



**MATEMATICKO-FYZIKÁLNÍ  
FAKULTA**  
Univerzita Karlova

## **BAKALÁŘSKÁ PRÁCE**

Kristýna Lhoťanová

# **Webová aplikace pro vyhledávání receptů**

Katedra softwarového inženýrství

Vedoucí bakalářské práce: doc. Mgr. Martin Nečaský, Ph.D.

Studijní program: Informatika

Studijní obor: Programování a vývoj software

Praha 2022

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne .....

Podpis autora

Poděkování.

Název práce: Webová aplikace pro vyhledávání receptů

Autor: Kristýna Lhořanová

Katedra: Katedra softwarového inženýrství

Vedoucí bakalářské práce: doc. Mgr. Martin Nečaský, Ph.D., Katedra softwarového inženýrství

Abstrakt: Cílem této práce je vyvinout webovou aplikaci pro vyhledávání receptů založenou na agregaci datových sad z existujících webových stránek s recepty a jejich obohacení o data ze znalostních grafů. Znalostní grafy byly zastoupeny projekty DBpedia a Wikidata, z nichž byla získána data o ingrediencích a kategoriích jednotlivých receptů. Data byla extrahována s pomocí knihovny Apify a v dokumentovém modelu uložena do databázového systému Apache CouchDB. Aplikace uživateli poskytuje různé možnosti filtrování výsledků včetně fasetového vyhledávání, k čemuž využívá platformu Apache Solr. Zaměřuje se zejména na vyhledávání dle ingrediencí. Jedná se o tzv. single-page aplikaci implementovanou pomocí JavaScriptové knihovny React pro uživatelské rozhraní a frameworku Express.js na straně serveru. Obě části aplikace jsou psány staticky typovaným jazykem TypeScript a komunikují spolu prostřednictvím REST API.

Klíčová slova: webová aplikace recept znalostní graf propojená data vyhledávání

Title: Web application for searching recipes

Author: Kristýna Lhořanová

Department: Department of Software Engineering

Supervisor: doc. Mgr. Martin Nečaský, Ph.D., Department of Software Engineering

Abstract: Abstract.

Keywords: web app recipe knowledge graph linked data search

# Obsah

<b>Úvod</b>	<b>2</b>
<b>1 Kontext vývoje</b>	<b>4</b>
1.1 Dostupné datové sady . . . . .	4
<b>2 Odkazy na literaturu</b>	<b>5</b>
2.1 Několik ukázek . . . . .	5
<b>3 Tabulky, obrázky, programy</b>	<b>6</b>
3.1 Tabulky . . . . .	6
3.2 Obrázky . . . . .	7
3.3 Programy . . . . .	7
<b>4 Formát PDF/A</b>	<b>12</b>
<b>Závěr</b>	<b>13</b>
<b>Seznam použité literatury</b>	<b>14</b>
<b>Seznam obrázků</b>	<b>15</b>
<b>Seznam tabulek</b>	<b>16</b>
<b>Seznam použitých zkratk</b>	<b>17</b>
<b>A Přílohy</b>	<b>18</b>
A.1 První příloha . . . . .	18

# Úvod

Vyhledávání relevantního obsahu je spolu s elektronickou komunikací jednou z klíčových funkcí internetu. S rostoucím množstvím dostupných informací se filtrování nalezených výsledků stává stále obtížnějším. Tvůrci webových stránek se často zaměřují spíše na uživatelsky přívětivé interaktivní rozhraní, zatímco optimalizace strojového vyhledávání jde stranou. Pro webové vyhledávače, jmenovitě např. Google, Bing nebo Yahoo, je pak náročné analyzovat obsah těchto stránek po sémantické stránce a tedy vyhodnotit, zda obsahují užitečné informace k zodpovězení dotazu uživatele.

V reakci na tuto problematiku vznikl tzv. *Sémantický Web* neboli Web dat jakožto rozšíření původního Webu dokumentů, tak jak jej známe z platformy *World Wide Web*. Sémantický Web lze vnímat jako globální databázi, nad kterou se lze pomocí speciálního jazyka *SPARQL* dotazovat podobně jako nad tradičními databázovými systémy. Data jsou poskytována v různých serializacích formátu RDF a mohou být přímo vnořena do HTML dokumentů nebo zpřístupněna v samostatných souborech. Tato strukturovaná data nazýváme *propojená* (v originále *Linked Data*). Umožňují snadnější hledání souvislostí mezi entitami z různých zdrojů na základě společných slovníků neboli ontologií (Consortium, 2015).

