International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 1, Jan-Feb 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

ONTOLOGY-BASED FRAMEWORK FOR CATEGORIZING MEDICAL DOCUMENTS

[1]S.Narmatha, [2]Dr.V.Maniraj
[1]Research Scholar, Dept. of Computer Science, AVVM Sri Pushpam College (Affiliated to Bharathidasan University, Tiruchirappalli), Poondi, Thanjavur, Tamilnadu, India.
(Mail Id-imnarmatha2792@gmail.com)
[2]Research Advisor, PG and Research Dept. of Computer Science, AVVM Sri Pushpam College (Affiliated to Bharathidasan University, Tiruchirappalli)., Poondi, Thanjavur, Tamilnadu, India
(Mail Id-manirajv61@gmail.com)

----------**********----------

ABSTRACT
The goal of this study is to explore the challenges involved in using domain ontology for classifying online content. We are applying the Medical Subject Headings (MeSH) thesaurus with the purpose of improving the classification of medical documents. As a result, we aim to develop a new representation based on our findings. We assessed and compared our method to traditional stem encoding using two popular data mining techniques, C4.5 and KNN. Our approach showed significantly better performance than the others. By incorporating concepts and hypernyms from domain ontology, we were able to significantly enhance the vectors crucial for effective data organization. Our analysis of the Ohsumed benchmark biological dataset confirmed the effectiveness of our method, with ontology-based categorization outperforming traditional stem representation by 30%.

----------**********----------

INTRODUCTION:

The widespread use of the internet has made the transmission of information more accessible, but the challenge of retrieving crucial data remains a significant issue. Ongoing research is focused on improving the automated retrieval of relevant information. Future developments in the Semantic Web will depend on these advancements, which require novel methods of data organization and the use of semantic annotations, typically via metadata. Ontologies, a widely used framework for expressing knowledge, help address this issue. These ontologies provide a logical structure based on the relevant domain, connecting and organizing concepts through various relationships. A key question in this context is: how does a conceptual model support the automated categorization of medical records? This study aims to explore that question. The organization of this paper is as follows: Section 2 reviews relevant literature, Section 3 explains the structure, methodology, and unique aspects of our approach, Section 4 introduces the Ohsumed benchmark and the MeSH medical ontology, allowing for a comparison with the bag-of-stems method, Section 5 provides a comprehensive summary of the experiments and results, and Section 6 concludes with reflections and future directions.

RELATEDWORK

Unfortunately, many document categorization methods still rely on the outdated bag-of-words approach. Researchers have long focused on developing conceptual representations of textual content, with one promising technique being the use of ontologies. The main drawback of the bag-of-words model is its inability to capture the relationships between terms. The significance of Amine's research, which advocates for enhancing text document clustering with an ontology (such as WordNet), is a prime example of this. Recently, G has shown that ontology-based text classification is a practical solution, demonstrating how bilingual websites can be categorized using a multilingual ontology. They propose creating a topic-specific ontology by combining linguistic and statistical methods, offering an alternative approach for extracting data from online documents. For an ontology to effectively represent meaning, web content must be categorized accurately and with relevance. This underscores the need for an algorithm designed to automate the classification of online content based on domain-specific ontologies, especially in the absence of a pre-existing knowledge base or learning algorithms.

A METHOD FOR CONCEPTUAL REPRESENTATION

To achieve effective document (text) classification, it is crucial to develop a method for encoding documents that allows for efficient processing while retaining only the information necessary for classification. The bag-of-words model is commonly used for this purpose, but several researchers have proposed ways to overcome its limitations. This paper presents a concept-based method aimed at improving and reducing the dimensionality of the representation vector, leading to two key improvements in our text classification system.

For ease of comparison, our approach will be implemented in two stages:

Conceptual Transformation: In this phase, we will convert selected terms into concepts using an appropriate matching and disambiguation strategy. This step will help incorporate contextual information and complex semantic relationships.

Hyperonym-based Enhancement: Building upon the conceptual representation, we will refine the representation vector further by incorporating hyperonyms. This will capture higher-level semantic relationships and enhance the overall accuracy of the representation.

