Тема: ***Solr* сервер и процесирање природног језика**

Наставник: **Бобан Стојановић**

Количина података на Интернету достигла је астрономске размере. Неки од тих података су прегледани и оцењени од стране хуманог фактора и због тога су високог квалитета, али већина веб садржаја је неконтролисана, сирова и пристрасна. Међутим, и поред тога, свака инфорамација може бити од корисити некој циљној групи. Додатна отежавајућа околност је да је већина садржаја није уређена у форматима који одговарају рачунарској обради, као што су XML и JSON, него у сировим сегментима текста написаним на природном језику.

Очигледно је да мануелна класификација, организација и претрага, услед претходно наведених разлога не долази у обзир. Такође, класични алгоритми за анализу и претрагу података имају велике потекшкоће да се носе са толиком количином неуређених информација. Срећом, заједница отвореног кода је развила изузено моћну бибилиотеку за индексирање и претрагу текстуалних података по имену *Lucene,* која се данас развија у оквиру *Apache Software Foundation*. *Lucene* користи принцип индексирања који се зове обрнути индекс, и проналази документе према терминима (речима) који се налазе у том документу. Поред тога, садржи и пакете које могу анализирати различите језике, уклањати честе речи (*stop words*) и сводити термине на корен речи (*stemming*).

Предмет рада требала би да буде инсталација, конфигурација и тестирање перформанси *Solr* сервера високих перформанси који функционише као омотач око *Lucene* језгра. Додатне функционалности којима *Solr* унапређује *Lucene* су могућности за имплементацију дистрибуираног система са подршком за балансирање оптерећења, распарчавање индекса и репликацију индекса. Како је тзв. *Big Data* инжењерство тренутно веома тражена информатичка дисциплина, овај рад би требало да представља пионирски подухват у том смеру и евентуално резултира будућим увођењем *Big Data* садржаја у редовну наставу на Мастер студијама информатике.

Литература


1. Michael McCandless, Eric Hatcher, Otis Gospodnetić, *Lucene in Action*, Manning Publications Co. www.manning.com, 2010.
2. Rafal Kuc, *Apache Solr 4 Cookbook*, Packt Publishing Ltd. www.packtpub.com, 2013.
3. Grant S. Ingersoll, Thomas S. Morton, Andrew L. Farris, *Taming Text*, Manning Publications Co. www.manning.com, 2010.
4. John Gantz, David Reinsel, *Extracting Value from Chaos*, www.emc.com