Thu 4th.Oct.2017

AI- Logo Detection and PDF Handling

Ninestars is the leading global digitization partner to the Media Intelligence industry.

Speaker:
Tony Prime

Company:
Ninestars Technologies

@NinestarsGlobal

49TH FIBEP WORLD MEDIA INTELLIGENCE CONGRESS
BERLIN_____4-6 October 2017

@_FIBEP
#FIBEP
#WMIC17

PDF handling credentials

- Clipping / Earned
- Ad & Brand monitoring / Paid
- Cloud-based delivery platform which Cloud Hybrid

34,000 Title

30 clients

1 clip per sec

SLA based TAT 30 mins from

Ninestars

FIBEP

WMIC BERLIN 2017

PDF Sources

Publisher Editorial systems
- High quality, fonts embedded, Multi layered, vectorized objects.
- No loss in quality when expanded

ePaper Systems
- Lower quality, partially embedded fonts, flattened
- Loss in quality when expanded

PDF's generated post image scans
- OCR based systems generated
- Quality based on scan quality
- Fonts not extractable.

Ninestars

FIBEP

WMIC BERLIN 2017

Challenges

- Publisher sourced text PDFs are multi layered. Need to flatten before processing.
- When fonts are not open, text extraction from the objects need workarounds and manual extraction. Alternates are to convert the PDFs to images and process them through OCR engines to extract text
- Reformating / Reflow of content. Need to retain the font structure and links within the PDF object structure

Ninestars

FIBEP

WMIC BERLIN 2017

Ninestars Clipping Platform – Text PDF Processing.

- 80+% automation and continuously improving
- Segmentation model based on PDF structure + proprietary recognition models
- Object and text layers recognized using integration with low level APIs
- Reliable text extraction using PDF SDKs
- Retain text quality and object definitions within PDF

Ninestars

FIBEP

WMIC BERLIN 2017

A.I. based Automation for Text PDFs.

Ninestars

Using AI techniques Ninestars is working on two key improvements

- Enhanced article boundary detection
- Greatly improved text extraction

49TH FIBEP

WMIC BERLIN 2017

Image based PDFs

- Text quality relies on scan quality and OCR engine recognition quality
- Difficult to segment odd shaped articles
- 80+% automation for rectangular shaped articles
- Based on OCR structure + proprietary segmentation recognition models

Ninestars

49TH FIBEP

WMIC BERLIN 2017

A.I. Automation - Image based PDFs

- Research on image based pattern recognition models
- Pattern recognition models for text areas and embedded images
- Employ Computer Vision technologies, Deep Learning and supervised training models
- Training segmentation models based on supervised machine learning
- Build enhanced OCR recognition capability for specific content layouts

Ninestars

FIBEP

WMIC BERLIN 2017

A.I. in M.I. space

Vision based A.I. (machine/deep learning)

Ninestars

FIBEP

WMIC
BERLIN
2017