Pre-processing Stage of the Texts:

At this point, we begin by refining the data to ensure we are working only with text suitable for classification. It is crucial to remove unnecessary elements from web content, such as images and HTML tags. A pre-processing step is essential to retain only the most relevant terms and ensure that duplicate instances of the same phrase are not mistakenly treated as distinct entities. This pre-processing is vital before using a lexicon and stemming algorithm, as it improves the relevance of the data and boosts categorization efficiency. A lot of content may not be directly related to the core message, and stop words are commonly used to handle this issue. Additionally, stemming is a pre-processing technique that refines text representation by reducing words to their root forms, thus enriching the representation vectors.

Stop Words: Stop words are commonly used words that hold little meaning in the context of text classification. Removing these words during pre-processing reduces the overall text size, leading to faster classification times. These words are prevalent in the dictionaries of most languages.

Stemming: The concept of stemming was introduced by Porter in 1980, primarily for the English language, with the aim of grouping words by their shared roots. This method aligns the classification process with human reading habits. For example, when encountering words like *walk*, *walker*, and *walking*, we understand that the document is likely centered around the concept of walking. An algorithm that does not apply stemming would treat each word individually. In contrast, using Porter's method links all related forms of a word to a single term, pointing to a common theme or topic. The stemming algorithm analyzes word structures to uncover their etymological roots. The success of applying Porter's stemming approach to works like Shakespeare's led to adaptations for other languages, with linguistic expertise playing an essential role in these developments.

Figure 1 shows an example of a cleaned and stemmed representation vector. This diagram likely illustrates how a text, such as "Infect Clinic Hemiplegia," is transformed into a numerical or vector representation, which encodes the semantic meaning and relationships between words. This vector allows the text to be processed and analyzed by algorithms.

……	Infect	clinic	Hemiplegia	.....

MAPPING TERMS TO CONCEPTS:

The Connection Between Words and Ideas Figure 2 presents an example illustrating how individual terms are mapped to their corresponding concepts. This process of linking terms to concepts is fundamental in natural language processing, as it allows for the representation of abstract ideas and the relationships between words.

In this context, Figure 2 is likely a visual aid that demonstrates the mapping concept. However, without seeing the actual figure, the explanation remains general. The ontology plays a vital role by associating terminology with the concepts they represent. For instance, the terms "appendicitis" (2) and "appendiceal" (1) are both linked to the broader concept of appendicitis, and their term frequencies are aggregated to reflect the frequency of the concept itself.
At this stage, three theoretical strategies for integrating or replacing terms with concepts can be outlined:

SUBSTITUTING WORDS WITH CONCEPTS:
This approach closely mirrors the first method but leverages the concept vector to represent all terms in MeSH (Medical Subject Headings), rather than simply reproducing the terms in the new representation. The term vector will only include terms not present in MeSH.

CONCEPT VECTOR ONLY:
This method differs from the previous one by using a new representation that omits any terms, including those in MeSH, and focuses exclusively on the concepts.

