UNIVERSITÉ DE CARTHAGE

L'ÉCOLE NATIONALE DE CARTHAGE

Département d'Informatique

intitulé:

Projet-Mise en place d'un système de classification d'un document pdf

Elaboré par :

Oumeima Liméme

Encadré par :

M. Houcemeddine Hermassi

Année Universitaire 2023/2024

3éme Informatique

Groupe B

Question 1 : Développement d' un outil de reconnaissance de document pdf écrit en langue arabe à base de OCR

J' ai installé tout d'abord les librairies que je vais utiliser pour extracter les textes écrits en arabes à partir de documents pdf en utilisant l' outil de reconnaissance **OCR**

```
## Install Tesseract OCR and required libraries

!pip install pydotplus
!pip install pytesseract
!pip install gTTS
!pip install PyPDF2
!pip install PyMuPDF
!pip install pytesseract
!apt-get install -y poppler-utils
!apt-get install -y tesseract-ocr
!apt-get install -y  libtesseract-dev
!pip install tesseract
!pip install pytesseract
!pip install pytesseract pdf2image
!pip install scikit-learn
!pip install pymupdf
!apt-get update
```

Il existe 2 types de pdf à partir desquels j'extrais le texte :

- Pdf écrit avec ordinateur
- Pdf sous forme d' image

J' ai importer les librairies et les donneés arabes que je vais l'utiliser pour pouvoir extracter les mots arabes

```
import pydotplus
from IPython.display import display, Image
import matplotlib.pyplot as plt
import pytesseract
from gtts import gTTS
import IPython.display as ipd
from requests import get  # to make GET request
```

```
## dowenload arabic traineddata
def download(url, file_name):
    # open in binary mode
    with open(file_name, "wb") as file:
        # get request
        response = get(url)
        # write to file
        file.write(response.content)

download("https://github.com/tesseract-ocr/tessdata/raw/master/ara.traineddata","ara.traineddata")
```

1-Pdf écrit avec ordinateur

```
import PyPDF2
import pytesseract
import cv2
import matplotlib.pyplot as plt
from PIL import Image

# Replace 'your_pdf.pdf' with the actual name of your PDF file
pdf_filename = "/kaggle/input/pdf-arabic/.pdf"

# Function to extract text from PDF using PyPDF2
def extract_text_from_pdf(pdf_filename):
    with open(pdf_filename, 'rb') as file:
        pdf_reader = PyPDF2.PdfReader(file)
        text = ""
        for page_number in range(len(pdf_reader.pages)):
            page = pdf_reader.pages[page_number]
            text += page.extract_text()
    return text

# Read text from the PDF
pdf_text = extract_text_from_pdf(pdf_filename)

# Check if there is text
if pdf_text:
    print("Text extracted from PDF:")
    print(pdf_text)
else:
    print(f"No text found in the PDF: {pdf_filename}")
```

```
Text extracted from PDF:
دعم الطاقة في منطقة الشرق ألوسط وشمال إفريقيا: دروس مستفادة لإلصالح
4 | إدارة الشرق ألوسط وآسيا الوسطى
  المنافع التي تتحقق من إصالح دعمالطاقة

يمكن أن يؤدي إصالح الدعم إلى إعطاء دفعة للنمو والحد
من الفقر وانعدام المساواة. فإعادة تخصيص الموارد التي
يحررها الدعم باتجاه زيادة اإلنفاق العام اإلنتاجي يمكن أن
يساعد على إعطاء دفعة للنمو في ألجل الطويل. وعالوة
على ذلك، يمكن أن يؤدي إلغاء الدعم، إذا اقترن بشبكة
لألمان االجتماعي جيدة التصميم وزيادة في اإلنفاق لصالح
الفقراء، إلى تحسينات كبيرة في رفاه الفئات منخفضة الخل
على المدى ألبعد. ويمكن أيضا أن يسهم إصالح الدعم في
خفض عجز الميزانية وأسعار الفائدة، مما يحفز استثمارات
القطاع الخاص ويعزز النمو.

و بإزالة التشوهات في عالمات ألسعار، يمكن أن يساعد
إصالح دعم الوقود على تحسين الجوافز العتماد
تكنولوجيات موفرة للطاقة. وتشير تقديرات أعدت باستخدام
المنهج التجريبي إلى أن زيادة االستثمار في التكنولوجيات
ألكثر كفاءة والموفرة للطاقة يمكن أن تعطي دفعة للنمو
4.على المدى البعيد % 2 بنسبة تصل إلى

وأخيرا،فمن شأن إلغاء دعم الطاقة أن يحقق منافع كبيرة في
مجالي البيئة والصحة عن طريق خفض التلوث المحلي.
...
الموجهة للمستحقين المنهج المفضل للتعويض. وعندما يتعذر
القيام بتحويالت نقدية بسبب محدودية القدرات اإلدارية،
يمكن التوسع في مبادرات أخرى، مثل برامج ألشغال العامة،
```

