testing_pdfplumber

October 16, 2021

[2]:
```
import pdfplumber
```

[5]:
```
page_count = 0

with pdfplumber.open('test_data/SDS DS 5559 Spring 2021 Syllabus.pdf') as pdf:
    for page in pdf.pages:
        page_count += 1
        print(page.extract_text())
        print(f"\n\n\n----> PAGE {page_count} <----\n\n\n")
```

```
Big Data Systems
Overview
Instructor Name and Contact Information:
Adam Tashman
Email:apt4c@virginia.edu
Subject Area and Catalog Number:Data Science DS 5110
Year and Term:Summer 2021
Class Title:Big Data Systems
Level:Graduate
Credit Type:Graded
Class Description
Increasingly, data scientists and data engineers areworking with datasets that
exceed the memory of a
single machine. This motivates the need for a differentparadigm of computing and
a different toolset.
This course will prepare you for this use case.
The focus of the course is learning Spark, an open-source,general-purpose
computing framework that is
scalable and blazingly fast. The fundamental datatypes and concepts will be
covered (e.g., resilient
distributed datasets, DataFrames). You will learnhow to use Spark for large-
scale analytics and machine
learning, among other topics. Tools for data storageand retrieval will be
covered, including AWS andthe
Hadoop ecosystem.
A team project is a large component of the course,whereby you will conduct an
end-to-end data science
project. This simulates the workflow of a professionaldata scientist, from



----> PAGE 1 <----
```

developing a hypothesis to
communicating with stakeholders.

----> PAGE 1 <----

After completing this course, you will have developedvaluable data science skills and experience
working with big data frameworks.
Required Text
Jules Damji, Brooke Wenig, Tathagata Das, and DennyLee. 2020.Learning Spark: Lightning-Fast Big Data
Analytics.2nd edition. Sebastopol: O'Reilly Media,Inc.
Tomasz Drabas, Denny Lee. 2017.Learning PySpark:Build data-intensive applications locally and deploy
at scale using the combined powers of Python and Spark2.0. Birmingham: Packt Publishing.
Learning Outcomes
Upon successful completion of this course, you willbe able to:
1. Execute distributed computing frameworks using MapReduceand Spark
2. Demonstrate knowledge of applications for big datastorage, retrieval, processing, and modeling
using Amazon AWS, Hive, and others from the Hadoopecosystem
3. Implement PySpark for prevalent data science tasks,including data analysis and machine
learning
4. Execute an end-to-end predictive modeling projectusing a large dataset
Delivery Mode Expectations
Web-based with weekly live meetings
Required Technical Resources and Technical Components
VPN app: Cisco AnyConnect
Class Speci c Information
Class Instruction and Activities
The topics covered in this course include the following:
Map Reduce Framework
Getting started in Spark
Fundamental objects in Spark: RDDs, Key Value Pairs,DataFrames
Running on a cluster
Machine Learning with MLlib Library
o Model tuning, training, validation
o Data preprocessing
o Pipelines
o Classical problems: classification, regression, clustering,recommendation

----> PAGE 2 <----

- HDFS for distributed data storage
- Hive for querying against big data
- Amazon AWS tools for computing, storage, and retrieval
- Streaming systems including Spark Streaming
- Workflows with Tensors, including Google TensorFlow
- GraphX

Class Requirements
Prior to taking this course, you should meet the followingprerequisites:

- At least one programming course
- Regression Analysis
- Machine Learning or Data Mining

The following are strongly recommended:

- Programming in Python (since PySpark will be usedin this course)
- At least one course in Probability

----> PAGE 3 <----

Evaluation Standards and Assessments
Quizzes Quizzes will assess student knowledge and applicationof topics covered in reading assignments and modules.
Participation Student participation and feedback are required andinclude contributions to the live sessions and meaningfulposts and responses in discussions.
Programming Assignments Programming assignments will be implemented in JupyterNotebooks
and provide hands-on experience writing/modifyingSpark code, while working with various datasets.
Final Project The final project is a large component of the courseand it includes
data collection, modeling, visualization, and presentation.
Your final letter grade will be determined by thefollowing scale:
A+ 100 98.0
A 97.999 93.0
A- 92.999 90.0
B+ 89.999 87.0
B 86.999 83.0
B- 82.999 80.0
C+ 79.999 77.0
C 76.999 73.0

