CONTENT

DRUG DISCOVERY FOR PROJECT APPLICATION USING MACHINE LEARNING

ABSTRACT

The advancements of information technology and related processing techniques have created a fertile base for progress in many scientific fields and industries. In the fields of drug discovery and development, machine learning techniques have been used for the development of novel drug candidates. The methods for designing drug targets and novel drug discovery now routinely combine machine learning and deep learning algorithms to enhance the efficiency, efficacy, and quality of developed outputs. The generation and incorporation of big data, through technologies such as high-throughput screening and high through-put computational analysis of databases used for both lead and target discovery, has increased the reliability of the machine learning and deep learning incorporated techniques. The use of these virtual screening and encompassing online information has also been highlighted in developing lead synthesis pathways. In this review, machine learning algorithms utilized in drug discovery and associated techniques will be discussed. The applications that produce promising results and methods will be reviewed.

Chapter 1

INTRODUCTION

Machine learning (ML), an essential component in AI, has been integrated into many fields, such as data generation and analytics. The basis of algorithm-based techniques, such as ML, requires a heavy mathematical and computational theory. ML models have been used in many promising technologies, such as deep learning (DL) assisted self-driving cars, advanced speech recognition, and support vector machine-based smarter search engines. The advent of these computer-assisted computational techniques, first explored in the 1950s, has already been used in drug discovery, bioinformatics, cheminformatics, etc. Drug discovery has been based on a traditional approach that focuses on holistic treatment. In the last century, the world's medical communities started to use an allopathic approach to treatment and recovery. This change led to the success of fighting diseases, but high drug costs ensued, becoming a healthcare burden. While quite diverse and specific to candidates, the cost of drug discovery and development has consistently and dramatically increased. The generalized components of early drug discovery include target identification and characterization, lead discovery, and lead optimization. Many computer-based approaches have been used for the discovery and optimization of lead compounds, including molecular docking, pharmacophore modeling, decision forests, and comparative molecular field analysis. ML and DL have become attractive and DL have become attractive approaches to drug discovery. The applications of ML and DL algorithms in drug discovery are not limited to a specific step, but for the whole process. In this article, we review the ML and DL algorithms that have been widely used in drug discovery.

1.1OBJECTIVE

- Drug discovery is the process by which drugs is aims at identifying a compound therapeutically useful in curing & treating disease.
- The process of drug discovery involves the identification of candidates, synthesis, characterization, screening & assays for therapeutic efficacy.
- A comparison of several machine learning techniques has been performed in order to obtain a satisfactory classifier for detecting drug target articles using semantic information .
- Our objective is to Identify the drug discovery for patients Using SVM and KNN algorithms.

1.2 Proposed Algorithms:

SVM

TfidfVectorizer transformer:

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction.

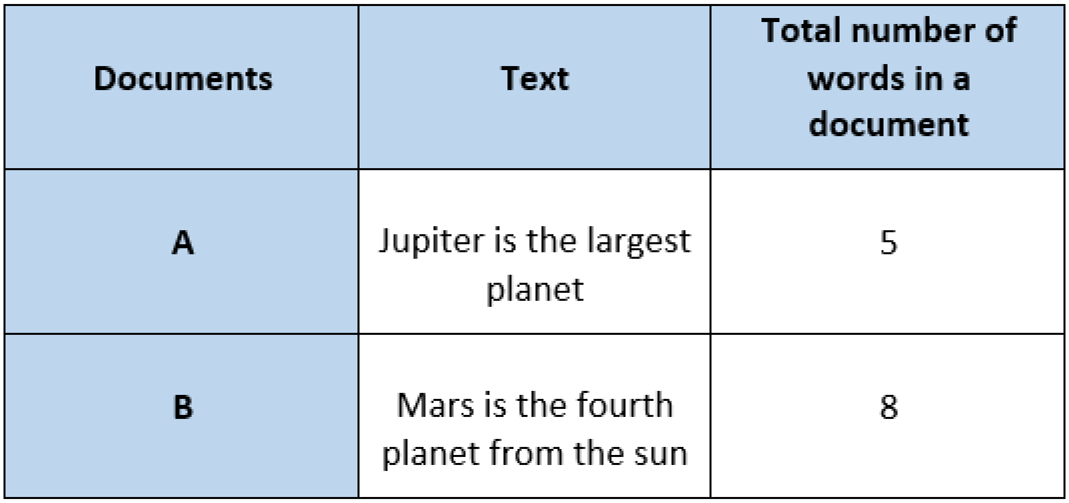
Term Frequency (TF)