2-Pdf sous forme d’ image

```
import urllib.request
arabic_data_url = "https://github.com/tesseract-ocr/tessdata/raw/main/ara.traineddata"
urllib.request.urlretrieve(arabic_data_url, "/kaggle/working/ara.traineddata")
import os
os.environ["TESSDATA_PREFIX"] = "/kaggle/working/"
import pytesseract
from PIL import Image
from pdf2image import convert_from_path
import urllib.request
import os
# Set the path to the Tesseract executable
pytesseract.pytesseract.tesseract_cmd = '/usr/bin/tesseract'

# Download Arabic language data
arabic_data_url = "https://github.com/tesseract-ocr/tessdata/raw/main/ara.traineddata"
urllib.request.urlretrieve(arabic_data_url, "/kaggle/working/ara.traineddata")

# Set TESSDATA_PREFIX
os.environ["TESSDATA_PREFIX"] = "/kaggle/working/"

# Replace 'your_pdf.pdf' with the actual name of your PDF file
pdf_filename = "/kaggle/input/other-dataset/economie tunisienne-histoire.pdf"

# Convert PDF pages to images
images = convert_from_path(pdf_filename, 500)  # 500 is the DPI (adjust as needed)

# Iterate through each image and perform OCR
for i, image in enumerate(images):
    # Save the image
    image_path = f"/kaggle/working/image_{i+1}.png"
    image.save(image_path, "PNG")
    # Perform OCR on the image with the 'ara' language
    arabic_text = pytesseract.image_to_string(image_path, lang='ara', config="--psm 6")
    # Print or use the extracted text as needed
    print(f"Text from Image {i + 1}:\n{arabic_text}")
```

```
Text from Image 1:
5 * مي  خصة ف تونس*« : خص ل . *
ي 5 [عدل]) تاريخ اقتصاد دودس

، بورصة تونس
التأميم (1956-1961) [عل] عدل
يُعيد الاستقلال؛ كان الهم الشاغل للحكومة التونسية يتمثل في تحرير الاقتصاد من مخلفات الاستعمار الفرنسي
والذي شجّع الفلاحة والاستخراج المنجمي مع إهمال تام للصناعة. وفي الفترة مابين سنة 6 و 1960 غادر أغلب الموظفين الفرنسيين وقُدّر
عددهم آنذاك ب 12000 الإدارة التونسية عائدين إلى فرنسا. ولتأكيد سيطرة الدولة على القطاعات الأساسية أسست الحكومة الشركة الوطنية للسكك
الحديدية سنة 1956؛ وأمّمت القطاع المصرفي وشركات الكهرباء والغاز والماء. تم أممت شركات النقل وشاركت ب 9650 في رأس مال شركة
الطيران تونيزار وأسست الشركة التونسية للملاحة. وبالتزامن مع ذلك أصبح الدينار التونسي بموجب القانون الصادر بتاريخ 18 أكتوبر 1958
العملة الرسمية للدولة التونسيّة. ولكن كل هذا لا يعكس نزعة اشتراكية بقدر ما يبين حرص الحكومة الناشئة على تعزيز سيطرتها مع اتباع سياسة
ليبرالية قائمة على تشجيع الإستثمار والتجارة الخارجية. ومن أجل ذلك منحت الحكومة امتيازات جبائية وتسهيلات في القروض في الخماسية التي تلت
الاستقلال لتحفيز مشاركة أكبر للقطاع الخاص.
```