V posledních letech termín Sémantický Web ustupuje do pozadí a často je místo něj zmiňován tzv. *znalostní graf* (anglicky *Knowledge Graph*). Začátky fenoménu znalostních grafů bychom mohli datovat do roku 2012, kdy společnost Google představila svůj znalostní graf pro vyhledávání obsahu na webu. K technologii znalostních grafů se brzy poté přihlásily další velké společnosti včetně firem Microsoft, IBM, Facebook, LinkedIn, Amazon, eBay, Airbnb nebo Uber. Grafový model totiž oproti tradičnímu relačnímu modelu nabízí flexibilnější správu dat z oblasti sociálních sítí, dopravních spojení, bibliografických citací a řady dalších odvětví. Výše zmíněné příklady znalostních grafů všechny spadají do kategorie komerčních znalostních grafů, které jsou určeny pro interní využití v rámci dané firmy. Protikladem jim jsou otevřené znalostní grafy poskytující data k volnému využití všem uživatelům internetu. Nejvýznamnějšími představiteli otevřených znalostních grafů jsou aktuálně DBpedia, Wikidata, Freebase a YAGO (Hogan a kol., 2021). První dva zmíněné projekty si představíme v této práci a integrujeme je s aplikací na vyhledávání receptů.

Oblast gastronomie je rozvěž vhodným kandidátem k zapojení do sítě znalostních grafů a propojených dat. Pro tvůrce webových aplikací je poměrně jednoduché publikovat obsah svých stránek ve formátu strukturovaných dat. Vhodným způsobem je např. vložení RDF reprezentace daných entit (receptů, uživatelů, recenzí) ve formátu JSON-LD<sup>1</sup> přímo do hlavičky jednotlivých HTML dokumentů. V takovém případě je žádoucí použít existující ontologie raději než definovat vlastní, byť by mohly být lépe strukturované a uzpůsobené dané doméně. Využití standardizovaných slovníků usnadňuje webovým vyhledávačům interpretaci stránky a je větší šance, že se aplikace dostane na vyšší příčky vyhledávaných výsledků.

Cílem této bakalářské práce je prozkoumat možnosti využití otevřených dat v doméně receptů, propojit je s daty publikovanými na různých webových strán-

---

<sup>1</sup>Koncovka *LD* v názvu JSON-LD odkazuje na pojem Linked Data.

kách shromažďujících recepty a prezentovat tyto výsledky uživateli ve formě vlastní webové aplikace. Zároveň v rámci této aplikace poskytnout užitečné možnosti filtrování agregovaných výsledků včetně fasetového vyhledávání. Proces sběru, konverze a uložení dat by měl být co nejvíce automatizovaný a snadno zreprodukovatelný. Práce se nevěnuje přidávání nových receptů prostřednictvím uživatelského rozhraní. Existujících webové stránky totiž obsahují velké množství dat, které lze díky bohaté historii v podobě hodnocení a recenzí lépe filtrovat. Navíc by bylo potřeba se vypořádat s automatickou kalkulací nutričních hodnot receptu z obsažených surovin, přičemž ne všechny ingredience dokážeme automaticky identifikovat a získat jejich nutriční hodnoty. V budoucnu by funkce nahrávání nových receptů měla být přidána spolu s více lokalizacemi aplikace, registrací uživatelů a celkovou personalizací obsahu pro přihlášené uživatele.

## Volba tématu

Příprava jídla je tématem každodenního života a na webových stránkách, které se této oblasti věnují, má velmi silnou komunitu. Většina z nás se chystání domácích pokrmů z ekonomických důvodů nevyhne, takže se hodí mít po ruce sadu receptů pro inspiraci. Typicky máme na recepty různé požadavky - někdo preferuje rychlejší postup, jiný se dívá po ceně ingrediencí nebo nutričních hodnotách. Občas dostaneme chuť na recept z řecké nebo italské kuchyně a jindy zkrátka chceme experimentovat a najít recept kombinující našich 5 oblíbených surovin. Některé ingredience z receptu nám mohou být neznámé, nebo si jen podle názvu nejsme jistí, zda máme na mysli tu správnou. V takovém případě musíme stránku s receptem opustit a dodatečné informace k ingredienci vyhledat jinde, pokud na ně aplikace přímo neodkazuje. Zde je příležitost zapojit otevřená data a namapovat názvy ingrediencí na jejich odpovídající entity ve znalostních grafech. Data pak můžeme začlenit do aplikace a nabídnout uživateli informace nad rámec samotného receptu, např. popisy a glykemické hodnoty surovin, ilustrační obrázky a podobně. Také můžeme identifikovat ingredience a tranzitivně recepty ze stejných kategorií. Oproti původní datové sadě tak vytvoříme nové vazby a poskytneme uživateli rozmanitější filtrování výsledků.

