Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-0000-00000

Michal Fašánek

Pokročilé spracovanie sekvenčných dát pomocou umelých neurónových sietí

Diplomová práca

Vedúci práce: Ing. Michal Barla, PhD

Január, 2016

Anotácia

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ

Študijný program:	Informačné systémy
Diplomová práca:	Pokročilé spracovanie sekvenčných dát pomocou umelých neurónových sietí
Autor:	Michal Fašánek
Vedúci práce:	Ing. Michal Barla, PhD

Január, 2016

Analýza dát z používateľského správania k zdrojom na webe je v súčasnosti populárna téma, vzhľadom na svoj potenciál zlepšovať služby poskytované návštevníkom webu. Najnovšie prístupy skúmajú aj možnosti aplikácie metód strojového učenia. Medzi týmito prístupmi si získavajú popularitu hlboké mnohovrstvové samoučiace sa neurónové siete a rôzne architektúry rekurentných neurónových sietí. Využívajú princípy učenia bez učiteľa, pomocou ktorých dokážeme v jednotlivých vzorkách dát identifikovať podstatné črty. V tejto práci sa zameriavam na možnosti využitia rekurentných neurónových sietí s dlhou krátkodobou pamäťou(LSTM) pri analýze dát z platobných brán pre používanie online spravodajských portálov. Takáto analýza poskytuje náhľad do používateľského správania a z toho vyplývajúce možnosti spätnej väzby voči návštevníkom online spravodajských portálov.

Annotation

Slovak University of Technology Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

Study program:	Informačné systémy
Master thesis:	Advanced processing of sequential data by artificial neural networks
Author:	Michal Fašánek
Supervisor:	Ing. Michal Barla, PhD
2016, January	

Data analysis of user behaviour in accessing web sources has become a popular topic, due to its potential in improvement of services offered to the visitors of web. Most recent approaches examinate possibilities of applying methods machine-learning. Among these approaches, deep belief networks and reccurent neural networks are gaining popularity. They work with principles of unsupervised learning to identify important patterns in given data. In this paper, I focus on using reccurent neural networks with long short-term memory(LSTM) to analyze data from paywall of online news portals. Such analysis provides insight into the user behaviour and resulting feedback possibilities to the users of online news portals.

Obsah

Kapitola 1

Úvod

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat.

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum. Typi non habent claritatem insitam; est usus legentis in iis qui facit eorum claritatem. Investigationes demonstraverunt lectores legere me lius quod ii legunt saepius.

Claritas est etiam processus dynamicus, qui sequitur mutationem consuetudium lectorum. Mirum est notare quam littera gothica, quam nunc putamus parum claram, anteposuerit litterarum formas humanitatis per seacula quarta decima et quinta decima. Eodem modo typi, qui nunc nobis videntur parum clari, fiant sollemnes in futurum.

Kapitola 2

Analýza

V tejto časti sa venujeme dôkladnej analýze podkladov. Jednotlivé časti sú popísané v rozsahu relevantnom pre túto prácu. Analýza je štrukturovaná na nasledovné časti:

- Problémová oblasť
- Dáta sprístupnené pre prácu
- Neurónové siete
- Výskum v danej oblasti

2.1 Problémová oblasť

V tejto práci sa zameriavame na predikciu úbytku zákazníkov(churn rate) pri predplatiteľských službách. V súčasnej dobe sa do popredia biznis prístupov stále viac dostávajú prístupy riadenia vzťahov zo zákazníkmi(customer relationship management). Ukazuje sa totiž, že na trhu s dostatočným pokrytím poskytovateľov cieľovej služby je niekoľkonásobne drahšie získať nového ako udržať si existujúceho zákazníka. Tento prístup však vyžaduje rozsiahlu znalosť dostupnej zákazníckej základne, ktorou poskytovateľ disponuje.

2.1.1 Predikcia úbytku zákazníkov

Predikcia úbytku zákazníkov sa venuje spracovaniu dostupných dát o zákazníckej aktivite, službách ktoré využívajú a vývoja ich správania v čase. Takéto dáta Výsledkom analýzy je štatistika poskytujúca informácie o jednotlivých zákazníkoch a ich šanci na presun k inému poskytovateľovi. Z týchto dát je následne odvoditeľné, aké percento zákazníkov odíde ku konkurencii a aký to bude mať dopad na finančné príjmy od ktorých je poskytovateľ závislý.

2.1.2 Moderovanie úbytku zákazníkov

Vo vzťahu k úbytku zákazníkov definuje CRM dva základné prístupy, ktorými je možné moderovať úbytok.

