

FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY
UNIVERZITA KOMENSKÉHO V BRATISLAVE

Záverečná správa

Anotátor pomenovaných entít
(APE)

Alexander Szendy
Michal Klein
2016/2017

Obsah

1	Úvod	4
1.1	Ciele Dokumentu	4
1.2	Rozsah projektu	4
1.3	Definície, skratky	4
1.4	Odkazy	4
1.5	Nasledujúce kapitoly	4
2	Všeobecný popis.....	5
2.1	Perspektíva projektu.....	5
2.2	Funkcie projektu	5
2.3	Charakteristika používateľov	5
2.4	Všeobecné obmedzenia.....	5
3	Špecifické požiadavky	6
3.1	Čistenie logu	6
3.2	Export logu.....	6
3.3	Práva používateľa	6
3.4	Anotovaný document	6
3.4.1	Načítanie dokumentu.....	6
3.4.2	Zmena obsahu anotovaného dokumentu	6
3.4.3	Zobrazovanie zmien v dokumente	6
3.4.4	Rozdelenie dokumentu	6
3.5	Anotácie.....	6
3.5.1	Pridávanie/mazanie/úprava anotácií.....	6
3.5.2	Prekryv anotácií.....	6
3.5.3	Referencie.....	6
3.5.4	Zmena referencií.....	7
3.5.5	Typy anotácií	7
3.5.6	Vyznačenie anotácií a referencií v texte	7
3.5.7	Zobrazovanie anotácií	7
3.6	Export	7
4	Diagramy.....	9
4.1	Entitno-relačný diagram	9
4.2	Use-case diagram	9
4.3	Stavový diagram	10
5	Používateľské rozhranie	11

5.1	Celé grafické rozhranie	11
6	Analýza technológií.....	13
6.1	Klient.....	13
6.1.1	Java a JavaFx.....	13
6.1.2	JavaFx CSS.....	13
6.1.3	RichTextFx	13
7	Dekompozícia	14
7.1	Diagram komponentov	14
8	Návrh	15
8.1	Úvod.....	15
8.2	Triedny diagram.....	15
9	Testovacie scenáre	16
9.1	Projekt	16
9.1.1	Načítanie dokumentu.....	16
9.1.2	Uloženie rozpracovaného projektu.....	16
9.1.3	Načítanie uloženého projektu	16
9.1.4	Export dokumentu	16
9.2	Anotácia	16
9.2.1	Zobrazovanie anotácií	16
9.2.2	Úprava kategórií	16
9.2.3	Úprava anotácie	17
9.3	Log	17
9.3.1	Zobrazenie popisu o vykonanej udalosti	17
10	Inštalačná príručka	18
10.1	Systémové požiadavky.....	18
10.2	Inštalácia.....	18
11	Používateľská príručka.....	19
11.1	Prvé spustenie Aplikácie	19
11.2	Načítanie textového súboru	19
11.3	Práca s textom (Anotovanie)	20
11.4	Export	23
11.5	Nastavenia textu.....	24
11.6	Log	24

1 Úvod

1.1 Ciele Dokumentu

Cieľom tohto dokumentu je čo najbližšie popísať požiadavky zadávateľa projektu bez bližšieho určenia technickej špecifikácie ako bude projekt APE implementovaný. Slúži ako zoznam funkcií, ktoré hotový produkt musí spĺňať a na základe ktorých bude aj vyhodnotený. Dokument je určený pre všetkých ľudí, ktorí prídu priamo alebo nepriamo do kontaktu s týmto systémom alebo jeho vývojom.

1.2 Rozsah projektu

Našou úlohou je vytvoriť offline verziu textového editoru pre jednoduchý text, v ktorom bude môcť používateľ anotovať jeden zvolený dokument.

1.3 Definície, skratky

- Anotácia – časť textu, ktorá bude vyznačená
- Typ/kategória anotácie – každá anotácia má určitý typ (mesto, osobu, štát, čas, ...)
- NE – named entity, pomenovaná entita, zvyčajne ide o konkrétnu osobu, čas, mesto, organizáciu, ...
- UI – umelá inteligencia

1.4 Odkazy

Nasledujú odkazy na podobný softvér:

- <https://gate.ac.uk/>
- <http://brat.nlplab.org/index.html>

Príklad textu, ktorý sa bude môcť anotovať:

- http://sagadb.org/egils_saga.en

1.5 Nasledujúce kapitoly

V nasledujúcich kapitolách bližšie špecifikujeme funkcie nášho systému, ako export dokumentov.

2 Všeobecný popis

2.1 Perspektíva projektu

APE bude mať využitie najmä v oblastiach UI, kde je často nevyhnutné program najskôr niečo naučiť – zvyčajne od človeka, ktorý rieši ten istý problém (v našom prípade ide o tzv. named entity recognition – či dané slovo/slová sú NE a ak áno, treba správne určiť kategóriu).

Jedným z využití je napríklad také, kde anotáciu objemnejších textových dokumentov bude možné rozdeliť medzi viacero ľudí, čím sa zníži časová aj pracovná vyťaženosť človeka.

2.2 Funkcie projektu

Aplikácia bude obsahovať jednu časť: klient. Primárna funkcia klienta (s grafickým rozhraním) bude samotná anotácia dokumentu a zobrazovanie zmien pri anotácii.

Podrobnejšie informácie sú uvedené v špecifických požiadavkách [3].

