Akshay Shirahatti

Interests Information retrieval, Distributed systems, NLP & statistical methods for large datasets.

Highlights

- 5+ years of experience working in the **Big Data** & **Information Retrieval** domain
- Currently working as **Big Data Engineer** at *DataSift*.
- Worked as Data Mining Engineer at *Mendeley*.
- Awarded **MSc in Computer Science** with *Distinction* from *University of Edinburgh*
- Experience in building large scale data processing pipelines

Skills

- Scala, Java, Python(scripts), Pig(prior experience), Linux shell scripting, R(scripts)
- Elasticsearch, Kafka, Hadoop, Apache Lucene, HBase, Solr, AWS, Chef, Apache Spark(Workshop)
- Git, Linux Operating System, Continuous Integration.

Experience

Big Data Engineer
DataSift Ltd, Reading, UK **2012 - present**

Technology Stack: Scala, Java, Kafka, Elasticsearch, Hadoop, HBase, MySQL, Akka, AWS

- Prototyped and productised a *highly-scalable*, *low-latency* data analysis platform to handle Facebook's real-time data stream.
- Developed performance/feasibility evaluation and QA tools to de-risk the pilot product.
- Delivered privacy-first, anonymised APIs to help extract insights from the Facebook Data.
- Helped build the archive of *twitter, bit.ly, tumblr etc* on *HDFS*.
  - Archiving over 2TB/day working with a 200+ nodes *Hadoop* cluster.
  - Custom *HDFS* compaction workflow to satisfy the deletes processing SLA
  - Fast sampled insights over the archive to estimate volumes and predict cost

Data Mining Engineer
Mendeley, London, UK **2010 - 2012**

Technology Stack: Java, Apache Solr, Lucene, Hadoop, HBase, Pig, MySQL, AWS, Voldemort

- Scaling the Mendeley search platform and improving the relevancy ranking model for documents, users, groups in the Mendeley ecosystem.
- Developed a solution(based on http://tinyurl.com/cd4klph) to aid exploratory tag based navigation of the Mendeley document corpus.
- Worked on an analytic product to compute and serve top papers/journals/authors in a discipline/domain using Apache Pig.

Education

M.S Computer Science (Distinction)
University of Edinburgh, Scotland, UK **2009 - 2010**

Relevant Academic Projects

- Text Retrieval for Systematic Reviews(Dissertation): To determine how *Latent Dirichlet Allocation (LDA)* compares with pseudo-relevance feedback methodologies like *Query expansion* and *Relevance-based language models* in producing the required level of precision and recall needed for Systematic Reviews. The dataset consisted of 10 million documents from PubMed corpus. Mallet toolkit was used for LDA implementation, query-document vector distance computation was done on Hadoop.
- Crawler for efficient content extraction from a set of hyper-linked news stories & also design search, relevance ranking, near-duplicate detection solutions for the corpus.

Key Modules: Text Technologies(Information Retrieval), Distributed Systems, Parallel Algorithms & Programming, Advanced Databases.

Bachelor of Engineering, Computer Science
University of Mumbai, India **2004 - 2008**