2.1.2.1 Reaktívny prístup

Motivácia zákazníka pre zotrvanie s pôvodnym poskytovateľom služby nastáva, až keď sa zákazník explicitne rozhodne pre prechod ku konkurenčnému poskytovateľovi. V tomto okamihu začína poskytovateľ na svojho zákazníka apelova výhodnými ponukami, zľavami alebo inými spôsobmi motivácie pre zotrvanie u poskytovateľa. Takýto prístup sa ukazuje ako ľahko zneužiteľný ostatnými zákazníkmi, ktorí by inak nemali motiváciu pre prechod ku konkurencii. Predikcia úbytku zákazníkov v tomto prístupe nemá nijakú významnú úlohu.

2.1.2.2 Proaktívny prístup

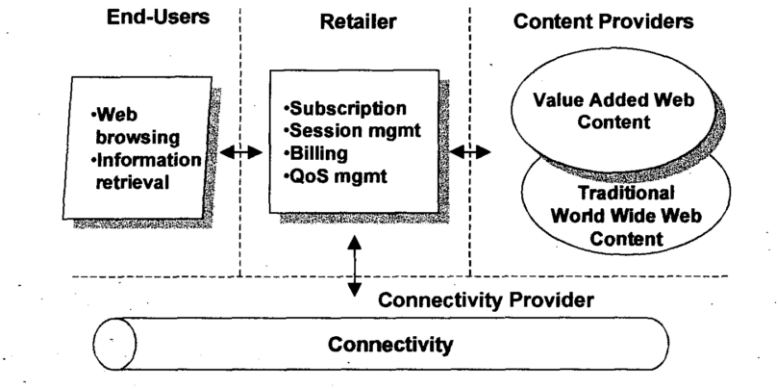
Pri úspešnej predikcii záujmu zákazníka o prechod ku konkurenčnému poskytovateľovi je možné efektívne jeho zámer smerovať pozitívnou motiváciou. Tento prístup však predpokladá vysokú úspešnosť predikčných metód. Pri nesprávnej identifikácii zákazníckeho správania je totiž možné nielen nezabrániť zákazníkom v presune ku konkurenčnému modelu, ale aj investícii finančných prostriedkov do skupiny zákazníkov, ktorá by naďalej generovala zisk aj bez významnejšej motivácie, resp. nevrátila by rozdielom v úbytku motivačné náklady, ktoré na ňu daný poskytovateľ vynaložil.

2.2 Dáta sprístupnené pre prácu

Pre túto prácu boli sprístupnené dáta z platobnej brány portálu pre online spravodajské denníky. Platobný portál poskytuje platformu pre periodiká, ktoré majú záujem o online funkcionalitu ale nemajú záujem implementovať vlastný platobný systém. Zákazníci tohto portálu tak získavajú rýchle riešenie pre možnosť vyhradenia exkluzívneho obsahu zo svojich online materiálov.

2.2.1 Exkluzívny obsah

Exkluzívny obsah je nástroj, ktorý množstvo poskytovateľov služieb využíva pri prechode na web. Umožňuje prístup k väčšiemu počtu potenciálnych zákazníkov, pričom poskytovateľovi ostáva možnosť oddeliť, čo bude prístupné každému od exkluzívneho obsahu určeného pre predplatiteľov.
Realizáciu exkluzívneho obsahu pomocou platobnej brány tretej strany umožňuje špecifikácia VAW(value added web). VAW aplikuje TINA(Telecommunications Information Networking Architecture) biznis model do klasického WWW(world wide web) prostredia. Určuje tak vzťahy medzi jednotlivými právnymi subjektami podľa obr. 2.1. Poskytovateľ služieb(spravodajské periodikum) tak môže poskytovať nielen klasický ale aj exkluzívny obsah bez toho, aby sa vo väčšej miere muselo zaoberať správou poskytovaných služieb a finančnou administratívou. Za tú zodpovedá sprostredkovateľ(platobný portál), ktorého úloha spočíva v správe exkluzívneho obsahu vo vzťahu ku koncovému používateľovi.



Obr. 2.1: *Základná schéma VAW*

2.2.2 Získavanie dát

Pri pokuse o prístup k exkluzívnemu obsahu stojí medzi používateľom a obsahom platobná brána portálu. Používateľovi bez predplatenej služby je zobrazená ponuka na platený prístup. Predplatitel prechádza cez bránu a je mu sprístupnený exkluzívny obsah. Pri všetkých aktivitách na portáli sú zaznamenávané používateľské údaje. Dostupné údaje sú vo forme záznamov - textových súborov priebežne generovaných používateľskou činnosťou. Bežná činnosť pri analýze záznamov z činnosti a práci s velkými dátami všeobecne je predspracovanie dát. Pri sledovaní činnosti používateľov sa generujú súbory so stovkami miliónov až miliardami záznamov. V súčasnosti nie je možné klasickými prístupmi spracovať takéto objemy dát bez predspracovania - filtrovania, segmentácie a čistenia dát. Spôsob predspracovania dát je z podstatnej časti ovplyvnený metódami, ktorými chceme dáta spracovať. Pri práci so záznamami je bežné deliť dáta na tzv. používateľské prístupy (user sessions). Používateľský prístup modeluje aktivitu - jeden prístup jedného používateľa. Všeobecne platí, že ak používateľ dosiahne v činnosti pauzu 30 a viac minút, jedná sa o samostatný nový prístup. Takto rozdelené záznamy poskytujú elasticitu pri spracovaní podľa špecifického času alebo podľa používatelov.

