Semester :VI______________ Subject :________DAV________ Academic Year: 2023 - 2024

CATEGORISING DOCUMENT BY TOPICS:

- Document classification is the act of labeling document using categories, depending on their content.
- Document classification can also be manual, for example in library science, or automated within the field of computer science
- It is used to sort and manage texts, videos or images.
- The advantage of classifying documents is that humans have full control over the procedure of classification , and they can also make decisions as to which categories to use.
- But when there is large handling of volumes of documents, then this process is very slow and monotonous.
- But it is mich faster, more cost-efficient, and more accurate, to carry out automatic document classification , when it is carried out by machine learning.

Document Classification Vs Text Classification

- Text Classification involves classifying text by performing text analysis techniques on text-based documents.
- Analyzing text at different levels can be done with text classification:

Document-level: Here one can get relevant information for a full document.
Paragraph- level : Here one can get the most important categories of just one paragraph.
Sentence – level : Relevant information of a single sentence can be obtained.

Working of Automatic Document Classification:

We consider three different approaches to document classification:

Supervised:

- Here, we define a set of tags, eg. Usability, pricing and customer service and manually tag a number of texts.
- This way machine learning models can start making predictions on their own.
- For example, a customer comment says, “”This Lap-Top is quite expensive”. This comment is to be tagged as ‘pricing’.
- More and more texts to be classified to have better confidence of the model.

Unsupervised:

- In this method, similar words or similar sentences in the documents are grouped together by a classifier. This is done at random, i.e. without any prior training.
- For example, the words Mohan, Printer, Geeta would be considered as sharing similar qualities, and are grouped in the same cluster.

Rule – based

- The method is based on linguistic rules and they give instructions to models.
- The rules and patterns based on morphology. Syntax, semantics and phonology, tag the texts.
- For example:
- (Update |OS| Bugs) -> Software
- Here, the model will tag any text that mentions these terms as ‘Software’.
- The main advantage of this model is that the performance of the model is constantly improving, that

way it provides higher quality and more accurate insights.

- But the disadvantages is that creating this type of system is compels, hard to scale and time consuming. And to analyze a new type of text, will have to add new rules or change existing one every time.

Document Classification through AI

For automated document classification, there are two steps for preparing the dataset and training the algorithm.

We mention them:

(1) To gather dataset

- The dataset should contain enough documents or examples for each category so that the algorithm can learn to differentiate between them.
- For example, if you want to classify document into five categories, for training a classifier, there should be at least 100-300 documents per category to obtain predictive capabilities. So that the total number of documents within the dataset for training this classifier would be more than 500. Note that more the data we use, more accurate the classifier will be.
- Also the quality of the data is crucial when training a classifier with machine learning.
- If the examples that are fed to the classifier are incorrectly tagged, the model will commit similar errors whenever making predictions.

Training the Algorithm

- Once we get the get data to train the model, we use that data to train a classification algorithm. There are many complex algorithms can we use, for example, Naïve Bayes and support vector machines.
- Knowing how to code, we can use open source tools such as scikit – learn or tensor flow to train these algorithms to classify the documents.

To wrap Up

- It is better to begin with 'machine learning' for effective document classification. Many classification tools make it easy to begin with AI for document classification.