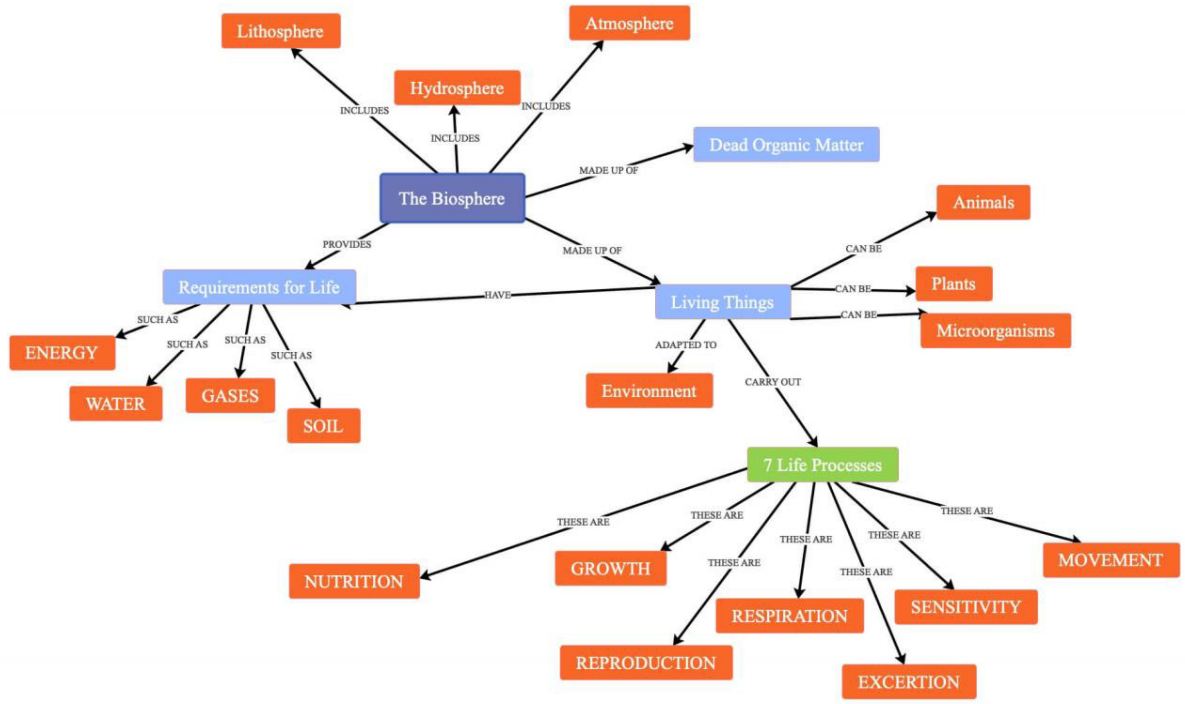


Figure 2. Representation of the correlation between terms and concepts.

METHODS TO ADDRESS AMBIGUITY:

Given that language is often subjective, assigning precise labels to concepts can be challenging. A single word can carry multiple meanings, making sense disambiguation crucial. Word Sense Disambiguation (WSD) is considered one of the toughest challenges in AI, often on par with the most complex issues in the field. While advanced WSD techniques may be difficult to implement, we will focus on two simpler approaches. Regardless of the specific concept, this method assumes that identifying central themes or ideas within a work can significantly enhance text representation. These themes naturally emerge, even as dimensionality increases. The calculation of concept frequencies will proceed as follows:

$$cf(d,c) = tf\{d, t \in T \mid c \in (ref_c(t))\}$$

The key idea behind this approach is that it emphasizes the most frequent meanings of a word. The ontology used will generate a ranked list of ideas, assigning more importance to the more common interpretations of a word and less importance to rare ones. This principle is widely followed by most ontologies. The process for calculating concept frequencies is outlined as follows:

$$cf(d,c)=tf\{d,t \in T|\ first(ref_c(t)) = c\}$$

UTILIZING HYPERNYMS:
Grasping the relationships between concepts is crucial for accurately reflecting the themes within texts. Simply substituting terms with concepts without considering their interrelationships may not lead to substantial improvements in performance. In fact, in some cases, it could be less effective than using the original terms. This observation aligns with findings in current research. To address this, we incorporate the frequency of each concept within a text and the frequency of its related hyponyms, exploring the hypernym relationships between concepts. The next step is to adjust the frequency of the concept vector component as follows:Where H(c) represents the hyponyms associated with a specific concept, c.

$$cf'(d,c) = \sum_{b \in H(c)} cf(d,b)$$

SELECTING AND REDUCING DESCRIPTORS:
The mapping process is performed on each document in the training corpus. The concepts derived from the ontology serve as descriptors for the vector that represents each document. The frequency of a concept within a specific category of the corpus indicates its significance. The goal here is to identify the most accurate descriptors for each category in comparison to others. Weighting a term highlights its importance within a given category. A basic approach might simply count the occurrences of the term within the category, but this method fails to consider the relevance of that category in comparison to others.

The FTIDF (Term Frequency-Inverse Document Frequency) method is more commonly applied and is a better weighting mechanism. It is used within the vector model framework, which involves:

«Frequency Term »* « Inverse document frequency »

FTIDF Formula:
FTIDF(Ci, Wj) = TF(Ci, Wj) * log (nbr_category / DF(Wj))

Where "FT: Frequency Term" refers to the number of occurrences of a specific term within the relevant category, and "IDF: Inverse Document Frequency" is calculated as the total number of categories divided by the number of categories containing the term in question.

$$\mathbf{FTIDF(Ci,Wj)} = TF(Ci,wj) * log \frac{\mathbf{nbr_{category}}}{\mathbf{DF(wj)}}$$

The dimensionality reduction selection process involves taking a set of features and creating a more focused subset that is effective for distinguishing between different categories. Managing the dimensions of the vector space is important for two reasons. First, when assessing the difficulty of a learning algorithm, both the volume of features and the number of learning examples must be considered. Reducing the number of index terms can improve the efficiency of these algorithms. Intuitively, having more features should lead to better classifiers, as...