Doména receptů navíc poskytuje spoustu prostoru pro zajímavá rozšíření se zapojením moderních technologií. Uplatnění by zde našlo například počítačové vidění s rozpoznáváním obrázků. S dostatečně velkou databází bychom díky němu mohli analyzovat fotografii hotového pokrmu a nalézt příslušný recept. Uspadnili bychom tak uživateli práci v situacích jako je návštěva restaurace, při které návštěvníkovi zachutnalo servírované jídlo a chtěl by si jej později připravit v domácích podmínkách. Uživatelé by také mohli ocenit výhody populárního *full-text* vyhledávání. Snadno by s ním objevili recepty na základě klíčových slov v popisku receptu, postupu či recenzích. V komerční sféře by se nabízelo propojení s online supermarketem, konkrétně zrychlení nákupu pomocí vyhledávání surovin k vybranému receptu. S tímto konceptem již na svých stránkách pracuje firma rohlik.cz, nabídka receptů a možnosti filtrování jsou ale omezené. Nepochybně by se hodilo integrovat také doporučovací systém pro ještě snadnější nalezení relevantních výsledků. Aplikace má velký prostor pro škálování objemu dat, přičemž datasety mohou být následně použity jako podklad pro strojové učení.

# 1. Kontext vývoje

V této kapitole si rozebereme různé alternativy datových sad pro naši aplikaci a představíme jejich výhody i nevýhody vzhledem k požadavkům aplikace. Dále se podíváme na existující webové stránky s recepty a provedeme diskuzi nad jejich funkcemi, možnými vylepšeními a rozšířeními.

## 1.1 Dostupné datové sady

V první fázi analýzy se zaměříme na veřejně dostupná zdrojová data s recepty, která by mohla posloužit jako podklad pro naši databázi. Jedním z nejdůležitějších projektů v této oblasti je Recipe1M+, strukturovaný korpus obsahující přes 1 milion receptů a 13 milionů souvisejících obrázků jídla. Aktuálně se jedná o největší veřejně dostupnou sadu receptů. Dataset je dostupný pouze přihlášeným uživatelům z ověřené organizace a je povoleno jej využívat pouze pro studium a výzkum. Z celkového počtu 1 milionu receptů obsahuje 50 000 receptů s nutričními informacemi (Marin a kol., 2019). V naší aplikaci plánujeme nutriční hodnoty zahrnout, takže bychom měli k dispozici 50 000 dokumentů s touto informací. Ostatní data jsou určena přednostně pro strojové zpracování prostřednictvím trénování modelů a pro finální zobrazení uživateli v rámci webové aplikace se nezdají být příliš vhodná. Celková velikost datové sady se navíc pohybuje v řádu stovek gigabytů, což je pro potřeby této práce příliš mnoho vzhledem k omezené výpočetní kapacitě.



## 2. Odkazy na literaturu

Odkazy na literaturu vytváříme nejlépe pomocí příkazů `\citet`, `\citep` atp. (viz L<sup>A</sup>T<sub>E</sub>Xový balíček `natbib`) a následného použití BibT<sub>E</sub>Xu. V matematickém textu obvykle odkazujeme stylem „Jméno autora/autorů (rok vydání)“, resp. „Jméno autora/autorů [číslo odkazu]“. V českém/slovenském textu je potřeba se navíc vypořádat s nutností skloňovat jméno autora, respektive přechylovat jméno autorky. Je potřeba mít na paměti, že standardní příkazy `\citet`, `\citep` produkují referenci se jménem autora/autorů v prvním pádě a jména autorek jsou nepřechýlena.

Pokud nepoužíváme bibT<sub>E</sub>X, řídíme se normou ISO 690 a zvyklostmi oboru. Jména časopisů lze uvádět zkráceně, ale pouze v kodifikované podobě.

