کار با فایل‌های PDF

بازگشت خوش آمدید، عامل. اغلب شما باید با فایل‌های PDF سروکار داشته باشید. در پایتون کتابخانه‌های بسیاری برای کار با PDF (https://www.binpress.com/tutorial/manipulating-pdfs-with-python/167) وجود دارد، هر کدام با مزایا و معایب خود، اما معروفترین آنها **PyPDF2** است. شما می‌توانید آن را با استفاده از دستور زیر نصب کنید (توجه کنید که حساس به حروف بزرگ و کوچک است، بنابراین باید مطمئن شوید که بزرگی و کوچکی حروف آن با تایپ شما همخوانی داشته باشد):

```
pip install PyPDF2
```

در نظر داشته باشید که هر فایل PDF نمی‌تواند با این کتابخانه خوانده شود. PDFهایی که بسیار مبهم هستند، دارای رمزگذاری ویژه‌ای هستند یا شاید فقط با یک برنامه خاص ساخته شده باشند که با PyPDF2 کار خوبی انجام نمی‌دهد، نمی‌توانند خوانده شوند. اگر خودتان در این موقعیت باشید، تلاش کنید از کتابخانه‌های موردی که در بالا لینک داده شده استفاده کنید، اما به خاطر داشته باشید که این نیز ممکن است کار نکند. دلیل این موضوع این است که به دلیل وجود بسیاری از پارامترهای مختلف برای PDF و عدم استاندارد بودن تنظیمات، ممکن است متن به جای رمزگذاری utf-8 به عنوان تصویر نمایش داده شود. در این زمینه باید به بسیاری از پارامترها توجه کرد.

در مورد PyPDF2، تنها قادر به خواندن متن از یک سند PDF است و قادر به دریافت تصاویر یا فایل‌های رسانه‌ای دیگر از یک PDF نخواهد بود.

کار با PyPDF2

بیایید با مبانی کتابخانه PyPDF2 آشنا شویم.

In [ ]:

```
!pip install PyPDF2
```

In [1]:

```
# note the capitalization
import PyPDF2
```

خواندن فایل‌های PDF

مشابه کتابخانه csv، ما یک فایل pdf را باز می‌کنیم و سپس یک شیء خواننده برای آن ایجاد می‌کنیم. توجه کنید که ما از روش دودویی خواندن، 'rb'، به جای فقط 'r' استفاده می‌کنیم.

In [2]:

```
# Notice we read it as a binary with 'rb'
f = open('Working_Business_Proposal.pdf','rb')
```

In [3]:

```
pdf_reader = PyPDF2.PdfReader(f)
```

In [4]:

```
len(pdf_reader.pages)
```

Out[4]:

```
5
```

In [5]:

```
page_number = 0
page_one = pdf_reader.pages[0]
```

In [ ]:

```
page_one
```

سپس متن را می‌توانیم استخراج کنیم:

In [7]:

```
page_one_text = page_one.extract_text()
```

In [8]:

```
page_one_text
```

Out[8]:

```
'Business Proposal The Revolution is Coming Leverage agile frameworks to p
rovide a robust synopsis for high level overviews. Iterative approaches to
corporate strategy foster collaborative thinking to further the overall va
lue proposition. Organically grow the holistic world view of disruptive in
novation via workplace diversity and empowerment. Bring to the table win-w
in survival strategies to ensure proactive domination. At the end of the d
ay, going forward, a new normal that has evolved from generation X is on t
he runway heading towards a streamlined cloud solution. User generated con
tent in real-time will have multiple touchpoints for offshoring. Capitaliz
e on low hanging fruit to identify a ballpark value added activity to beta
test. Override the digital divide with additional clickthroughs from DevOp
s. Nanotechnology immersion along the information highway will close the l
oop on focusing solely on the bottom line. Podcasting operational change m
anagement inside of workflows to establish a framework. Taking seamless key
performance indicators offline to maximise the long tail. Keeping your eye
on the ball while performing a deep dive on the start-up mentality to deri
ve convergence on cross-platform integration. Collaboratively administrate
empowered markets via plug-and-play networks. Dynamically procrastinate B2
C users after installed base benefits. Dramatically visualize customer dire
cted convergence without revolutionary ROI. Efficiently unleash cross-media
information without cross-media value. Quickly maximize timely deliverable
s for real-time schemas. Dramatically maintain clicks-and-mortar solutions
without functional solutions. BUSINESS PROPOSAL!1'
```

