Python course materials

# Working with PDF Files

Welcome back Agent. Often you will have to deal with PDF files. There are [many libraries in Python for working with PDFs](https://www.binpress.com/tutorial/manipulating-pdfs-with-python/167), each with their pros and cons, the most common one being **PyPDF2**. You can install it with (note the case-sensitivity, you need to make sure your capitilization matches):

pip install PyPDF2

Keep in mind that not every PDF file can be read with this library. PDFs that are too blurry, have a special encoding, encrypted, or maybe just created with a particular program that doesn’t work well with PyPDF2 won’t be able to be read. If you find yourself in this situation, try using the libraries linked above, but keep in mind, these may also not work. The reason for this is because of the many different parameters for a PDF and how non-standard the settings can be, text could be shown as an image instead of a utf-8 encoding. There are many parameters to consider in this aspect.

As far as PyPDF2 is concerned, it can only read the text from a PDF document, it won’t be able to grab images or other media files from a PDF. \_\_\_

## Working with PyPDF2

Let’s being showing the basics of the PyPDF2 library.

# note the capitalization  
import PyPDF2

## Reading PDFs

Similar to the csv library, we open a pdf, then create a reader object for it. Notice how we use the binary method of reading , ‘rb’, instead of just ‘r’.

# Notice we read it as a binary with 'rb'  
f = open('Working\_Business\_Proposal.pdf','rb')

pdf\_reader = PyPDF2.PdfFileReader(f)

pdf\_reader.numPages

5

page\_one = pdf\_reader.getPage(0)

We can then extract the text:

page\_one\_text = page\_one.extractText()

page\_one\_text

'Business Proposal\n The Revolution is Coming\n Leverage agile frameworks to provide a robust synopsis for high level \noverviews. Iterative approaches to corporate strategy foster collaborative \nthinking to further the overall value proposition. Organically grow the \nholistic world view of disruptive innovation via workplace diversity and \nempowerment. \nBring to the table win-win survival strategies to ensure proactive \ndomination. At the end of the day, going forward, a new normal that has \nevolved from generation X is on the runway heading towards a streamlined \ncloud solution. User generated content in real-time will have multiple \ntouchpoints for offshoring. \nCapitalize on low hanging fruit to identify a ballpark value added activity to \nbeta test. Override the digital divide with additional clickthroughs from \nDevOps. Nanotechnology immersion along the information highway will \nclose the loop on focusing solely on the bottom line. Podcasting operational change management inside of workßows to \nestablish a framework. Taking seamless key performance indicators ofßine \nto maximise the long tail. Keeping your eye on the ball while performing a \ndeep dive on the start-up mentality to derive convergence on cross-\nplatform integration. \nCollaboratively administrate empowered markets via plug-and-play \nnetworks. Dynamically procrastinate B2C users after installed base \nbeneÞts. Dramatically visualize customer directed convergence without \nrevolutionary ROI. \nEfÞciently unleash cross-media information without cross-media value. \nQuickly maximize timely deliverables for real-time schemas. Dramatically \nmaintain clicks-and-mortar solutions without functional solutions. \nBUSINESS PROPOSAL\n!1'

f.close()

### Adding to PDFs

We can not write to PDFs using Python because of the differences between the single string type of Python, and the variety of fonts, placements, and other parameters that a PDF could have.

What we can do is copy pages and append pages to the end.

f = open('Working\_Business\_Proposal.pdf','rb')  
pdf\_reader = PyPDF2.PdfFileReader(f)

first\_page = pdf\_reader.getPage(0)

pdf\_writer = PyPDF2.PdfFileWriter()

pdf\_writer.addPage(first\_page)

pdf\_output = open("Some\_New\_Doc.pdf","wb")

pdf\_writer.write(pdf\_output)

f.close()

Now we have copied a page and added it to another new document!