### 2.1 Několik ukázek

Mezi nejvíce citované statistické články patří práce Kaplana a Meiera a Coxe (??). ? napsal článek o t-testu.

Prof. Anděl je autorem učebnice matematické statistiky (viz ?). Teorii odhadu se věnuje práce ?. V případě odkazů na specifickou informaci (definice, důkaz, ...) uvedenou v knize bývá užitečné uvést specificky číslo kapitoly, číslo věty atp. obsahující požadovanou informaci, např. viz ?, Věta 4.22 nebo (viz ?, Věta 4.22).

Mnoho článků je výsledkem spolupráce celé řady osob. Při odkazování v textu na článek se třemi autory obvykle při prvním výskytu uvedeme plný seznam: ? představili koncept EM algoritmu. Respektive: Koncept EM algoritmu byl představen v práci Dempstera, Lairdové a Rubina (?). Při každém dalším výskytu již používáme zkrácenou verzi: ? nabízejí též několik příkladů použití EM algoritmu. Respektive: Několik příkladů použití EM algoritmu lze nalézt též v práci Dempstera a kol. (?).

U článku s více než třemi autory odkazujeme vždy zkrácenou formou: První výsledky projektu ACCEPT jsou uvedeny v práci Genbergové a kol. (?). V textu *nenapíšeme*: První výsledky projektu ACCEPT jsou uvedeny v práci ?.

## 3. Tabulky, obrázky, programy

Používání tabulek a grafů v odborném textu má některá společná pravidla a některá specifická. Tabulky a grafy neuvádíme přímo do textu, ale umístíme je buď na samostatné stránky nebo na vyhrazené místo v horní nebo dolní části běžných stránek. L<sup>A</sup>T<sub>E</sub>X se o umístění plovoucích grafů a tabulek postará automaticky.

Každý graf a tabulku očíslovujeme a umístíme pod ně legendu. Legenda má popisovat obsah grafu či tabulky tak podrobně, aby jim čtenář rozuměl bez důkladného studování textu práce.

Na každou tabulku a graf musí být v textu odkaz pomocí jejich čísla. Na příslušném místě textu pak shrneme ty nejdůležitější závěry, které lze z tabulky či grafu učinit. Text by měl být čitelný a srozumitelný i bez prohlížení tabulek a grafů a tabulky a grafy by měly být srozumitelné i bez podrobné četby textu.

Na tabulky a grafy odkazujeme pokud možno nepřímou v průběhu běžného toku textu; místo „*Tabulka 3.1 ukazuje, že muži jsou v průměru o 9,9 kg těžší než ženy*“ raději napíšeme „*Muži jsou o 9,9 kg těžší než ženy (viz Tabulka 3.1)*“.

### 3.1 Tabulky

U **tabulek** se doporučuje dodržovat následující pravidla:

- Vyhybat se svislým linkám. Silnějšími vodorovnými linkami oddělit tabulku od okolního textu včetně legendy, slabšími vodorovnými linkami oddělovat záhlaví sloupců od těla tabulky a jednotlivé části tabulky mezi sebou. V L<sup>A</sup>T<sub>E</sub>Xu tuto podobu tabulek implementuje balík `booktabs`. Chceme-li výrazněji oddělit některé sloupce od jiných, vložíme mezi ně větší mezeru.
- Neměnit typ, formát a význam obsahu políček v tomtéž sloupci (není dobré do téhož sloupce zapisovat tu průměr, onde procenta).
- Neopakovat tentýž obsah políček mnohokrát za sebou. Máme-li sloupec *Rozptyl*, který v prvních deseti řádcích obsahuje hodnotu 0,5 a v druhých deseti řádcích hodnotu 1,5, pak tento sloupec raději zrušíme a vyřešíme to jinak. Například můžeme tabulku rozdělit na dvě nebo do ní vložit popisné řádky, které informují o nějaké proměnné hodnotě opakující se v následujícím oddíle tabulky (např. „*Rozptyl = 0,5*“ a níže „*Rozptyl = 1,5*“).
- Čísla v tabulce zarovnávat na desetinnou čárku.

Efekt	Odhad	Směrod. chyba <sup>a</sup>	P-hodnota
Abs. člen	−10,01	1,01	—
Pohlaví (muž)	9,89	5,98	0,098
Výška (cm)	0,78	0,12	< 0,001

Pozn: <sup>a</sup> Směrodatná chyba odhadu metodou Monte Carlo.

