Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE

Martin Smrt

Shannonův test

Katedra softwarového inženýrství

Vedoucí bakalářské práce: Mgr. Jan Lánský

Studijní program: Informatika, Obecná informatika

2008

Na tomto místě bych rád poděkoval vedoucímu práce Mgr. Janu Lánskému, který mne k tématu práce přivedl a během jejího řešení mi ochotně poskytoval cenné rady a náměty. Dále bych chtěl poděkovat svým rodičům, kteří mi během celého mého dosavadního studia poskytovali podporu a tolik potřebné zázemí.

Prohlašuji, že jsem svou bakalářskou práci napsal(a) samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 8. srpna 2008 Martin Smrt

Obsah

Název práce: Shannonův test
Autor: Martin Smrt
Katedra (ústav): Katedra softwarového inženýrství
Vedoucí bakalářské práce: Mgr. Jan Lánský
e-mail vedoucího: lansky@ksi.ms.mff.cuni.cz

Abstrakt: Předložená práce ze zabývá tvorbou aplikace pro měření entropie psaného textu a představuje výsledky experimentu, který byl s využitím této aplikace proveden. Vychází z díla amerického matematika C. E. Shannona a využívá jednu z metod, kterou Shannon použil ve vlastním výzkumu. Obsahem práce je popis testovací aplikace a prezentace zjištěných výsledků. Zatímco Shannon pracoval jen s písmeny, tato práce obsahuje výsledky zjištěné pro písmena, slabiky a celá slova. Odhady entropie jsou k dispozici pro český jazyk a pro angličtinu.

Klíčová slova: Claude E. Shannon, entropie, online aplikace

Title: Shannon's test
Author: Martin Smrt
Department: Department of Software Engineering
Supervisor: Mgr. Jan Lánský
Supervisor's e-mail address: lansky@ksi.ms.mff.cuni.cz

Abstract: In the present work we focus on creating an application for the purpose of measuring entropy of written language. The work presents results yielded from actual use of the application. It is based upon the work of an American mathematician Claude E. Shannon and exploits one of the methods which he used for estimating entropy. The content of the work describes the application and presents results of the experiment. While Shannon only focused on letters in his work, this thesis compares results of working with letters, syllables and words. Entropy estimates are given for Czech and English languages.

Keywords: Claude E. Shannon, entropy, online application

Kapitola 1

Úvod

Tato práce se zaměřuje na jednu z metod odhadování entropie psaného textu. S pojmem entropie se můžeme setkat v několika oborech lidské činnosti, kromě informatiky například také v biologii či v termodynamice, význam tohoto termínu je však ve všech zmíněných případech zcela odlišný. Entropie v informatice může být definována jako „*střední hodnota míry informace potřebné k odstranění neurčitosti, která je dána konečným počtem vzájemně se vylučujících jevů,*“ jak je uvedeno v [1].

1.1 Seznámení s pojmem entropie

Entropie tedy vyjadřuje míru nejistoty informace. Obecnou entropii informace definoval Američan Claude E. Shannon v roce 1948 ve svém díle [2]. Podle něj je entropie H diskrétní náhodné veličiny, která nabývá hodnot $\{x_1, ..., x_n\}$, definována výrazem:

$$H = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{1.1}$$

Volba základu logaritmu ovlivňuje jednotku výsledné hodnoty, při základu 2 je jednotkou entropie bit.

Pro snadné pochopení lze entropii vysvětlit na příkladu hodu mincí. Výsledek hodu může nabývat dvou hodnot (panna a orel), jejichž pravděpodobnosti jsou rovny p a $q = (1 - p)$. Dosazením do vzorce 1.1 dostáváme $H = -(p \log p + q \log q)$. Pokud hod bude probíhat se spravedlivou mincí, u které je pravděpodobnost padnutí obou stran shodná, bude výsledná entropie mít hodnotu 1 bit.

To znamená, že právě 1 bit postačuje k poskytnutí úplné informace o tom, jaká strana mince padla. Tento jednoduchý případ je zcela zřejmý, pomáhá však demonstrovat možný způsob náhledu na otázku entropie. Pokud bychom minci upravili tak, aby častěji padala jedna ze stran, bude se s rostoucí pravděpodobností padnutí jedné strany mince výsledná entropie snižovat. Pro teoretickou minci, na které vždy padá jen jedna strana, pak dosáhne nuly.

1.2 Predikce textu

V roce 1951 Shannon publikoval článek [3], ve kterém se zabýval entropií psaného anglického textu. Entropie zde vyjadřuje průměrnou míru informace, kterou obsahuje každé písmeno textu. Shannon ve své práci prezentoval několik způsobů měření entropie. Kromě statistických výpočtů, prováděných nad vybranými vzorky anglického textu v souladu s výše uvedenou teorií, připravil experiment spočívající v odhadování textu. Metoda odhadování využívá faktu, že každý člověk, mluvící určitým jazykem, „*disponuje ohromnou znalostí statistik (daného) jazyka,*“ jak je uvedeno v [3]. Znalost slov, frází a gramatiky umožňuje mimo jiné kontrolu chyb v textu a doplňování nedokončených sdělení, např. při běžné konverzaci.

Cílem této práce bylo vytvoření testovací aplikace a provedení experimentu, inspirovaného výzkumem Claudea Shannona, s texty psanými v českém jazyce. Zatímco Shannon pracoval pouze na úrovní jednotlivých písmen, tato práce měla za úkol zaměřit se kromě písmen také na slabiky a celá slova. Dále v textu bude pro písmena, slabiky a slova používán souhrnný termín entita.

1.3 Struktura práce

Kapitola 2 popisuje základní úvahy vedoucí k vytvoření testovací aplikace. Představuje také slabikové kódování jako motivaci ke zkoumání entropie.

Kapitola 3 je programátorskou dokumentací vytvořené aplikace a věnuje se především problematice návrhu aplikace a řešení některých praktických problémů vývoje aplikace.

Kapitola 4 slouží jako uživatelská dokumentace, která vysvětluje práci s vytvořenou aplikací.

Kapitola 5 představuje výsledky experimentu, provedeného s pomocí vytvořené aplikace.