Medzi najdôležitejšie dostupné údaje z platobného portálu patria:

- IP adresa

- Používateľský účet
- Časový rozsah prístupu
- Prehliadaný obsah
- Aktivácia/prerušenie predplatného

2.3 Neurónové siete

Koncept neurónových sietí vznikol v 40. rokoch minulého storočia inšpiráciou biologickými neurónovými sieťami v mozgu. Cieľom bolo prekonať bariéru medzi tým, čo je pre ľudský mozog ľahko riešitelné ale ťažko formálne definovatelné matematickými pravidlami. Tieto problémy, ktoré riešime intuitívne, pri pokuse o formálnu špecifikáciu ukazujú, aké množstvo znalostí používame v každodennom živote. Ako vhodný príklad slúži vizuálne rozoznávanie objektov, ktoré je pre osobu samozrejmé, no až v posledných rokoch zaznamenávame prvé úspechy v tejto problematike.

2.3.1 Štruktúra

Podobne ako v mozgu, základ neurónovej siete tvoria neuróny a prepojenia medzi nimi. Neuróny sú organizované vo vrstvách, ktoré sa delia na 3 základné typy:

2.3.1.1 Vstupná vrstva

Reprezentuje dáta, ktoré podsúvame sieti pre interpretáciu. Dáta musia byť pred posunutím vstupnej vrstve často predspracované, aby bola sieť schopná interpretovať ich. Počet neurónov na vstupnej vrstve je ovplyvnený množstvom dát, ktoré máme na vstupe. V sieti existuje iba jediná vstupná vrstva.

2.3.1.2 Výstupná vrstva

Interpretácia dát neurónovou sieťou. Výstupnú vrstvu je možné nazvať „výsledok"siete. Jedná sa o jedinú vrstvu s obvykle jediným neurónom.

2.3.1.3 Skrytá vrstva

2.3.2 Učenie neurónovej siete

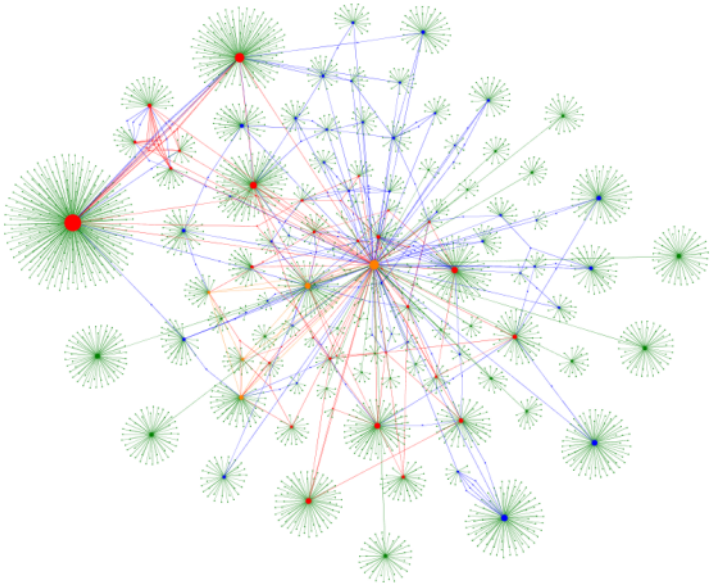
2.3.3 Hyperparametre

2.3.4 Pokročilé modely

2.4 Výskum v danej oblasti

2.5 Časť

V tejto časti sa venujeme



Obr. 2.2: *Name figure*

2.5.1 Číslovaný zoznam

1. cieľ 1
   (a) cieľ 1.a

(b) cieľ 1.b

2. cieľ 2
3. cieľ 3

2.5.2 Citácia

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat [**?**].

2.5.3 Návestia & Referencie

Viď. sekcia 2.5.4.
Viď. ukážka **??**.
Viď. číslovanie 2.1.
Viď. tabuľka **??**.