Tabulka 3.1: Maximálně věrohodné odhady v modelu M.

- V tabulce je někdy potřebné používat zkratky, které se jinde nevyskytují. Tyto zkratky můžeme vysvětlit v legendě nebo v poznámkách pod tabulkou. Poznámky pod tabulkou můžeme využít i k podrobnějšímu vysvětlení významu některých sloupců nebo hodnot.

## 3.2 Obrázky

Několik rad týkajících se obrázků a grafů.

- Graf by měl být vytvořen ve velikosti, v níž bude použit v práci. Zmenšení příliš velkého grafu vede ke špatné čitelnosti popisků.
- Osy grafu musí být řádně popsány ve stejném jazyce, v jakém je psána práce (absenci diakritiky lze tolerovat). Kreslíme-li graf hmotnosti proti výšce, nenecháme na nich popisky **ht** a **wt**, ale osy popíšeme *Výška [cm]* a *Hmotnost [kg]*. Kreslíme-li graf funkce  $h(x)$ , popíšeme osy  $x$  a  $h(x)$ . Každá osa musí mít jasně určenou škálu.
- Chceme-li na dvourozměrném grafu vyznačit velké množství bodů, dáme pozor, aby se neslily do jednolitě černé tmy. Je-li bodů mnoho, zmenšíme velikost symbolu, kterým je vykresluje, anebo vybereme jen malou část bodů, kterou do grafu zaneseme. Grafy, které obsahují tisíce bodů, dělají problémy hlavně v elektronických dokumentech, protože výrazně zvětšují velikost souborů.
- Budeme-li práci tisknout černobíle, vyhneme se používání barev. Čáry rozlišujeme typem (plná, tečkovaná, čerchovaná, ...), plochy dostatečně rozdílnými intenzitami šedé nebo šrafováním. Význam jednotlivých typů čar a ploch vysvětlíme buď v textové legendě ke grafu anebo v grafické legendě, která je přímo součástí obrázku.
- Vyhýbejte se bitmapovým obrázkům o nízkém rozlišení a zejména JPEGům (zuby a kompresní artefakty nevypadají na papíře pěkně). Lepší je vytvářet obrázky vektorově a vložit do textu jako PDF.

## 3.3 Programy

Algoritmy, výpisy programů a popis interakce s programy je vhodné odlišit od ostatního textu. Jednou z možností je použití L<sup>A</sup>T<sub>E</sub>Xového balíčku **fancyvrb** (fancy verbatim), pomocí něhož je v souboru **makra.tex** nadefinováno prostředí **code**. Pomocí něho lze vytvořit např. následující ukázky.

```
> mean(x)
[1] 158.90
> objekt$prumer
[1] 158.90
```

Menší písmo:

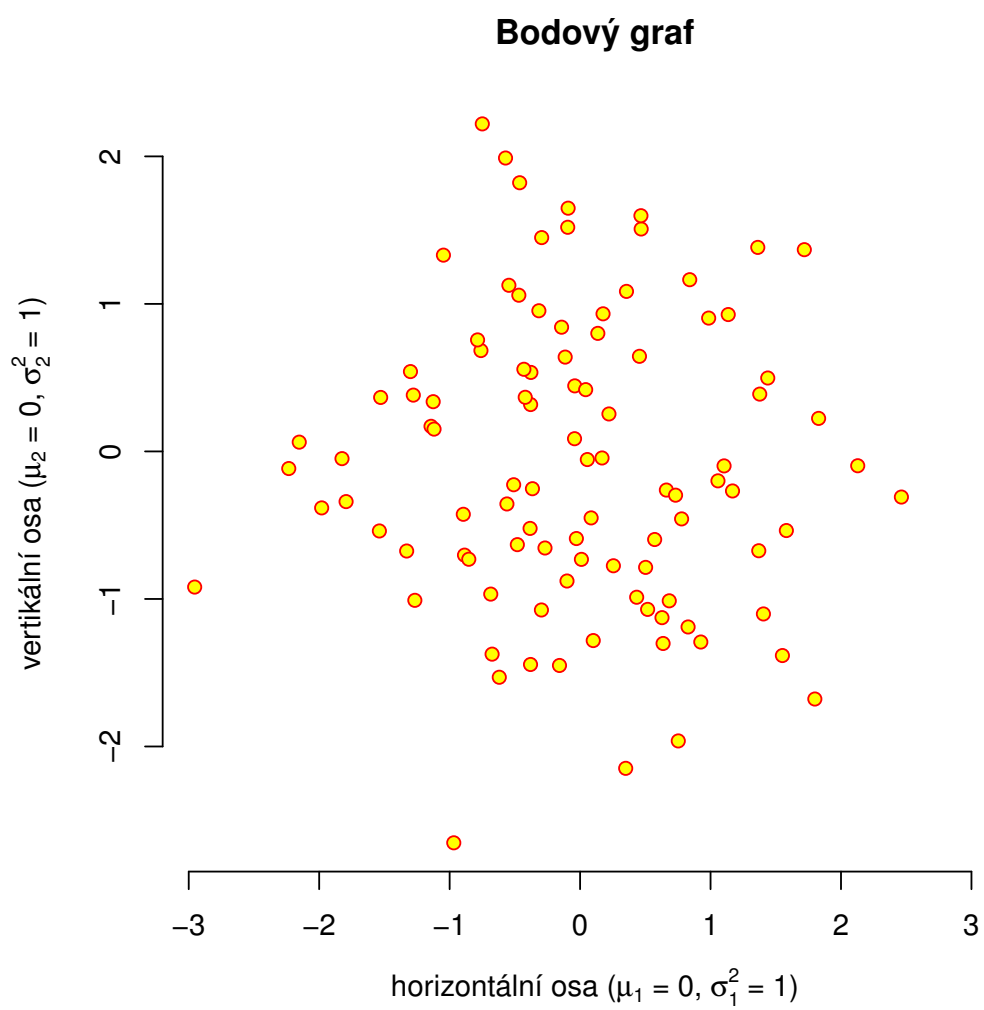
```
> mean(x)
[1] 158.90
> objekt$prumer
[1] 158.90
```

Bez rámečku:

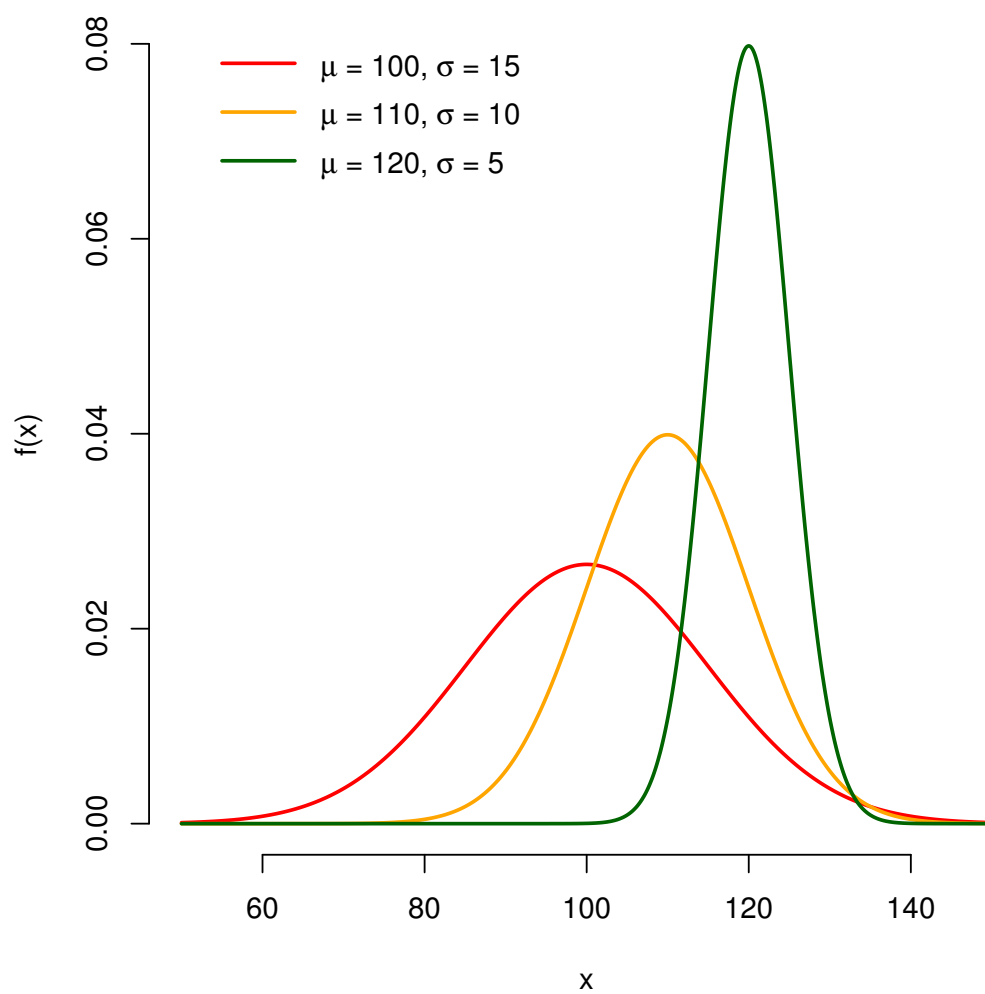
```
> mean(x)
[1] 158.90
> objekt$prumer
[1] 158.90
```