Question 2 : Classer le document selon le contenu en se référant à des labels

```
!pip install scikit-learn
!pip install tensorflow
```

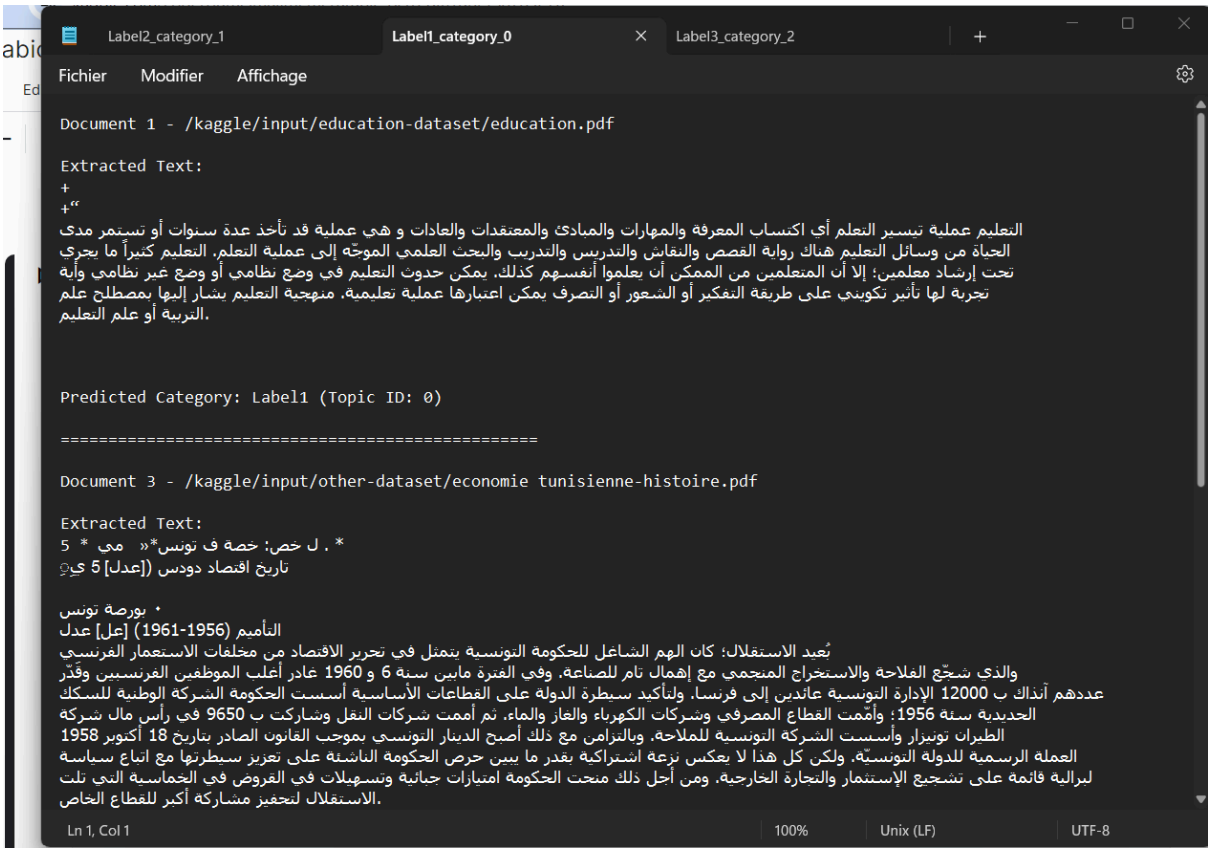
```
import os
from gensim import corpora, models
from pdf2image import convert_from_path
import pytesseract

# Function to extract text from PDF using pdf2image and pytesseract
def extract_text_from_pdf(pdf_path):
    images = convert_from_path(pdf_path, 500)  # Adjust DPI as needed
    text = [pytesseract.image_to_string(image, lang='ara', config="--psm 6") for image in images]
    return ' '.join(text)

# Function to predict category for a document
def predict_category(pdf_file, lda_model, dictionary):
    text = extract_text_from_pdf(pdf_file)
    bow_vector = dictionary.doc2bow(text.split())

    # Get the topic distribution for the document
    topic_distribution = lda_model[bow_vector]

    # Choose the topic with the highest probability as the predicted category
    predicted_category = max(topic_distribution, key=lambda x: x[1])[0]

    return predicted_category

# Function to map topic IDs to predefined labels
def map_topic_to_label(topic_id):
    # You should replace the following with your actual labels
    labels = {
        0: "Label1",
        1: "Label2",
        2: "Label3",
        3: "Label4",
        4: "Label5",
    }
    return labels.get(topic_id, "Unknown")
```

