiać (establish)
- indeks (index)
- sesja (session)
- WIG (index name)
- NIF (index of invest funds)
- WIG20 (index name)
- notowanie (quotation)
- podstawowy (basic)
- giełda (stock market)
- punkt (point)
- warszawskiej (Warsaw adj)
- fundusz (fund)
- zwyżka (increase)
- zniżka (decrease)
- notowany (quoted)

# Some inherent problems in LSA IR

- With m terms, n documents and c non-zero elements per column, time complexity of the SVD: O(m*n*c)
- In practice: *„83,098-term by 79,316-document matrix as the input for SVD, which projects vectors into a 300-dimensional space. The SVD computation consumes 1.7GB memory and takes 57 minutes to complete on a 2GHz Pentium 4 machine.”* *(Tang et al. ‘On Scaling Latent Semantic Indexing for Large Peer-to-Peer Systems” 2004)*
- Long response time: project the query onto the new search space, compare with all vectors (dense matrix)

# Some inherent problems in LSA IR

- As with any statistical method, only works well when data is representative of a domain
- Improvement over raw term comparison not as great in heterogeneous collections
- Erroneous associations introduce unrelated documents into the ranking

⇨ try to analyse the derived term associations to see whether some of these problems can be reduced

# Deviance indicator

- try to discover why documents enter the ranking
- see whether a document that has more associations to query terms than matching terms is likely to be deviant
- for each document in the ranking define 2 sets:
  - M – terms matching the terms in query
  - A – terms associated with terms in query
- calculate the deviance indicator as:

$$di = (|A| - |M|)/(|A| + |M|)$$

# Deviance indicator

- di value can be in range [*−1, 1], where:*
  - *di = 1 – no matching terms*
  - *di = 0 – equal number of matching and associated* terms
  - *di = −1 – no associated terms*

- expect documents whose di > 0 to be deviant
- match against human intuition

# Experiment setting

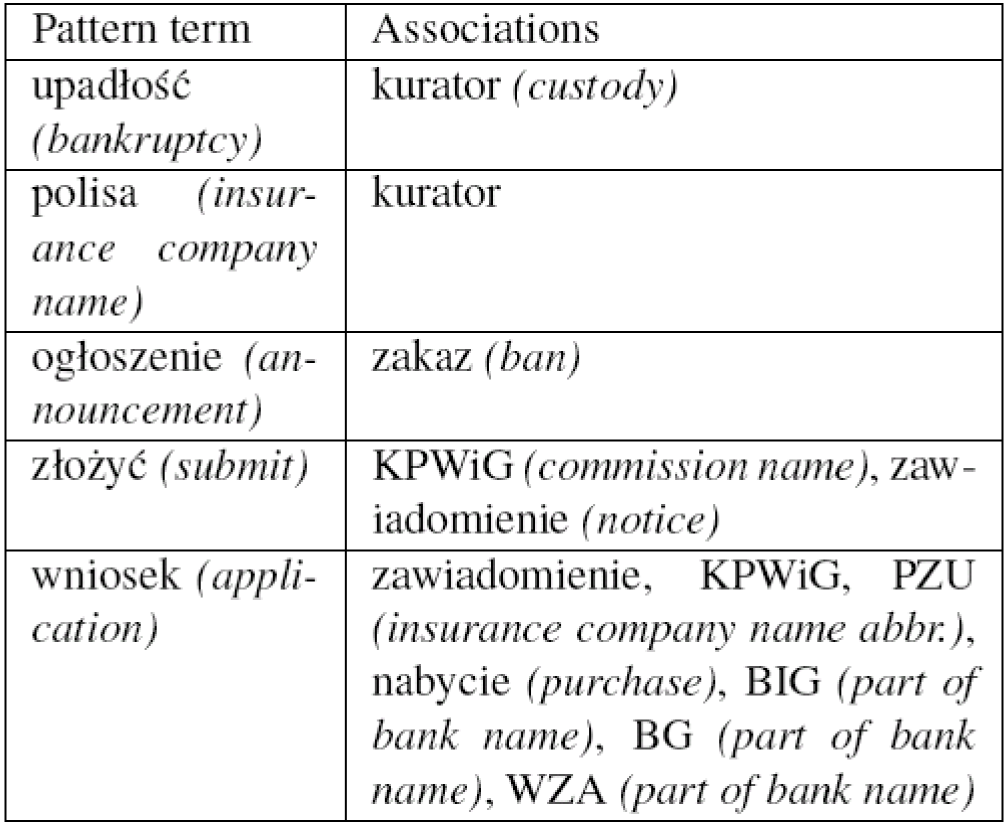
- corpus of news items in Polish, 16 thousand texts
- query is a document from the corpus
- Platypus – prototype LSA engine
- preprocessing involves:
  - word form conflation using an inflection dictionary
  - stemming of unknown words
  - abbreviation expansion
  - cutting of most and least frequent words
- lexicon: ~10 000 terms
- 300 LSA factors

# Example

*„Polisa SA will submit another bankruptcy application in court.”*

- 62% *„After the Polisa SA bankruptcy announcement, the interests of people insured by this company and injured by Polisa customers will be represented by custody.”* -0.67
- 62% *„The Warsaw Business Court, on application of the National Insurance Supervisory Commission and company board, announced bankruptcy of Polisa SA.”* -1.0
- 57% *„The District Public Prosecutor’s Office in Warsaw received notice of an offence by PZU Życie SA. The notice was submitted by the Polish Securities and Exchange Commission (PSaEC). It is concerned with exceeding the 5 percent WZA BIG BG SA vote limit by PZU Życie. PSaEC states that it posesses materials concerning the fact that PZU SA deliberately did not notify the commission about purchasing the actions of BIG BG.”* 0.5
- 55% *„The 16th Business Department of the District Court ordered a ban on selling BIG BG actions by PZU SA and PZU Życie SA and their dependent subjects.”* 0.43

# Associations of pattern terms



| Pattern term | Associations |
|---|---|
| upadłość *(bankruptcy)* | kurator *(custody)* |
| polisa *(insurance company name)* | kurator |
| ogłoszenie *(announcement)* | zakaz *(ban)* |
| złożyć *(submit)* | KPWiG *(commission name)*, zawiadomienie *(notice)* |
| wniosek *(application)* | zawiadomienie, KPWiG, PZU *(insurance company name abbr.)*, nabycie *(purchase)*, BIG *(part of bank name)*, BG *(part of bank name)*, WZA *(part of bank name)* |

# Conclusions

- A news corpus is a specific text collection: proper names, limited vocabulary, concise language
- The proposed method makes use of these features, since they allow to assume some consistency
- Requires quantitative evaluation
- Associations derived from the LSA space can be used for other purposes

# Thank you
# for your attention!!!