Užší rámeček:

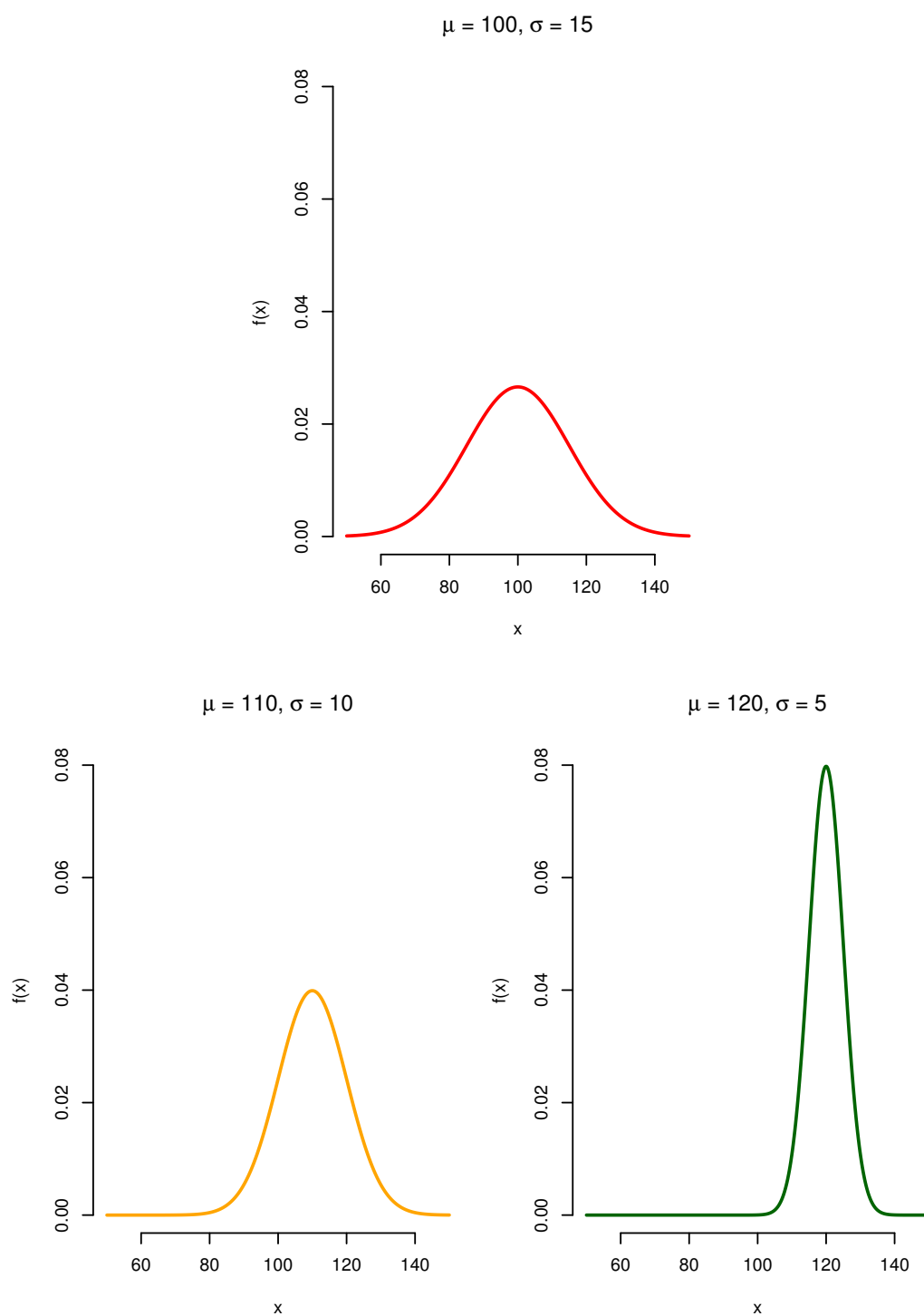
```
> mean(x)
[1] 158.90
> objekt$prumer
[1] 158.90
```