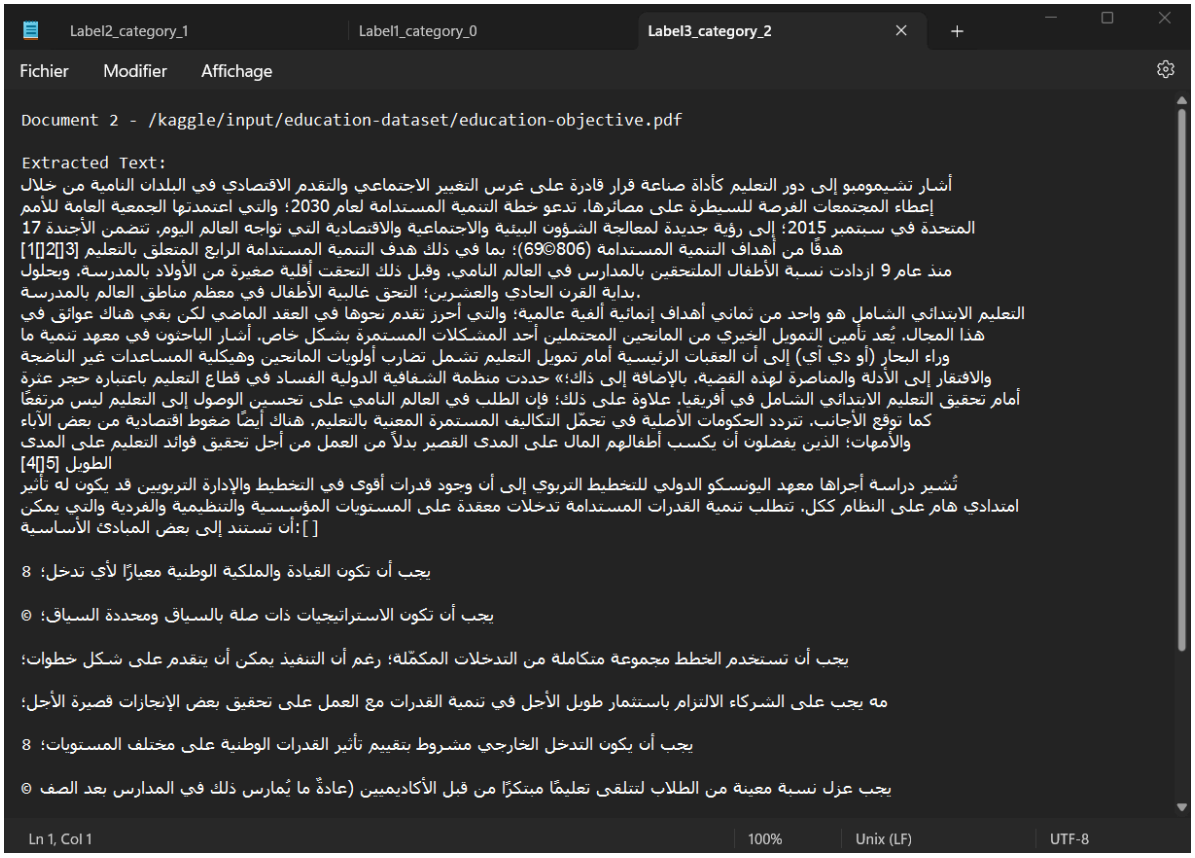
```
# Directory containing initial set of PDF files
initial_pdf_directory = "/kaggle/input/origin"
initial_pdf_files = [os.path.join(initial_pdf_directory, file) for file in os.listdir(initial_pdf_directory) if file.endswith(".pdf")]

# Directory containing new PDF files
new_pdf_directory = "/kaggle/input/autre-data"
new_pdf_files = [os.path.join(new_pdf_directory, file) for file in os.listdir(new_pdf_directory) if file.endswith(".pdf")]

# Combine all PDF files
all_pdf_files = initial_pdf_files + new_pdf_files

# Extract text from each PDF and combine into a single list
all_text = [extract_text_from_pdf(pdf_file) for pdf_file in all_pdf_files]

# Tokenize the text and remove common words
tokenized_text = [text.split() for text in all_text]

# Create a dictionary representation of the documents
dictionary = corpora.Dictionary(tokenized_text)

# Convert the tokenized documents into a bag of words representation
corpus = [dictionary.doc2bow(tokens) for tokens in tokenized_text]

# Train the LDA model
lda_model = models.LdaModel(corpus, num_topics=5, id2word=dictionary, passes=15)

# Print the topics discovered by LDA
print(lda_model.print_topics())

# Create a directory to store categorized text files
output_directory = "/kaggle/working/arabic_categories"
os.makedirs(output_directory, exist_ok=True)

# Write the predicted categories for each document to separate text files
for i, pdf_file in enumerate(all_pdf_files):