2.3 Charakteristika používateľov

Používateľov nerozlišujeme, a teda máme iba jeden typ:

- normálny používateľ (ďalej len používateľ)

Používateľ používa klienta a jeho cieľom je anotovanie.

2.4 Všeobecné obmedzenia

Anotovať bude možné len jednoduché textové dokumenty. Dokumenty budú môcť mať rozličné konce riadkov ako aj kódovania (minimálne akceptované budú UTF-8 a ASCII).

3 Špecifické požiadavky

3.1 Čistenie logu

Log bude môcť používateľ aj zmazať. Log sa vždy viaže na aktívnu skupinu a pri jej zmene sa sám zmaže.

3.2 Export logu

Používateľ bude mať možnosť aktuálny log aj exportovať.

3.3 Práva používateľa

Používateľ má práva:

- anotovania – umožňuje pridávať anotácie
- reanotovania – umožňuje meniť/mazať anotácie
- editovania – umožňuje pridávať/meniť/mazať typy a podtypy anotácií

3.4 Anotovaný document

3.4.1 Načítanie dokumentu

Používateľ po spustení bude môcť otvoriť nejaký textový súbor, v akceptovanom formáte, ako aj zvoliť možnosť automatického načítania tohto súboru pri ďalšom spustení.

3.4.2 Zmena obsahu anotovaného dokumentu

Obsah dokumentu nemožno upravovať iným spôsobom ako anotovaním.

3.4.3 Zobrazovanie zmien v dokumente

Zmeny vykonané používateľom sa okamžite zobrazia v texte dokumentu.

3.4.4 Rozdelenie dokumentu

Dlhé texty, ktoré sa nezmestia na zobrazenie do okna, sa budú dať aj scrollovať.

3.5 Anotácie

3.5.1 Pridávanie/mazanie/úprava anotácií

Vyznačený text v dokumente môže používateľ anotovať stlačením kláves, kde si bude môcť vybrať napr.: typ anotácie (napr. “mesto”), popis, odkazy, Anotácie môžu byť zmenené alebo zmazané.

3.5.2 Prekryv anotácií

Ľubovoľná časť textu môže byť anotovaná aj viackrát.

3.5.3 Referencie

Používateľ bude môcť špecifikovať, či je daná anotácia referencia a na ktorú predošlú anotáciu sa vzťahuje.

3.5.4 Zmena referencií

Referenciám možno iba zmeniť objekt, na ktorý referujú (rodič), prípadne ich vymazať. Referencia dedí od rodiča všetky jeho vlastnosti, pri zmene rodiča je patričná zmena vykonaná na všetkých jeho referenciách.

3.5.5 Typy anotácií

Používateľ bude môcť vytvárať nové typy a podtypy anotácií, alebo ich iným spôsobom upravovať (farba, tag, názov,...) či mazať.

3.5.6 Vyznačenie anotácií a referencií v texte

Anotácie v textovom editore budú farebne vyznačené podľa daného typu/kategórie. Farbu rôznych kategórií si môže užívateľ zvoliť. Prekryv anotácií bude nejakým spôsobom rozlíšený (napr. konkrétnou farbou alebo zmiešaním farieb jednotlivých prekrývajúcich sa anotácií). Referencie budú mať vždy rovnakú farbu, ako anotácia, na ktorú odkazujú.

3.5.7 Zobrazovanie anotácií

Používateľ si bude môcť určiť, ktoré typy anotácií sa mu budú zobrazovať v texte (napr. sa rozhodne, že anotácie typu “osoba” nebudú momentálne vyznačené). Táto zmena je len lokálna – nezobrazuje sa ostatným používateľom v skupine.

3.6 Export

Používateľ bude môcť exportovať dokument svojej aktívnej skupiny. Každý riadok vyexportovaného dokumentu bude mať nasledovný formát:

- číslo riadku/poradie slovo X:Z,... A-B:... odkazy

Kde **X** bude buď písmeno ‘B’ (begin – označuje začiatok anotácie) alebo ‘I’ (pokračovanie anotácie) prípadne nič, **Z** bude zoznam tagov daného typu anotácie (napr. PER pre persons, tagy si môže meniť používateľ s editačnými právami, viaceré tagy budú oddelené čiarkou, referencie budú mať nemenný tag). **A:B** je interval $<A, B)$, ktorý určuje riadok slova alebo slov, na ktoré sa referencia vzťahuje, viacero intervalov je oddelených čiarkou. Pri anotáciách, ktoré nie sú referencie a pri neanotovanom texte je A aj B prázdne. Viaceré intervaly budú oddelené čiarkou. Nasledovať bude zoznam odkazov, pre danú anotáciu (pri viacslovných anotáciách je to vždy v riadku prvého anotovaného slova).

Interpunkcia je takisto na samostatnom riadku.

Samotná dvojbodka v 3. stĺpci označuje, že slovo nie je anotované a vo 4. stĺpci, že slovo nie je referencia (kvôli lepšiemu parsovaniu).