## Simple Example

Let’s try to grab all the text from this PDF file:

f = open('Working\_Business\_Proposal.pdf','rb')  
  
# List of every page's text.  
# The index will correspond to the page number.  
pdf\_text = []  
  
pdf\_reader = PyPDF2.PdfFileReader(f)  
  
for p in range(pdf\_reader.numPages):  
   
 page = pdf\_reader.getPage(p)  
   
 pdf\_text.append(page.extractText())

pdf\_text

['Business Proposal\n The Revolution is Coming\n Leverage agile frameworks to provide a robust synopsis for high level \noverviews. Iterative approaches to corporate strategy foster collaborative \nthinking to further the overall value proposition. Organically grow the \nholistic world view of disruptive innovation via workplace diversity and \nempowerment. \nBring to the table win-win survival strategies to ensure proactive \ndomination. At the end of the day, going forward, a new normal that has \nevolved from generation X is on the runway heading towards a streamlined \ncloud solution. User generated content in real-time will have multiple \ntouchpoints for offshoring. \nCapitalize on low hanging fruit to identify a ballpark value added activity to \nbeta test. Override the digital divide with additional clickthroughs from \nDevOps. Nanotechnology immersion along the information highway will \nclose the loop on focusing solely on the bottom line. Podcasting operational change management inside of workßows to \nestablish a framework. Taking seamless key performance indicators ofßine \nto maximise the long tail. Keeping your eye on the ball while performing a \ndeep dive on the start-up mentality to derive convergence on cross-\nplatform integration. \nCollaboratively administrate empowered markets via plug-and-play \nnetworks. Dynamically procrastinate B2C users after installed base \nbeneÞts. Dramatically visualize customer directed convergence without \nrevolutionary ROI. \nEfÞciently unleash cross-media information without cross-media value. \nQuickly maximize timely deliverables for real-time schemas. Dramatically \nmaintain clicks-and-mortar solutions without functional solutions. \nBUSINESS PROPOSAL\n!1',  
 'Completely synergize resource taxing relationships via premier niche \nmarkets. Professionally cultivate one-to-one customer service with robust \nideas. Dynamically innovate resource-leveling customer service for state of \nthe art customer service. \nObjectively innovate empowered manufactured products whereas parallel \nplatforms. Holisticly predominate extensible testing procedures for reliable \nsupply chains. Dramatically engage top-line web services vis-a-vis \ncutting-edge deliverables. Proactively envisioned multimedia based expertise and cross-media \ngrowth strategies. Seamlessly visualize quality intellectual capital without \nsuperior collaboration and idea-sharing. Holistically pontiÞcate installed \nbase portals after maintainable products. \nPhosßuorescently engage worldwide methodologies with web-enabled \ntechnology. Interactively coordinate proactive e-commerce via process-\ncentric "outside the box" thinking. Completely pursue scalable customer \nservice through sustainable potentialities. \nCollaboratively administrate turnkey channels whereas virtual e-tailers. \nObjectively seize scalable metrics whereas proactive e-services. \nSeamlessly empower fully researched growth strategies and interoperable \ninternal or "organic" sources. \nCredibly innovate granular internal or "organic" sources whereas high \nstandards in web-readiness. Energistically scale future-proof core \ncompetencies vis-a-vis impactful experiences. Dramatically synthesize \nintegrated schemas with optimal networks. Interactively procrastinate high-payoff content without backward-\ncompatible data. Quickly cultivate optimal processes and tactical \narchitectures. Completely iterate covalent strategic theme areas via \naccurate e-markets. Globally incubate standards compliant channels before scalable beneÞts. \nQuickly disseminate superior deliverables whereas web-enabled \nBUSINESS PROPOSAL\n!2',  
 'applications. Quickly drive clicks-and-mortar catalysts for change before \nvertical architectures. \nCredibly reintermediate backend ideas for cross-platform models. \nContinually reintermediate integrated processes through technically sound \nintellectual capital. Holistically foster superior methodologies without \nmarket-driven best practices. Distinctively exploit optimal alignments for intuitive bandwidth. Quickly \ncoordinate e-business applications through revolutionary catalysts for \nchange. Seamlessly underwhelm optimal testing procedures whereas \nbricks-and-clicks processes. \nSynergistically evolve 2.0 technologies rather than just in time initiatives. \nQuickly deploy strategic networks with compelling e-business. Credibly \npontiÞcate highly efÞcient manufactured products and enabled data. \nDynamically target high-payoff intellectual capital for customized \ntechnologies. Objectively integrate emerging core competencies before \nprocess-centric communities. Dramatically evisculate holistic innovation \nrather than client-centric data. Progressively maintain extensive infomediaries via extensible niches. \nDramatically disseminate standardized metrics after resource-leveling \nprocesses. Objectively pursue diverse catalysts for change for \ninteroperable meta-services. \nProactively fabricate one-to-one materials via effective e-business. \nCompletely synergize scalable e-commerce rather than high standards in \ne-services. Assertively iterate resource maximizing products after leading-\nedge intellectual capital. Distinctively re-engineer revolutionary meta-services and premium \narchitectures. Intrinsically incubate intuitive opportunities and real-time \npotentialities. Appropriately communicate one-to-one technology after \nplug-and-play networks. Quickly aggregate B2B users and worldwide potentialities. Progressively \nplagiarize resource-leveling e-commerce through resource-leveling core \nBUSINESS PROPOSAL\n!3',  
 'competencies. Dramatically mesh low-risk high-yield alignments before \ntransparent e-tailers. \nAppropriately empower dynamic leadership skills after business portals. \nGlobally myocardinate interactive supply chains with distinctive quality \nvectors. Globally revolutionize global sources through interoperable \nservices. Enthusiastically mesh long-term high-impact infrastructures vis-a-vis \nefÞcient customer service. Professionally fashion wireless leadership rather \nthan prospective experiences. Energistically myocardinate clicks-and-\nmortar testing procedures whereas next-generation manufactured \nproducts. \nDynamically reinvent market-driven opportunities and ubiquitous \ninterfaces. Energistically fabricate an expanded array of niche markets \nthrough robust products. Appropriately implement visionary e-services vis-\na-vis strategic web-readiness. \nCompellingly embrace empowered e-business after user friendly \nintellectual capital. Interactively actualize front-end processes with \neffective convergence. Synergistically deliver performance based \nmethods of empowerment whereas distributed expertise. \nEfÞciently enable enabled sources and cost effective products. \nCompletely synthesize principle-centered information after ethical \ncommunities. EfÞciently innovate open-source infrastructures via \ninexpensive materials. Objectively integrate enterprise-wide strategic theme areas with \nfunctionalized infrastructures. Interactively productize premium \ntechnologies whereas interdependent quality vectors. Rapaciously utilize \nenterprise experiences via 24/7 markets. Uniquely matrix economically sound value through cooperative \ntechnology. Competently parallel task fully researched data and enterprise \nprocess improvements. Collaboratively expedite quality manufactured \nproducts via client-focused results. \nBUSINESS PROPOSAL\n!4',  
 'Quickly communicate enabled technology and turnkey leadership skills. \nUniquely enable accurate supply chains rather than frictionless \ntechnology. Globally network focused materials vis-a-vis cost effective \nmanufactured products. \nBUSINESS PROPOSAL\n!5']