    predicted_category = predict_category(pdf_file, lda_model, dictionary)
    label = map_topic_to_label(predicted_category)
```

```python
    # Create a text file for the category and write the extracted text
    category_file_path = os.path.join(output_directory, f"{label}_category_{predicted_category}.txt")
    with open(category_file_path, 'a', encoding='utf-8') as category_file:
        category_file.write(f"Document {i + 1} - {pdf_file}\n\n")
        category_file.write(f"Extracted Text:\n{all_text[i]}\n\n")
        category_file.write(f"Predicted Category: {label} (Topic ID: {predicted_category})\n\n")
        category_file.write("=" * 50 + "\n\n")

print(f"Results saved in the '{output_directory}' directory.")
```

```
[(0, '0.020*""عملية"*0.008 + "التونسي"*0.008 + "الاقتصاد"*0.010 + "أو"*0.010 + "مع"*0.010 + "التعليم"*0.010 + "من"*0.012 + "على"*0.014 + "في
إلى"*0.006 +"'), ("غير"*0.007 + "ذلك"*0.007 + "الجسم"*0.010 + "الطعام"*0.010 + "التي"*0.010 + "الإنسان"*0.013 + "على"*0.013 + "من"*0.026' ,1
الغذائية"*0.007 + "يحتاج"*0.007 +"'), ("يجب"*0.012 + "التعليم"*0.014 + "إلى"*0.016 + "أن"*0.019 + "من"*0.021 + "على"*0.023 + "في"*0.029' ,2
والتي"*0.006 + "العالم"*0.008 + "المستدامة"*0.008 +"'), ("إلى"*0.001 + "ذلك"*0.001 + "أن"*0.001 + "على"*0.001 + "في"*0.001 + "من"*0.002' ,3
الإنسان"*0.001 + "التعليم"*0.001 + "التي"*0.001 + "مع"*0.001 +"'), (+ "من"*0.001 + "التعليم"*0.001 + "على"*0.001 + "أن"*0.001 + "في"*0.002' ,4
أو"*0.001 + "يمكن"*0.001 + "ذلك"*0.001 + "يجب"*0.001 + "إلى"*0.001"')]
```

