**O čom je projekt?**

Webová aplikácia, ktorá umožňuje transformáciu naskenovaných faktúr do formy digitálnych dát pomocou OCR technológie. Výstupom by bol napríklad json, prípadne nejaký editovateľný dokument.

# Prvá verzia

### **Návrhová fáza:**

**Prieskum OCR technológii, použiteľných v jazyku Python.**

* priorita: nevyhnutné
* časový odhad: 5 hodín
* skutočný strávený čas: 3 hodiny
* náročnosť implementácie: vieme odhadnúť
* riesenie: použitie OCR technológie Tesseract
* stav: hotovo

**Prieskum predtrénovaných modelov OCR neurónovej siete.**

* priorita: nevyhnutné
* časový odhad: 1.5 hodín
* skutočný strávený čas: 0.75 hodiny
* náročnosť implementácie: vieme odhadnúť približne
* riesenie: Tesseract poskytuje predtrénovaný model, ktorý je postačujúci pre projekt
* stav: hotovo

**Prieskum možnosti trénovania vlastných modelov.**

* priorita: nevyhnutné
* časový odhad: 2 hodiny
* skutočný strávený čas: 0.75 hodiny
* náročnosť implementácie: vieme odhadnúť približne
* riesenie: Tesseract poskytuje možnosť trénovať vlastné modely
* stav: hotovo

**Prieskum webových technológii, host providerov.**

* priorita: nevyhnutné
* časový odhad: 1.5 hodín
* skutočný strávený čas: 1h
* náročnosť implementácie: vieme odhadnúť
* riesenie: Framework Django umožní jednoduché spojenie s Tessaractom, obe sú písané v pythone.
* stav: hotovo

**Prieskum poskytovateľov domén.**

* priorita: prinášajú veľkú hodnotu
* časový odhad: 0.5 hodiny
* skutočný strávený čas: 0.5 hodiny
* náročnosť implementácie: vieme odhadnúť približne
* riesenie: freenom.com poskytuje domény zadarmo na 3 mesiace
* stav: hotovo

**Návrh rozloženia GUI/dizajn stránky.**

* priorita: prinášajú veľkú hodnotu
* časový odhad: 10 hodín
* skutočný strávený čas: zatial 3h
* náročnosť implementácie: vieme odhadnúť približne
* riesenie: Bude sa jednať o 1 page application
* stav: rozpracovaný

**Prieskum invoice templates.**

* priorita: prinášajú veľkú hodnotu
* časový odhad: 2 hodiny
* skutočný strávený čas: zatial 1h
* náročnosť implementácie: vieme odhadnúť približne
* riesenie: zatiaľ máme funkčný 1 template a návrh ako by sa mohli pridávať/dynamicky vytvárať ďalšie
* stav: rozpracovaný

**Prieskum css technológii/templates pre dizajn stránky.**

* priorita: prinášajú veľkú hodnotu
* časový odhad: 5 hodín
* skutočný strávený čas: zatial 1h
* náročnosť implementácie: vieme odhadnúť približne
* riesenie: Django ponúka na tvorbu stránok vlastný template engine, ktorý dokáže zobraziť dynamické dáta. Bootstrap5 nám poskytne jednotlivé elementy, z ktorých bude stránka poskladaná.
* stav: hotovo

### 

### **Implementačná fáza:**

**Testovanie/vyber OCR technológii.**

* priorita: nevyhnutné
* časový odhad: 8 hodín
* skutočný strávený čas: 5 hodín
* náročnosť implementácie: vieme odhadnúť približne
* riesenie: Testovali sme Tesseract
* stav: hotovo

**Načítanie faktúry ako súboru do formuláru, jeho odoslanie na backend.**

* priorita: nevyhnutné
* časový odhad: 5 hodín
* skutočný strávený čas: 3 hodiny
* náročnosť implementácie: vieme odhadnúť približne
* riesenie: Využívame asynchronný XMLhttp request, do ktorého vložíme súbor a odošleme na daný endpoint.
* stav: hotovo

**Transformácia naskenovanej faktúry do stringu.**

* priorita: nevyhnutné
* časový odhad: 10 hodín
* skutočný strávený čas: 3.5 hodiny
* náročnosť implementácie: vieme odhadnúť približne
* riesenie: Tesseract umožnuje extrahovať text z faktúry do stringu, snaží sa zachovať formát faktúry v stringu pomocou tab a newline-ov
* stav: hotovo

**Automatická transformácia vyťažených dát zo stringu do dict/json.**

* priorita: nevyhnutné
* časový odhad: 11 hodín
* skutočný strávený čas: 3 hodiny
* náročnosť implementácie: vieme odhadnúť približne
* riesenie: pomocou predom definovaných súradnic pre položky faktúry ide vytvoriť json, ktorý obsahuje hodnoty nascanovanej faktúry
* stav: hotovo

# Druhá verzia

### **Implementačná fáza:**

**Manuálny editor/kontrola správnosti vyťažených dát.**

* priorita: zákazníkov by to potešilo
* časový odhad: 18 hodín
* skutočný strávený čas: -
* náročnosť implementácie: nevieme odhadnúť

**Export výsledného jsonu.**

* priorita: nevyhnutné
* časový odhad: 2 hodiny
* skutočný strávený čas: -
* náročnosť implementácie: vieme odhadnúť

**Export v pdf formáte podľa nejakého dopredu určeného template.**

* priorita: zákazníkov by to potešilo
* časový odhad: 10 hodín
* skutočný strávený čas: -
* náročnosť implementácie: nevieme odhadnúť

**Pdf template editor.**

* priorita: zákazníkov by to potešilo
* časový odhad: 18 hodín
* skutočný strávený čas: -
* náročnosť implementácie: nevieme odhadnúť

**Implementácia navrhnutého dizajnu stránky.**

* priorita: zákazníkov by to potešilo
* časový odhad: 12 hodín
* skutočný strávený čas: -
* náročnosť implementácie: vieme odhadnúť približne

**Nasadenie stránky na server.**

* priorita: nevyhnutné
* časový odhad: 3 hodiny
* skutočný strávený čas: -
* náročnosť implementácie: vieme odhadnúť približne

Plus dorábanie rozpracovaných kartičiek z prvej verzie.

Projekt sa ladí a testuje pomerne tažko kedže je technológia OCR viac menej mágia, ale priložili sme testy, ktoré kontrolujú správnosť extrahovaných dát pri pôvodnej verzii, ale aj pri fotkách faktúry z rôznych uhlov. Pri fotkách OCR zatiaľ nefunguje na 100%.