Week 11 Research Notebook

From the reading:
In the paper, the author has analysed the structure of computer science research articles by examining the organization and rhetorical moves present in various sections, such as introductions, results, and conclusions. The author has compared these moves and structures to other models proposed by researchers like Swales, Cooper, and Hughes.

The analysis of these structural elements contributes to our understanding of how computer science research articles are written, and how they might differ from articles in other disciplines. This kind of text analysis is particularly useful for academic writing, as it can help researchers and students better understand the conventions and expectations within their field, ultimately improving their own writing and comprehension of academic texts.

In summary, the paper and the discussion above are related to text analysis because they involve the examination of the structure, rhetorical moves, and patterns in academic research articles, with the goal of better understanding the conventions and characteristics of computer science papers.

From the research topic "Universities":

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.
URL: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Summary: This is the original paper that introduced the Latent Dirichlet Allocation (LDA) model. It presents LDA as a generative probabilistic model for collections of discrete data, particularly text corpora. The authors demonstrate the effectiveness of the model in discovering latent topics in a large dataset of scientific abstracts.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.
URL: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.108.8490

Summary: This paper introduces Latent Semantic Analysis (LSA), a technique for representing the contextual usage of words in documents by a lower-dimensional semantic space. LSA uses singular value decomposition (SVD) to reduce the dimensionality of the document-term matrix, resulting in a more compact and noise-resistant representation. The authors demonstrate the effectiveness of LSA in information retrieval and other text-related tasks.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.
URL: https://arxiv.org/abs/1301.3781

Summary: This paper introduces the Word2Vec algorithm, a method for learning high-quality word embeddings using a shallow neural network architecture. The authors propose two model architectures, the Continuous Bag-of-Words (CBOW) model and the Skip-gram model, for learning word embeddings from large text corpora. The paper demonstrates that the learned word embeddings capture syntactic and semantic relationships between words and can be used as input features for various natural language processing tasks.

Salton, G., & McGill, M. J. (1986). Introduction to modern information retrieval. McGraw-Hill.
URL: https://dl.acm.org/doi/book/10.5555/576283

Summary: This book provides an overview of information retrieval techniques, including the Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme. It explains the importance of term weighting in information retrieval and the rationale behind the TF-IDF formula. The book covers various techniques and models for indexing and searching text documents, providing a solid foundation for understanding text analytics.

These papers and book provide foundational knowledge on LDA, LSA, word embeddings, and TF-IDF, which are essential for understanding the various feature extraction techniques used in text analytics.