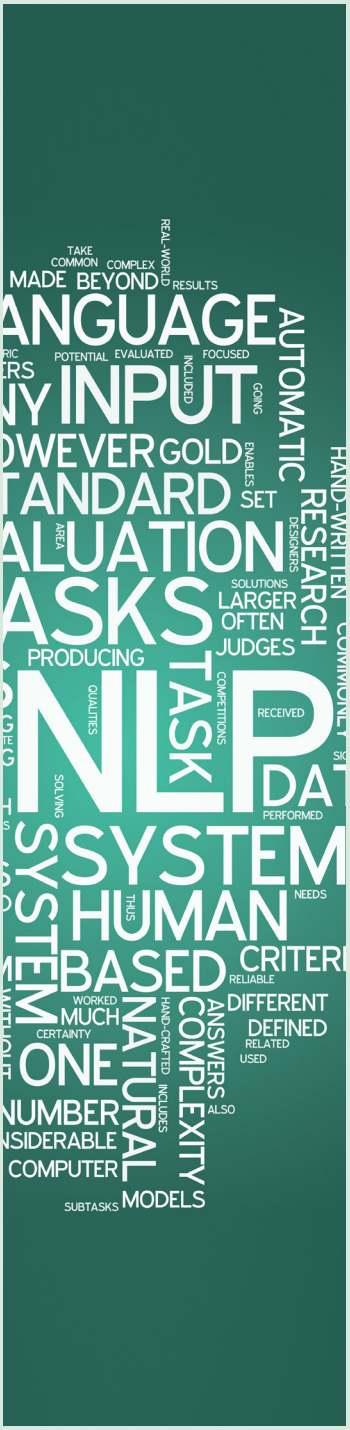
WHITE PAPER



MAKING CONTENT SERVICES SMARTER USING NLP

Abstract

There is an exponential growth in unstructured content due to digitization, IoT, regulatory compliance, etc. Because of this growth, organizations are facing challenges, on how to capture, ingest and manage huge volume of unstructured content effectively in Content Services platform. Traditional approaches are becoming ineffective. Need of the hour is to address these challenges differently and bring some "intelligence", so that Content Services platforms can manage the unstructured content smartly.

This paper discusses how Natural Language Processing (NLP) can make a difference in Content Services by addressing challenges which are being faced. Here, we will first introduce NLP and then briefly describe how NLP can be used in Content Services while capturing, ingesting and managing.

Infosys®
Navigate your next

1 Natural Language Processing (NLP)

1.1 What is NLP

Natural language processing is a subfield of artificial intelligence, dealing with the interactions between computers and human languages, in particular on how to program computers to process and analyze large amounts of natural language data. NLP includes text preprocessing, sentence segmentation, tokenization, POS tagging, Named Entity Recognition (NER), chunking, parsing, co-reference resolution and text categorizer. In this white paper we will focus on NER and categorizer.

1.1.1 Named Entity Recognition (NER)

Named Entity Recognition, is for identifying the entities in unstructured content. Entities such as name, person, location, time, money are few of the common entities which can be extracted. Apart from these, models can be trained to extract business specific entities such as account number, tax file number specific to the business. Core capability of NER is, classification and extraction of the named entities. NER can extract static as well as dynamic entities.

1.1.2 Categorizer

Core capability of categorizer is classification. It helps in categorizing a chunk of unstructured text into predefined categories. For example, categorizer models can be used to categorize news articles into different genre, sentiment analysis of messages, identifying the custom document types for an organization, etc. using the predefined category phrases, categorizer can best classify the content.

1.2 Terminology

1.2.1 Entity

Entity is a singular, identifiable and separate object. For example; name of a person, organization, city, account number, TFN number can be considered as an entity.

1.2.2 Entity Relationship

The semantic relationship between two different entities are called as entity relationship. For example; relationship between a person name and an organization is "employed by".

1.2.3 Annotation

Annotation is a process, in which, the information of interest is tagged with an entity name. For example, John is tagged with the entity "person_name".