Kapitola 2

Analýza problematiky

Základním cílem této práce je vytvoření testovací aplikace pro měření entropie zkoumaných textů. Aplikace by měla být dostupná širokému spektru uživatelů, aby bylo možné pro test získat dostatečné množství účastníků. Cílem není provedení jednorázového výzkumu v laboratorních podmínkách, ale poskytnutí nástroje pro dlouhodobé testování.

Aplikace by měla umožňovat snadné přidávání textů, na kterých bude výzkum prováděn. Jednotlivé texty je třeba rozdělit na požadované entity, konkrétně písmena, slabiky a slova. Samotný test pak spočívá v postupném odhadování po sobě jdoucích částí textu.

2.1 Komprese textů jako motivace

Jak již bylo řečeno v úvodu, entropii textu lze chápat jako míru počtu bitů potřebných k zakódování jednoho znaku. Výsledkem testu budou odhady entropie pro každou dílčí entitu textu, ze kterých pak lze počítat entropii celkovou. Ukázkový výsledek testu přibližuje následující tabulka:

Písmeno	B	u	d	o	v	y		n	a	v	a	z	o	v	a	l	y	
Pokusů	3	5	1	1	1	2	1	3	1	11	1	1	1	1	1	1	1	1

Tabulka 2.1: Příklad výsledku testu prováděného po písmenech

V prvním řádku je uvedena hádaná entita, v druhém řádku pak příslušný počet pokusů, které testující uživatel potřeboval k jejímu uhodnutí. Prázdná buňka v prvním řádku znamená, že hádanou entitou byla mezera. Lze snadno nahlédnout, že počet pokusů udává míru entropie patřičné entity z pohledu účastníka testu. K uhodnutí prvního znaku potřeboval 3 pokusy, k zakódování by tedy byly potřeba 3 bity (dvě zamítavé odpovědi na předchozí tipované entity a třetí kladná odpověď). U znaků, které uživatel uhodl napoprvé, tak stačí k zakódování 1 bit.

Odhady entropie nacházejí využití v úvahách o možnostech kódování a komprese textu. Princip kódování pomocí entropie dobře ilustruje myšlenka identických dvojčat, jak je popsána v [3]. Pakliže bychom z tabulky 2.1 uchovali jako zprávu pouze druhý řádek, tzv. redukovaný text, zdánlivě tím ztratíme informace o obsahu kódovaného textu. Pokud bychom ovšem následně identickému dvojčeti uživatele z testu zadali stejný úkol hádání textu, bude nám informace o potřebném počtu pokusů plně dostačovat k rozkódování původního sdělení. Víme totiž, v kolikátém pokusu si druhé dvojče tipne správně, a žádnou další informaci tedy k obnovení textu nepotřebujeme.

Teoretická identická dvojčata lze v praxi nahradit stejně naprogramovanými stroji, z nichž jeden bude text kódovat a druhý bude provádět zpětné rozkódování. Tak by bylo možné sestavit komunikační kanál, po kterém se bude přenášet zakódovaná podoba textu. v jednoduché variantě by stroje mohly odhadovat znaky podle obecné četnosti výskytu v daném jazyce, lepší výsledky by pak dávaly stroje disponující znalostmi podrobnějších statistik jazyka. Entropie anglického jazyka činí v případě, kdy pracujeme se statistickými vlastnostmi jazyka v rozsahu maximálně osmi po sobě jdoucích znaků, přibližně 2,3 bitu/znak [3].

Tradiční metody komprese textu pracují buď s kódováním jednotlivých písmen, nebo celých slov. Ukazuje se ovšem, že zajímavou alternativou by mohla být také komprese po slabikách [4]. Komprese po jednotlivých písmenech je vhodná pro kratší texty, slovní komprese naopak nalézá uplatnění u textů dlouhých. Rozhodujícím faktorem pro výběr vhodné metody je velikost tabulky přípustných hodnot jednotlivých elementů a její poměr k celkové délce textu. Vzhledem k tomu, že množina slabik je v každém jazyce výrazně menší množiny všech slov, mohla by slabiková komprese nabídnout zajímavé výsledky především pro středně dlouhé texty.

2.2 Testovací aplikace

Primárním výstupem testu budou vždy počty pokusů, které uživatel potřebuje k uhodnutí jednotlivých entit. Aplikace tedy musí pokusy počítat. Zároveň je žádoucí, aby účastník testu zkoumaný text neznal, protože jinak by byly výsledky značně zkresleny. To samozřejmě nelze zcela garantovat, ovšem vhodným výběrem textů lze riziko snížit. Použity by měly být takové texty, které nejsou snadno dohledatelné na Internetu ani nejsou notoricky známé. Zároveň je třeba uplatnit omezení, podle kterého může každý uživatel s konkrétním textem pracovat jen jednou, bez ohledu na druh entit zkoumaných v proběhlém testu.

Prostředí aplikace by mělo uživateli v řešení testu pomáhat. Základní pomůckou budou statistiky, informující o vlastnostech zkoumaného textu. Statistiky poskytují seznam všech v textu se vyskytujících entit a informace o četnosti jejich výskytu. K orientaci ve statistikách by mohla pomoci také funkce, která uživateli bude na základě zadaných znaků navrhovat odpovídající entity. Jako tip bude možné zadat pouze entitu, která se textu skutečně vyskytuje.

Shannon ve svém experimentu seskupil interpunkci a mezery mezi slovy do jedné entity, čímž vznikla rozšířená anglická abeceda o 27 znacích [3]. Ekvivalentním způsobem bude tvořena množina přípustných entit i v dalších zkoumaných variantách. V průběhu testu je nutné zjistit, zda uživatel slovo považuje za dokončené, ovšem hádání interpunkce by již bylo zbytečně složité. Všechny oddělovače slov tak mohou být zastoupeny jednou tipovanou entitou, například tečkou. Ta pak může zastupovat i případná nepísmenná slova, jejichž hádání by také bylo neúměrně složité. Z důvodu zachování konzistentního chování aplikace by mělo být nutné tuto nepísmennou entitu hádat ve všech variantách testu, včetně hádání po slovech.