Nasledujú príklady riadkov exportovaného dokumentu:

- 2 Bratislava B:CIT : en.wikipedia.org/wiki/Bratislava

Viacslovný anotovaný text:

- 3 Barack B:PER : en.wikipedia.org/wiki/Barack_Obama
- 5 Obama I :

Referencia:

- 7 He B:REF 1-2

Neanotované slovo:

- 11 foo : :

Viacrát anotované jedno slovo:

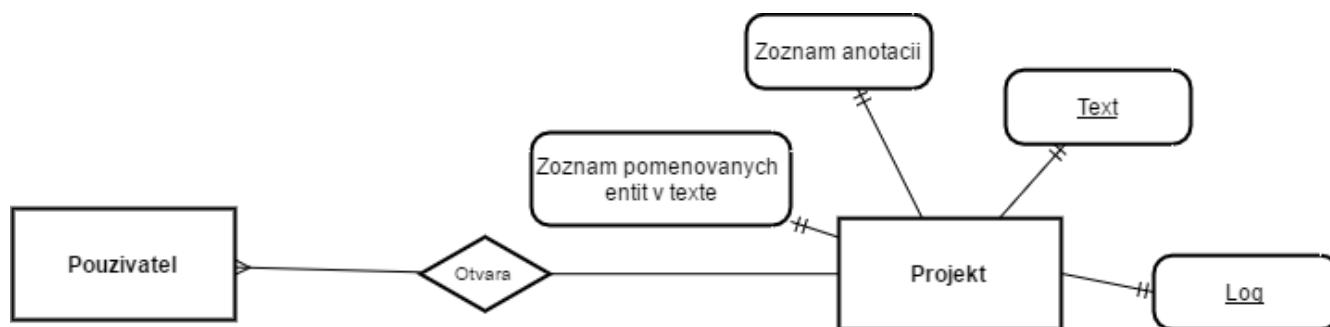
- 13 bar B:NAME1,NAME2 :

A nakoniec referencia odkazujúca na entitu, ktorá je na viacerých miestach:

- 14 It B:REF 123-124:2-3

4 Diagramy

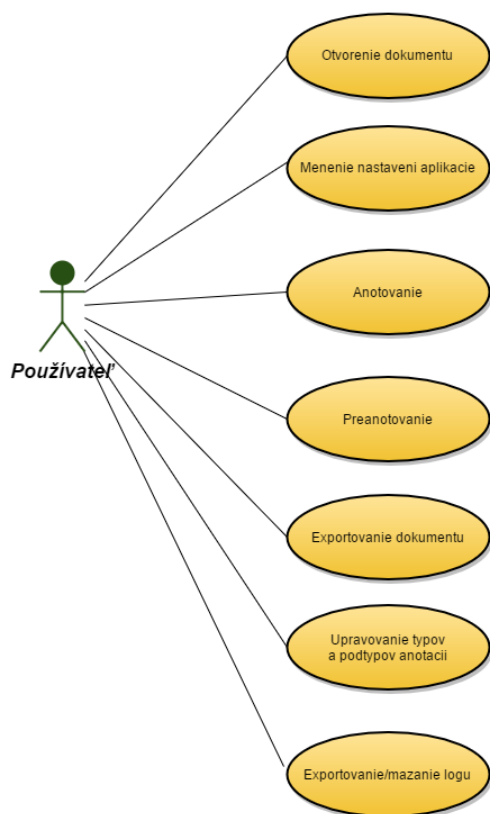
4.1 Entitno-relačný diagram



Obr. 4. 1 Entitno-relačný diagram

Entitno-relačný diagram (Obr. 4.1) zobrazuje entity nachádzajúce sa v projekte a relácie medzi nimi.

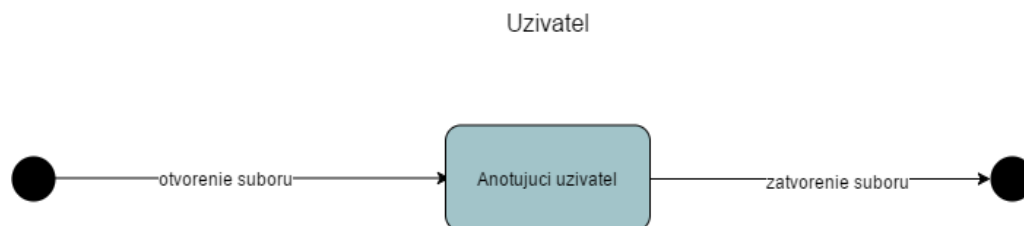
4.2 Use-case diagram



Obr. 4. 2 Use-Case diagram

Use-case diagram (Obr. 4.2) zobrazuje možnosti použitia projektu. Nachádza sa v ňom jedna entita používateľ'a, ktorá môže vykonávať rôzne funkcie.