In [9]:

```
f.close()
```

اضافه کردن به فایل‌های PDF

ما نمی‌توانیم از طریق پایتون به فایل‌های PDF بنویسیم به دلیل تفاوت‌های میان نوع رشته یکپارچه پایتون و انواع قلم‌ها، قرارگیری‌ها و سایر پارامترهایی که یک فایل PDF می‌تواند داشته باشد.

آنچه که می‌توانیم انجام دهیم، کپی کردن صفحات و الحاق صفحات به انتها است.

In [11]:

```
f = open('Working_Business_Proposal.pdf','rb')
pdf_reader = PyPDF2.PdfReader(f)
```

In [12]:

```
page_number = 0
page_one = pdf_reader.pages[0]
```

In [13]:

```
pdf_writer = PyPDF2.PdfWriter()
```

In [14]:

```
pdf_writer.add_page(page_one);
```

In [15]:

```
pdf_output = open("Some_New_Doc.pdf","wb")
```

In [16]:

```
pdf_writer.write(pdf_output)
```

Out[16]:

```
(False, <_io.BufferedWriter name='Some_New_Doc.pdf'>)
```

In [17]:

```
f.close()
```

حالا ما یک صفحه را کپی کرده و آن را به یک سند جدید اضافه کرده‌ایم!

مثال ساده

In [18]:

```
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.PdfReader(f)

for p in range(len(pdf_reader.pages)):

    page = pdf_reader.pages[p]

    pdf_text.append(page.extract_text())

```

In [19]:

```
1 pdf_text
```

Out[19]:

['Business Proposal The Revolution is Coming Leverage agile frameworks to provide a robust synopsis for high level overviews. Iterative approaches to corporate strategy foster collaborative thinking to further the overall value proposition. Organically grow the holistic world view of disruptive innovation via workplace diversity and empowerment. Bring to the table win-win survival strategies to ensure proactive domination. At the end of the day, going forward, a new normal that has evolved from generation X is on the runway heading towards a streamlined cloud solution. User generated content in real-time will have multiple touchpoints for offshoring. Capitalize on low hanging fruit to identify a ballpark value added activity to beta test. Override the digital divide with additional clickthroughs from DevOps. Nanotechnology immersion along the information highway will close the loop on focusing solely on the bottom line. Podcasting operational change management inside of workflows to establish a framework. Taking seamless key performance indicators offline to maximise the long tail. Keeping your eye on the ball while performing a deep dive on the start-up mentality to derive convergence on cross-platform integration. Collaboratively administrate empowered markets via plug-and-play networks. Dynamically procrastinate B2C users after installed base benefits. Dramatically visualize customer directed convergence without revolutionary ROI. Efficiently unleash cross-media information without cross-media value. Quickly maximize timely deliverables for real-time schemas. Dramatically maintain clicks-and-mortar solutions without functional solutions. BUSINESS PROPOSAL!1',
 'Completely synergize resource taxing relationships via premier niche markets. Professionally cultivate one-to-one customer service with robust ideas. Dynamically innovate resource-leveling customer service for state of the art customer service. Objectively innovate empowered manufactured products whereas parallel platforms. Holisticly predominate extensible testing procedures for reliable supply chains. Dramatically engage top-line web services vis-a-vis cutting-edge deliverables. Proactively envisioned multimedia based expertise and cross-media growth strategies. Seamlessly visualize quality intellectual capital without superior collaboration and idea-sharing. Holistically pontificate installed base portals after maintainable products. Phosfluorescently engage worldwide methodologies with web-enabled technology. Interactively coordinate proactive e-commerce via process-centric "outside the box" thinking. Completely pursue scalable customer service through sustainable potentialities. Collaboratively administrate turnkey channels whereas virtual e-tailers. Objectively seize scalable metrics whereas proactive e-services. Seamlessly empower fully researched growth strategies and interoperable internal or "organic" sources. Credibly innovate granular internal or "organic" sources whereas high standards in web-readiness. Energistically scale future-proof core competencies vis-a-vis impactful experiences. Dramatically synthesize integrated schemas with optimal networks. Interactively procrastinate high-payoff content without backward-compatible data. Quickly cultivate optimal processes and tactical architectures. Completely iterate covalent strategic theme areas via accurate e-markets. Globally incubate standards compliant channels before scalable benefits. Quickly disseminate superior deliverables whereas web-enabled BUSINESS PROPOSAL!2',
 'applications. Quickly drive clicks-and-mortar catalysts for change before vertical architectures. Credibly reintermediate backend ideas for cross-platform models. Continually reintermediate integrated processes through technically sound intellectual capital. Holistically foster superior methodologies without market-driven best practices. Distinctively exploit optimal alignments for intuitive bandwidth. Quickly coordinate e-business applications through revolutionary catalysts for change. Seamlessly underwhelm optimal testing procedures whereas bricks-and-clicks processes. Synergistically evolve 2.0 technologies rather than just in time initiatives. Quickly deploy strategic networks with compelling e-business. Credibly pontificate highly efficient manufactured products and enabled data. Dynamically target high-payoff intellectual capital for customized technologies. Objectively