2.5.4 Príklady

```
<table class="metric_index">
  <tr>
    <th>Lines of code</th>
    <th>Value</th>
  </tr>
  <% if (filenum and modulenum) then %>
    <tr>
      <td class="name">Number of files</td>
      <td class="value"><%=filenum%></td>
    </tr>
    <tr>
      <td class="name">Number of modules</td>
      <td class="value"><%=modulenum%></td>
    </tr>
    <tr>
  <% end %>
  <tr>
    <td class="name">Lines Total</td>
    <td class="value"><%=LOC.lines%></td>
  </tr>
  <!--
              skryty zdrojovy kod
    podobne zobrazenie ostatnych metrik riadkov
  -->
```

```
</table>
```

Ukážka 2.1: *Príklad 1*

```
local parser  = require 'leg.parser'
local rules = require 'metrics.rules'
-- << skryty zdrojovy kod >> --
local capture_table = {}
grammar.pipe(LOC_capt.captures, AST_capt.captures)
grammar.pipe(block_capt.captures, LOC_capt.captures)
-- << viacero rovnakych volani s tabulkami captures inych modulov >> --
grammar.pipe(capture_table, cyclo_capt.captures)
local lua = lpeg.P(grammar.apply(parser.rules, rules.rules, capture_table))
local patt = lua / function(...)
  return {...}
end
local result = patt:match(code)[1]
```

Ukážka 2.2: *Názov*

Ukážka 2.3: *Manager*

```
int a;
```

Number of males	51
Number of woman	57
Gender not given	27
Average age	21,83

Tabuľka 2.1: *Information about users*

Kapitola 3

Dizajn

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat.

3.1 Časť

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum. Typi non habent claritatem insitam; est usus legentis in iis qui facit eorum claritatem. Investigationes demonstraverunt lectores legere me lius quod ii legunt saepius. Claritas est etiam processus dynamicus, qui sequitur mutationem consuetudium lectorum. Mirum est notare quam littera gothica, quam nunc putamus parum claram, anteposuerit litterarum formas humanitatis per seacula quarta decima et

quinta decima. Eodem modo typi, qui nunc nobis videntur parum clari, fiant sollemnes in futurum.

Kapitola 4

Výsledky

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat.

4.1 Časť

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum. Typi non habent claritatem insitam; est usus legentis in iis qui facit eorum claritatem. Investigationes demonstraverunt lectores legere me lius quod ii legunt saepius. Claritas est etiam processus dynamicus, qui sequitur mutationem consuetudium lectorum. Mirum est notare quam littera gothica, quam nunc putamus parum claram, anteposuerit litterarum formas humanitatis per seacula quarta decima et

quinta decima. Eodem modo typi, qui nunc nobis videntur parum clari, fiant sollemnes in futurum.

Kapitola 5

Záver

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi sit amet arcu. Fusce pharetra dapibus elit. Duis malesuada. Proin at elit vitae quam cursus tristique. Quisque fermentum. Praesent dictum. Nullam vehicula. Nunc pharetra dolor ut velit. Sed pulvinar, est sed congue tempor, nibh arcu cursus enim, quis consequat magna lacus sed pede. In sagittis. Etiam volutpat, velit id tincidunt egestas, augue ligula auctor eros, sit amet viverra sapien tortor at odio. In diam libero, fringilla ut, adipiscing condimentum, ultricies at, dui. Phasellus vitae risus.

Pellentesque vulputate ante ut diam. Sed adipiscing malesuada odio. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Nam a leo. Praesent velit. Aenean vehicula accumsan quam. Nulla dolor lorem, imperdiet a, ullamcorper hendrerit, ultrices at, urna. Integer placerat ligula id purus. Sed id nisl. Pellentesque tincidunt neque in lacus. In non quam et felis suscipit viverra.

Príloha A

Technická dokumentácia

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat.

A.1 Implementácia

Modul abc Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum.

Modul def Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation

ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum. Typi non habent claritatem insitam; est usus legentis in iis qui facit eorum claritatem. Investigationes demonstraverunt lectores legere me lius quod ii legunt saepius. Claritas est etiam processus dynamicus, qui sequitur mutationem consuetudium lectorum. Mirum est notare quam littera gothica, quam nunc putamus parum claram, anteposuerit litterarum formas humanitatis per seacula quarta decima et quinta decima. Eodem modo typi, qui nunc nobis videntur parum clari, fiant sollemnes in futurum.

Príloha B

Používateľská dokumentácia

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat.

B.1 Inštalácia

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

B.1.1 Spustenie aplikácie

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure

dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

Príloha C

Electronic medium

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat:

/Application

- implementácia opisovaného riešenia

/Documentation

- bakalárska práca spolu s anotáciami v slovenskom a anglickom jazyku

/Documentation/Latex

- latex zdrojové súbory dokumentácie

/Documentation/BibTeX

- BibTeX súbor s použitými referenciami

/Documentation/Resources

- dostupné použité zdroje

/Resources

- vstupne/testovacie dáta opisované v dokumente

/Source/Dependencies

- inštalačné súbory pre knižnice, ktoré potrebuje aplikácia

read.me - popis obsahu média v slovenskom a anglickom jazyku