2.3 Pokusné texty

Stejně jako během Shannonových výzkumů budou v testu použity beletrické texty. Odborná literatura nebo novinové články zpravidla vykazují větší entropii [3]. Pro zkoumané texty byl zvolen požadovaný rozsah přibližně 50 kB čistého textu, který by měl zajistit náležitou obtížnost testu při zachování rozumných nároků na aplikaci. Kratší texty by usnadnily test v tom ohledu, že by uživatel vybíral z výrazně omezeného počtu entit, jejichž složení by se zdaleka nepodobalo souhrnným vlastnostem jazyka. Bylo by tak při znalosti předchozího textu snadnější vybrat odpovídající entitu. Naopak delší texty by zvýšily hardwarovou náročnost aplikace, především z hlediska práce se statistikami entit v prostředí webového prohlížeče během samotného testu. Přitom samotný vliv na výsledky testu by již při případném zvýšení délky použitých textů neměl být výrazný. Primárním cílem práce je zkoumání vlastností českých textů, pro ilustraci a porovnání zjištěných hodnot pak poslouží texty anglické.

Zkoumané texty budou rozděleny do tří částí, označených jako A, B a C. Část A je tvořena úvodní pasáží textu v rozsahu několik vět až jednoho odstavce a bude uživateli poskytnuta na začátku testu. Část B je určena k hádání a bude ji tvořit vždy 10 slov, bez započtení nepísmenných slov, interpunkce a prázdných znaků. Část C pak tvoří zbytek textu.

Z částí B a C budou tvořeny souhrnné statistiky, které uživateli mohou sloužit jako nápověda v průběhu testu. Kromě přehledu všech dostupných entit bude k dispozici také počet jejich výskytu v celém textu.

Kapitola 3

Programátorská dokumentace

3.1 Výběr platformy

Z hlediska výběru platformy pro realizaci testovací aplikace byl rozhodující fakt, že test by měl být dostupný širokému spektru uživatelů. Z toho pohledu se jako nejvhodnější provedení jeví forma webové aplikace. Uživatelé díky tomu nejsou nuceni cokoliv instalovat do svého počítače a mohou se pohybovat v důvěrně známém prostředí webové prohlížeče. To může mít značný pozitivní vliv na ochotu oslovených respondentů test absolvovat.

Mezi dostupnými platformami pro vývoj online aplikace bylo zvoleno řešení postavené na open-source technologiích. Programovou část obstarává skriptovací jazyk PHP, jako úložiště dat slouží databázový systém MySQL. Tato kombinace je pravděpodobně nejrozšířenější platformou pro tvorbu online aplikací. Výhodou je především snadná přenositelnost aplikací a široká nabídka serverů, na kterých je možné takové aplikace provozovat s nízkými či dokonce nulovými náklady. Tvorba aplikace v tomto prostředí je také relativně snadná, pro robustní řešení s vysokými nároky na výkon či bezpečnost však nejde o optimální platformu. Nicméně pro potřeby vytvářené testovací aplikace jde o dostupnou a plně postačující variantu.

3.2 Zpracování textů

Rozdělení textu na požadované entity je prováděno jednorázově při založení nového testu. Jde o časově náročnou operaci, jejíž provádění při každém spuštění textu by zpomalovalo odezvu aplikace a z hlediska uživatelů by působilo negativně. Uložení textů v databázi po jednotlivých entitách je sice prostorově náročnější, zato však umožňuje snadné a efektivní provádění všech potřebných operací prostřednictvím standardních databázových prostředků. Implementace vlastního formátu pro ukládání dat by buď vyžadovala využití pokročilých datových struktur, nebo by se negativně projevila na výkonu aplikace.

Jeden z textů použitých pro testování o velikosti 50 kB byl rozdělen na 88118 entit, z toho 47169 písmen, 24435 slabik a 16514 slov. Vzhledem k tomu, že k testování není nutné používat větší počet unikátních testů než řádově jednotky, jsou kapacitní možnosti moderních databázových systémů v tomto ohledu zcela dostačující.

Algoritmus dělení textu na entity začíná jediným průchodem přidávaným textem, během kterého postupně načítá jednotlivé znaky. Z nich vytváří vytváří písmenné entity a zároveň hlídá hranice mezi slovy. Každé nalezené slovo je poté uloženo jako slovní entita a následně rozděleno na jednotlivé slabiky. Nepísmenná slova jsou zároveň považována za slabiku.

Pro dělení na slabiky je využit algoritmus Universal Middle-left, popsaný v [5]. Pro český jazyk vykazuje zhruba 94% a v případě jazyka anglického zhruba 93% správnost dělení vzhledem k přirozenému jazyku.

Případné rozšíření aplikace pro testování dalších jazyků je nutné provést prostřednictvím úprav ve zdrojovém kódu, protože především pravidla pro dělení textu na slabiky se v jednotlivých jazycích liší. Přidání dalšího jazyka tedy vyžaduje naprogramování dělícího algoritmu pro požadovaný jazyk.

3.3 Problematika kódování a znakových sad

Zásadním faktorem pro správné fungování aplikace je korektní práce s národními znakovými sadami. Od prvotního přidání pokusného textu, přes zobrazování dat uživateli až po vyhodnocování výsledků musí všechny části aplikace správně pracovat s použitým kódováním. Pro webové aplikace, budované na kombinaci technologií PHP a MySQL, se jako vhodné kódování jeví kódování UTF-8 z rodiny Unicode. Je dobře podporováno jak na straně serverových technologií, tak v internetových prohlížečích [6].

Správného zobrazení textu v prohlížečích a tím i vhodného kódovaní dat vstupujících přes webové rozhraní lze dosáhnout definicí znakové sady ve zdrojovém kódu stránky pomocí zápisu

```
<meta http-equiv="Content-Type: text/html; charset=utf8" />
```

při současném uvedení kódování v hlavičce načítaného dokumentu. V PHP slouží k odeslání hlavičky informující o kódování souboru následující příkaz [7]:

```
Header("Content-Type: text/html; charset=utf8");
```

Korektní zpracovávání řetezců zakódovaných pomocí kódování UTF-8 zajišťuje knihovna řetězcových funkcí Multibyte String [7]. Ta poskytuje univerzální funkce pro práci s vícebytovými znakovými sadami, před použitím některé z funkcí je tak ještě nutno použité kódování nastavit příkazem

```
mb_internal_encoding("UTF-8");
```

V databázi MySQL je pro práci s texty v českém jazyce definováno porovnávání (angl. collation) utf8_czech_ci. To zajišťuje správné abecední řazení podle českých pravidel [8], bohužel ale není pro práci s českými texty dostačující. Při porovnávání totiž chybně nerozlišuje znaky s diakritikou a znaky bez diakritiky [9], což je chování nepřípustné pro účely této práce.