print(pdf\_text[3])

competencies. Dramatically mesh low-risk high-yield alignments before   
transparent e-tailers.   
Appropriately empower dynamic leadership skills after business portals.   
Globally myocardinate interactive supply chains with distinctive quality   
vectors. Globally revolutionize global sources through interoperable   
services. Enthusiastically mesh long-term high-impact infrastructures vis-a-vis   
efÞcient customer service. Professionally fashion wireless leadership rather   
than prospective experiences. Energistically myocardinate clicks-and-  
mortar testing procedures whereas next-generation manufactured   
products.   
Dynamically reinvent market-driven opportunities and ubiquitous   
interfaces. Energistically fabricate an expanded array of niche markets   
through robust products. Appropriately implement visionary e-services vis-  
a-vis strategic web-readiness.   
Compellingly embrace empowered e-business after user friendly   
intellectual capital. Interactively actualize front-end processes with   
effective convergence. Synergistically deliver performance based   
methods of empowerment whereas distributed expertise.   
EfÞciently enable enabled sources and cost effective products.   
Completely synthesize principle-centered information after ethical   
communities. EfÞciently innovate open-source infrastructures via   
inexpensive materials. Objectively integrate enterprise-wide strategic theme areas with   
functionalized infrastructures. Interactively productize premium   
technologies whereas interdependent quality vectors. Rapaciously utilize   
enterprise experiences via 24/7 markets. Uniquely matrix economically sound value through cooperative   
technology. Competently parallel task fully researched data and enterprise   
process improvements. Collaboratively expedite quality manufactured   
products via client-focused results.   
BUSINESS PROPOSAL  
!4

Excellent work! That is all for PyPDF2 for now, remember that this won’t work with every PDF file and is limited in its scope to only text of PDFs.