4.3 Stavový diagram



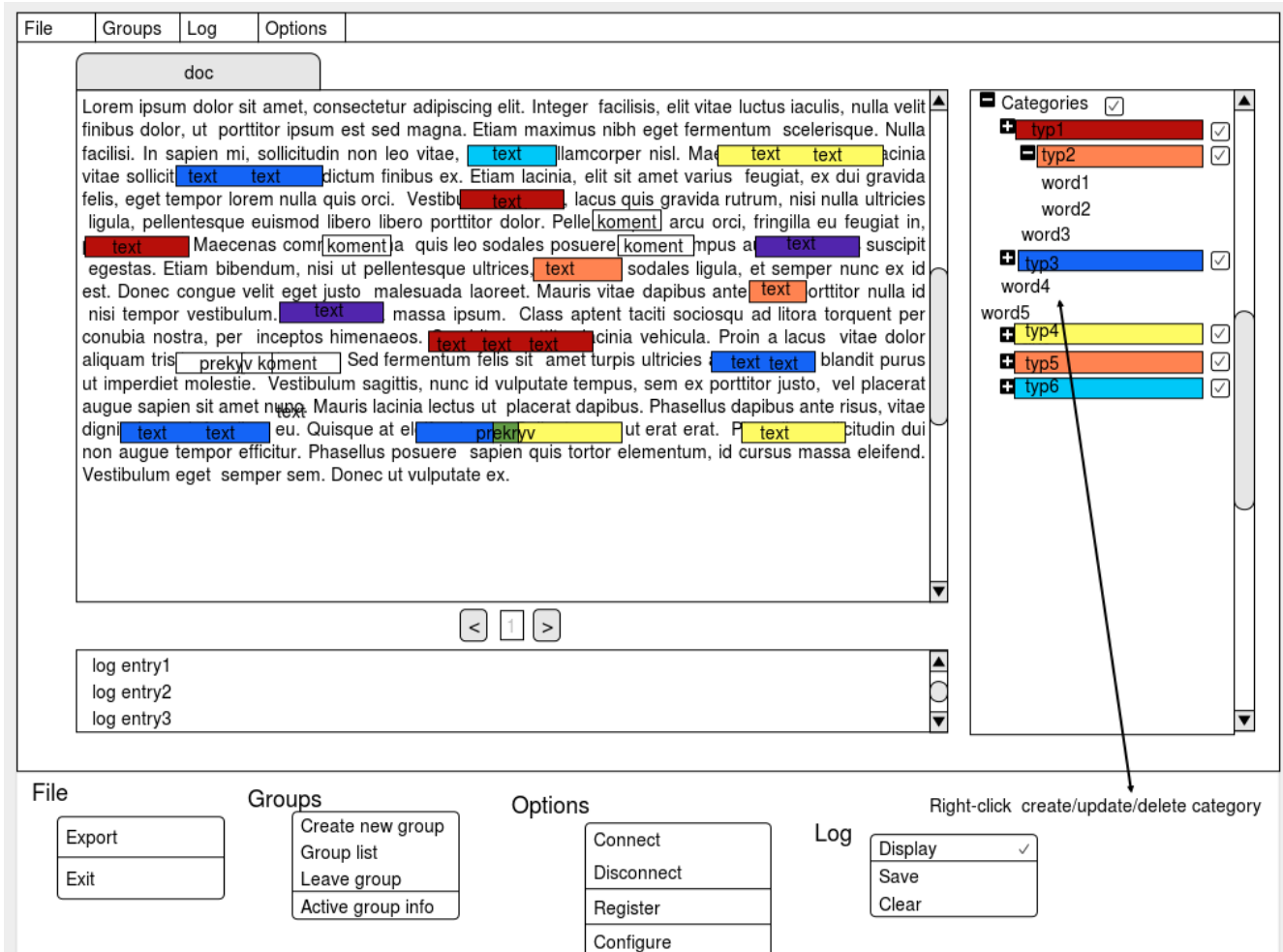
Obr. 4. 3 Stavový diagram

Stavový diagram (Obr. 4.3) zobrazuje stav, ktorý môže užívateľ nadobúdať po otvorení dokumentu. Počas tohto stavu užívateľ upravuje dokument anotovaním.

5 Používateľské rozhranie

V tejto časti je zobrazený približný náčrt designu a jeho funkčnosti pre výslednú aplikáciu.

5.1 Celé grafické rozhranie



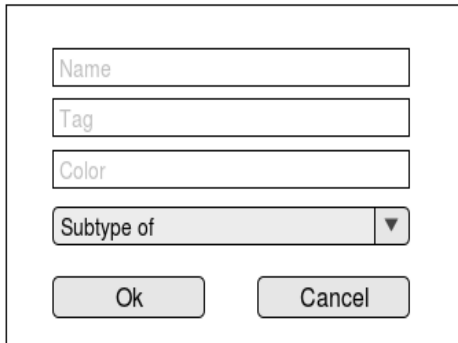
Obr. 5. 1 Grafické rozhranie aplikácie

Obr. 5. 1 zobrazuje náčrt celého grafického rozhrania zobrazeného po spustení aplikácie. Rozhranie obsahuje komponenty : Menu (Vľavo hore), Dokument (Vľavo v strede), Log (Vľavo dole), Kategórie (Vpravo).


Menu obsahuje záložku

- File
 - o Export – umožňuje exportovať dokument do textového súboru, podľa pravidiel uvedených v sekcii 3.6 tohto dokumentu
- do výslednej podoby tejto záložky boli pridané možnosti :
 - o Vytvorenia projektu, ktorá vytvorí prázdny dokument s textom,
 - o Otvorenia projektu, ktorá načíta už uložený projekt z pamäte.
 - o Uloženia projektu, ktorá uloží aktuálny stav projektu

- Groups – túto položku výsledná aplikácia neobsahuje
- Options
 - o Configure – nastavenie fontu a zobrazenia potvrdzovacích dialógov.
 - ostatné nie sú implementované
- Log
 - o Display- možnosť výberu zobrazenia – v okne/ pod dokumentom
 - o Save- možnosť ukladať históriu
 - o Clear- zmazanie histórie udalostí



Obr. 5. 2 New category



Obr. 5. 3 New annotation

- Annotations (Pridaná záložka do výslednej podoby rozhrania)
 - o New category (Obr. 5.2) – vytvorenie novej kategórie
 - o New annotation (Obr. 5.3)– vytvorenie novej anotácie textu
 - o New coreference – vytvorenie referencie na anotáciu

6 Analýza technológií

6.1 Klient

6.1.1 Java a JavaFx

Klient bude naprogramovaný v jazyku Java s využitím platformy JavaFx, ktorá slúži na tvorbu grafického rozhrania nielen pre desktopové aplikácie.