Řešení spočívá v použití takového porovnávání, které diakritiku rozlišuje, jakým je například univerzální utf8_bin. To bylo v databázi použito jako výchozí porovnávání pro sloupec obsahující texty jednotlivých entit. Pak je ale nutné při práci s daty explicitně vyžadovat použití českého porovnávání v příkazech, jejichž výstup by měl odpovídat korektnímu českému abecednímu řazení. K tomu slouží direktiva COLLATE, jejíž použití demonstruje následující příklad:

```
SELECT obsah FROM entity ORDER BY obsah COLLATE utf8_czech_ci;
```

Každé spojení s databází je vhodné zahájit provedením následujícího příkazu, který zajistí použití správné znakové sady na straně databázového serveru [10]:

```
SET NAMES utf8;
```

3.4 Průběh testu

Po spuštění testu vybere aplikace takový text, se kterým daný uživatel dosud nepracoval. Pokud již vyčerpal všechny dostupné texty, nemůže být test spuštěn. Uživateli je zobrazena úvodní část hádaného textu a klientskému skriptu, který obsluhuje poskytování návrhů během hádání a práci se statistikami, jsou poskytnuta potřebná data o výskytech jednotlivých entit. Ukládání výsledků do databáze probíhá průběžně během samotného testu. Při každém uživatelově tipu zjišťuje klientský skript správnost zadané entity dotazem na server, kde tak může být uložena informace o dalším využitém pokusu. Při nesprávném pokusu je pouze zvýšen čítač pokusů, při správném tipu je entita označena jako uhodnutá.

Počítání pokusů využívá příkazu databáze MySQL `INSERT ... ON DUPLICATE UPDATE`. Ten je k dispozici až od verze MySQL 4.1 [10]. Při prvním pokusu na každé entitě je přidán záznam do tabulky výsledků, při dalších pokusech je již pouze zvyšován čítač.

3.5 Klientská část

Interaktivní část testu zajišťuje skriptovací jazyk JavaScript s využitím postupů souhrnně označovaných jako AJAX [1]. Ty umožňují, aby stránka byla průběžně upravována na základě komunikace se serverem, což je využito pro ověřování správnosti zadaných tipů. Tento přístup poskytuje dvě hlavní výhody – interaktivní chování testu lze zajistit bez zdlouhavého opakovaného načítání celé stránky, přesto je ale celé řešení ochráněno proti potenciálním podvodům během testu. Komunikace se serverem je realizováno na základě skriptu popsaného v [12].

Server na začátku testu klientskému skriptu neposkytuje žádnou informaci o správném řešení, což eliminuje nebezpečí podvodů uživateli, kteří by byli schopni klientský skript modifikovat. Jedinou možností, jak zjistit správné řešení, je postupné dotazování serveru na správnost jednotlivých tipů. Obsluhu těchto dotazů na straně serveru má na starosti skript `test-data.php`. V případě správného tipu vrací server kromě potvrzení správnosti také skutečný obsah hádané entity. Díky tomu je možné doplňovat na straně klienta uhodnuté entity přesně v takové podobě, v jaké se vyskytují v původním textu. Ačkoliv tedy z hlediska hádajícího uživatele test nerozlišuje velikost písmen, uhodnuté znaky doplňuje ve správné velikosti. Hlavním důvodem tohoto řešení je pak správné doplnění interpunkce, která je pro účely hádání seskupena do jedné speciální entity. Do uhodnutého textu musí být doplněna v přesném znění, aby měl uživatel jednoznačnou informaci o struktuře hádaného textu.

Řazení statistik v uživatelském rozhraní dle požadavků uživatele využívá postup popsaný v [13].

3.6 Instalace

Systémové požadavky jsou následující:

- Webový server Apache s funkčním modulem mod_rewrite
- PHP s nainstalovanou knihovnou Multibyte String (mbstring)
- MySQL verze 4.1 nebo vyšší

Soubory potřebné k instalaci jsou k dispozici v adresáři `instalace` na přiloženém CD . Nejprve je nutné založit databázi a nastavit uživatelský účet s právem zápisu do databáze. Strukturu databáze lze založit spuštěním skriptu, uloženého v souboru `struktura-db.sql`. V podadresáři `www` se nachází soubory určené k nahrání do kořenového adresáře webového serveru, například prostřednictvím protokolu FTP. Web využívá soubor `.htaccess` pro změnu potřebných nastavení jednotlivých adresářů,

[1] Zkratka doslovně znamená Asynchronous JavaScript And XML [11]. Ve skutečnosti zde server nevrací odpověď ve formátu XML, označení AJAX se však vžilo obecně pro aplikace, ve kterých JavaScript získává data ze serveru bez nutnosti obnovovat celou stránku.

jeho použití tedy musí být povoleno prostřednictvím direktivy `AllowOverride` v nastavení webového serveru. V souboru `includes/db.inc.php` je třeba upravit údaje pro přístup k databázovému serveru dle následujícího vzoru:

```
mysql_connect("server", "uživatel", "heslo");
```

Administrační rozhraní nevyužívá komponenty ostatních částí webu a je potřeba, aby adresář admin byl samostatně zabezpečen proti neoprávněnému přístupu. Výchozí nastavení využívá HTTP autentizaci a je proto nutné zde vytvořit soubor `.htpasswd`, obsahující informace o oprávněných uživatelích. Pokud je k dispozici terminál serveru, lze tento soubor vytvořit spuštěním příkazu

```
htpasswd -c .htpasswd přihlašovací_jméno
```

Příkaz je nutné spustit přímo v adresáři admin, nebo upravit cestu k souboru .htpasswd. Případní další uživatelé se následně přidávají příkazem

```
htpasswd .htpasswd jméno_dalšího_uživatele
```

V závislosti na nastavení webového serveru může být nutné zabezpečení adresáře nastavit v administračním rozhraní poskytovatele webhostingu, případně je možné soubor .htpasswd vytvořit k tomu určenými utilitami a na server jej nahrát prostřednictvím FTP.

Po provedení výše uvedených kroků bude testovací web funkční, nebude ale obsahovat žádné testy. Popis přidání textů pro test je součástí uživatelské dokumentace.

