Legal Tech I

Jan Černý, Tomáš Knap

crimczech

Charles University

Soubory a datové formáty

 CC BY-NC-ND 4.0

Co je potřeba ke splnění kurzu

1. Zápočet

Úkoly

- v poštovní sově se budou objevovat úkoly, na řešení každého bude čas nejméně do pondělka před další přednáškou. Úkol lze odevzdat i po stanoveném termínu, nejpozději však druhé pondělí následující po jeho zadání. Za takto opožděně odevzdaný úkol lze získat nejvýše 50 % maximálního počtu bodů. Za úkol odevzdaný později je 0
- pro zápočet musíte získat alespoň 90 bodů za úkoly, celkem lze získat 100
- za pomoc s předmětem můžete získat bonusové body (nahlašování chyb, bugů, překlepů…)

Testy

- testy se budou psát na papír
- musíte úspěšně projít všemi testy - hranice úspěšného složení testu je 70 %

2. Písemná zkouška

Ke zkoušce budete připuštěni po splnění požadavků na zápočet uvedených výše.

CC BY-NC-ND 4.0

Anonymní feedback



CC BY-NC-ND 4.0

Ukázkové soubory

```
smlouva_navrh.docx
smlouva_podepsana.pdf
prehled_smluv.xlsx
export_smluv.csv
poznamky.md
```

- Ve kterém souboru budeme navrhovat změny?
- Ze kterého snadno spočítáme celkovou hodnotu smluv?
- Co z toho dokáže program rovnou přečíst jako text?

> Stejné informace mohou mít různé podoby. Záleží, co s nimi chceme dělat.

CC BY-NC-ND 4.0

Výběr formátu

1. Budeme ho hlavně **číst**, **upravovat**, nebo **zpracovávat programem**?
2. Potřebujeme zachovat vzhled, nebo především jednotlivé údaje?
3. V čem ho otevře příjemce?
4. Co se nesmí při předání nebo převodu ztratit?

Dnes budeme porovnávat možnosti a zkoušet jejich volbu na konkrétních situacích.

CC BY-NC-ND 4.0

Soubory a formáty

CC BY-NC-ND 4.0

Přípona a formát souboru

- Soubor obsahuje uložená data a má svůj název a umístění
- **Přípona** je konec názvu, například `.pdf` nebo `.csv`
- **Formát** určuje, jak jsou data uvnitř uspořádaná
- Aplikace musí pravidlům daného formátu rozumět

Přejmenování `smlouva.pdf` na `smlouva.txt` text ze souboru nevytáhne.

CC BY-NC-ND 4.0

Textový a binární soubor

Textový soubor	Binární soubor
Obsah je uložen jako posloupnost znaků	Bajty mají význam podle konkrétního formátu
Při správném kódování ho čteme v textovém editoru	Pro smysluplné zobrazení potřebujeme vhodný nástroj
Například TXT, CSV, JSON, XML, Markdown	Například PNG, ZIP nebo DOCX

Na disku jsou v obou případech bajty. Rozdíl je v tom, jak je máme číst.

Textový soubor nemusí být obyčejný odstavec. Může obsahovat tabulku nebo značky.

CC BY-NC-ND 4.0

Text v různých formátech

Soubor	Co uvidíme ve vhodné aplikaci	Co čekat v textovém editoru
TXT	Písmena a řádky	Stejný text
Markdown	Nadpisy a odrážky	Text s jednoduchými značkami
DOCX	Formátovaný dokument	Bez rozbalení ne souvislý text dokumentu
PDF	Hotové stránky	Směs interních zápisů, nikoli běžný text stránky

- PDF může uvnitř obsahovat text i komprimovaná nebo obrazová data
- To, že člověk na obrazovce vidí slova, neznamená snadnou extrakci textu

CC BY-NC-ND 4.0

Kontejnerové formáty a ZIP

- DOCX obsahuje části dokumentu v XML a další soubory, například obrázky
- XLSX podobně obsahuje informace o listech, buňkách a dalších prvcích
- Tyto formáty používají balíček založený na ZIPu
- ZIP se hodí také pro předání více souborů najednou

> Zabalení do ZIPu nemění PDF na text ani tabulku na dokument.