6.1.2 JavaFx CSS

JavaFx CSS je založená na webovej verzii CSS. Používa sa na prispôsobenie vzhľadu aplikácií. V našom prípade je použitá na zvýraznenie anotácií ich zafarbením na príslušnú farbu.

6.1.3 RichTextFx

RichTextFx poskytuje API pre JavaFx a obsahuje rôzne užitočné funkcie, ako pridávanie štýlu k rôznym častiam textu, číslovanie riadkov, zobrazenie popup-ov, ... a bude tvoriť základ nášeho editora.

7 Dekompozícia

7.1 Diagram komponentov



Obr. 7. 1 Diagram komponentov

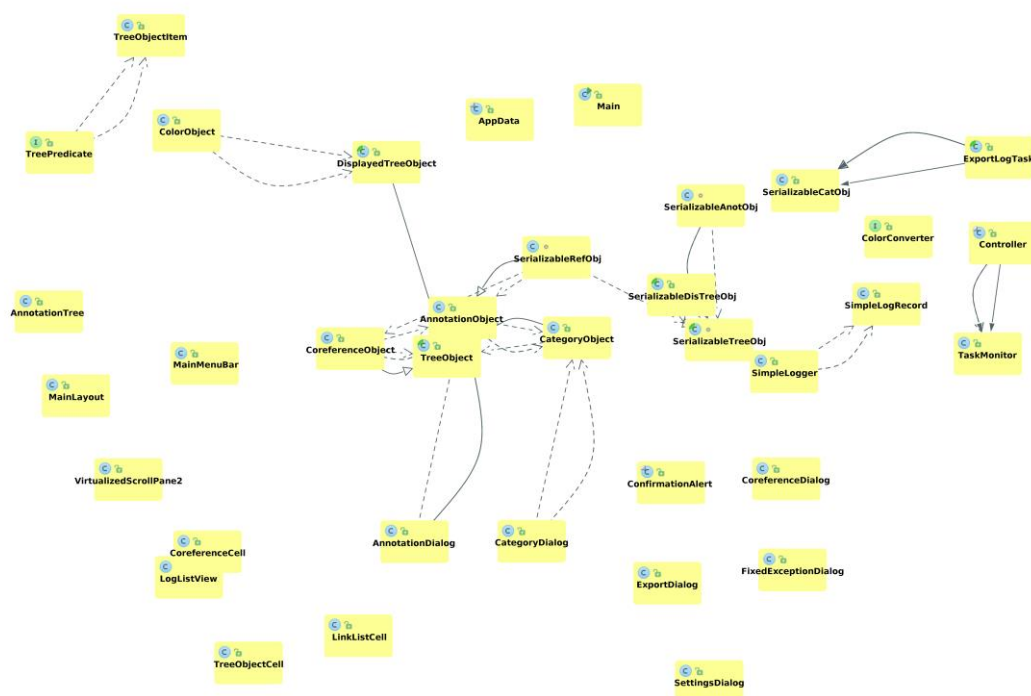
Obr. 7.1 obsahuje jeden komponent, ktorý slúži na prácu s dokumentom.

8 Návrh

8.1 Úvod

Systém je vyvíjaný v jazyku Java, ktorý používa grafické rozhranie JavaFx. JavaFx sa stará o všetky zobrazovania okien v tejto aplikácii. Ďalšiou využitou architektúrou je RichTextFx, ktorá upravuje zobrazenie textu dokumentu rôznymi štýlmi.

8.2 Triedny diagram



Obr. 8. 1 Class diagram

Tiedny diagram (Obr. 8.1) zobrazuje hlavné triedy, ktoré sú použité v našom systéme. Obsahuje napríklad triedy pre zobrazenie rôznych dialógových okien, Väčšina tried priamo súvisí so správnym fungovaním celého rozhrania- rôzne zobrazenia, rozloženia okien. Ďalšia časť tried pracuje s dátami, ako sú napríklad polia kategórií, anotácií, či samotný text dokumentu.

9 Testovacie scenáre

9.1 Projekt

9.1.1 Načítanie dokumentu

- 1 Vstup:** výber textového dokumentu
- 2 Výstup:** zobrazenie textu dokumentu v okne
- 3 Stav:** otestované

9.1.2 Uloženie rozpracovaného projektu

- 4 Vstup:** výber adresára na uloženie súboru
- 5 Výstup:** súbor úspešne uložený
- 6 Stav:** otestované

9.1.3 Načítanie uloženého projektu

- 7 Vstup:** výber adresára na uloženie súboru
- 8 Výstup:** text s anotáciami a všetkými predošlými nastaveniami sa zobrazí v aplikácii
- 9 Stav:** otestované

9.1.4 Export dokumentu

- 10 Vstup:** výber adresára na uloženie súboru
- 11 Výstup:** súbor úspešne uložený v správnom formáte
- 12 Stav:** otestované