It is a measure of the frequency of a word (w) in a document (d). TF is defined as the ratio of a word’s occurrence in a document to the total number of words in a document. The denominator term in the formula is to normalize since all the corpus documents are of different lengths.

$$TF(w,d) = \frac{occurences\ of\ w\ in\ document\ d}{total\ number\ of\ words\ in\ document\ d}$$

Image by Author

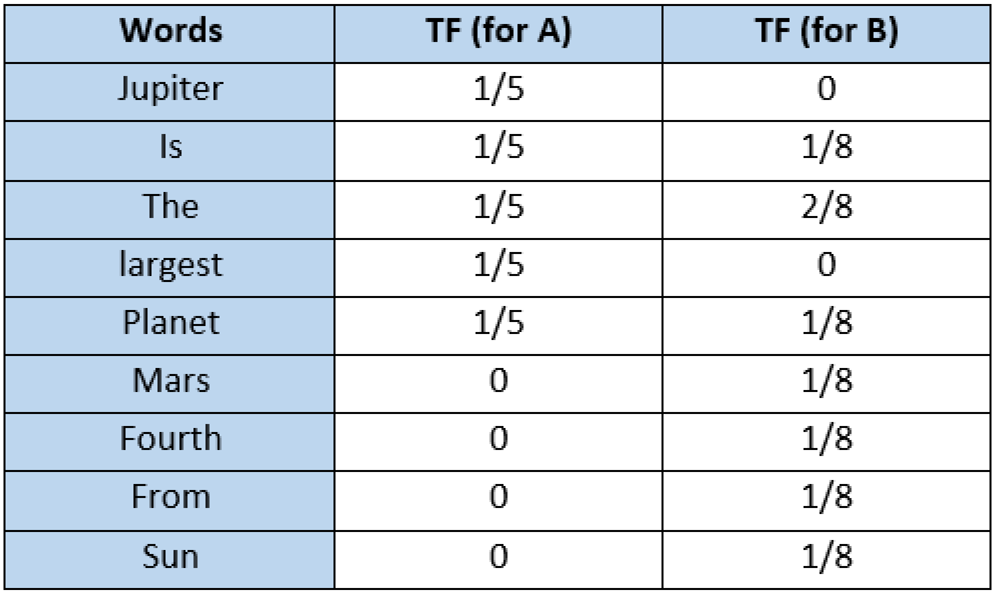
EXAMPLE



Documents	Text	Total number of words in a document
A	Jupiter is the largest planet	5
B	Mars is the fourth planet from the sun	8

Image by Author

The initial step is to make a vocabulary of unique words and calculate TF for each document. TF will be more for words that frequently appear in a document and less for rare words in a document.



Words	TF (for A)	TF (for B)
Jupiter	1/5	0
Is	1/5	1/8
The	1/5	2/8
largest	1/5	0
Planet	1/5	1/8
Mars	0	1/8
Fourth	0	1/8
From	0	1/8
Sun	0	1/8

Image by Author

Inverse Document Frequency (IDF)