C- 72.999 70.0
D+ 69.999 67.0
D 66.999 63.0
D- 62.999 60.0
F 59.999 0

----> PAGE 4 <----

Class Schedule
There are two sections on Wednesdays:
Section 1: 7pm - 8pm Eastern Time
Section 2: 8:15pm - 9:15pm Eastern Time
Communication & Student Response Time
Discussion boards are set up in each module and designedto be a place where students can reach out to
peers and instructors and ask questions related tocontent and technology. Students are encouraged to
check the discussion boards daily for updates andcorrespondence. Specific queries regarding your
progress should be addressed to me via email and youwill receive a response within 24 hours.
Throughout our time together, the sooner you informme of any problem (personal or academic) that
may affect your attendance or performance, the betterthe chance we have of solving it together.
Assignments
Quizzes (30% of grade)
All quizzes are multiple choice, with full pointsawarded for a correct answer, and no points awardedfor
an incorrect answer.
Participation (10% of grade)
Student participation makes the course more interactiveand enriching. For each module, participation
points are either earned in full, or not earned. Pointscan be earned in any of the following ways:
 Contributing to the live discussion with questions,answers, or comments.
 Contributing a meaningful post in a discussion. Postingadd-ons such as "I agree" would not be
meaningful, but answering a question would be meaningful.
 Posting to an assignment forum. These forums allowfor student exchange of ideas and support.
Programming Assignments (30% of grade)
Programming assignments will include exercises indata analysis, pipeline development, and machine

learning. The outline of the exercise will be sketchedout by the instructor, and students will fill inthe
missing pieces, as well as modify and run the code.
Final Project (30% of grade)
The final group project will include forming a hypothesis,data acquisition and analysis, programming in
PySpark, writing a report, and presenting to the class.There will be an ungraded assignment in each
module to help build toward the final project.

----> PAGE 5 <----

The final project will consist of three components, each worth one-third of the final project grade:
1. Code
2. Presentation
3. Paper
Grading
Assignment Percent of Grade Due
Attendance & Participation 10 All modules
Programming Assignments 30 Modules 1-7, 9-11
Quizzes 30 Modules 1-11
Final Project 30 Module 12
Spirit of the Course
Students must attend weekly live sessions and completethe final project as a team. I encourage you to
post on the forums and exchange ideas. For the programmingassignments and quizzes, you should
submit your own work.
Electronic Submission of Assignments
All assignments must be submitted electronically throughCollab by the specified due dates and
times. It is crucial to complete all assigned work-failureto do so will likely result in failing the class.
For late assignments, 10% of the total grade willbe deducted per day, where the day means 11:59 p.m.
Eastern time cutoff. After five days late, it willbe marked as 0 points.
Technical Support
Technical Speci cations: Computer Hardware
Operating system: Microsoft Windows 8.1 (64-bit) orMac OS X 10.10 Minimum hard drive free space:
100 GB, SSD recommended Minimum processor speed: Intel4th Gen Core i5 or faster Minimum RAM: 4 GB
Technical Support Contacts

UVaCollab: collab-support@virginia.edu

----> PAGE 6 <----

UVA Policies
SDS Grading Policies
The standing of a graduate student in each courseis indicated by one of the following grades:
A+, A, A-; B+,B, B-; C+, C, C-; D+, D, D-; F. B-is the lowest satisfactory grade for graduate credit.
Attendance
Students are expected to attend all class sessions.Instructors establish attendance and
participation requirements for each of their courses.Class requirements, regardless of delivery
mode, are not waived due to a student's absence fromclass. Instructors will require students to
make up any missed coursework and may deny creditto any student whose absences are excessive.
Instructors must keep an attendance record for eachstudent enrolled in the course to document
attendance and participation in the class.
University Email Policies
Students are expected to check their of cial UVA emailaddresses on a frequent and consistent basis
to remain informed of University communications, ascertain communications may be time sensitive.
Students who fail to check their email on a regularbasis are responsible for any resulting
consequences.
Mid-Term and End-of-Class Evaluations
Students may be expected to participate in an onlinemid-term evaluation. Students are expected to
complete the online end-of-class evaluation. As thesemester comes to a close, students will
receive an email with instructions for completingthis. Student feedback will be very valuable to
the school, the instructor, and future students. Weask that all students please complete these
evaluations in a timely manner. Please be assuredthat the information you submit online will be
anonymous and kept con dential.
University of Virginia Honor System
All work should be pledged in the spirit of the HonorSystem at the University of

```
Virginia. The
instructor will indicate which assignments and activitiesare to be done
individually and which
permit collaboration. The following pledge shouldbe written out at the end of
all quizzes,
examinations, individual assignments and papers: "I pledge that I have neither
given nor received
help on this examination (quiz, assignment, etc.)."The pledge must be signed by
the student. For
more information, visitwww.virginia.edu/honor.
Special Needs
It is my goal to create a learning experience thatis as accessible as possible.
If you anticipate
any issues related to the format, materials, or requirementsof this course,
please meet with me
outside of class so we can explore potential options.Students with disabilities
may also wish to



----> PAGE 7 <----



work with the Student Disability Access Center to discuss a range of options to
removing barriers
in this course, including of cial accommodations.Please visit their website for
information on
this process and to apply for services online:
sdac.studenthealth.virginia.edu.If you have already
been approved for accommodations through SDAC, pleasesend me your accommodation
letter and
meet with me so we can develop an implementation plantogether.



----> PAGE 8 <----
```