Obrázek 3.1: Náhodný výběr z rozdělení  $\mathcal{N}_2(\mathbf{0}, I)$ .



Obrázek 3.2: Hustoty několika normálních rozdělení.



Obrázek 3.3: Hustoty několika normálních rozdělení.

## 4. Formát PDF/A

Opatření rektora č. 13/2017 určuje, že elektronická podoba závěrečných prací musí být odevzdávána ve formátu PDF/A úrovně 1a nebo 2u. To jsou profily formátu PDF určující, jaké vlastnosti PDF je povoleno používat, aby byly dokumenty vhodné k dlouhodobé archivaci a dalšímu automatickému zpracování. Dále se budeme zabývat úrovní 2u, kterou sázíme  $\text{\LaTeX}$ .

Mezi nejdůležitější požadavky PDF/A-2u patří:

- Všechny fonty musí být zabudovány uvnitř dokumentu. Nejsou přípustné odkazy na externí fonty (ani na „systémové“, jako je Helvetica nebo Times).
- Fonty musí obsahovat tabulku ToUnicode, která definuje převod z kódování znaků použitého uvnitř fontu to Unicode. Díky tomu je možné z dokumentu spolehlivě extrahovat text.
- Dokument musí obsahovat metadata ve formátu XMP a je-li barevný, pak také formální specifikaci barevného prostoru.

Tato šablona používá balíček `pdfx`, který umí  $\text{\LaTeX}$  nastavit tak, aby požadavky PDF/A splňoval. Metadata v XMP se generují automaticky podle informací v souboru `prace.xmpdata` (na vygenerovaný soubor se můžete podívat v `pdfa.xmpi`).

Validitu PDF/A můžete zkontrolovat pomocí nástroje VeraPDF, který je k dispozici na <http://verapdf.org/>.

Pokud soubor nebude validní, mezi obvyklé příčiny patří používání méně obvyklých fontů (které se vkládají pouze v bitmapové podobě a/nebo bez unicodových tabulek) a vkládání obrázků v PDF, které samy o sobě standard PDF/A nesplňují.

Další postřehy o práci s PDF/A najdete na <http://mj.ucw.cz/vyuka/bc/pdfaq.html>.



# Závěr

# Seznam použité literatury

CONSORTIUM, T. W. W. W. (2015). Semantic Web. URL <https://www.w3.org/standards/semanticweb/>.

HOGAN, A., BLOMQVIST, E., COCHEZ, M., D'AMATO, C., DE MELO, G., GUTIÉRREZ, C., KIRrane, S., LABRA GAYO, J. E., NAVIGLI, R., NEUMAIER, S., NGONGA NGOMO, A.-C., POLLERES, A., RASHID, S. M., RULA, A., SCHMELZEISEN, L., SEQUEDA, J. F., STAAB, S. a ZIMMERMANN, A. (2021). *Knowledge Graphs*. Number 22 in Synthesis Lectures on Data, Semantics, and Knowledge. Morgan & Claypool. ISBN 9781636392363. doi: 10.2200/S01125ED1V01Y202109DSK022. URL <https://kgbook.org/>.

MARIN, J., BISWAS, A., OFLI, F., HYNES, N., SALVADOR, A., AYTAR, Y., WEBER, I. a TORRALBA, A. (2019). Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Trans. Pattern Anal. Mach. Intell.*

# Seznam obrázků

3.1	Náhodný výběr z rozdělení $\mathcal{N}_2(\mathbf{0}, I)$ . . . . .	9
3.2	Hustoty několika normálních rozdělení. . . . .	10
3.3	Hustoty několika normálních rozdělení. . . . .	11

# Seznam tabulek

3.1	Maximálně věrohodné odhady v modelu M. . . . .	6
-----	--	---

# Seznam použitých zkratek

## A. Přílohy

### A.1 První příloha