DIMENSIONALITY REDUCTION AND FEATURE SELECTION:
Although more features may provide more information, it isn't always beneficial for the learning algorithm to handle a large number of features, some of which may be irrelevant. This is where dimensionality reduction

becomes essential. The basic assumption here is that a term that is strongly associated with a specific category will be more useful for distinguishing it from other categories. During the dimensionality reduction process, terms with low χ^2 values are discarded.

The χ^2 (chi-squared) technique is a supervised method that considers both term frequencies in each category and the interactions between terms and categories. To apply this method, we first identify each category and then extract the K most important features that characterize it in relation to the other categories.

COMPUTING χ^2 USING MATHEMATICS:
Here's a simplified mathematical formula to compute χ^2:

$$X^2(D,t,c) = \frac{(N_{11}+N_{10}+N_{01}+N_{00}) * (N_{11}N_{00}-N_{10}+N_{01})^2}{(N_{11}+N_{01}) * (N_{11}+N_{10}) * (N_{10}+N_{00}) * (N_{01}+N_{00})}$$

Where:

χ^2 (chi-squared) represents the association between a term and a category

Ntc is the number of documents in category c, with a value of 0.1 (to ensure sparse data is accounted for)

N11 (n_0_1) is the total number of documents containing the term and categorized under c

Understanding the Components of χ^2 Calculation:

In the χ^2 (chi-squared) method used for feature selection, counts related to the presence or absence of specific terms within documents are crucial for analysis:

N10: The number of documents that contain the particular term but do not belong to the corresponding category.

N01: The count of documents that belong to the category but do not contain the specific term.

N00: This refers to documents that neither contain the term nor belong to the category.

Key Features of the χ^2 Method:

Supervised Approach: The χ^2 method is considered supervised as it relies on pre-classified data, which helps it utilize this information for selecting features.

Multivariate Nature: It is a multivariate approach because it assesses not just the individual contributions of each feature but also their interactions with each other and the categories.