9.2 Anotácia

9.2.1 Zobrazovanie anotácií

- 13 Vstup:** odznačená voľba na zobrazenie anotácie v texte
- 14 Výstup:** v texte nebudú farebne vyznačené anotácie danej kategórie
- 15 Stav:** otestované

16 Vstup: označená voľba na zobrazenie anotácie v texte

17 Výstup: farebné vyznačenie sa znovu zobrazí

18 Stav: otestované

9.2.2 Úprava kategórií

19 Vstup: správne vyplnený formulár pre vytvorenie novej/úpravu kategórie

20 Výstup: kategória úspešne pridaná do stromu s kategóriami, alebo upravená

21 Stav: otestované

9.2.3 Úprava anotácie

22 Vstup: označený text + vyplnený formulár pre pridanie/zmazanie/úpravu anotácie

23 Výstup: zobrazenie zmien v texte dokumentu a pridanie/ zobrazenie zmeny v strome kategórií

24 Stav: otestované

9.3 Log

9.3.1 Zobrazenie popisu o vykonanej udalosti

25 Vstup: Ľubovoľná zmena v projekte

26 Výstup: zobrazenie popisu zmeny s aktuálnym časom

27 Stav: otestované pre rôzne vykonané zmeny

10 Inštalčná príručka

10.1 Systémové požiadavky

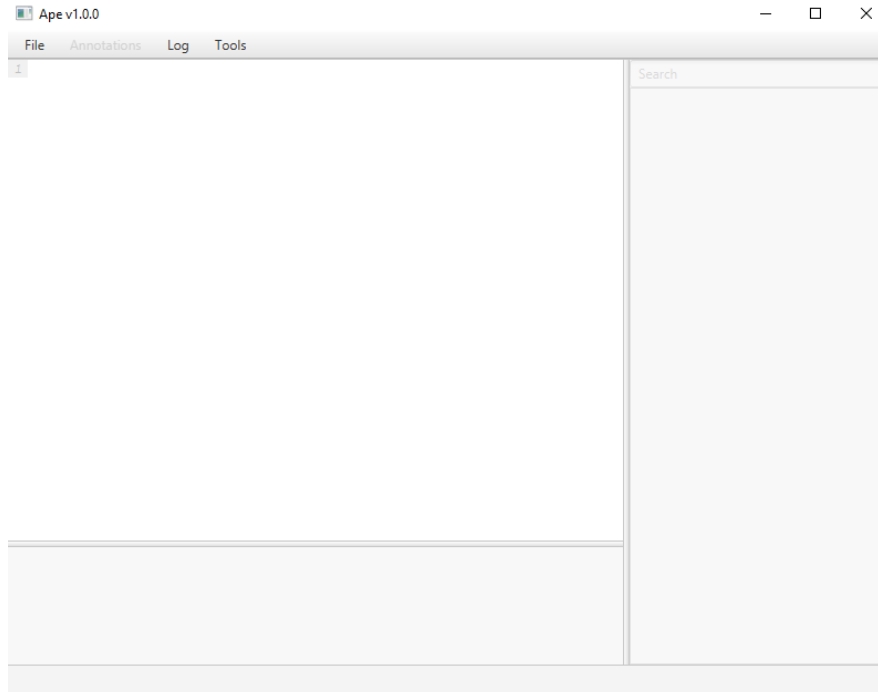
Java Runtime Environment \geq 8, s JavaFx

10.2 Inštalácia

Stačí rozzipovať súbor ape.zip. Aplikácia sa spúšťa súborom ape.jar. Pre správne fungovanie by mal byť priečinok libs spolu s aplikáciou .jar v jednom priečinku.

11 Používateľská príručka

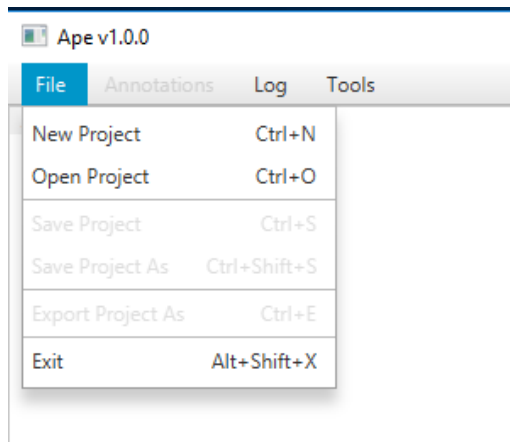
11.1 Prvé spustenie Aplikácie



Obr. 11. 1 Prázdné okno

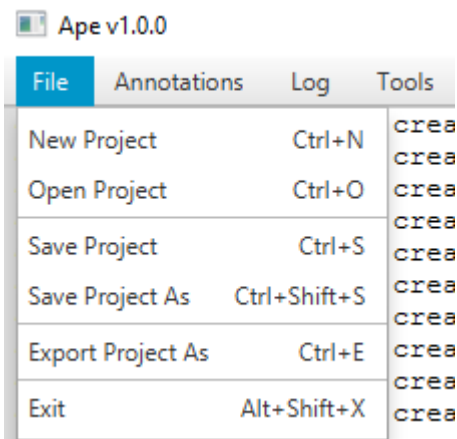
Obr. 11.1 zobrazuje ako vyzerá aplikácia po spustení, väčšina funkcií je zablokovaná. Pre viac funkcií je potrebné načítať súbor pomocou položky v menu: File -> New project (zobrazené na Obr. 11.2 nižšie).

