Objective: to create game xml file from PDF using OCR (all sports -

Identify pages to extract

Assess usability /strategy per game

Extract pages

AWS Textract and Jaisys correction

Business analyst finalize

**Stage 0: Extract relevant pages and create raw text files**

0.1 Identify and extract play by play, roster and schedule pages and create a 'new pdf' for each game, roster and schedule.

 Identify degraded documents   JAISYS

Step 1: Jaisys qualifies pdf for stat extraction, runs pdf-text program and corrects

Action: classification of qualified pdfs needed (based on availability/clarity of POSS-(DOWN-TYG)-SPOT information

Step 2: Loop 1 program run and output to Excel spreadsheet

~~Step 3: loop 1 output is validated (alerts cleaned up).  Sentence type, Parent and Child are identified.~~

Step 4: Prep for Loop 2 design, build and test

Input 1: Parent-child map to stat extraction template (Parent linked stats, Parent linked flags, Child specific stats, child specific flags?)

Input 2: Validation cases identified and logic documented

Input 3: Rosters with POS (POS needed for player validation)

Step 5: Loop 2 Stage 1 tested

Step 6: Loop 2 Stage 2 (generation of xml file) tested

Additional notes

Step 1: identify page numbers and content.  Example page 9-13,25-28,38-41 from Purdue 1992 pdf for game files (games play by play pages, rosters and season schedule)

Step 2: prep pages for AWS Textract (example remove border and contrast (Frank's preprocessing))

0.2 Create raw text files from 'new pdf' files using APIs developed by Taheer

0.3 Validate usability.  Jaisys

0.4 Usable document correction. Jaisys

0.5 List correction requirements (Frank)

0.6 List game header details as well as text conventions

Discussion - training for Jaisys/-- Finalize text output requirements

List common issues and fixes

I find that if we draw lines to create a table for a roster (using available pdf editors), Textract works much better at outputting a table

Also, when loading the csv in to Excel, the weight col (and most other columns) should be labeled as text (not as general)  otherwise the weight is converted to a date by Excel

**Stage 1: Sentence parsing**Sentence types:

NON PLAY:  need a 'dictionary' approach (analysis) but includes

game header,

TOSS,

End of Quarter and score,

SCORE,

Scoring play summary,

Play by play format descriptor,

quarter start and end,

drive start/possession,

Overtime