Information Extraction from video webinar recordings

Team P41
Artem Filipenko
1st year MSc student

Timofei Ganjušev
1st year MSc student

Karl Erik Karindi
1st year MSc student
The client

Mission is to transform huge content blocks (e.g. video, audio) into neat packages that are easy to search, access and share.
The problem

• Massive growth in educational video content
• No easy way to generate keywords from videos
• Videos are not as easily discoverable and searchable
The problem

• Massive growth in educational video content
• No easy way to generate keywords from videos
• Videos are not as easily discoverable and searchable

Idea:

• Let’s generate these keywords automatically!
The problem

• Massive growth in educational video content
• No easy way to generate keywords from videos
• Videos are not as easily discoverable and searchable

Idea:
• Let’s generate these keywords automatically!
• and do it quickly (less than 5 minutes for 1 hour)
The pipeline

Part 1
Slide detection

Part 2
OCR

Part 3
Key-phrase detection
The pipeline

Part 1
Slide detection
1. Slide detection

Original Idea:

- Use pre-existing tool
- 5-6 years old and unmaintained
Slide detection

- Original results:
  - 40-minute video
    - About 2-3 minutes for slide-detection
  - Really slow
Slide detection

• Original results:
  • 40-minute video
    • About 2-3 minutes for slide-detection
  • Really slow

So why not do slide detection ourselves?
Optimized slide detection

• Improved slide detection times:
  • 20-minute video
    • 2 seconds
  • 40-minute video
    • 15 seconds
  • 60-minute video
    • 30 seconds

• Key points:
  • We don’t need to look at every frame
  • Not every change in video is a new slide
Optimized slide detection

• Improved slide detection times:
  • 20-minute video
    • 2 seconds
  • 40-minute video
    • 15 seconds
  • 60-minute video
    • 30 seconds

• Key points:
  • We don’t need to look at every frame
  • Not every change in video is a new slide
The pipeline

Part 1
Slide detection

Part 2
OCR
2. OCR

• Recognizing characters from images
• Extracting text information
  • size
  • location
  • slide number

• Wanted to use Google Vision API...
2. OCR

- Recognizing characters from images
- Extracting text information
  - size
  - location
  - slide number

- Wanted to use Google Vision API...
- ... but it’s not free
- Used PyTesseract instead
Interactive Brokers

in conjunction with

Drawing Capital

present:

Opportunities in the Innovation Economy

Sean van der Wal
Sagar Joshi

Jugal Lodaya
Managing Partners at Drawing Capital

November 4, 2020
The pipeline

Part 1
Slide detection

Part 2
OCR

Part 3
Key-phrase detection
3. Key-phrase detection

• What words or phrases best describe what a video is about?

• Tested multiple text-based keyphrase extractors

Keybert, TopicalPageRank, Multipartite, TopicRank, TextRank, Rake, PositionRank
3 models that performed best

• KeyBert
  • Used SpaCy for text preprocessing
  • Time expensive
  • Results too generic

• MultipartiteRank
  • Actually pretty good results
  • Fast

• TopicRank
  • Best performance
  • Fast
Engineering Design Webinar
20 minute video – 12 seconds
Fintech Disruption in Leasing Webinar
41 minute video – 17 seconds

['fintech journey', 'fintech', 'questions', 'technology platform', 'change management', 'equipment leasing', 'advisors', 'domain knowledge', 'evaluation', 'models', 'options', 'operations initiatives', 'chrome', 'leader global', 'leases']

Embedded AI, Machine Learning and Analytics Webinar
60 minute video – 55 seconds

['innovation', 'sap leonardo', 'design thinking', 'data', 'digital disruption', 'technology', 'empathy', 'systems', 'intelligence', 'industry', 'processes', 'next', 'machine learning', 'many companies', 'enterprise']
Lessons learned / value received

• Video processing skills
• Different techniques for OCR
• Experience with different extractor models
• Teamwork

• Fairly satisfied with the results
• Fun project to work on
Thank you for listening!
Any questions?