Kapitola 4

Uživatelská dokumentace

Aplikace je určena k měření entropie psaného textu. Měření probíhá v online aplikaci, přístupné prostřednictvím webového prohlížeče, a má formu hry, ve které je úkolem hráče postupné odhadování dalšího pokračování textu.

4.1 Registrace a přihlášení

Pro přístup k testu je nutné založení uživatelského účtu a přihlášení. Registrace je vázána na unikátní adresu elektronické pošty, kromě hesla je vyplnění dalších údajů nepovinné a slouží jen pro účely vyhodnocení experimentu. Ovládací prvky nutné k provedení registrace i k přihlášení jsou k dispozici v pravé horní části každé stránky aplikace.

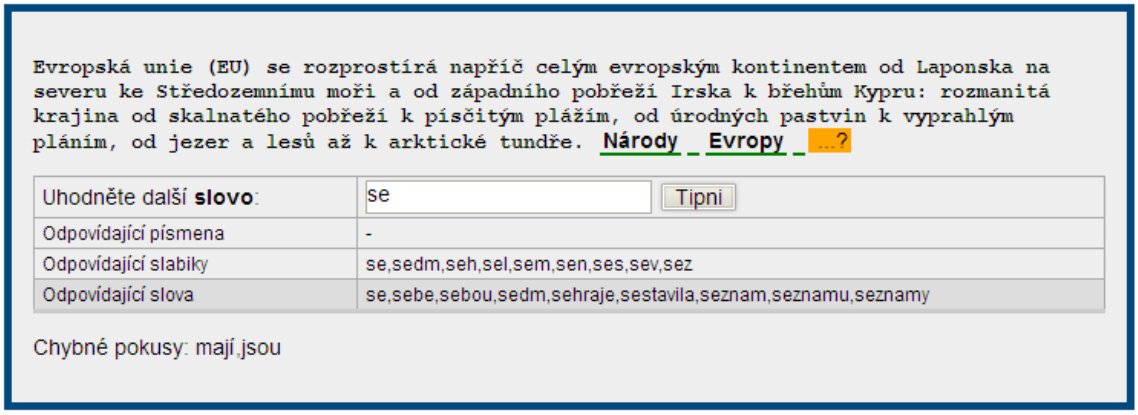
Po přihlášení přibude v hlavní nabídce položka Testy. Po jejím zvolení se zobrazí stránka s přehledem dostupných testů. S každým pokusným textem může jeden uživatel pracovat pouze jednou, bez ohledu na zvolenou délku entit. Informace o počtu zbývajících testů je umístěna na začátku přehledu. Pokud uživateli zbývá alespoň jeden dostupný text, může si zvolit jeden z typů entit a kliknutím na odkaz spustit test.

4.2 Práce s testem

Samotný test probíhá v orámovaném prostoru stránky. Zde je zobrazena úvodní pasáž textu, jehož pokračování uživatel hádá. Symbol zvýrazněného otazníku pomáhá ujasnit návaznost odhadovaného textu na zobrazený úvod. Cílem je pouze správné doplnění textu s využitím co možná nejmenšího počtu pokusů, čas potřebný k dokončení nehraje roli.

Ovládací prvky testu tvoří vstupní textové pole, určené k zadávání tipovaných entit, a tlačítko pro potvrzení tipu. Po zadání tipu je možné jej potvrdit jak kliknutím

na tlačítko, tak stiskem klávesy Enter. Tipovat lze pouze takové entity, které se v textu skutečně vyskytují, což aplikace po každém zadaném znaku kontroluje. Pokud zadanému textu žádná entita požadovaného typu neodpovídá, nelze tip potvrdit.



Obrázek 4.1: Průběh testu

Po stisknutí tlačítka odešle aplikace dotaz na server, kde se uživateli započítá provedený tip. Zpět putuje informace o správnosti tipu, která je vzápětí uživateli zobrazena. Pokud byl tip správný, doplní se uhodnutá entita k úvodnímu textu a hádání pokračuje další entitou. V pravém sloupci je k dispozici tabulka uhodnutých entit a potřebného počtu pokusů. Nesprávný tip je zařazen do seznamů nevyhovujících entit a uživateli je nabídnut další pokus. Pokud je entita označena jako nesprávná, nelze ji až do uhodnutí správného řešení tipnout znovu.

Aplikace nerozlišuje velikost písmen, tipy lze tedy zadávat v libovolné velikosti. Záleží naopak na diakritice, entity musí být z tohoto hlediska zadány v přesném znění. Tečka zastupuje veškerou interpunkci, mezery mezi slovy i případná nepísmenná slova, jako jsou např. čísla či speciální znaky. Na konci každého slova tak stačí zadat tečku a aplikace text patřičně doplní. Místo tečky lze zadat také čárku či mezeru.

V tabulce pod vstupním polem se zobrazují návrhy jednotlivých entit, které odpovídají textu zadanému v tipovacím poli. Hádaný typ entit je pak pro lepší orientaci zvýrazněn odlišnou barvou pozadí. Pokud není možné odpovídající entity vypsat do jednoho řádku tabulky, zobrazí se číselná informace o jejich celkovém počtu.

Uživateli jsou v průběhu testu k dispozici statistiky zkoumaného textu. V prostoru pod samotným testem jsou v tabulkách vypsány všechny entity jednotlivých druhů společně s počtem jejich výskytů v celém textu. Pomocí tlačítek lze tabulky seřadit buďto právě podle počtu výskytů jednotlivých entit, nebo abecedně. S ohledem na výkon počítače může seřazení trvat až několik vteřin [1], během kterých nelze s testem dále pracovat.

[1] Seřazení statistik pro test, tvořený textem o velikosti 52 kB (93325 entitami všech druhů), trvalo na počítači s procesorem Intel Core 2 T5500, 1 GB RAM a operačním systémem MS Windows XP v prohlížeči Internet Explorer 7 přibližně 7 sekund, prohlížeč Opera 9.5 zvládl stejný úkol v čase přibližně 2,5 sekundy.

Pomocí zaškrtávacího pole je možné zapnout označování neplatných pokusů v tabulkách statistik. Zapnutí této funkce může také na pomalejších počítačích zvýšit prodlevy v průběhu testu, konkrétně během vyhodnocování správnosti odeslaného tipu. Pokud je funkce aktivní, neplatné pokusy jsou v tabulce označeny červenou barvou a přeškrtnutím. Samotné zaškrtnutí volby nezpůsobí označení dosud zadaných tipů, to proběhne až při dalším zadání nebo po seřazení statistik prostřednictvím k tomu určených tlačítek.

