**Visual Workbench :**

Overview of problem statement:

Language Model must understand the financial keywords related to the User stories and extract them.

For example in below user story model must extract the keywords like earned profit, unearned profit, claim types.

“show me expense report for last 3 years. Also show me my earned profit, unearned profit across claim type.”

To get the required predictions we are using Spark nlp, BERT and Glove

Spark NLP, BERT Overview :

1. Spark NLP is an open-source natural language processing library, built on top of Apache Spark and Spark ML.
2. It provides an easy API to integrate with ML Pipelines and it is commercially supported by John Snow Labs.

BERT – (Bidirectional Encoder Representations from Transformers) & Glove Models :

* pre-trained on a lot of data.
* accounts for a word’s context.
* open-source.

Motivation :

Spark NLP library provides simple, performant as well as accurate NLP notations for machine learning pipelines which can scale easily in a distributed environment. Being used in enterprise projects, built natively on Apache Spark and TensorFlow and offering an all-in-one state of the art NLP solutions

Requirements and Constraints:

Spark NLP is built on top of Apache Spark 3.x. For using Spark NLP you need:

* Java 8 and 11
* Apache Spark 3.3.x, 3.2.x, 3.1.x, 3.0.x
* Python 3.9

Complete SetUp file at below link :

[SetUp.docx](https://hexawareonline.sharepoint.com/:w:/r/sites/MLVisualWorkbench/Shared%20Documents/General/Visual%20Workbench%20Master%20Backup/Backend/Features/NLP%20by%20Abhilash/SetUp.docx?d=w174d3a7bfbac440ea6d24148bafc04df&csf=1&web=1&e=BwHhdd)

Methodology :

1. Detailed approach:

Model must understand the 4 Segments of User stories. Below are the Segments.

Category, Goal, Measure and Timeline

2.Data :

User stories are provided by the client or need to use Synthetic data.

Ex : show me expense report for last 3 years. Also show me my earned profit, unearned profit across claim type.

Measure – earned profit , unearned profit

Category – claim type

Time line – years

Data Flow :

CoNLL dataset

JsonL file of Doccano

Doccano

Tool

User Stories

Annotation

3. Techniques:

NLP – Spark-NLP

4. Model training :

Steps to Follow :

1.Collected User stories Should be annotated and convert it into CoNLL data format, So that each word is recognized by the Model.

* To Annotate we use DOCCANO tool and DOCCANO transformer.

http://172.25.125.45:8001/

<https://github.com/doccano/doccano>

<https://github.com/doccano/doccano-transformer>

* We use the below script to convert exported DOCCANO JsonL file to CoNLL format

Scripts/vwb\_convertJsonlToCoNLL

o/p :- test.dataset

2.Creating the Model – Glove100 and BERT

* CoNLL Dataset to be used

1. Dataset/vwb\_converted\_dataset
2. Dataset/vwb\_test\_dataset\_file

* Below are the scripts to be used

1. Scripts/VisualWorkbench\_GloveModel
2. Scripts/VisualWorkbench\_nlp\_NER\_BERT\_WithLabelledData

* Model Name : Glove :- Models/Vwb\_NER\_glove\_e5\_b32

BERT :- Models/Vwb\_NER\_BERT\_e5\_b32\_28112022

3. Running Model :

* Running the model to get the correct predictions. We can use any model either Glove or BERT
* Below are the script to be used

Scripts/vwb\_run\_model

5.Implementation and Deployment:

1. Creating space in LINUX server to deploy the model. So that everyone can access it.

2. Below are the files to be run in the server.

* App/app
* App/App\_model

3. Below is the UI to access the current model

* http://172.25.125.45:8080/#/NLPPredictions