Komprese šetří místo. Sama o sobě neznamená šifrování.

CC BY-NC-ND 4.0

Kódování textu

- Formát říká, jak je obsah uspořádaný
- Kódování říká, jak z uložených bajtů získat znaky
- Pro naše textové soubory budeme používat **UTF-8**
- Při chybném kódování může zmizet česká diakritika nebo načtení selže

Problém	Co ověřit
České znaky vypadají nesmyslně	Kódování zdroje a nastavení při načtení
CSV má všechny údaje v jednom sloupci	Oddělovač při importu

Formát CSV sám nezaručuje UTF-8 ani čárku jako oddělovač.

CC BY-NC-ND 4.0

Procvičení formátů

1. Kolega přejmenoval sken z PDF na TXT. Půjde teď hledat slova?
2. Soubor obsahuje českou smlouvu. Musí to být textový formát?
3. JSON otevřeme v textovém editoru. Přestal tím být JSONem?
4. Složku smluv zabalíme do ZIPu. Jsou tím smlouvy automaticky zašifrované?

Máte tři minuty. U každé odpovědi vysvětlete důvod.

CC BY-NC-ND 4.0

Řešení procvičení

1. Přejmenování skenu nezíská text. Potřebujeme rozpoznání textu.
2. Smlouva může být například v DOCX, PDF nebo TXT.
3. JSON zůstává JSONem, pokud jeho obsah dál splňuje pravidla formátu.
4. ZIP může být bez šifrování. Samotné zabalení ochranu nezaručuje.

Vždy se ptejme, co se změnilo uvnitř souboru.

CC BY-NC-ND 4.0

Texty a dokumenty

CC BY-NC-ND 4.0

TXT

Hodí se pro vytěžený text, jednoduché poznámky a vstup do programu.

Výhody	Omezení
Otevře ho mnoho aplikací	Neuchová běžné formátování dokumentu
Snadno se prohledává a zpracovává	Neobsahuje styly, komentáře ani vložené obrázky
Není vázaný na jeden kancelářský program	Nadpis a běžný odstavec nemusí být odlišeny

TXT se hodí, když nás zajímají slova. Pro připomínkování smlouvy bude obvykle pohodlnější DOCX.

CC BY-NC-ND 4.0

DOCX

Hodí se pro návrhy smluv, společné úpravy a dokumenty s formátováním.

Výhody	Omezení
Styly, tabulky, obrázky a poznámky	Pro zpracování potřebujeme nástroj pro tento formát
Komentáře a sledování změn	Vzhled se může mezi aplikacemi lišit
Pohodlná práce pro běžného příjemce	Je potřeba kontrolovat, jaké revize a údaje předáváme

Při předání je dobré rozlišit pracovní dokument a verzi určenou jen ke čtení.

CC BY-NC-ND 4.0

OpenDocument

Formát	Obdoba v Microsoft Office
ODT	DOCX, textový dokument
ODS	XLSX, tabulkový sešit
ODP	PPTX, prezentace

- **OpenDocument (ODF)** je otevřená alternativa k Microsoft Office, specifikace je standard ISO/IEC 26300
- Výchozí formát LibreOffice, otevře ho i Word nebo Google Dokumenty
- Uvnitř je to také ZIP s XML, podobně jako DOCX

Stejný vzhled všude nezaručuje: Word může změnit formátování, Apple Pages ODT neotevře.

CC BY-NC-ND 4.0

PDF

Hodí se pro čtení, tisk a předání dokumentu se zachovaným rozvržením.

Výhody	Omezení
Zachovává podobu jednotlivých stránek	Úpravy bývají méně pohodlné než v DOCX
Příjemce ho běžně otevře	Tabulka nemusí být uložená jako skutečná tabulka
Vhodné pro zobrazení a tisk	Text může být obtížné získat nebo může úplně chybět

PDF lze upravit. Samotná přípona nedokazuje původ ani správnost obsahu.