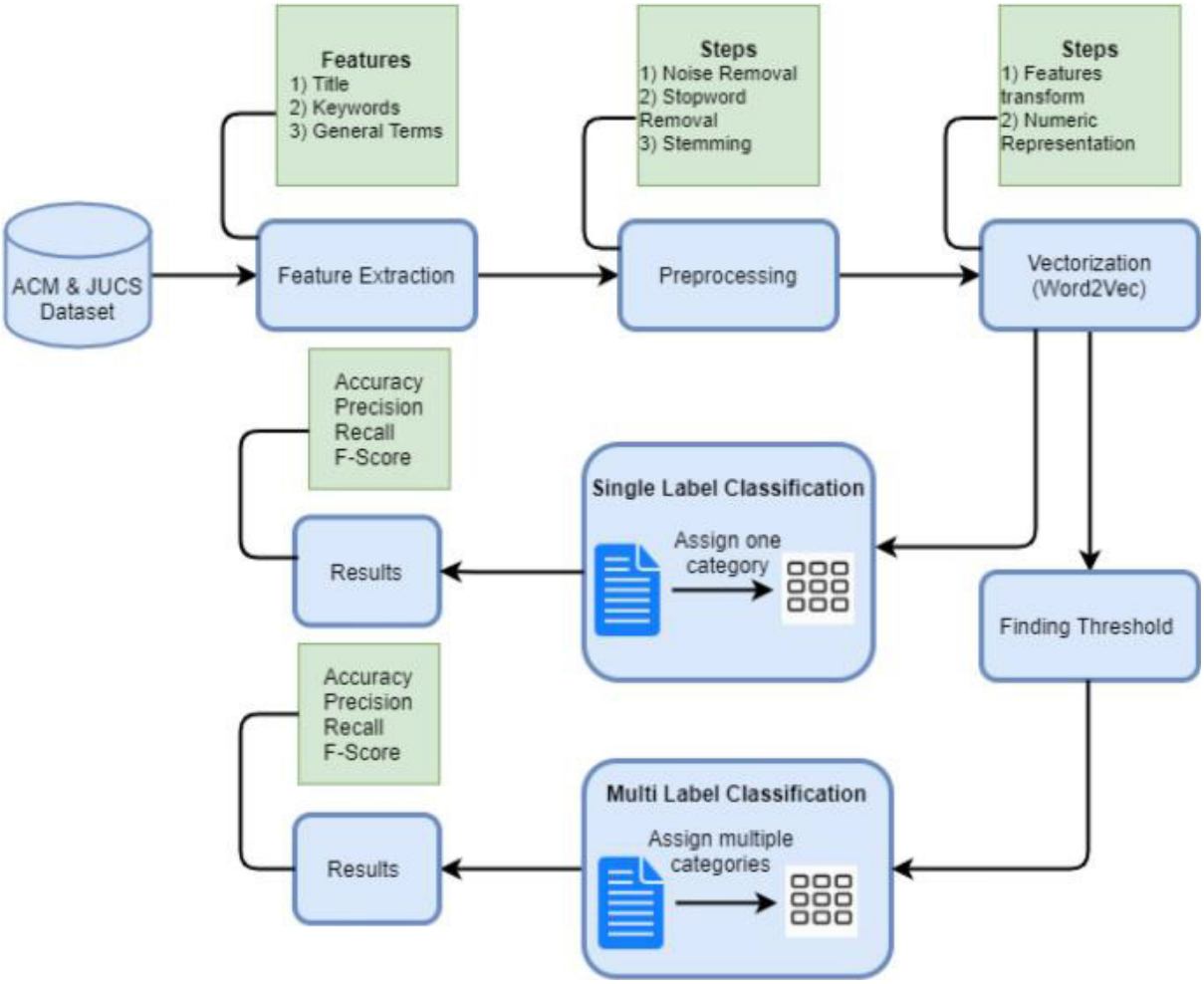
Focus on Interactions: The method's primary goal is to explore the relationships between features and categories, helping to identify the most relevant terms for classification.

Linear Complexity: Despite its intricate analysis, the method remains efficient with linear complexity regarding the number of calculations, making it scalable for large datasets.

In essence, this method is based on category-specific data, emphasizes term interactions, and remains computationally efficient, making it an effective tool for feature selection in classification tasks.

BUILDING AND EVALUATING A TEXT CLASSIFICATION MODEL:

Once the texts are categorized and preprocessing is done, the next step involves building a machine learning model using the concept matrix derived from the vectors. This process follows the typical approach for supervised classification tasks. We will evaluate the performance of two widely used techniques: C4.5 and K-Nearest Neighbors (KNN). These popular methods are chosen not for their technical details but for their intellectual significance in the analysis. After training the model, we can classify new texts by inputting their concept vectors into the trained model. A simplified diagram in Figure 3 illustrates the entire workflow, from text preprocessing to classification.



ASSESSMENT OF THE METHODOLOGY:

The OHSUMED Dataset:

In 2000, we utilized the OHSUMED dataset within the TREC9 Task-Filtering framework. This dataset includes 270 medical articles, published or summarized between 1987 and 1991, each containing six sections: the title (.T), summary (.W), author (.A), source (.S), publication date (.P), and MeSH-indexed concepts (.M).

Categories and Document Counts:

EVALUATION METHODOLOGY:

This section outlines experimental results using the **"F-measure"**, which is the harmonic mean of recall and precision, defined as:

$$F = \frac{2 * recall * precision}{recall + precision}$$

In text categorization, the two common metrics, recall and precision, are used to assess the algorithm's effectiveness within a particular category. These metrics are incorporated into the F-measure formula:

$$Precision = \frac{True\ positive}{True\ positive + false\ positive}$$

$$recall = \frac{True\ Positive}{True\ positive + false\ negative}$$

We refer to the **MeSH Ontology** in our research, using the biomedical thesaurus developed by the National Library of Medicine (NLM). The **Medical Subject Headings (MeSH)** thesaurus, established to help in indexing and retrieving medical information, has evolved over time, containing thousands of descriptors classified into categories like anatomy, diseases, organisms, and more, with up to 11 hierarchical levels.
For evaluating the proposed methodology, we used the OHSUMED dataset and two data mining algorithms: **C4.5** (a decision tree algorithm) and **KNN** (K-Nearest Neighbors). Both algorithms have proven effective for document classification. Our comparison is based on the representation of stems, and the methodology was tested across eight different groups from the OHSUMED corpus.