11.2 Načítanie textového súboru



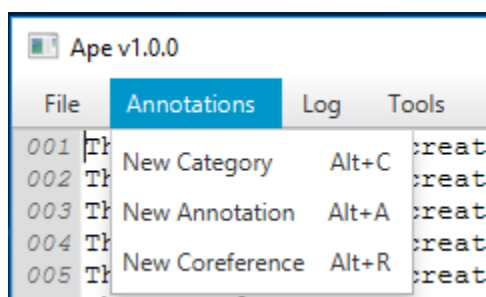
Obr. 11. 2 Kontextové menu File

Po kliknutí na New Project sa zobrazí nové okno, kde možno vybrať nejaký textový súbor (.txt) pre ďalšie spracovanie. Po potvrdení výberu sa text obsiahnutý vo vybratom súbore zobrazí na ploche nižšie a zároveň sa odomknú všetky funkcie, ktoré boli zablokované (viz Obr. 11.3).



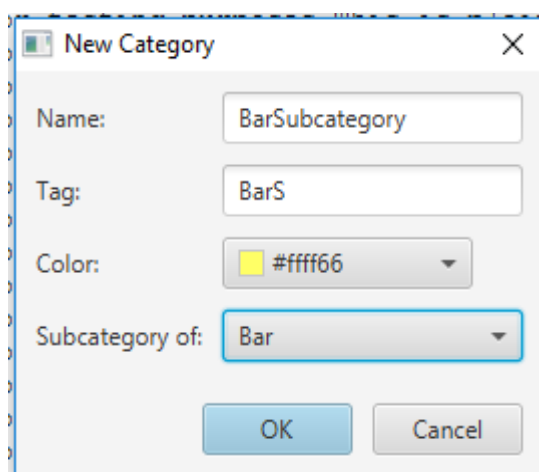
Obr. 11. 3 Odomknuté funkcie menu

11.3 Práca s textom (Anotovanie)



Obr. 11. 4 Anotácie

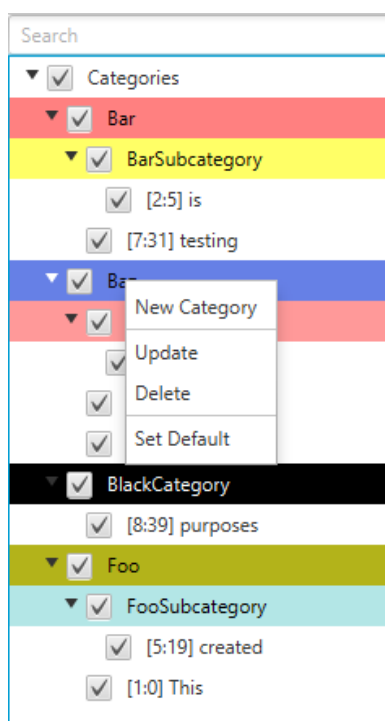
Obr. 11.4 zobrazuje obsah kontextového menu po kliknutí na voľbu Annotations.



Obr. 11. 5 Nová kategória

Pre neskoršie anotovanie je nutné vytvoriť rôzne kategórie, ktoré budú reprezentovať typ slova/ viacerých slov v texte. Tag slúži na reprezentovanie

danej anotácie pri exportovaní projektu do textového súboru, teda by mal byť tiež unikátny. Vybratie farby je tiež dôležité, pre rozlíšenie medzi ostatnými anotáciami priamo v texte. Je možné tiež vytvárať pod-kategórie už vytvorených kategórií. Všetky tieto veci sa dajú nastaviť pri kliknutí v menu na Annotations-> New Category, kde sa zobrazí nové okno. (Obr. 11.4 a Obr. 11.5)



Obr. 11. 6 Kontextové menu po kliknutí na jednu z kategórií

Zmena kategórie je možná po kliknutí pravým tlačítkom myši na danú kategóriu a zvolením možnosti Update.

(Obr. 11.6) V tomto kontextovom menu je tiež možnosť vytvoriť novú kategóriu, zmazať vybranú, alebo nastaviť vybranú kategóriu ako základnú, ktorá sa defaultne zobrazí pri anotácii.

New Annotation

Text: This

Category: Foo New Category

Urls:
<https://www.java.com/en/download/>
 FooBarBazURL

Add URL

Description:
 Plain text description

OK Cancel

Obr. 11. 7 Nová anotácia

Vytvorenie anotácie je možné po vybratí textu a buď vybratí v menu (Obr. 11.4) možnosť New Annotation, alebo použitím klávesovej skratky Alt+A. Zobrazí sa formulár na vyplnenie (Obr. 11.7). Tu je možnosť vytvoriť novú kategóriu, alebo vybrať z existujúcich. Tiež sa dá vložiť URL, ktorá nejak súvisí s vybratým textom, a dá sa vložiť aj popis k danej anotácii.