CC BY-NC-ND 4.0

PDF s textem a skenované PDF

PDF s textem	PDF obsahující jen sken
Slova jsou uložená jako textové objekty	Stránka je uložená jako obrázek
Často lze text vybrat a vyhledávat	Bez další vrstvy text nevyhledáme
Extrakce může zaměnit pořadí sloupců	Nejdřív potřebujeme OCR

- **OCR** rozpoznává znaky z obrazu a vytváří text
- Výsledek může zaměnit číslice, diakritiku nebo pořadí částí stránky
- Sken může mít i doplněnou textovou vrstvu, kterou je potřeba zkontrolovat

CC BY-NC-ND 4.0

Převod do TXT

Představte si smlouvu s nadpisy, tabulkou plateb, komentáři a obrázkem podpisu.

Původní informace	Co čekat po prostém textovém exportu
Slova v odstavcích	Obvykle zůstanou
Velikost písma a barvy	Ztratí se
Řádky a sloupce tabulky	Mohou se rozpadnout nebo změnit pořadí
Komentáře a revize	Záleží na způsobu exportu
Obrázek podpisu	V prostém textu nebude

Textový výstup je další pracovní podoba. Původní dokument si ponecháme.

CC BY-NC-ND 4.0

Úkol na výběr formátu

1. Kolega má ve smlouvě navrhnout změny a vysvětlit je komentáři.
2. Klient má pohodlně přečíst hotový dokument se stejným rozvržením stránek.
3. Program má v tisíci smlouvách najít výskyty určitého slova.

U každé situace navrhněte vhodný formát a jednu jeho nevýhodu.

Máte čtyři minuty. Můžete navrhnout i dvě různé verze stejného dokumentu.

CC BY-NC-ND 4.0

Řešení výběru formátu

Situace	Volba	Co pohlídat
Úpravy a komentáře	DOCX	Revize, verze a kompatibilitu aplikací
Čtení hotového dokumentu	PDF	Čitelnost, správný export a úplnost
Hromadné hledání slov	TXT získaný ze zdrojů	Správnost extrakce a vazbu na originál

Pro jednu smlouvu můžeme potřebovat pracovní DOCX, výsledné PDF i text pro zpracování.

CC BY-NC-ND 4.0

Tabulky a přehledy

CC BY-NC-ND 4.0

XLSX

Hodí se pro rozpočty, výpočty a přehledy, které lidé průběžně upravují.

Výhody	Omezení
Více listů, vzorce a grafy	Význam údajů může být schovaný ve vzhledu
Formátování a typy buněk	Sloučené buňky komplikují automatické čtení
Pohodlné filtrování a výpočty	Některé funkce se mezi aplikacemi liší

Tabulka přehledná na pohled nemusí být dobře připravená pro program.

CC BY-NC-ND 4.0

CSV

```
id,typ,castka,mesto
S01,nájemní,18000,Praha
S02,kupní,250000,Brno
S03,o dílo,120000,Praha
```

Hodí se pro exporty, importy a pravidelné zpracování řádků.

- Výhodou je jednoduchost a podpora v mnoha programech
- Nemá listy, barvy ani výpočtový model jako tabulkový sešit
- Musíme znát oddělovač, kódování a význam sloupců

CC BY-NC-ND 4.0

Export do CSV

V pracovním sešitu	V běžném exportu do CSV
Více listů	Exportujeme jednu tabulku, obvykle aktivní list
Vzorec pro součet	Zpravidla se uloží jeho výsledná hodnota
Barvy, grafy a šířky sloupců	Neuchovají se
Poznámka vyjádřená barvou	Bez samostatného sloupce může zmizet její význam

Když potřebujete znovu upravovat výpočty, ponechte si také původní XLSX.

CC BY-NC-ND 4.0

Datové typy při importu

Hodnota	Co může znamenat	Riziko při automatickém převodu
`0012`	ID smlouvy	Změní se na `12`
`01/02/2026`	Datum v různých zápisech	Není jasný den a měsíc
`1,5`	Desetinné číslo	Čárka může být zaměněna za oddělovač
Prázdná buňka	Neznámá částka	Nesmí se bez rozmyslu změnit na nulu

- Při importu ověřte typy sloupců a náhled dat
- Pro výměnu dat si dohodněte jednotný zápis a význam prázdných hodnot

CC BY-NC-ND 4.0

Oddělovače a uvozovky v CSV

```
id,nazev,castka
S04,"Novák, s.r.o.",50000
```

- Tento záznam má tři hodnoty, nikoli čtyři
- Uvozovky říkají, že čárka uvnitř patří do názvu
- Proto pro CSV použijeme čtečku daného formátu
- Při špatně nastaveném importu ověřujeme oddělovač i uvozovky

Ruční rozdělení podle každé čárky může rozbít správně uložená data.