la Résultat: **La classification est effectuée en fonction du contexte du pdf, les pdf ayant le même contenu sont classés ensemble**


Label2_category_1
Label1_category_0
Label3_category_2
Fichier
Modifier
Affichage
Document 1 - /kaggle/input/education-dataset/education.pdf
Extracted Text:
+
+“
التعليم عملية تيسير التعلم أي اكتساب المعرفة والمهارات والمبادئ والمعتقدات والعادات و هي عملية قد تأخذ عدة سنوات أو تستمر مدى الحياة من وسائل التعليم هناك رواية القصص والنقاش والتدريس والتدريب والبحث العلمي الموجّه إلى عملية التعلم. التعليم كثيراً ما يجري تحت إرشاد معلمين؛ إلا أن المتعلمين من الممكن أن يعلموا أنفسهم كذلك. يمكن حدوث التعليم في وضع نظامي أو وضع غير نظامي وأية تجربة لها تأثير تكويني على طريقة التفكير أو الشعور أو التصرف يمكن اعتبارها عملية تعليمية. منهجية التعليم يشار إليها بمصطلح علم التربية أو علم التعليم.
Predicted Category: Label1 (Topic ID: 0)
==========
Document 3 - /kaggle/input/other-dataset/economie tunisienne-histoire.pdf
Extracted Text:
5 * ل خص: خصة ف تونس*« مي * 5 .
تاريخ اقتصاد دودس ([عدل] 5 ي
• بورصة تونس
التأميم (1956-1961) [عل] عدل
بُعيد الاستقلال؛ كان الهم الشاغل للحكومة التونسية يتمثل في تحرير الاقتصاد من مخلفات الاستعمار الفرنسي
والذي شجّع الفلاحة والاستخراج المنجمي مع إهمال تام للصناعة. وفي الفترة مابين سنة 6 و 1960 عادر أغلب الموظفين الفرنسيين وقدّر
عددهم آنذاك ب 12000 الإدارة التونسية عائدين إلى فرنسا. ولتأكيد سيطرة الدولة على القطاعات الأساسية أسست الحكومة الشركة الوطنية للسكك
الحديدية سنة 1956؛ وأمّمت القطاع المصرفي وشركات الكهرباء والغاز والماء. ثم أممت شركات النقل وشاركت ب 9650 في رأس مال شركة
الطيران تونيزار وأسست الشركة التونسية للملاحة. وبالتزامن مع ذلك أصبح الدينار التونسي بموجب القانون الصادر بتاريخ 18 أكتوبر 1958
العملة الرسمية للدولة التونسيّة. ولكن كل هذا لا يعكس نزعة اشتراكية بقدر ما يبين حرص الحكومة الناشئة على تعزيز سيطرتها مع اتباع سياسة
ليبرالية قائمة على تشجيع الإستثمار والتجارة الخارجية. ومن أجل ذلك منحت الحكومة امتيازات جبائية وتسهيلات في القروض في الخماسية التي تلت
الاستقلال لتحفيز مشاركة أكبر للقطاع الخاص.
Ln 1, Col 1
100%
Unix (LF)
UTF-8


