Prítomní:	Bc. Róbert Horváth	*Vedúci:*	Ing. Dušan Zeleník
	Bc. Peter Jurčík	*Miesto:*	Softvérové štúdio
	Bc. Peter Macko	*Dátum a čas:*	02.04.2012 16:00
	Bc. Peter Sládeček (zápis)	*Trvanie:*	180 minút
	Bc. Maroš Ubreži (diskusia)		
	Bc. Matúš Vacula		
	Bc. Vladimír Ruman		

Prehľad splnenia úloh z predchádzajúceho stretnutia

ID	Popis	Pridelená	Predpokladané ukončenie	stav	Skutočné ukončenie
3313	logovanie	Peter Sládeček	26.03.2012	ukončená	02.04.2012
3314	spájanie N-gramov	Matúš Vacula, Róbert Horváth	26.03.2012	ukončená	02.04.2012
3315	vytvorenie šablóny dokumentácie	Peter Jurčík	02.04.2012	ukončená	02.04.2012
3316	nájdenie vhodných testovacích viet v korpuse	Maroš Ubreži	02.04.2012	ukončená	02.04.2012
3317	upravenie základného skóre v ElasticSearch	Vladimír Ruman	02.04.2012	ukončená	02.04.2012

Základné body stretnutia

Na stretnutí boli prediskutované nasledujúce body:

- Stav úloh z minulého stretnutia
- Termín ďalšieho stretnutia
- Problém N-gramov
- Návrat k verzii so synonymami
- Úlohy na ďalší týždeň

Stav úloh z minulého stretnutia

Všetky úlohy prenášané z minulého stretnutia sa podarilo členom tímu uzavrieť. Robo pracoval na optimalizácii N-gramov. Objavil problém, ktorý bol následne na stretnutí riešený. Matúš v rámci tejto úlohy nahradil pôvodný anglicko – slovenský slovník novšou verziou. Peter S. pridal do zdrojového kódu prekladača logovanie, ktorého výstup v XML súbore je mapovaný do webového rozhrania. Vladovi na základe analýzy ElasticSearch providerov vyplynulo, že poskytované riešenia nie sú vhodné. Vytvoril teda vlastného providera, ktorého implementoval do

knižníc ElasticSearch-u. Peter J. prezentoval šablónu dokumentácie, ktorú bude nasledujúci týždeň potrebné odovzdať. Maroš hľadal vhodné testovacie vstupy pre náš prekladač. Zistil, že vo vytváraných N-gramoch boli objavené duplicity, ktoré by bolo vhodné odstrániť.

Termín ďalšieho stretnutia

Nakoľko na pravidelný čas nášho stretávania pripadol štátny sviatok, bolo potrebné si určiť iný termín. Členovia tímu spolu s vedúcim sa dohodli, že sa stretnú **v stredu 11.04.2012 o 17:00hod**.

Problém N-gramov

Na stretnutí sa intenzívne riešil novoobjavený problém N-gramov. Postupne sa prišlo k záverom:

- dáta v slovníku majú príliš veľa významov => treba nový slovník
- máme možno príliš malý korpus => slovníkový origin vracia lepšie výsledky
- niektoré slová sú zle prekladané z dôvodu prevodu slov na malé písmená (slovo *„it“* neprekladá iba ako *„to“* ale aj ako *„informačné technológie“*) => väčší počet requestov

Postupne sme analyzovali možné východiská z tejto situácie. Skúmali sme:

1. *upravenie hodnotenia pre kratšie vety* – pokus o ich zvýhodnenie pred dlhými súvetiami.
2. *implementáciu Matúšovho slovníka* – priniesol nám asi o 30% viac záznamov ako predchádzajúca verzia. To spôsobilo i väčší počet N-gramov, ktoré boli vytvárané z relevantných prekladov.
3. *kombinácia N-gramov a synoným* – efektívnosť algoritmu by klesla, pretože synonymá sú aplikované iba na strane response, nie na strane requeste. Treba odskúšať, efektívnosť na strane request.
4. *Vladove skórovanie na synonymá* – použiť bez term frequency. Mohlo by opraviť zlé skórovanie ElasticSearch-u, ktoré „pokazí“ naše ohodnotenie cez N-gramy.
5. *dĺžka slova odstráni spojky a zlepší prekryv* – otázkou je, či by tento parameter mohol ovplyvniť skóre prekryvu. Problémom je však napríklad spojka *„alebo“* a sloveso *„byť“*.
6. *ElasticSearch a časť nenachádzajúca sa v prekryve* – pôvodne sme chceli hľadať neprekrývajúcu časť vo výsledku, ktorý vráti ElasticSearch. Po diskusii sme však usúdili, že v danom výsledku sa nemusia nachádzať presné preklady slov, čím by sme ešte viac zhoršili súčasné riešenie.

Návrat k verzii so synonymami

Z vyššie spomenutej analýzy sme dospeli k záveru, že je vhodné sa opäť vrátiť k verzii, ktorá využíva synonymá bez term frequency. Je potrebné dodatočne upraviť Vladom navrhnuté skórovanie tak, aby bolo úplne nezávislé od ElasticSearch-u. Synonymá, ktoré vyjadrujú rovnaký význam, by bolo vhodné zoskupiť do jednej skupiny tak, aby sme vedeli lepšie zoptimalizovať počet requestov. Otázne je, či ich bude možné dostať z ElasticSearch-u, alebo bude potrebné navrhnúť vlastnú HashMapu, v ktorej budú umiestnené. Na stretnutí sme sa dohodli aj na penalizovaní viet na základe ich dĺžky.

Úlohy na ďalší týždeň

Vzhľadom na odovzdávanie ďalšieho väčšieho celku dokumentácie sme sa dohodli, že vyššie spomenuté výsledky analýzy budú zapracované až v nasledujúcom šprinte. Členovia tímu majú teda za úlohu spísať dokumentácie z riešení, na ktorých pracovali. Peter J. ich následne spojí do jedného celku tak, aby boli v ďalší deň pripravené na odovzdanie vedúcemu tímu. Úlohou Maroša je i prečistiť duplicity v N-gramov pre prípad, že by sme sa ich v budúcnosti snažili zapracovať do nášho riešenia.

Prehľad úloh vyplývajúcich z diskusie

ID	Popis	Pridelená	Predpokladané ukončenie	stav	Skutočné ukončenie
3318	odstránenie duplicity N-gramov	Maroš Ubreži	11.04.2012	nová	
3319	vytvorenie jednotlivých častí dokumentácie šprintov	členovia tímu	11.04.2012	nová	
3320	pospájanie dokumentácie do finálnej verzie	Peter Jurčík	11.04.2012	nová	

Spracoval: Bc. Peter Sládeček

Overil: Ing. Dušan Zeleník