FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY
UNIVERZITA KOMENSKÉHO V BRATISLAVE

Katalóg požiadaviek pre Anotátor pomenovaných entít (APE)

Alexander Szendy
Juraj Kočan
Marián Glatzner
Michal Klein
2016/2017

Obsah

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 najmä pre ľuďí, ktorí budú tento projekt implementovať.

1.2 Rozsah projektu

Našou úlohou je vytvoriť kolaboratívny textový editor pre jednoduchý text, v ktorom budú môcť viacerí používatelia spoločne anotovať jeden a ten istý dokument v reálnom čase naraz – zmeny vykonané jedným používateľom sa budú zobrazovať aj ostatným užívateľom.

1.3 Definície, skratky

- Skupina – ľudia, ktorí spoločne anotujú nejaký dokument
- 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, prístupové práva pre používateľov a skupiny.

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).
Keďže pôjde o kolaboratívny editor zobrazujúci zmeny v reálnom čase, anotáciu objemnejších textových dokumentov bude možné rozdeliť medzi viacero ľuďí, čím sa zníži časová aj pracovná vyťaženosť človeka.

2.2 Funkcie projektu

Aplikácia bude rozdelená na dve časti: server a klient. Server bude slúžiť na spravovanie informácií týkajúcich sa skupín a šírenie komunikácie medzi používateľmi, zatiaľčo 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 delíme na dva typy:

- správca
- normálny používateľ (ďalej len používateľ)

Správca slúži len na inštaláciu servera a jeho nasledovné spravovanie, nemá funkcie ako obyčajný používateľ. Používateľ sa pripája na server cez klienta a jeho cieľom je anotovanie.
Práva používateľa sú určené podľa momentálne aktívnej skupiny. Každý používaťeľ má najviac 1 aktívnu skupinu. Rozlišujeme medzi 3 základnými právami (možno ich kombinovať):

- anotačné – umožňuje pridávať anotácie
- reanotačné – umožňuje preanotovať (meniť a mazať) anotácie iných a komentovať
- editačné – umožňuje pridávať/meniť/mazať typy a podtypy anotácií v danej skupine

Administrátor skupiny má všetky horeuvedené práva a takisto môže meniť štruktúru skupiny (pridať/meniť/odobrať: užívateľa alebo ich práva, heslo skupiny, názov, default práva).

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 Registrácia a prihlasovanie

Po spustení aplikácie sa používatelia budú môcť zaregistrovať. Ak je používateľ už zaregistrovaný, tak pomocou prihlasovacieho mena a hesla sa bude môcť pripojiť na server. Používateľ si bude môcť určiť, či sa hneď po spustení aplikácie chce pripojiť na server na základe konfiguračného súboru.

3.2 Používatelia

Všetko dôležité ohľadom používateľov je spomenuté v [2.3].

3.3 Anotovaný dokument

Dokumenty na vstupe budú zväčša články z časopisov.Dokument sa viaže ku skupine vždy v pomere 1:1 – jedna skupina anotuje práve jeden dokument. Dokument uploaduje pri tvorbe skupiny jej tvorca a pre danú skupinu je nemenný – skupina nemôže zmeniť anotovaný dokument. Dlhé texty sa budú dať aj scrollovať a rozsiahle dokumenty budú rozdelené(podľa počtu viet) na viacero častí/stránok medzi ktorými bude možné listovať. Zmeny (anotácie) vykonané používateľom sú okamžite uložené a zobrazené aj iným používateľom v skupine. Obsah dokumentu nemožno upravovať iným spôsobom ako anotovaním a komentovaním.

3.4 Skupiny

Zakladateľ skupiny (admin) môže pri jej tvorbe zvoliť nasledovné položky:

- názov skupiny
- dokument, ktorý bude skupina anotovať
- zoznam používateľov v skupine aj s ich prístupovými právami
- default práva
- heslo pre vstup do skupiny (optional)

Default práva, ak sú dané, sa nastavia používateľovi, ktorý sa chce pripojiť do skupiny, do ktorej zoznamu však nepatrí. Ak default práva nie sú zadané, do skupiny so môžu pripojiť len ľudia v jej zozname.
Po prihlásení na server si používateľ môže zvoliť svoju aktuálnu skupinu zo zoznamu všetkých skupín alebo vytvoriť vlastnú. Používateľ má možnosť sa od skupiny aj odpojiť alebo sa prepnúť do inej, ktorá sa stane jeho aktívnou. Aktívna skupina je vždy maximálne jedna. Skupiny môže vytvárať ľubovolný prihlásený používateľ a po jej vytvorení sa automaticky stáva jej administrátorom. Skupina má vždy práve jedného administrátora a ten je nemenný.

3.5 Anotácie

Každá anotácia má nasledovné atribúty:

- typ (napr. “osoba”), ktorý ju jednoznačne určuje
- farba, ktorou je text znázornený v editore
- popis
- zoznam podtypov
- zoznam odkazov/linky (nie pre referencie)

Anotácie v textovom editore budú farebne vyznačené podľa daného typu/kategórie. Farbu rôznych kategórií si môže užívateľ zmeniť. Používateľ bude môcť rozšíriť informáciu aj o odkazy, ktoré by sa mali týkať danej NE (napr. pre nejakú známu osobnosť by to mohol byť odkaz na wikipédiu). Referencia je anotovaný text, ktorý odkazuje na nejakú predošlú anotáciu (napr. môže to byť zámeno). Používateľ bude môcť špecifikovať, či daná časť textu je referencia a na akú anotáciu sa bude odkazuje (rodič). Referencie dedia od rodiča typ anotácie a ak sa rodič zmaže/zmení, zmažú/zmenia sa aj všetky referencie, ktoré naň odkazujú. Typ referencie bude vždy závisieť na type jej rodiča.
Anotácie môžu byť zmenené/zmazané používateľmi na základe ich práv. Anotácie sa budú môcť aj prekrývať – prekryv bude nejakým spôsobom rozlíšený (napr. konkrétnou farbou alebo farba prekryvu bude určená zmiešaním farieb jednotlivých prekrývajúcih sa anotácií).

3.6 Zobrazovanie anotácií

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

3.7 Komentáre

Používateľ s právom na reanotáciu bude pridať komentár do textu. Po nadídení myšou/stlačením kláves nad okomentovaným textom sa komentár zobrazí používateľovi. Komentáre môžu byť priamo zmazané len ich autorom. Okomentovaný text bude v dokumente nejakým spôsobom vizuálne rozlíšený.

3.8 Hľadanie v texte

Aplikácia bude umožňovať používateľom hľadať slová v texte.

3.9 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). **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.
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:PER 1:2

Neanotované slovo:

- 11 foo : :

Viackrát anotované jedno slovo:

- 13 bar B:NAME1,NAME2 :

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

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

3.10 Aktivita v dokumente

Aktivita v skupine (anotovanie textu a rôzne iné zmeny) sa všetkým používateľom budú zobrazovať pomocou logu, kde budú všetky dôležité informácie ako čas zmeny, užívateľ, ktorý danú zmenu vykonal, samotná zmena – komentár, zmazanie, anotácia, ...).