Label2_category_1
Label1_category_0
Label3_category_2
Fichier
Modifier
Affichage
Document 2 - /kaggle/input/education-dataset/education-objective.pdf
Extracted Text:
أشار تشيموميو إلى دور التعليم كأداة صناعة قرار قادرة على غرس التغيير الاجتماعي والتقدم الاقتصادي في البلدان النامية من خلال
إعطاء المجتمعات الفرصة للسيطرة على مصائرها. تدعو خطة التنمية المستدامة لعام 2030؛ والتي اعتمدتها الجمعية العامة للأمم
المتحدة في سبتمبر 2015؛ إلى رؤية جديدة لمعالجة الشؤون البيئية والاجتماعية والاقتصادية التي تواجه العالم اليوم. تتضمن الأجندة 17
هدفًا من أهداف التنمية المستدامة (69©806)؛ بما في ذلك هدف التنمية المستدامة الرابع المتعلق بالتعليم [3][2][1]
منذ عام 9 ازدادت نسبة الأطفال الملتحقين بالمدارس في العالم النامي. وقبل ذلك التحقت أقلية صغيرة من الأولاد بالمدرسة. ويحلول
بداية القرن الحادي والعشرين؛ التحق غالبية الأطفال في معظم مناطق العالم بالمدرسة.
التعليم الابتدائي الشامل هو واحد من ثماني أهداف إنمائية ألفية عالمية؛ والتي أحرز تقدم نحوها في العقد الماضي لكن بقي هناك عوائق في
هذا المجال. يُعد تأمين التمويل الخيري من المانحين المحتملين أحد المشكلات المستمرة بشكل خاص. أشار الباحثون في معهد تنمية ما
وراء البحار (أو دي آي) إلى أن العقبات الرئيسية أمام تمويل التعليم تشمل تضارب أولويات المانحين وهيكلية المساعدات غير الناضجة
والافتقار إلى الأدلة والمناصرة لهذه القضية. بالإضافة إلى ذلك؛» حددت منظمة الشفافية الدولية الفساد في قطاع التعليم باعتباره حجر عثرة
أمام تحقيق التعليم الابتدائي الشامل في أفريقيا. علاوة على ذلك؛ فإن الطلب في العالم النامي على تحسين الوصول إلى التعليم ليس مرتفعًا
كما توقع الأجانب. تتردد الحكومات الأصلية في تحمّل التكاليف المستمرة المعنية بالتعليم. هناك أيضًا ضغوط اقتصادية من بعض الآباء
والأمهات؛ الذين يفضلون أن يكسب أطفالهم المال على المدى القصير بدلاً من العمل من أجل تحقيق فوائد التعليم على المدى
الطويل [5][4]
تُشير دراسة أجراها معهد اليونسكو الدولي للتخطيط التربوي إلى أن وجود قدرات أقوى في التخطيط والإدارة التربويين قد يكون له تأثير
امتدادي هام على النظام ككل. تتطلب تنمية القدرات المستدامة تدخلات معقدة على المستويات المؤسسية والتنظيمية والفردية والتي يمكن
أن تستند إلى بعض المبادئ الأساسية:[ ]
8 يجب أن تكون القيادة والملكية الوطنية معيارًا لأي تدخل؛
© يجب أن تكون الاستراتيجيات ذات صلة بالسياق ومحددة السياق؛
يجب أن تستخدم الخطط مجموعة متكاملة من التدخلات المكمّلة؛ رغم أن التنفيذ يمكن أن يتقدم على شكل خطوات؛
مه يجب على الشركاء الالتزام باستثمار طويل الأجل في تنمية القدرات مع العمل على تحقيق بعض الإنجازات قصيرة الأجل؛
8 يجب أن يكون التدخل الخارجي مشروط بتقييم تأثير القدرات الوطنية على مختلف المستويات؛
© يجب عزل نسبة معينة من الطلاب لتتلقى تعليمًا مبتكرًا من قبل الأكاديميين (عادةً ما يُمارس ذلك في المدارس بعد الصف
Ln 1, Col 1
100%
Unix (LF)
UTF-8