1.2.4 Labeled Content

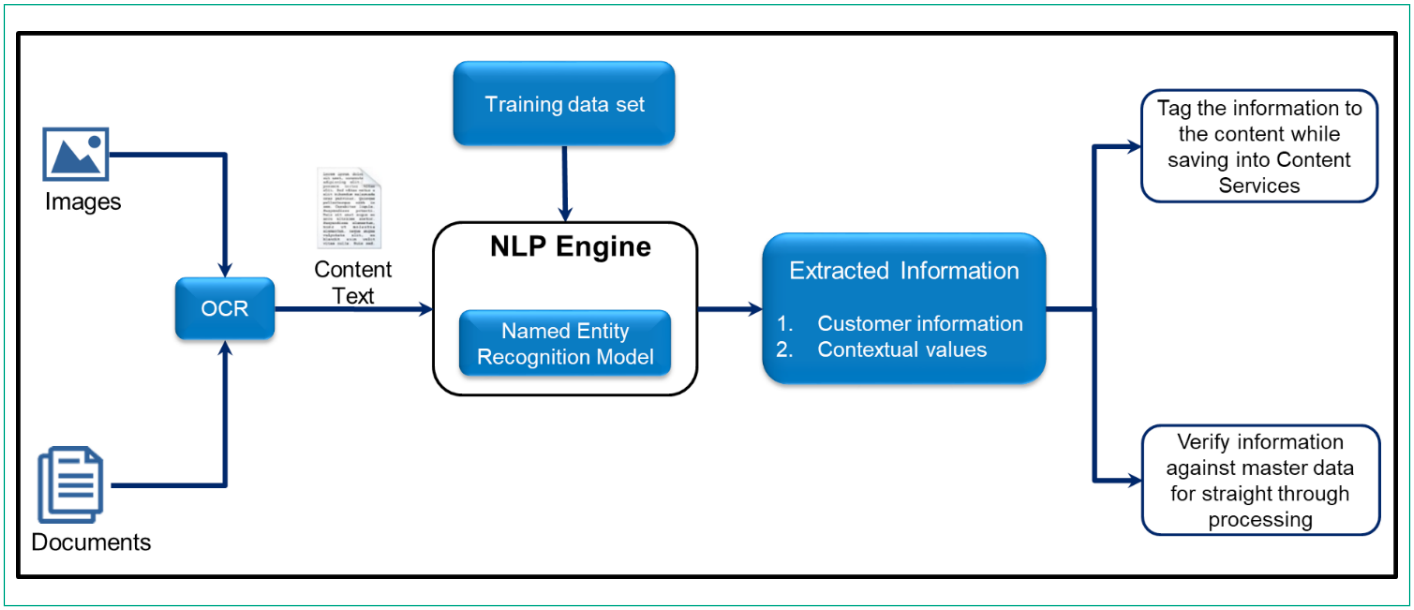
Labeled content is the content, whose classification category and/or entities are already known and which can serve as "training content" for training the AI models.

2 How NLP Can Make Content Services Smarter

2.1 Content Capture

Capturing content in digital format is one of the first steps in Content Services. Along with the content, information such as customer name, zip code, type of the document is also captured. Current capture process is based on the location of the information. In this process, templates are fed to capture tool and the tool is trained to extract information from a particular location, for that template. This location based information extraction is no longer productive due to huge variation in content layouts.

This issue can be solved by context based extraction. Using NLP, an AI model can be trained which can then be used for extraction. This process is also called as Cognitive Capture. Extracted information then can be directly tagged to the content while saving into the Content Services or it can be used for verification during straight through processing, etc.



To implement the information extraction, as a first step, AI model needs to be trained. NLP engines accepts annotated text files as training data set. Once the information in text document (or labeled content) is annotated, it can be used to train NLP engine. More the number of annotated files, accommodating variations in the interested information, more will be the accuracy of the AI model.

Once the AI model is trained, it can be used to capture the information from new or unseen documents. The new or unseen documents may need to be OCRed if required. Capture process implemented using NLP will not depend on layout of the content and still be able to capture appropriate information even if there is change in the layout, as long as the information is in the same context.

2.1.1 Implementation Scenarios

Below described are the some of the different implementation scenarios in the context of cognitive capture

2.1.1.1 Extracting metadata while Capture

Captured information can be tagged to the content while saving into Content Services. The tagged information will give identity to the content, which helps in locating and reusing the content.