CC BY-NC-ND 4.0

Úkol na předání tabulky

Kolega má sešit se třemi listy. Obsahuje vzorce, barevně označené neuhrazené částky a graf.

1. Chcete pokračovat v jeho výpočtech.
2. Chcete každý měsíc automaticky sečíst neuhrazené částky.

Jaký soubor potřebujete v každé situaci? Co musí kolega před exportem doplnit?

> Máte tři minuty. Samotná odpověď „pošli CSV“ nestačí.

CC BY-NC-ND 4.0

Řešení předání tabulky

Další práce	Co si vyžádáme
Pokračování ve výpočtech	XLSX se všemi potřebnými listy a vzorci
Pravidelný součet	Jednu pravidelnou tabulku, například CSV

Do exportu přidáme **stav úhrady jako údaj**, nikoli jen barvu buňky.

Dohodneme názvy sloupců, měnu, zápis částek a to, co znamená chybějící hodnota.

CC BY-NC-ND 4.0

Strukturovaná data

CC BY-NC-ND 4.0

Vnořené údaje

U jedné smlouvy chceme uložit

- Identifikátor, částku a měnu
- Dvě nebo více smluvních stran
- Seznam příloh
- Údaj, zda je kontrola dokončená

> Seznam uvnitř záznamu se přirozeně vyjadřuje v JSONu nebo XML.

Tabulky to také zvládnou, například rozdělením do více propojených tabulek.

CC BY-NC-ND 4.0

Ukázka JSONu

```
{
  "id": "S03",
  "castka": 120000,
  "mena": "CZK",
  "zkontrolovano": false,
  "strany": ["Alfa s.r.o.", "Beta s.r.o."]
}
```

- Názvy polí vysvětlují, co jednotlivé hodnoty představují
- Částka je číslo, `strany` jsou seznam
- JSON je textový formát, který lze číst i bez speciální aplikace

CC BY-NC-ND 4.0

Použití JSONu

Hodí se pro výměnu dat mezi programy, odpovědi webových služeb a vnořené záznamy.

Výhody	Omezení
Přehledné pojmenované údaje a seznamy	Pro ruční práci může být méně pohodlný než tabulka
Rozlišuje text, číslo a logickou hodnotu	Není určený k rozvržení dokumentu na stránky
Dobře se načítá do programu	Názvy polí a jejich význam si musíme dohodnout

Platný JSON může obsahovat věcně chybnou částku.

CC BY-NC-ND 4.0

Ukázka XML

```xml
<smlouva id="S03">
  <castka mena="CZK">120000</castka>
  <zkontrolovano>false</zkontrolovano>
  <strany>
    <strana>Alfa s.r.o.</strana>
    <strana>Beta s.r.o.</strana>
  </strany>
</smlouva>
```

- Obsah je ohraničen pojmenovanými značkami
- Další údaje mohou být v atributech, zde například `mena`
- XML je také textový formát, i když na první pohled působí složitěji
- Stejný princip má HTML webových stránek: značky jako `<p>` nebo `<a href="…">`, jejichž názvy a význam předem určuje specifikace HTML

CC BY-NC-ND 4.0

Použití XML

Hodí se pro zavedené výměnné formáty a dokumenty s bohatou vnitřní strukturou.

Výhody	Omezení
Pojmenované části a jejich vnoření	Zápis bývá delší a hůře se čte
Umí kombinovat běžný text se značkami	Ruční úpravy vyžadují opatrnost
Lze ověřovat podle dohodnutého schématu	Je potřeba znát pravidla konkrétního formátu

Pokud systém požaduje konkrétní XML, podobný JSON ho nenahradí.

CC BY-NC-ND 4.0

Význam dat

```
{"castka": 120000, "datum": "2026-10-27"}
```

1. V jaké měně je částka?
2. Je to cena s DPH, nebo bez DPH?
3. Je datum dnem podpisu, účinnosti, nebo exportu?

Navrhněte za dvě minuty přesnější názvy nebo další pole.

