Doctoral School of Information and Biomedical Technologies
Polish Academy of Sciences (TIB PAN)

SUBJECT: **Towards Semantic Measures of Information Content - Integrating LinkGraph with Seminstructured Textual Information**

SUPERVISOR: M. Kłopotek, professor, Institute of Computer Science, PAS.
m.klopotek@ipipan.waw.pl

DESCRIPTION:

The access to textual information was never easy in the history of mankind. There was time when text documents were rare and now we have to handle a giant stream of various documents. However, in each case, creation of overviews of document collections is vital. And the understanding of the contents of these documents is vital.

Understanding the value of information from the human point of view is one of the most important concepts of the modern IT dealing with the processing of information on the Internet. It has not only significant from theoretical but also from an engineering point of view, since the present interpretation of information content, based on Shennon entropy, seems to be completely inadequate. In particular, the pressing issue is to provide information that corresponds to human understanding and allow for an automatic assessment of information in a manner that satisfies a human person.

The elements of such assessment include, among others, research in such areas as: document clustering, sentiment analysis, concept hierarchy extraction and spam detection.

Subject indexes and classification/categorization were used for centuries for such purposes.

Earlier, it was possible to prepare both a dictionary of categories and the assignment of documents to this dictionary manually. Now, automation is needed more and more urgently in both tasks. While there exist human made elaborated category dictionaries for general purposes, specialized collections of documents require local specialized methods of splitting the documents into reasonable groups and appropriate group labelling. The latter task seems to require engagement of automated methodologies.

Clustering methodologies and accompanying group membership explanation technologies may prove here as helpful supporting tools. They rely on so-called embeddings that exhibit varying degrees of semantic in-depth analysis.

Many types of embedding spaces were developed since the very first proposal of term vector space (TVS). Despite its straightforwardness, TVS has two notable drawbacks: (1) the document is regarded as a collection (or "bag") of words, which leads to the loss of context and the relationships among terms (2) the dimensionality can be as high as dozens of thousands, even for a moderate-size collection of several thousand documents.

Therefore, new embedding approaches, such as Word2vec, Doc2Vec, GloVe, BERT and many others, were developed to take relationships between terms into account.

Usually, we do not need only the clusters of textual documents, but also a characterization of the contents in terms e.g. of keywords. Hereby frequently we are interested in explaining why a document belongs to a cluster that is why the algorithm put a document into this cluster. We shall term this "cluster membership explanation".

Cluster membership explanation seems to be simple if we cluster the documents directly in the Term Vector Spaced via e.g. k-means algorithm because the cluster center coordinates tell us the importance of words/terms for cluster membership.

However, both when clustering in the modern embedding spaces, like word2vec, doc2vec, GloVe, BERT and other transformer methods, we get results that are hard to explain as the relationship between the embedding space coordinates and document words/terms gets lost which is counterproductive with respect to the librarian's goal to assign some automated subject index.
Furthermore, both the clustering and the choice of keywords should take into account the neutrality or emotions (like being sure what is stated) on the author(s) as well as the general value of the documents ("better documents" should influence the clustering more).
The aim of the project is to carry out research in the following areas:

- Document clustering
- Clustering explanation
- Spam detection
- Sentiment analysis

Synergic effects between the mentioned area should be explored.

A candidate should contact prospective supervisor before submission of the documents.

BIBLIOGRAPHY:

[1] Zhou, Lina, 2007: „Ontology learning: state of the art and open issues", Information Technology and Management", 2007/09 pp. 241- 252 UR - https://doi.org/10.1007/s10799-007-0019-5DO - 10.1007/s10799-007-0019-5

[2] Kaity, Mohammed and Balakrishnan, Vimala, 2019: "An automatic non-English sentiment lexicon builder using unannotated corpus", The Journal of Supercomputing. 2019/75, pp. 2243—2268

[3] Roffo, Giorgio: Learning to rank and ranking to learn. On the role of ranking in pattern recognition applications. arXiv http://arxiv.org/abs/1706.05933

[4] Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, Bartłomiej Starosta, Dariusz Czerski, Piotr Borkowski: A Method for Handling Negative Similarities in Explainable Graph Spectral Clustering of Text Documents. https://arxiv.org/abs/2504.12360

[5] Bartłomiej Starosta , Mieczysław A. Kłopotek , Sławomir T. Wierzchoń , Dariusz Czerski , Marcin Sydow , Piotr Borkowski: Explainable Graph Spectral Clustering of text documents , *PLOS ONE* , 2025 volume 20(2), pages e0313238 DOI https://doi.org/10.1371/journal.pone.0313238