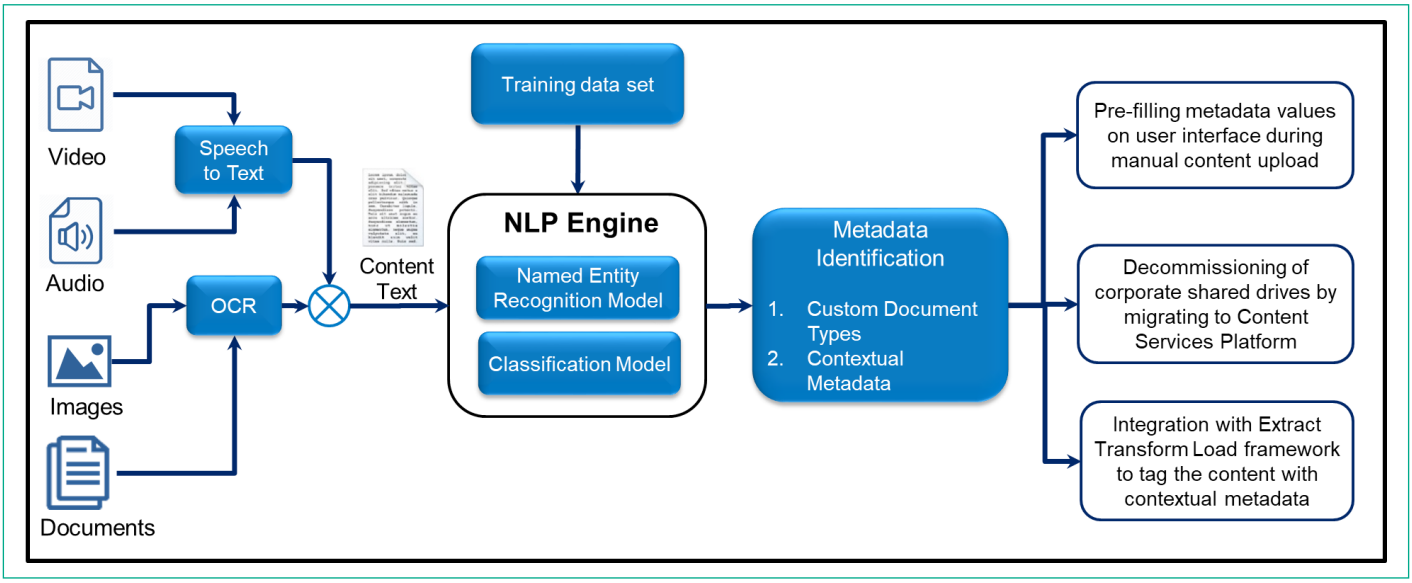
2.1.1.2 Data Verification for Straight through Processing

Data captured can be verified with master data management system for automatic processing, such as claim processing, etc.

2.2 Content Ingestion

Content Ingestion is a process where content is loaded into Content Services platform, either through a batch process or manually through a user interface. Biggest challenge here is identifying the metadata to tag with. For batch process metadata needs to be made available to the ingestion process, through an XML, CVS, database, etc. For manual ingestion or upload, user needs to think what metadata the file should have, and then accordingly enter the metadata on user interface. It takes significant amount of time to finalize what metadata the content should have. This manual process is also prone to mistakes.

NLP can help here by cognitively tagging the metadata to the content during ingestion. NLP engine can be trained to identify the metadata to be tagged, based on the context of the content. The identified metadata can be tagged with the content during ingestion. This can be illustrated by following diagram:



As a first step, the AI model is to be trained with labeled documents to retrieve metadata information from the document. Since it is context based metadata identification, both models, NER as well as Categorizer, can be used.

For training NER models, the process is same, that is to annotate the documents and then use the annotated documents for training.

Categorizer model is supported by some of the NLP tools such as Apache OpenNLP. To train a categorizer model, first the categories need to be defined. For example, to implement metadata identification through classification, a category needs to be mapped to a metadata value. As next step, from labelled content, important business terms/phrases, which helps in classifying the content to the category, need to be listed against that category. This list of category mapped to business terms can be fed to NLP engine, which then build a categorizer model. Based on the occurrences of the business terms/phrases in the new or unseen content, it can be classified into a category (which is being mapped to a metadata value).

New or unseen content is subjected to either NER or Categorizer or both the models. Use of both models will need more time to train, will give higher accuracy and of course, less throughput.

With appropriate plug-ins, text from audio and video files can be generated and subjected to the AI models for metadata identification and then ingested into the Content Services Platform.

2.2.1 Implementations scenarios

Described below are the different implementation scenarios in the context of cognitive content ingestion.

2.2.1.1 During Batch Ingestion

In any batch ingestion process, appropriate metadata values are needed to complete the process. One of the major challenges is the non-availability of the metadata values, which are being expected by Content Service Platform. NLP can help in bridging the expectation, by extracting context based values. AI models can be trained which can identify the metadata values as

a part of ingestion process and then load the content with appropriate metadata values. With NLP tool, as a part of ingestion pipeline, content from sources where the metadata values are not available, can be ingested into Content Services Platform.