CC BY-NC-ND 4.0

Markdown

CC BY-NC-ND 4.0

Zápis Markdownu

```
# Kontrola smlouvy

Prověřit **smluvní pokutu**.

- Doplnit přílohu
- Ověřit částku
```

- `#` označuje nadpis, hvězdičky zvýraznění a pomlčka položku seznamu
- Soubor zůstává čitelný i v obyčejném textovém editoru
- Náhled tyto značky převede do formátovaného zobrazení

CC BY-NC-ND 4.0

Náhled Markdownu

Kontrola smlouvy

Prověřit **smluvní pokutu**.

- Doplnit přílohu
- Ověřit částku

> Změnil se způsob zobrazení. Ve zdrojovém souboru jsou pořád text a značky.

CC BY-NC-ND 4.0

Použití Markdownu

Hodí se pro poznámky, návody, dokumentaci a jednoduché texty sdílené mezi nástroji.

Výhody	Omezení
Jednoduchý a čitelný zdroj	Omezená kontrola přesného rozvržení
Snadno se porovnávají změny textu	Některé aplikace podporují jiné rozšíření značek
Lze převést do HTML nebo PDF	Obrázky mohou být uložené zvlášť

Pro složité připomínkování smlouvy může být DOCX pohodlnější.

CC BY-NC-ND 4.0

Zdroj v LaTeXu

```
\documentclass{article}
\usepackage{amsmath}
\begin{document}
\section*{Smlouva o dílo}
\subsection*{Čl.~3 \quad Smluvní pokuta}
Za každý den prodlení zaplatí zhotovitel smluvní pokutu
ve výši \textbf{500~Kč}. Celková výše pokuty činí
\[
  P = 500 \cdot d \ \text{Kč}, \qquad P \le 0{,}1 \cdot C,
\]
kde $d$ je počet dní prodlení a $C$ cena díla.\footnote{Srov.~§\,2048
občanského zákoníku.}
\end{document}
```

- Stejný princip jako Markdown: text a značky, jen značek je víc a jsou přísnější
- `\` uvozuje příkaz, `{ }` jeho obsah, `$ $` matematiku

CC BY-NC-ND 4.0

Výsledek v LaTeXu

Smlouva o dílo

Čl. 3 Smluvní pokuta

Za každý den prodlení zaplatí zhotovitel smluvní pokutu ve výši **500 Kč**. Celková výše pokuty činí

$$P = 500 \cdot d \text{ Kč}, \qquad P \leq 0{,}1 \cdot C,$$

kde d je počet dní prodlení a C cena díla.[1]

[1] Srov. § 2048 občanského zákoníku.

- Sazbu řeší program: zalomení, mezery, číslování i poznámky pod čarou
- **Hodí se pro** odborné texty se vzorci, citacemi a pevnou typografií
- Výstupem je obvykle PDF; zdroj se dobře porovnává a verzuje jako text

Pro běžnou smlouvu s připomínkami zůstane pohodlnější DOCX.

CC BY-NC-ND 4.0

Prezentace v Marpu

- Text a ukázky jsou zapsané v souboru `.md`
- Marp podle pravidel a tématu vytvoří jednotlivé slidy
- Výsledné PDF slouží k promítání nebo čtení
- Pro další úpravy si ponecháme zdroj, téma i použité obrázky

PDF je výstup. Markdown je v tomto případě pracovní zdroj.

CC BY-NC-ND 4.0

Úkol na předání prezentace

1. Má si prezentaci jen přečíst.
2. Má v ní přepsat příklady a vytvořit novou verzi.

Stačí v obou případech poslat stejné PDF?

Minutu přemýšlejte. Potom vysvětlete rozdíl mezi výstupem a zdrojovými soubory.

CC BY-NC-ND 4.0

Základní práce v Pythonu

CC BY-NC-ND 4.0

Cesty k souborům

```
projekt/
├── analyza.py
├── data/
│   └── smlouva.txt
└── vystupy/
```

- Cesta `data/smlouva.txt` říká, ve které složce soubor hledat
- Relativní cesta začíná v aktuálním pracovním adresáři
- Ten nemusí být stejný jako složka se skriptem
- V dalších příkladech pracujeme ve složce `projekt`

CC BY-NC-ND 4.0

Vstupní text

Soubor `data/smlouva.txt` je uložený v UTF-8.

```
Smlouva o dílo
Cena díla činí 120000 Kč.
Smluvní pokuta činí 500 Kč za den.
Změna smluvní pokuty vyžaduje dodatek.
```

Chceme zjistit, zda se v něm objevuje slovo **pokuta**.