FINDINGS:
Our research indicates that the ontology-based representation provides superior classification accuracy. This finding is particularly noteworthy, showing a **30% improvement in performance**, which marks a significant advancement. Moreover, enhancing the representation vector by incorporating **hypernyms** and related techniques has resulted in substantial gains in performance. This supports the idea that the **KNN algorithm** is an optimal choice for classification tasks, especially when combined with the χ^2 reduction technique.
The observed improvement in performance could be attributed to the compatibility between the KNN algorithm and the χ^2 method. This relationship may hold the key to more effective classification and representation optimization, warranting further exploration.

CONCLUSION AND FUTURE WORK:

The primary goal of our methodology was to enhance the classification process using domain-specific ontology. We chose the medical domain due to its relevance and the growing interest in applying data mining techniques within this field. After testing our approach on a benchmark corpus and evaluating it with two prominent algorithms, KNN and C4.5, we first used the standard **Stems representation** for comparison, then presented the results and their interpretations. We subsequently integrated **MeSH concepts and hypernyms** into the document representation framework.
Our results validate the effectiveness of the approach, revealing a **30% improvement in performance** over initial expectations. These findings suggest that subject- and ontology-based document tagging is a highly efficient strategy for classification tasks. Looking ahead, there are several areas to explore. One promising direction is to investigate **hyperonymy**, which allows us to extend the analysis to multiple levels within the MeSH ontology, offering potential for improved performance. Ultimately, the goal is to achieve a level of generalization that could make the methodology more widely applicable.
Additionally, we could apply the same framework with a **multilingual MeSH ontology**, exploring the classification of medical literature in various languages, thereby expanding the scope and relevance of our research.

REFERENCE:

1.Hassan, Sahar, Franck Hétroy, and Olivier Palombi. "Ontology-guided mesh segmentation." *FOCUS K3D Conference on Semantic 3D Media and Content*. 2010.
2.Osborne, John D., et al. "Interpreting microarray results with gene ontology and MeSH." *Microarray Data Analysis: Methods and Applications* (2007): 223-241.
3.Trieschnigg, Dolf, et al. "MeSH Up: effective MeSH text classification for improved document retrieval." *Bioinformatics* 25.11 (2009): 1412-1418.
4.Elberrichi, Zakaria, Belaggoun Amel, and Taibi Malika. "Medical Documents Classification Based on the Domain Ontology MeSH." *arXiv preprint arXiv:1207.0446* (2012).
5.Yoo, Illhoi, and Xiaohua Hu. "Biomedical ontology mesh improves document clustering qualify on medline articles: A comparison study." *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*. IEEE, 2006.

6. Ayeldeen, Heba, Aboul Ella Hassanien, and Ali Ali Fahmy. "Evaluation of semantic similarity across MeSH ontology: a Cairo University thesis mining case study." 2013 12th Mexican International Conference on Artificial Intelligence. IEEE, 2013.
7. Gope, Hira Lal, et al. "Medical document classification from OHSUMED dataset." IJCSN International Journal of Computer Science and Network 3.4 (2014): 215-219
8. Katris, Nikolaos, Richard Sutcliffe, and Theodore Kalamboukis. "Using a Cross-Language Information Retrieval System based on OHSUMED to Evaluate the Moses and KantanMT Statistical Machine Translation Systems." Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016
9. Sayed, Mahmoud F., and Douglas W. Oard. "Jointly modeling relevance and sensitivity for search among sensitive content." Proceedings of the 42nd international ACM SIGIR
conference on research and development in information retrieval. 2019.
10. Gupta, Vishwa, and Sitesh Kumar Sinha. "Comparison of machine learning methods for OHSUMED-F Data Set: A Cardiovascular Diseases Simulation Study." NeuroQuantology 20.15 (2022): 385