2.2.1.2 Manual content upload

When users upload content manually into Content Services Platform, using a user interface, they have to enter metadata values. In order to do this, users have to read through the document, listen or view the content and then decide what metadata values this content can have. From users' perspective, this process is very time consuming and prone to errors.

NLP can solve this challenge. A trained NER or Categorizer model can identify what metadata values the content can have. Users are first asked to upload content which can be analyzed by the AI model and the suggested metadata that can be prefilled on user interface. Users need to just verify it and submit it. This saves time for the user.

2.2.1.3 Shared Drive Migration

Most of the organization store content on shared network drives. Content, which is stored on shared drive are not tagged with any metadata or organized in any directory structure. Due to this, the content cannot be located and reused. More than that, content laying on shared drive can be a risk on adherence to compliance and regulatory standards. To overcome these challenges and to manage the content efficiently, it needs to be migrated to a Content Service platform.

To migrate the content of shared drives, a trained NER or Categorizer model can be plugged-in into the migration framework. NLP model can identify the metadata of content and it can be tagged to the content while loading the content into the content service platform.

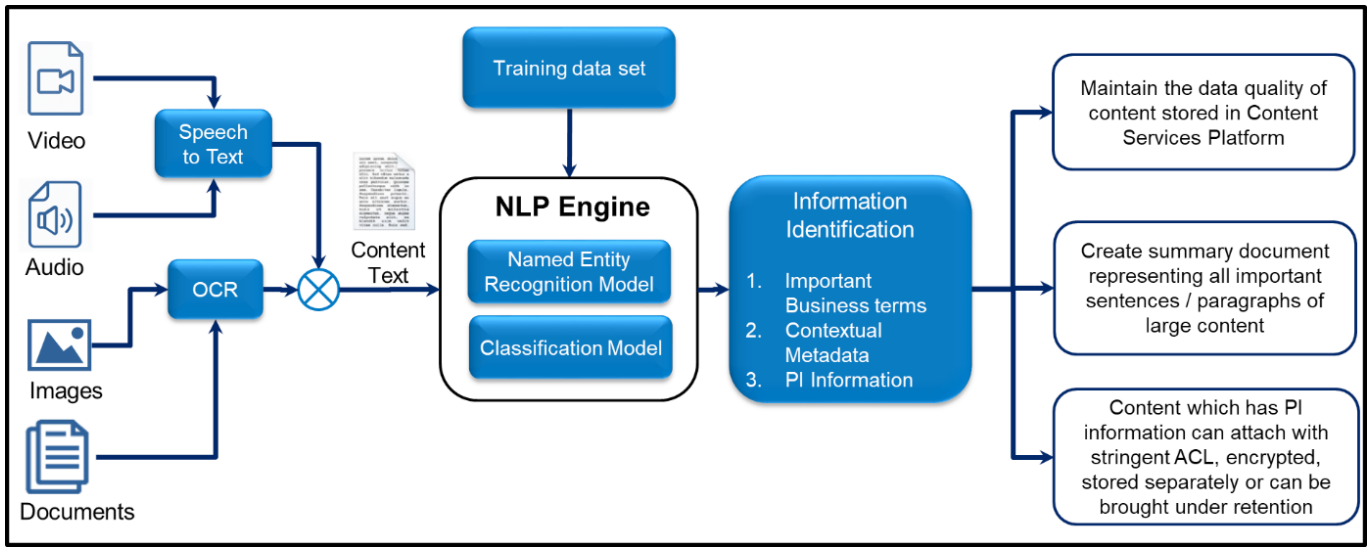
2.3 Content Management

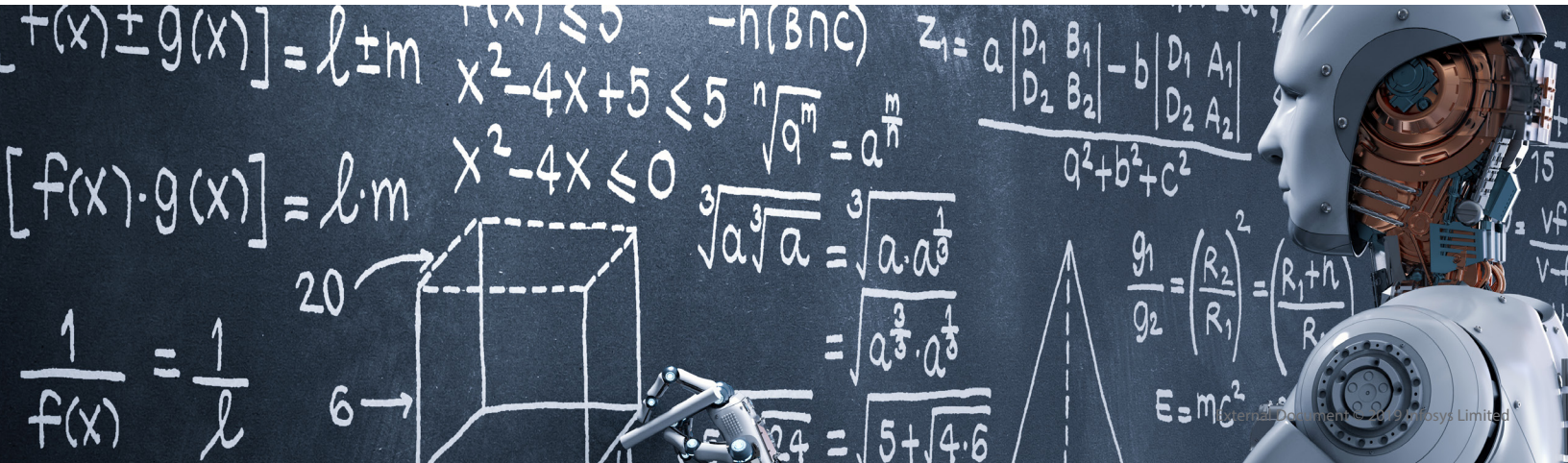
Unstructured content stored on Content Services Platform needs to be managed effectively and efficiently, in order to locate and reuse the content, to enable the use of the content in business processes, etc. This can be achieved by maintaining the proper identity of content. In Content Services Platform, identity of a document is determined by the metadata values the content is tagged with.

Due to various compliance regulation, such as GDPR, identifying Personally Identifiable (PI) Information from the content is becoming a mandatory business function. NLP based trained AI models can, identify PI information from the content.

Going through huge content to understand the context the content is a daunting task. It would be great for users, if a summary of the content is available.

Following diagram illustrate three implementations which shows how NLP can add value while managing content.





2.3.1 Implementation scenarios

Described below are implementation scenarios in the context of cognitive content management.

2.3.1.1 Maintaining Metadata Quality