In [20]:

```
print(pdf_text[3])
```

```
integrate emerging core competencies before process-centric communities. D
ramatically evisculate holistic innovation rather than client-centric dat
a. Progressively maintain extensive infomediaries via extensible niches. D
ramatically disseminate standardized metrics after resource-leveling proce
sses. Objectively pursue diverse catalysts for change for interoperable me
ta-services. Proactively fabricate one-to-one materials via effective e-bu
siness. Completely synergize scalable e-commerce rather than high-standard
s in services. Assertively iterate resource-maximizing products after le
ading-edge intellectual capital. Distinctively re-engineer revolutionary m
etaservices and premium architectures. Intrinsicly incubate intuitive o
pportunities and real-time potentialities. Appropriately communicate one-t
o-one technology after plug-and-play networks. Quickly aggregate B2B users
and worldwide potentialities. Progressively plagiarize resource-leveling e
-commerce through resource-leveling core BUSINESS PROPOSAL!3',
 'Competently mesh low-risk high-yield alignments before tr
ansparent e-tailers. Appropriately empower dynamic leadership skills after
 business portals. Globally myocardinate interactive supply chains with dis
tinctive quality vectors. Globally revolutionize global sources through int
eroperable services. Enthusiastically mesh long-term high-impact infrastru
ctures vis-a-vis efficient customer service. Professionally fashion wirele
ss leadership rather than prospective experiences. Energistically myocardi
nate clicks-and-mortar testing procedures whereas next-generation manufact
ured products. Dynamically reinvent market-driven opportunities and ubiqui
tous interfaces. Energistically fabricate an expanded array of niche marke
ts through robust products. Appropriately implement visionary services v
is-a-vis strategic web-readiness. Competingly embrace empowered e-busines
s after user-friendly intellectual capital. Interactively actualize front-
end processes with effective convergence. Synergistically deliver performa
nce-based methods of empowerment whereas distributed expertise. Efficientl
y enable enabled sources and cost-effective products. Completely synthesize
 principle-centered information after ethical communities. Efficiently innov
ate open-source infrastructures via inexpensive materials. Objectively int
egrate enterprise-wide strategic theme areas with functionalized infrastru
ctures. Interactively productize premium technologies whereas interdepende
nt quality vectors. Rapaciously utilize enterprise experiences via 24/7 ma
rkets. Uniquely matrix economically sound value through cooperative techno
logy. Competently parallel task fully researched data and enterprise proces
s improvements. Collaboratively expedite quality manufactured products vi
a client-focused results. BUSINESS PROPOSAL!4',
 'Quickly communicate enabled technology and turnkey leadership skills. Un
iquely enable accurate supply chains rather than frictionless technology.
Globally network-focused materials vis-a-vis cost-effective manufactured p
```

کار عالی! تمام مطالب مربوط به PyPDF2 برای الان تمام شد، به خاطر داشته باشید که این روش تا هر فایل PDF کار نمی‌کند و محدود به استخراج متن از فایل‌های PDF است.