Pokud uživatel test úspěšně dokončí, ovládací prvky se stanou neaktivními a zobrazí se informace o dokončení testu. Zároveň se zobrazí odkaz, vybízející k návratu na stránku s přehledem testů.

4.3 Administrace testů

Správa testů a vyhodnocování výsledků se provádí váři admin, který je chráněn proti neoprávněnému přístupu. Po otevření adresáře `/admin/` ve webovém prohlížeči a po úspěšném přihlášení se zobrazí stránka, obsahující dvě volby. První odkaz slouží k prohlížení dosavadních výsledků testů, druhý odkaz pak směřuje na stránku určenou k administraci testů.

Stránka výsledků obsahuje souhrnné informace o všech probíhajících testech. Pro jednotlivé texty jsou zde vypsány průměrné hodnoty počtu pokusů potřebných k uhodnutí příslušných entit a zároveň je k dispozici volba podrobného náhledu na výsledky konkrétního testu. Zde jsou pak vypsány všechny proběhlé testy až do úrovně jednotlivých pokusů. Pokud je počet pokusů uveden v hranatých závorkách, znamená to, že uživatel ani po uvedeném počtu pokusů entitu neuhodl.

Stránka administrace testů nabízí dvě funkce. Formulář umístěný v horní části stránky umožňuje založení zcela nového testu. Pro přidání je třeba vyplnit název testu, který slouží pouze pro pozdější vyhodnocování a uživatelům se nezobrazuje, a vybrat jazyk testu. Následně se kliknutím na tlačítko test přidá do seznamu existujících testů. Poté je ještě třeba do testu přiřadit pokusný text.

Níže na stránce je umístěna tabulka obsahující výpis všech dosud založených testů. U testů s nulovým počtem přiřazených entit je k dispozici odkaz na stránku, kde je možné vložit text určený pro příslušný test. K dispozici není žádná volba pro úpravy existujících testů, aby byla zajištěna konzistence získaných výsledků s aktuální podobou testu.

Přidání experimentálního textu vyžaduje vyplnění dvou formulářových polí. Do prvního z nich se zadává úvodní část textu, která bude uživateli zobrazena na začátku experimentu jako nápověda. Kompletní zbytek část textu se vkládá do druhého pole, testovací část již aplikace oddělí automaticky. Přidání textu je, vzhledem k jeho rozdělování na tři druhy entit a jejich postupné přidávání do databáze, časově náročnou operací. U rozsáhlejších textů nebo v případě limitujícího nastavení serveru tak

může během zpracovávání vypršet čas povolený pro běh PHP skriptu. Pokud k tomu dochází, je třeba v nastavení serveru zvýšit hodnotu povolené doby běh skriptu.

Pokud dojde během ukládání textu do databáze k chybě nebo je třeba nahradit již vložený text, musejí být před opakovaným zadáním z databáze odstraněny již vložené entity. To lze provést spuštěním následujícího SQL dotazu:

```
DELETE FROM entity WHERE testy_idtesty = Idtestu;
```

Za Idtestu je zde třeba dosadit unikátní číselný identifikátor testu, který se zobrazuje v první řádce stránky pro přidání entit či v tabulce s přehledem testů. Tato operace by neměla být provedena v případě, že byl test již vyplněn některým z uživatelů, protože by tím byl pokus znehodnocen.

Kapitola 5

Výsledky testu

Testu se zúčastnilo 42 uživatelů, kteří celkem vyplnili 89 testů. Zcela dokončeno bylo pouze 25 testů, většinu tedy uživatelé opustili předčasně. Dle údajů, které o sobě uživatelé uvedli, bylo nejmladšímu účastníkovi testu 20 let a nejstaršímu pak 66 let. Testy zkoumající texty v anglickém jazyce vyplňovali především rodilí mluvčí, kteří byli za tím účelem osloveni.

Test probíhal na webové adrese `http://shannon.martinsmrt.com/`.

5.1 Entropie českého textu

V jednom z testů byl použit text z díla Ohyzdný duch od Williama S. Burroughse [14]. Uživateli byla známa tato úvodní pasáž textu:

„Kapitán Mission si přes rameno přehodil dvouhlavňovou křesadlovku, kterou měl vždy nabitou dvěma kulemi. K pasu si připevnil pochvu s námořnickým tesákem. Pak vzal své zavazadlo a kráčel kolonií, zastavuje se chvílemi na kus řeči s tím či oním osadníkem. Podařilo se jim najít vynikající rudý cihlářský jíl a budovali nyní jednopatrové cihlové stavby s verandami v patře, podpírané sloupy z mohutných kmenů tropických stromů.“

Následující tabulka ukazuje výsledek testu vybraného uživatele. První řádek vždy obsahuje hádanou entitu, na druhém řádku je uveden počet pokusů, které uživatel potřeboval k uhodnutí.

Písmeno	B	u	d	o	v	y		n	a	v	a	z	o	v	a	l	y	
Pokusů	3	5	1	1	1	2	1	3	1	11	1	1	1	1	1	1	1	1

Písmeno	j	e	d	n	a		n	a		d	r	u	h	o	u		a	
Pokusů	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Písmeno	t	v	o	ř	i	l	y		ř	a	d	u		d	o	m	ů	
Pokusů	12	17	2	1	1	1	1	1	14	1	1	2	1	11	1	2	8	1

Tabulka 5.1: Příklad výsledku testu prováděného po písmenech

Pokud lze z jednoho testu vyvozovat závěry, pak je zde zřejmá vyšší náročnost tipování na začátku slova a naopak vysoká úspěšnost uživatele v souvislých pasážích, kde znalost předchozího textu umožňila velmi přesné odhadování následujících entit.

Tabulka B.1 v příloze B ukazuje souhrnné výsledky testů zaměřených na písmena provedených na tomto textu. První sloupec obsahuje hádanou entitu, druhý sloupec ukazuje relativní četnost výskytu této entity v celém textu a ve třetím sloupci je uveden průměrný počet pokusů, který všichni účastníci testu potřebovali k uhodnutí příslušné entity. Do průměru jsou započteny jen ty pokusy, které vedly k úspěšnému uhodnutí, nedokončené tipy nejsou započítány. Výsledná hodnota entropie je aritmetickým průměrem dílčích průměrných počtů pokusů a činí 2,89 bitu/znak. Souhrnný výpočet přes všechny použité texty v českém jazyce udává výsledný odhad hodnot entropie ve výši 2,98 bitu/znak.

