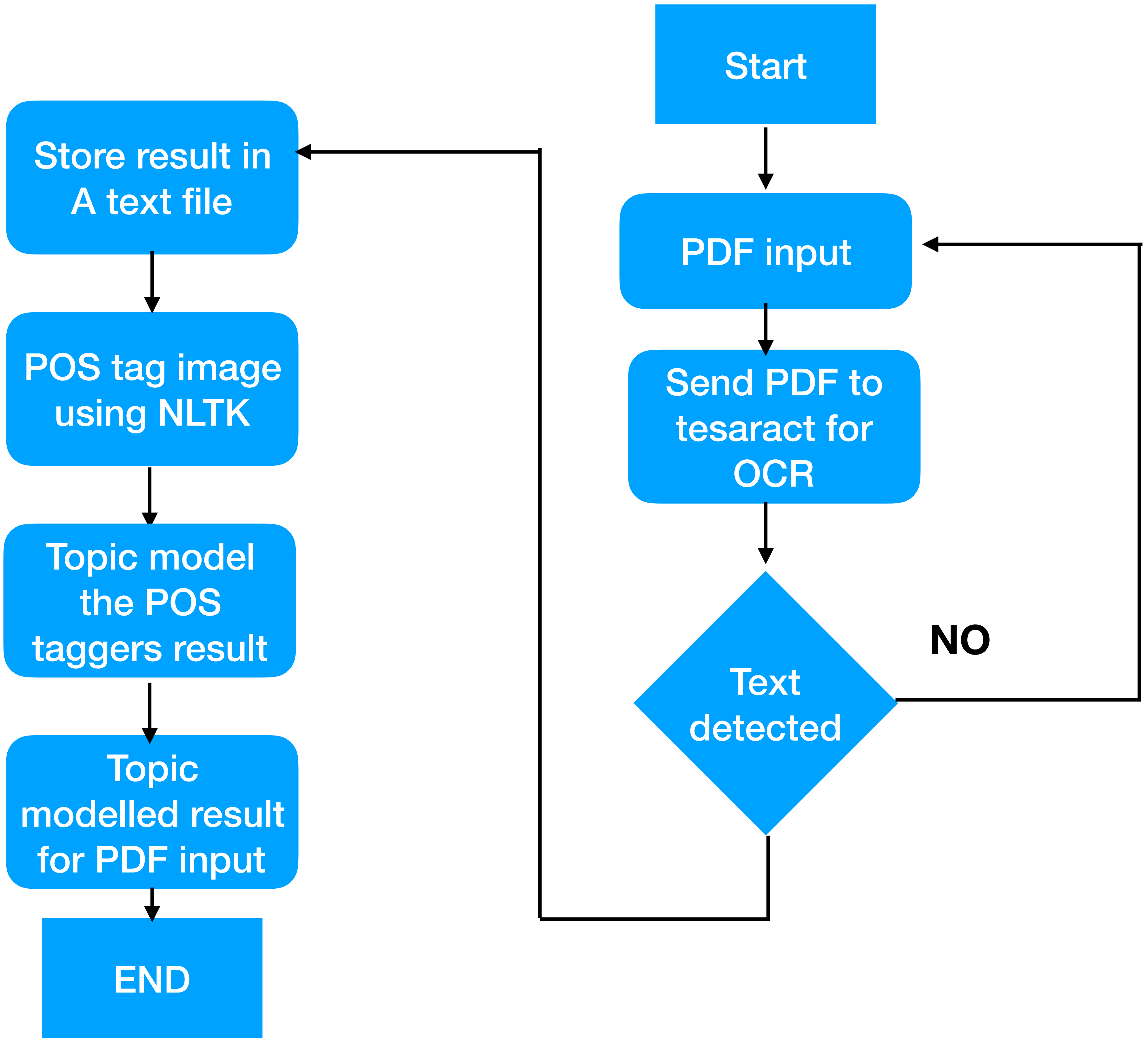
Detecting Text Based Image with OCR


Start
PDF input
Send PDF to tesaract for OCR
Text detected
NO
Store result in A text file
POS tag image using NLTK
Topic model the POS taggers result
Topic modelled result for PDF input
END

Teseract OCR

- Teseract is used to train a model to recognise PDF input and convert the PDF content into textual data

- Tesseract is fast and easy to train any type of input (i.e.) text, plain text ,pdf or image input

- Tesseract can be trained to recognise over a 100 languages and various fonts as well as hand written text , so it can be easily scaled to other applications

POS tagging

- POS (PART OF SPEECH) tagging is done using NLTK where the output spit out form tesseract is fed as input

- The pos tagger constructs a parse tree and the relevant and irrelevant elements can be ignored this makes the topic modelling easier

Topic Modelling

- Topic modelling is done by LDA , a model is trained and the POS tagged input is given, the topic is found and given as the final output.