[6]:
```
import os
```

[8]:
```
os.listdir('./test_data')
```

[8]:
```
['SDS DS 5559 Spring 2021 Syllabus.pdf',
 'DS 6014 Syllabus Spring 2021.pdf',
```

```
 'nlp_cheatsheet.pdf',
 'syllabus.pdf']
```

[10]:
```
for f in os.listdir('./test_data'):
    print('./test_data/'+f)
```

```
./test_data/SDS DS 5559 Spring 2021 Syllabus.pdf
./test_data/DS 6014 Syllabus Spring 2021.pdf
./test_data/nlp_cheatsheet.pdf
./test_data/syllabus.pdf
```

[26]:
```
p_text = []

for f in os.listdir('./test_data'):

    page_count = 0
    with pdfplumber.open('./test_data/'+f) as pdf:

        doc_text = []

        for page in pdf.pages:
            page_count += 1
            dt = page.extract_text()
            doc_text.append(dt)
            p_text.append(t)
    p_text.append(doc_text)
```

[28]:
```
p_text[0]
```

[28]:
```
'DS 6001: PRACTICE AND APPLICATION OF DATA SCIENCE  FALL SEMESTER 2021\n
Article: https://www.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108?\nnee
dAccess=true \n  John W. Tukey "Exploratory Data Analysis: Past, Present, and
Future", pages 1-7: https://\napps.dtic.mil/sti/pdfs/ADA266775.pdf \n  "Surfing
the Data Pipeline with Python", chapter 10 \n• Live session: Tuesday, November
9, 2021, 7p-8m (sec. 1, 17400) or 8:15p-9:15p (sec. 2, 17404) \n• Reading quiz
and lab assignment due date: Sunday, November 14, 2021, 11:59pm \nModule 11:
Static Visualizations \n• Readings:  \n  Textbook: Molin "Visualizing Data with
Pandas and Matplotlib", "Plotting with Seaborn and \nCustomization Techniques"
https://www.oreilly.com/library/view/hands-on-data-analysis/\n9781789615326 \n
Textbook: Wilke, chapters 2, 17, 29 https://serialmentor.com/dataviz/ \n
"Surfing the Data Pipeline with Python", chapter 11 \n• Live session: Tuesday,
November 16, 2021, 7p-8m (sec. 1, 17400) or 8:15p-9:15p (sec. 2, 17404) \n•
Reading quiz and lab assignment due date: Sunday, November 21, 2021, 11:59pm
\nModule 12: Interactive Visualizations \n• Readings:  \n  Browsing the Plotly
Gallery to see what is possible and how to code different graphs:
https://\nplotly.com/python/plotly-fundamentals/ \n  Working through the Dash
tutorial: https://dash.plotly.com/installation \n  Some thoughts on how to make
```

```
an effective UX design:
https://www.toptal.com/designers/data-\nvisualization/dashboard-design-best-
practices \n  "Surfing the Data Pipeline with Python", chapter 12 \n• Live
session: Tuesday, November 30, 2021, 7p-8m (sec. 1, 17400) or 8:15p-9:15p (sec.
2, 17404) \n• Reading quiz and lab assignment due date: Sunday, December 5,
2021, 11:59pm\n12'
```