Testy prováděné na slabikách a slovech se během testování ukázaly jako neúnosně náročné. Ze 40 zahájených testů byly dokončeny pouhé dva, jeden pro slabiky a jeden pro slova. Většinu uživatelů odradilo již několik prvních pokusů, během kterých si pravděpodobně uvědomili náročnost testu a ztratili zájem dále pokračovat.

Výsledná entropie slabik získaná výpočtem z jediného dokončeného textu činí 8,88 bitu/slabiku. Pokud započteme i částečně dokončené testy pro všechny zkoumané texty, činí odhad entropie 9,02 bitu/slabiku.

Jediný dokončený test zkoumající entropii slov poskytuje výsledný odhad entropie ve výši 15,32 bitu/slovo. Nedokončené testy zde nemá smysl brát v úvahu. Uživatelé v nich uhodli v nejlepším případě pouze několik slov, následně však test přerušili po velkém počtu pokusů, které nevedly k uhádnutí slova.

Výsledky obou dokončených textů pro slabiky a slova jsou k dispozici v příloze B. Test prováděný po slabikách je uveden v tabulce B.2 a test prováděný po slovech v tabulce B.3. V obou případech obsahuje první sloupec vždy hádanou entitu, druhý sloupec vyjadřuje počet pokusů potřebných k uhodnutí. Obě tabulky vypovídají o značné náročnosti testů, absolvování uvedeného počet pokusů vyžaduje velkou dávku trpělivosti. V těch částech textu, kde panuje velká nejednoznačnost možného dalšího pokračování, musí uživatel vyzkoušet mnoho variant. Test provedený po slabikách v tabulce B.2 zároveň ukazuje, že uváděné statistiky byly pro uživatele užitečnou nápovědou. Ačkoliv v češtině existuje velké množství slov začínajících předponou vy-, uživatel po uhodnutí první slabiky doplnil všechny další slabiky slova „vyvolili“

na první pokus, přestože nejde o hojně používané slovo. V testu zaměřeném na slova v tabulce B.3 také naleznema slova uhodnutá na první pokus. Zde uživatel pravděpodobně uplatnil zejména znalost předchozího textu v kombinaci s obecným citem pro jazyk a povědomím o běžně používaných frázích.

5.2 Entropie anglického textu

Testy probíhaly také s texty psanými v anglickém jazyce. Relevantní výsledky se podařilo získat pouze pro testy s písmeny, na dalších testech se projevil nedostatek účastníků a také jejich nízká motivace k plnění náročného úkolu, jakým dokončení testu pro slabiky nebo slova bylo. Žádný z testů pro slabiky a slova nebyl v anglickém jazyce dokončen.

Odhad entropie spočítaný na základě všech vyplněných testů pro písmenné entity činí v anglickém jazyce 2,60 bitu/znak, což je hodnota nižší než jaký je získaný odhad entropie písmen pro český jazyk. Ačkoliv je relevance výsledných hodnot vzhledem k nízkému počtu účastníků testu velmi malá, tento výsledek potvrzuje očekávání. Zřejmou příčinou vyšší entropie českého textu je větší počet znaků české abecedy oproti abecedě anglické. Roli pravděpodobně hraje také složitější české tvarosloví.

V testu byl použit také jeden text, který byl k dispozici v českém i anglickém překladu. Šlo o text informačního materiálu Evropské unie, věnovaného cestování po Evropě [15]. Zjištěná entropie činí 2,47 bitu/znak v případě českého překladu a 1,68 bitu/znak pro anglickou verzi.

Kapitola 6

Závěr

Primárním cílem této práce bylo vytvoření testovací aplikace pro zkoumání entropie psaného textu. Zamýšlený experiment byl realizován a aplikace v jeho průběhu prokázala svou funkčnost. Povedlo se vytvořit systém pro zpracování pokusných textů a navrhnout funkční uživatelské rozhraní. To je přístupné prostřednictvím webového prohlížeče a užití aplikace tak není omezeno na konkrétní softwarovou platformu, ani nevyžaduje jakékoliv zásahy do uživatelova počítače.

Po vytvoření aplikace byl proveden experiment, jehož výstupem měly být odhady entropie dílčích entit českého jazyka a angličtiny. Dílčími entitami rozumíme písmena, slabiky a slova. Během testu se ukázalo, že obtížnost odhadování textu po slabikách a slovech je vyšší, než jsou uživatelé obecně ochotni akceptovat.

Vzhledem k tomu, že účastníci se testu věnovali ve svém volném čase pouze na základě osobní či zprostředkované prosby a nebyli k dokončení nijak výrazně motivováni, opouštěli velmi často experiment předčasně. Prostředí webové aplikace umožňuje velmi jednoduché okamžité opuštění textu.

Množství nedokončených testů snižuje statistickou věrohodnost získaných výsledků. Do budoucna je tedy potřeba najít způsoby, jak účastníky testu motivovat k jeho dokončení, a zároveň nabídnout další mechanismy, které by uživatelům mohly zadaný úkol usnadnit.

Možnosti pro další vylepšení testovací aplikace lze spatřit jak na straně uživatelského rozhraní, tak v samotné programové realizaci. Účastníkům testu by v plnění úkolu mohly pomoci další informace o vlastnostech hádaného textu, propojené s pokročilými nástroji pro práci s těmito statistikami. Aplikace by mohla být rozšířena tak, aby ve výchozí podobě podporovala širší spektrum jazyků pro provádění testů. V současné podobě by přidání dalšího jazyka vyžadovalo nezanedbatelné zásahy do programového kódu.

Ačkoliv nejsou naměřené údaje z výše uvedených důvodů zcela spolehlivé, výsledky experimentu v obecné rovině souhlasí s předběžnými odhady. Potvrdila se očekávaná vyšší entropie českého textu oproti anglickému, způsobená jednak vyšším počtem znaků české abecedy, jednak komplexnějším českým tvaroslovím.

Literatura

[1] Benešová, J., et al (1999): *Všeobecná encyklopedie v osmi svazcích*. Diderot, Praha.