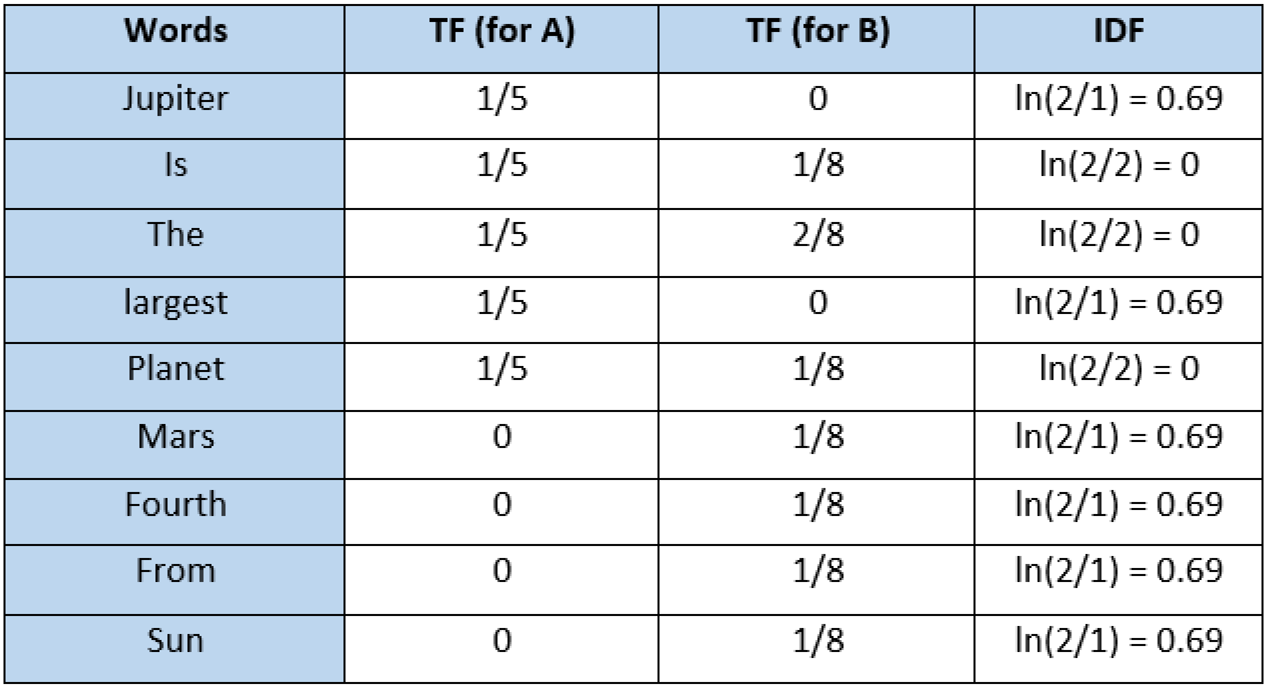
It is the measure of the importance of a word. Term frequency (TF) does not consider the importance of words. Some words such as' of', 'and', etc. can be most frequently present but are of little significance. IDF provides weightage to each word based on its frequency in the corpus D.

IDF of a word (w) is defined as

$$IDF(w, D) = \ln(\frac{Total\ number\ of\ documents\ (N)\ in\ corpus\ D}{number\ of\ documents\ containing\ w})$$

Image by Author

In our example, since we have two documents in the corpus, N=2.



Words	TF (for A)	TF (for B)	IDF
Jupiter	1/5	0	ln(2/1) = 0.69
Is	1/5	1/8	ln(2/2) = 0
The	1/5	2/8	ln(2/2) = 0
largest	1/5	0	ln(2/1) = 0.69
Planet	1/5	1/8	ln(2/2) = 0
Mars	0	1/8	ln(2/1) = 0.69
Fourth	0	1/8	ln(2/1) = 0.69
From	0	1/8	ln(2/1) = 0.69
Sun	0	1/8	ln(2/1) = 0.69

Image by Author

Term Frequency — Inverse Document Frequency (TFIDF)