Update Annotation

Text: [8:39] purposes

Category: BlackCategory New Category

Urls:
 Add URL

Description:
 Coreferences:

OK Cancel

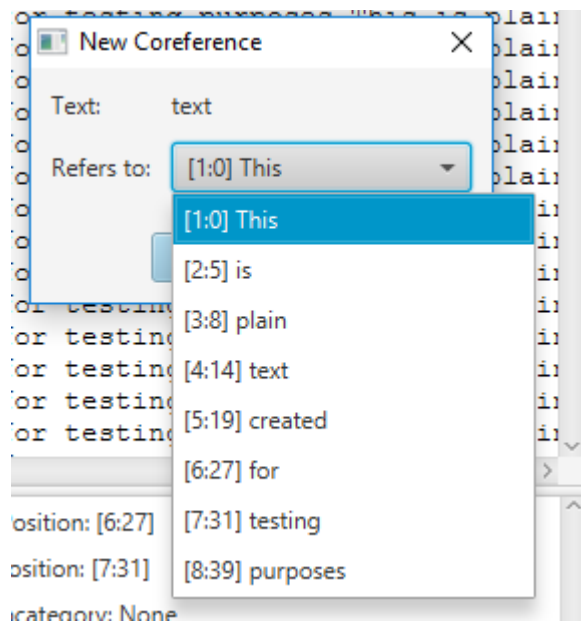
Search

- Categories
 - Bar
 - BarSubcategory
 - [2:5] is
 - [7:31] testing
 - Baz
 - BazSubcategory
 - [4:14] text
 - BlackCategory
 - [8:39] purposes
 - Foo
 - FooSubcategory
 - [5:19] created

Obr. 11. 8 Update anotácie

Možnosť upraviť už vytvorenú anotáciu môžete zobraziť po kliknutí pravým tlačítkom myši na anotáciu, ktorú chcete modifikovať (prípadne zmazať). (Obr. 11.8)

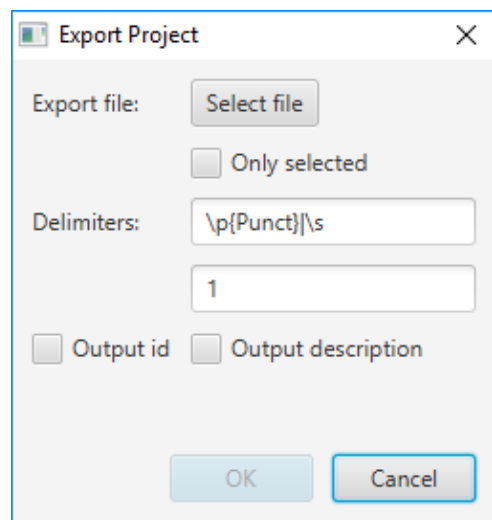
Rovnako ako pri vytváraní anotácie, je možné zadať všetky údaje znovu.



Obr. 11. 9 Referencia

Pridanie referencie sa vykonáva výberom textu a následným zvolením možnosti v menu Annotations->Coreference (Obr. 11.4), prípadne stlačením kláves Alt+R. Zobrazí sa nové okno, kde je možné zvoliť, na ktorú anotáciu daný text referuje.

11.4 Export

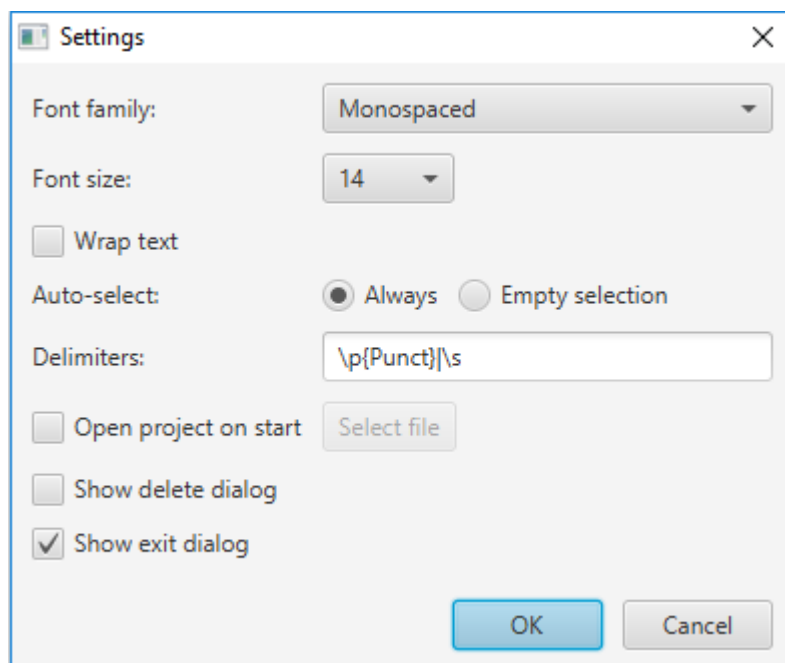


Obr. 11. 10 Export

Export dokumentu sa vykoná výberom v menu File->Export Project as (Obr. 11.3).

Zobrazí sa nové okno kde sa dá nastaviť formátovanie, ako aj výber miesta na disku, kam sa exportovaný dokument uloží.

11.5 Nastavenia textu



Obr. 11. 11 Nastavenia

Nastaviť rôzne formátovanie zobrazenia textu je možné v menu Tools->Options. Tiež je možné vybrať, ktoré potvrdzovacie dialogy sa budú zobrazovať. (Obr. 11.11)

11.6 Log

Po každej akcii sa do poľa vľavo dole pridá výpis druhu akcie s aktuálnym časom. Vymazanie, prípadne export tohto textu je možné vykonať zvolením menu položky Log.