In Content Service Platform, metadata values represent the identity of the content and using the metadata values, the content will be located, accessed and reused. If a content is tagged with inaccurate values or some values are missing, then either inaccurate content will be referred and reused, or correct content will not be located. Hence maintaining the quality of metadata (or identity) of content is extremely important.

NLP based trained AI models can help us in this. It cannot only make sure that the content is tagged with corrected metadata, but also tag the content with new metadata values or missing values. Implementation steps, for maintaining the metadata quality, is similar to that of metadata identification but the post processing is different. Process of maintaining metadata quality can run in background, as batch process and make sure the metadata values are accurate.

2.3.1.2 Intelligent Summarization

Content stored in Content Services Platform can be large. It will be time consuming for users to go through the entire content and understand the content completely. Legal contracts, agreements etc. falls in these category of content. The challenge of large content can be addressed by NLP. Using NLP, an AI model can be trained which identifies important business terms in the content and then create the summary of the content using the sentences or paragraph around these important business terms. The summarized document will be the new document and can be linked to the original content. Once summary is created, users can just go through the summary document to understand the complete content, thereby saving lot of time. Intelligent summarization process is suited for batch process as against to real time processing.

2.3.1.3 PII Identification

Personally Identifiable Information belongs to an individual and there are regulations, such as GDPR, which mandates the protection of PI information. If PII is not protected it might lead to data breach, loss of customer trust, and incur high penalty and legal issues. Identifying PI information in unstructured content is a difficult and challenging task.

An NLP model can be trained which can identify PI information in a content. Already, trained models are available, which helps in identifying standard PI information such as customer name, zip code etc. These AI models can be contextualized to identify PI information from a content.

Once content with PI information is identified, it can be encrypted or the content with PI information can be attached stringent ACL or the content can be brought under retention or all of these can be done to ensure the content with PI information is properly protected.

2.3.2 More Implementation Scenarios

There can be additional NLP implementation scenarios. Analyzing the risk in contract or agreements is one such scenario. NLP can be trained to identify risky terms or phrases and based on the occurrences of risky terms, the contract or agreement can be classified as risky.