[2] Shannon, C. E. (1948): A Mathematical Theory of Communication. *Bell System Technical Journal* **27**, 379–423, 623–656.

[3] Shannon, C. E. (1951): Prediction and Entropy of Printed English. *Bell System Technical Journal* **30**, 50–64.

[4] Lánský, J. (2006): Slabiková komprese textových dat, In: Vojtáš, P., Skopal, T., (eds.), *Datakon 2006, sborník z konference*. Masarykova univerzita, Brno, 209–218.

[5] Lánský, J. (2005): *Slabiková komprese*. Diplomová práce. Katedra softwarového inženýrství MFF UK, Praha.

[6] Wood, A. (2007): *Unicode and Multilingual Web Browsers*, `http://www.alanwood.net/unicode/browsers.html`

[7] *Dokumentace jazyka PHP*, `http://www.php.net/docs.php`

[8] Vláda České republiky (1995): *Usnesení Vlády České republiky o technickém standardu státního informačního systému České republiky*, dostupné na serveru `http://kormoran.vlada.cz/`

[9] *Popis chyby v databázovém systému MySQL*, dostupný na `http://bugs.mysql.com/bug.php?id=32404`

[10] *Dokumentace databázového systému MySQL*, `http://dev.mysql.com/doc/`

[11] Garrett, J. J. (2005): *Ajax: A New Approach to Web Applications*, `http://www.adaptivepath.com/ideas/essays/archives/000385.php`

[12] *Výukový text věnovaný technologii AJAX*, dostupný na `http://www.w3schools.com/Ajax/Default.Asp`

[13] Ondra, L. (2004): Seřazení tabulky JavaScriptem snadno a rychle, *Interval.cz*, `http://interval.cz/clanky/serazeni-tabulky-javascriptem-snadno-a-rychle/`

[14] Burroughs, W. S.: *Ohyzdný duch*. Votobia, Olomouc, 1995. Elektronická podoba získána ze serveru http://literatura.jinak.cz/

[15] Evropská komise (2008): *Cestování po Evropě*, brožura dostupná na http://ec.europa.eu/publications/booklets/eu_glance/72/index_cs.htm

Příloha A

Obsah přiloženého CD

Součástí této práce je přiložený kompaktní disk. Kromě elektronické verze samotné práce obsahuje vytvořenou testovací aplikaci a úplné výsledky provedených testů.

- Adresář `instalace` obsahuje soubory potřebné ke spuštění testovací aplikace na webovém serveru, který splňuje požadavky uvedené v programátorské dokumentaci.
- Adresář `prace` obsahuje elektronickou verzi této práce ve formátu PDF.
- Adresář `texty` obsahuje veškeré texty použité pro testování. Soubor `seznam.txt` obsahuje přehled textů, v dalších souborech již je pouze přesné znění textu, které bylo použito pro test.
- Adresář `vysledky` obsahuje výsledky všech proběhlých testů ve formátu CSV, rozdělené v souborech podle jednotlivých testů.

Příloha B

Vybrané výsledky testů

V této příloze jsou uvedeny tabulky, doplňující text kapitoly 5 věnované výsledkům testů.

Tabulka B.1 představuje souhrnné výsledky pro jeden z testů prováděných po písmenech. První sloupec obsahuje hádanou entitu, druhý sloupec ukazuje relativní četnost výskytu této entity v celém textu a ve třetím sloupci je uveden průměrný počet pokusů, který všichni účastníci testu potřebovali k uhodnutí příslušné entity. Do průměru jsou započteny jen ty pokusy, které vedly k úspěšnému uhodnutí, nedokončené tipy nejsou započítány.

Tabulky B.2 a B.3 obsahují výsledky jediných dokončených testů pro slabiky, resp. slova. V prvním sloupci je vždy uvedena hádaná entita, druhý sloupec obsahuje počet pokusů potřebných k uhodnutí.

Entita	**Relativní četnost**	**Entropie**
B	1,40 %	10,70
u	2,74 %	5,11
d	3,00 %	1,00
o	6,84 %	1,22
v	3,45 %	1,22
y	1,74 %	1,67
	16,40 %	3,67
n	5,00 %	7,78
a	5,72 %	2,11
v	3,45 %	6,67
a	5,72 %	2,89
z	2,12 %	1,00
o	6,84 %	1,11
v	3,45 %	1,00
a	5,72 %	1,11
l	4,08 %	1,00
y	1,74 %	1,22
	16,40 %	1,00
j	1,69 %	14,89
e	6,41 %	1,56
d	3,00 %	2,44
n	5,00 %	1,11
a	5,72 %	2,00
	16,40 %	1,56

Entita	**Relativní četnost**	**Entropie**
n	5,00 %	1,44
a	5,72 %	1,00
	16,40 %	1,00
d	3,00 %	1,00
r	2,84 %	1,00
u	2,74 %	1,00
h	2,25 %	1,00
o	6,84 %	1,00
u	2,74 %	1,00
	16,40 %	1,00
a	5,72 %	5,11
	16,40 %	3,22
t	4,09 %	7,11
v	3,45 %	12,38
o	6,84 %	2,63
ř	1,05 %	1,14
i	3,68 %	1,00
l	4,08 %	1,00
y	1,74 %	1,00
	16,40 %	1,00
ř	1,05 %	17,86
a	5,72 %	1,14
d	3,00 %	1,00
u	2,74 %	1,14
	16,40 %	1,00
d	3,00 %	6,71
o	6,84 %	1,00
m	2,98 %	1,29
ů	0,35 %	2,14
	16,40 %	1,00
s	3,99 %	5,14
	16,40 %	1,29

Tabulka B.1: Entropie písmen pro jeden z textů

Entita	**Entropie**
Byl	66
	6
jsem	2
	1
též	23
	1
jed	33
ním	1
	1
z	1
	1
nich	1
,	1
a	24
	1
o	26
ni	8
	1
mě	9
	1
vy	10
vo	1
li	1
li	1
,	1

Tabulka B.2: Výsledek testu prováděného po slabikách (text `cze06.txt`)

Entita	**Entropie**
Když	24
	1
se	26
	1
naskytne	74
	1
případ	1
	1
vyžadující	10
	1
schopnosti	119
,	1
jež	7
	1
některý	19
	1
z	1
	1
nás	1

Tabulka B.3: Výsledek testu prováděného po slovech (text `cze02.txt`)