[29]:
```
p_text[1]
```

[29]:
```
'DS 6001: PRACTICE AND APPLICATION OF DATA SCIENCE  FALL SEMESTER 2021\n
Article: https://www.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108?\nnee
dAccess=true \n  John W. Tukey "Exploratory Data Analysis: Past, Present, and
Future", pages 1-7: https://\napps.dtic.mil/sti/pdfs/ADA266775.pdf \n  "Surfing
the Data Pipeline with Python", chapter 10 \n• Live session: Tuesday, November
9, 2021, 7p-8m (sec. 1, 17400) or 8:15p-9:15p (sec. 2, 17404) \n• Reading quiz
and lab assignment due date: Sunday, November 14, 2021, 11:59pm \nModule 11:
Static Visualizations \n• Readings:  \n  Textbook: Molin "Visualizing Data with
Pandas and Matplotlib", "Plotting with Seaborn and \nCustomization Techniques"
https://www.oreilly.com/library/view/hands-on-data-analysis/\n9781789615326 \n
Textbook: Wilke, chapters 2, 17, 29 https://serialmentor.com/dataviz/ \n
"Surfing the Data Pipeline with Python", chapter 11 \n• Live session: Tuesday,
November 16, 2021, 7p-8m (sec. 1, 17400) or 8:15p-9:15p (sec. 2, 17404) \n•
Reading quiz and lab assignment due date: Sunday, November 21, 2021, 11:59pm
\nModule 12: Interactive Visualizations \n• Readings:  \n  Browsing the Plotly
Gallery to see what is possible and how to code different graphs:
https://\nplotly.com/python/plotly-fundamentals/ \n  Working through the Dash
tutorial: https://dash.plotly.com/installation \n  Some thoughts on how to make
an effective UX design:
https://www.toptal.com/designers/data-\nvisualization/dashboard-design-best-
practices \n  "Surfing the Data Pipeline with Python", chapter 12 \n• Live
session: Tuesday, November 30, 2021, 7p-8m (sec. 1, 17400) or 8:15p-9:15p (sec.
2, 17404) \n• Reading quiz and lab assignment due date: Sunday, December 5,
2021, 11:59pm\n12'
```

[34]:
```
# make dataframe with 1 column of files in test_data directory
import pandas as pd

test = pd.DataFrame({'file_name':os.listdir('./test_data')})
test
```

[34]:

	file_name
0	SDS DS 5559 Spring 2021 Syllabus.pdf
1	DS 6014 Syllabus Spring 2021.pdf
2	nlp_cheatsheet.pdf
3	syllabus.pdf

[38]:
```
test['col2'] = test['file_name'].map(lambda x: x + 'TEST')
```

[43]:
```
# write function to extract text from one pdf
def get_pdf_text(path, dat_dir='test_data'):

    page_count = 0

    with pdfplumber.open(dat_dir+'/'+path) as pdf:
        for page in pdf.pages:
            page_count += 1
            page_text = page.extract_text()
            return page_text+f'\n\n\n----> PAGE {page_count} <----\n\n\n'
```

[44]:
```
test['pdf_text'] = test['file_name'].map(lambda x: get_pdf_text(path=x))
```

[45]:
```
test
```

[45]:
```
                            file_name  \
0  SDS DS 5559 Spring 2021 Syllabus.pdf
1      DS 6014 Syllabus Spring 2021.pdf
2                   nlp_cheatsheet.pdf
3                         syllabus.pdf

                                     col2  \
0  SDS DS 5559 Spring 2021 Syllabus.pdfTEST
1      DS 6014 Syllabus Spring 2021.pdfTEST
2                   nlp_cheatsheet.pdfTEST
3                         syllabus.pdfTEST

                                            pdf_text
0  Big Data Systems\nOverview\nInstructor Name an…
1  Bayesian Machine Learning \nOverview \nInstruc…
2  © CFA Institute. For personal use only. Not fo…
3  UNIVERSITY OF VIRGINIA  FALL 2021\nDS 6001: Pr…
```

[49]:
```
test.iloc[0,2]
```

[49]:
```
'Big Data Systems\nOverview\nInstructor Name and Contact Information:\nAdam
Tashman\nEmail:apt4c@virginia.edu\nSubject Area and Catalog Number:Data Science
DS 5110\nYear and Term:Summer 2021\nClass Title:Big Data
Systems\nLevel:Graduate\nCredit Type:Graded\nClass Description\nIncreasingly,
data scientists and data engineers areworking with datasets that exceed the
memory of a\nsingle machine. This motivates the need for a differentparadigm of
computing and a different toolset.\nThis course will prepare you for this use
case.\nThe focus of the course is learning Spark, an open-source,general-purpose
computing framework that is\nscalable and blazingly fast. The fundamental
datatypes and concepts will be covered (e.g., resilient\ndistributed datasets,
DataFrames). You will learnhow to use Spark for large-scale analytics and
machine\nlearning, among other topics. Tools for data storageand retrieval will
```

```
be covered, including AWS andthe\nHadoop ecosystem.\nA team project is a large
component of the course,whereby you will conduct an end-to-end data
science\nproject. This simulates the workflow of a professionaldata scientist,
from developing a hypothesis to\ncommunicating with stakeholders.\n\n\n---->
PAGE 1 <----\n\n\n'
```

[50]:
```
test = test[['file_name', 'pdf_text']]
```

[51]:
```
test.to_csv('test_pdf_df.csv')
```

[ ]:
```

```