> Jde o hledání znaků. Program tím neposuzuje právní význam ustanovení.

CC BY-NC-ND 4.0

Načtení textu ze souboru

```
with open("data/smlouva.txt", encoding="utf-8") as soubor:
    text = soubor.read()

print("pokuta" in text.lower())
```

```
True
```

- `open()` otevře soubor a `read()` načte jeho obsah
- `with` po skončení bloku soubor zavře
- `text` je obyčejný řetězec, se kterým už umíme pracovat

CC BY-NC-ND 4.0

Zápis do souboru

```
with open("vystupy/poznamka.txt", "w", encoding="utf-8") as soubor:
    soubor.write("Prověřit smluvní pokutu.\n")
```

V souboru `vystupy/poznamka.txt` bude

```
Prověřit smluvní pokutu.
```

- `"w"` vytvoří nový soubor nebo přepíše obsah existujícího
- `write()` zapíše text, `\n` ukončí řádek
- Složka `vystupy` musí předem existovat

CC BY-NC-ND 4.0

Režimy otevření souboru

Režim	Co udělá	Příklad použití
`"r"`	Čte existující soubor	Načtení smlouvy
`"w"`	Vytvoří soubor nebo vyprázdní existující	Nově vypočítaný přehled
`"a"`	Přidává na konec, případně vytvoří soubor	Postupně doplňované poznámky

Když program spustíme podruhé, chceme přehled vytvořit znovu, nebo k němu připojit další řádky?

Režim `"w"` vyprázdní existující soubor už při otevření.

CC BY-NC-ND 4.0

Atomický zápis

Při přímém přepisování souboru může pád programu zanechat jen část nového obsahu.

Atomický zápis neboli *atomic write* tomu předchází ve dvou krocích.

1. Nový obsah celý zapíšeme do dočasného souboru ve stejné složce a soubor zavřeme.
2. Hotovým souborem pak jedním krokem nahradíme původní soubor.

> Pod původním názvem je dostupná buď celá stará, nebo celá nová verze.

Hodí se například při pravidelné aktualizaci přehledu smluv nebo nastavení programu.
Když selže zápis do dočasného souboru, původní přehled zůstane zachovaný.

CC BY-NC-ND 4.0

Úkol na pět minut

Přečtěte si kód a doplňte vhodný režim do místa `DOPLNIT`.

```
with open("data/smlouva.txt", encoding="utf-8") as vstup:
    with open("vystupy/pokuty.txt", "DOPLNIT", encoding="utf-8") as vystup:
        for radek in vstup:
            if "pokut" in radek.lower():
                vystup.write(radek)
```

1. Výstup se má při každém spuštění vytvořit znovu. Co doplníte?
2. Které dva řádky ze smlouvy budou ve výstupu?
3. Co se změní, když použijeme režim pro přidávání?

CC BY-NC-ND 4.0

Řešení výběru textu

```
with open("data/smlouva.txt", encoding="utf-8") as vstup:
    with open("vystupy/pokuty.txt", "w", encoding="utf-8") as vystup:
        for radek in vstup:
            if "pokut" in radek.lower():
                vystup.write(radek)
```

```
Smluvní pokuta činí 500 Kč za den.
Změna smluvní pokuty vyžaduje dodatek.
```

Použijeme `"w"`. S `"a"` by druhé spuštění připojilo stejné řádky ještě jednou.

CC BY-NC-ND 4.0

Načítání různých formátů

Soubor	Jak k němu v Pythonu přistoupíme
TXT nebo zdroj Markdownu	Načteme text
CSV	Čtečka `csv` rozpozná sloupce a uvozovky
JSON	Čtečka `json` vytvoří slovníky, seznamy a hodnoty
XML	Použijeme nástroj, který rozumí značkám
DOCX, XLSX nebo PDF	Potřebujeme nástroj pro konkrétní formát

`open()` v textovém režimu samo nepřevede libovolný dokument na jeho text.