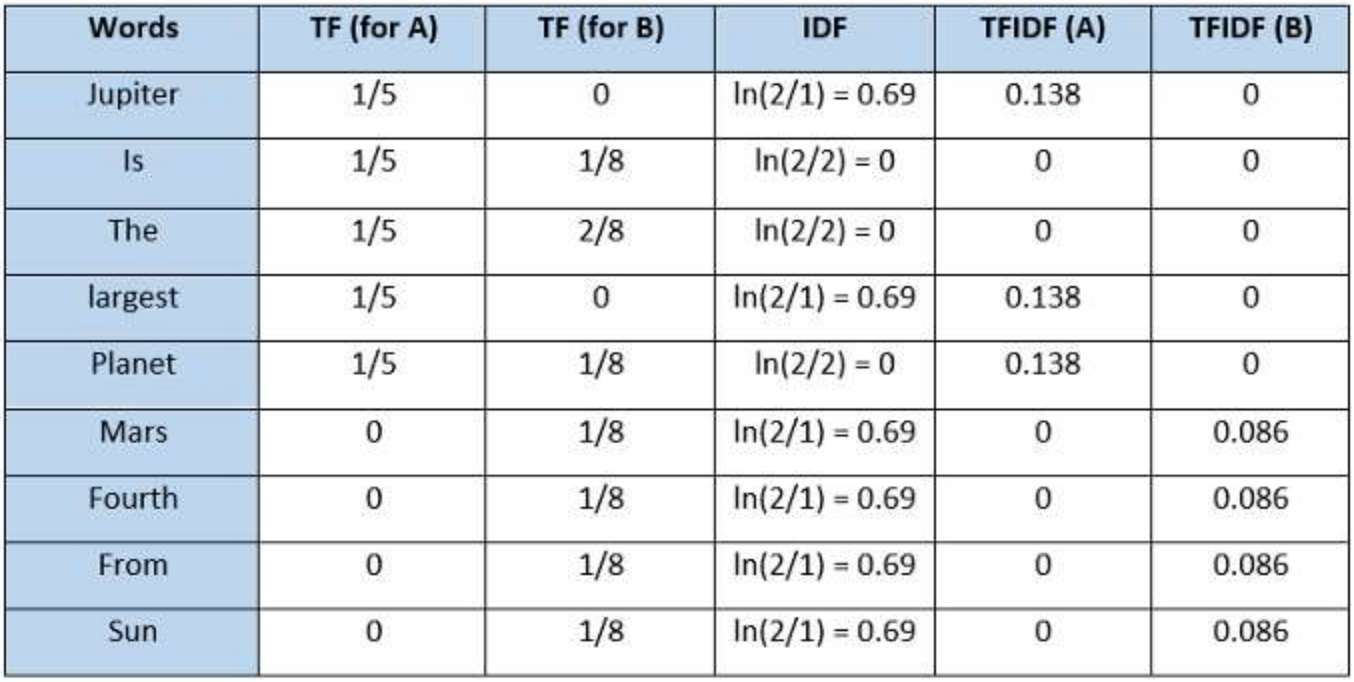
It is the product of TF and IDF.

TFIDF gives more weightage to the word that is rare in, the corpus (all the documents).

TFIDF provides more importance to the word that is more frequent in the document.

$$TFIDF\ (w, d, D) = TF(w, d) * IDF(w, D)$$

Image by Author



Words	TF (for A)	TF (for B)	IDF	TFIDF (A)	TFIDF (B)
Jupiter	1/5	0	ln(2/1) = 0.69	0.138	0
Is	1/5	1/8	ln(2/2) = 0	0	0
The	1/5	2/8	ln(2/2) = 0	0	0
largest	1/5	0	ln(2/1) = 0.69	0.138	0
Planet	1/5	1/8	ln(2/2) = 0	0.138	0
Mars	0	1/8	ln(2/1) = 0.69	0	0.086
Fourth	0	1/8	ln(2/1) = 0.69	0	0.086
From	0	1/8	ln(2/1) = 0.69	0	0.086
Sun	0	1/8	ln(2/1) = 0.69	0	0.086

After applying TFIDF, text in A and B documents can be represented as a TFIDF vector of dimension equal to the vocabulary words

Chapter 2

LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, then the next step is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system. The major part of the project development sector considers and fully survey all the required needs for developing the project. For every project Literature survey is the most important sector in software development process. Before developing the tools and the associated designing it is necessary to determine and survey the time factor, resource requirement, man power, economy, and company strength. Once these things are satisfied and fully surveyed, then the next step is to determine about the software specifications in the respective system such as what type of operating system the project would require, and what are all the necessary software are needed to proceed with the next step such as developing the tools, and the associated operations.

Applications of machine learning in drug discovery and development

Drug discovery and development pipelines are long, complex and depend on numerous factors. Machine learning (ML) approaches provide a set of tools that can improve discovery and decision making for well-specified questions with abundant, high-quality data. Opportunities to apply ML occur in all stages of drug discovery. Examples include target validation, identification of prognostic biomarkers and analysis of digital pathology data in clinical trials. Applications have ranged in context and methodology, with some approaches yielding accurate predictions and insights. The challenges of applying ML lie primarily with the lack of interpretability and repeatability of ML-generated results, which may limit their application. In all areas, systematic and comprehensive high-dimensional data still need to be generated. With ongoing efforts to tackle these issues, as well as