Classifying a content in a particular taxonomy is also one such implementation. Each level of taxonomy can be identified with specific terms or contextual phrases. An NLP model can be trained which will classify new or unseen content to a taxonomy term.

Similarly, relevancy of a search result can be improved by query enrichment. A search query before executing against a search engine, can be analyzed. Entities or context from the query can be identified and then the query can be enriched with additional selective criteria to get relevant search results.

There can be many other NLP implementations in Content Services Platforms. Only few are discussed here.

3 High Level Implementation Steps

This section discusses a high level process for implementing scenarios which are discussed above.

3.1 Getting Labeled Content

To train AI model, labeled content is required. Most of the time, labeled content is collected from production environment and because of this, analyzing the content and building AI model from the content needs to be done in secured way. Giving access only to few developers, masking the content (if possible), are some of the ways to secure the labeled content.

3.2 Processing training data

Labeled content needs to be processed before it can be used for training. What processing is required, depends on which AI model will be trained.

For training NER model, an annotation tool is required. There are annotation tools such as Brat, tagtog, prodigy etc., using which the entities in the labeled content, can be annotated and relationship between the different entities can be established. Depending on the file format supported by the tool, the labeled content needs to be converted into the different format. For example, an audio file needs to be converted into the text file using "speech-to-text" plug-in, and then it can be used for annotation by the annotation tool. During the annotation, entities are identified and tagged with an entity name. For example, an employee name can be identified and tagged to entity name "Person". For accurate results, relationship between the entities can be established. For example, there are two entities, employee name as "person" and a bank name as "organization" in one sentence. One of the possible relationships between these entities, can be "works at". Relationship helps in extracting the information accurately. More the annotation of content, more will the training and hence, more will be accuracy.

For training categorizer model, business terms, phrases or sentences which helps in classifying the content in that category are identified. Terms identified for different categories are listed in plain text file. This file is then used for training the categorizer model.

If AI model need to be trained for more than one language, then it is better to train language specific AI models.

3.3 Building AI models

For training NER model, once annotation of labeled content is done, the files need to be converted into a machine readable format, which NLP engine understand. Most of the annotation tools offers APIs to convert the annotation format, into the machine readable format. Once converted, these files are then fed to NLP engine, which in turn, internally build an AI model.

For training categorizer model, the text file, listing category mapped to business terms, phrases or sentences, is fed to NLP engine. The NLP engine then internally build an AI model.

3.4 Testing AI Model

From the labeled content, a sub-set can be used to test the accuracy of the trained AI model. The content whose classification or entities to be extracted, is already known and is subjected to the AI model. If the AI model is able to correctly identify the category or extract the information, the AI model is trained with sufficient accuracy and hence can be deployed for production use. If the AI model does not perform adequately in this test run, then the AI model need to be re-trained with the additional labeled content.

3.5 Post Processor

NLP engine, with trained AL model, can classify or extract the entities from new or unseen content. To process the classification or extraction result of the NLP, a post processer is required. For example, in case of metadata identification, a categorizer model will give the list of metadata value with their probabilities. Post processor should select the metadata value with highest probability. Processing logic of Post processor depends on the implementation scenario.

3.6 Execution

Once the AI model is trained and tested and post processor is developed, the pipeline is ready to classify or extract the information from new or unseen documents. The new content can be subjected to NER model or categorizer model or both depending upon the use case.

Conclusion

With above discussion, it is evident that NLP can play very crucial role in addressing some of the key challenges encountered in Content Services Platform. There can be many other challenges, not discussed here, whichcan be addressed by NLP. Using the capabilities offered by NLP, Content Services can help users not only save their time but also, accurately locating and reusing the content.

To know more about this paper, please reach out to digital@infosys.com

Authors:

A senior technology architect within the Digital practice of financial unit at Infosys, Girish Pande, has around 19 years of experience in Information Technology. He has played key role in architecting and implementing end to end Content Services, Enterprise Search, NLP and Automation solutions for various clients across the globe. He is an M.Tech. in Industrial Engineering and Operations Research from IIT Bombay.

He can be reached at girish_pande@Infosys.com

A Technology Architect with the Digital Practice unit at Infosys, Yamuna Sri Kannaian, has around 13 years of experience in Information Technology. She has played various key roles and has wide experience in architecting, designing and implementing Web applications, Content Services and NLP solutions. She holds a B. Tech. degree on Information Technology from Anna University.

She can be reached at Yamuna_kannaian@infosys.com

For more information, contact askus@infosys.com

Infosys®
Navigate your next