CC BY-NC-ND 4.0

Volba formátu v praxi

CC BY-NC-ND 4.0

Přehled formátů

Potřebujeme	Obvyklá volba
Společně upravovat dokument	DOCX
Předat hotové stránky ke čtení	PDF
Zpracovat samotná slova	TXT
Udržovat výpočty a několik listů	XLSX
Předat pravidelnou tabulku dat	CSV
Předat vnořené záznamy	JSON, případně požadované XML
Psát jednoduché poznámky a návody	Markdown

Rozhodující jsou i požadavky příjemce a možnosti jeho programu.

CC BY-NC-ND 4.0

Závěrečný úkol

Máte návrhy smluv v DOCX, několik skenů v PDF a přehled částek v XLSX.

Potřebujete

1. Dopsat připomínky k návrhům.
2. Najít v dokumentech zmínky o smluvní pokutě.
3. Vytvořit pravidelný přehled částek pro další program.
4. Předat kolegovi srozumitelný popis vašeho postupu.

Ve dvojici navrhněte za čtyři minuty soubory, převody a jednu kontrolu výsledku.

CC BY-NC-ND 4.0

Řešení závěrečného úkolu

1. Návrhy upravíme v DOCX a ponecháme potřebné komentáře.
2. Z dokumentů získáme text. U skenů použijeme OCR a výsledek ověříme.
3. Pro pravidelný přehled vyexportujeme dohodnuté sloupce do CSV.
4. Popis postupu napíšeme například v Markdownu, příjemci můžeme přidat PDF.

- U textu ponecháme vazbu na původní soubor
- U tabulky ověříme počet záznamů, částky, měnu a ID
- Původní dokumenty a pracovní sešit uchováme

CC BY-NC-ND 4.0

Kontrola před předáním

- Otevře ho příjemce a zvládne v něm požadovanou práci?
- Potřebuje jen výstup, nebo i zdrojové soubory?
- Neztratily se při převodu důležité údaje?
- Odpovídá diakritika, počet záznamů a několik kontrolních hodnot?
- Nezaměnili jsme prázdnou hodnotu, nulu a neznámý údaj?

Správný formát usnadní další práci. Správnost obsahu musíme ověřit zvlášť.

CC BY-NC-ND 4.0

Materiály k další práci

- NPFL125, inspirace pro kódování a datové formáty
- Python, čtení a zápis souborů
- Microsoft, export tabulky do textového formátu
- OpenDocument, otevřený formát kancelářských dokumentů
- Microsoft, struktura dokumentu DOCX
- W3C, formát XML
- CommonMark, Markdown a jeho pravidla

CC BY-NC-ND 4.0

Děkujeme za pozornost

CC BY-NC-ND 4.0

Doplňující ukázky

CC BY-NC-ND 4.0

CSV v Pythonu

```
import csv

with open("data/smlouvy.csv", encoding="utf-8", newline="") as soubor:
    smlouvy = list(csv.DictReader(soubor))

print(smlouvy[0]["id"])
print(int(smlouvy[0]["castka"]) + 1000)
```

```
S01
19000
```

`DictReader` použije hlavičku jako klíče slovníků. Částky z našeho CSV načítá jako text.

CC BY-NC-ND 4.0

JSON v Pythonu

```python
import json

with open("data/smlouva.json", encoding="utf-8") as soubor:
    smlouva = json.load(soubor)

print(smlouva["castka"] + 1000)
print(smlouva["strany"][1])
```

```
121000
Beta s.r.o.
```

Částka je už číslo. Ze seznamu stran vybíráme druhý prvek, tedy index `1`.

CC BY-NC-ND 4.0

Binární čtení

- V textovém režimu Python převádí bajty na znaky podle kódování
- V binárním režimu dostaneme původní bajty bez tohoto převodu
- Textový soubor můžeme otevřít i binárně, jeho formát se tím nezmění
- Přečtení bajtů PDF ještě nezíská odstavce dokumentu

Režim otevření říká, jak chceme se souborem pracovat. Nemění jeho formát.

CC BY-NC-ND 4.0