TOPIC MODELING FOR TECHNOLOGY SCOUTING

Sean P. McKenna

Introduction

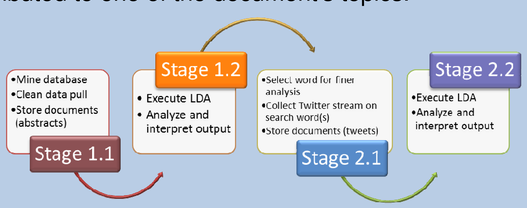
Determining where to invest time and resources is a key challenge faced by many Booz Allen clients, not to mention Booz Allen itself. It is a process that generally falls on the shoulders of Subject Matter Experts and others in Business Intelligence, Marketing, and Leadership roles. The process can be time-consuming, subjective, and inconsistent. Enter Topic Modeling (TM) – a blend of machine learning and natural language processing that seeks to uncover hidden topical patterns that pervade a collection of documents. By applying TM to an appropriate corpus of text (e.g., scientific abstracts or articles), one can begin to create a picture of the "technical landscape," and, even better, observe how that landscape evolves over time.

Approach

The essence of TM is data mining of text documents relevant to a particular topic of interest. In this case, the term "documents" can refer to anything from journal abstracts to web/blog pages to tweets. A key to TM is to identify the most appropriate source of documents for a given topic. Journal abstracts are a reasonable place to start for many topics, thus, selecting a journal citation database (e.g., INSPEC, JSTOR, Web of Science, IEEE *Xplore)* is typically the first step.

Using advanced text analytics (Latent Dirichlet Allocation, LDA), topics can be extracted from an initial corpus. From these results, other keywords/topics may emerge that warrant further investigation. This can be achieved by a second step in which a new corpus is used – one that has finer granularity, e.g., in time/content. For instance, scraping certain websites or blogs, newsfeeds, or Twitter content may be appropriate secondary, low-latency sources of information. These types of sources would contain more up-to-date material, thus providing a better picture of the current landscape. This second corpus is analyzed in a manner similar to the first, and a final sense of the topic of interest can be achieved.

LDA is a generative probabilistic modeling approach that can uncover thematically-related terms in a document corpus. Essentially, it posits that each document is a mixture of a finite set of topics, and that each word appearing in the document can be attributed to one of the document's topics.



Results

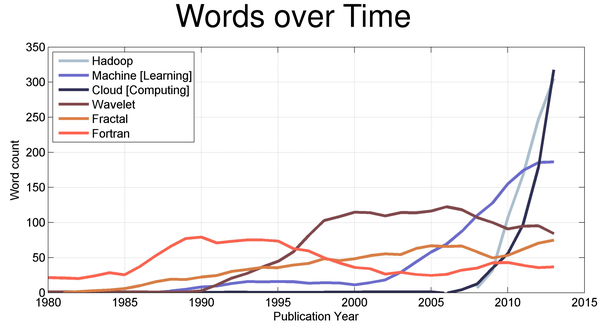
As a way to explore the efficacy of the TM approach, the term "Big Data" was selected. The intent was to see if TM could produce insight into the current state of Big Data and what might be currently trending in this field. For the first pass, IEEE *Xplore* was chosen as the database and, not including 2014, 1,314 records were extracted (dominated by conference publications). The number of topics was set to eight, and from the LDA output, we created word clouds (below) showing the top ten most likely words in each topic. From these topics and topic words, "Hadoop" was chosen for deeper investigation using the real-time Twitter stream. LDA was used again, this time with four topics, and the results are shown below-right. These results provide a finer sense of that particular technology, illuminating a variety of Hadoop providers and implementations, arguably providing a valuable sense of the current market space. Lastly, using the IEEE abstracts, we were also able to look at how terms trended over time (bottom-left).

"Big Data" Word Clouds

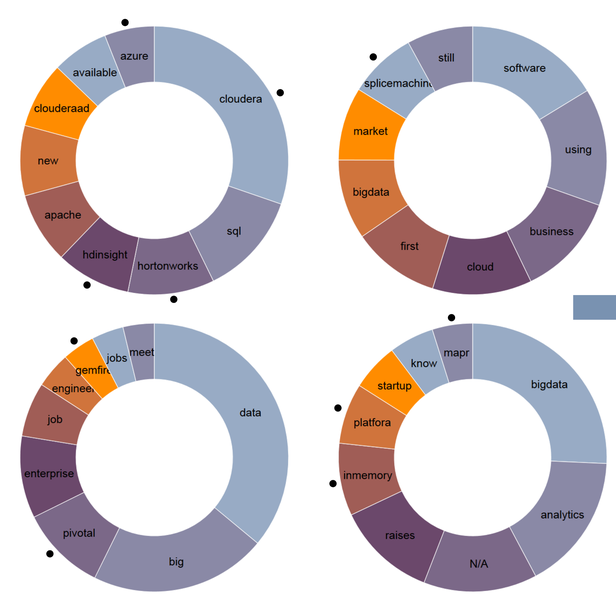
"Big Data" Topics

	Topic
A	Databases, storage, graphs
B	Analytics, [data] mining
C	Machine [learning], complex systems, sets, quer[ies]
D	[Map]Reduce, problems, techniques and process[es]
E	Social networks, (large) network analysis
F	Cloud computing, [something-as-a]-service
G	**Hadoop**, MapReduce, algorithms, parallel processing
H	Modeling, [machine] learning, approaches and methods

Words over Time



Hadoop Topics



Summary

This study has illustrated how a TM tool can be used to shed light on a particular topic of interest in a largely automated fashion. As an example, we looked at "Big Data," and using the TM methodology were able to gain a broad sense of the field, capturing a high-level representation of the major aspects of the Big Data space. Drilling down into "Hadoop" yielded a finer sense of that particular technology.

It should be noted that different subject keywords likely will necessitate different databases and different secondary sources. Twitter worked well for this case study, but would probably not be appropriate for a subject like ballistic missile analysis. Furthermore, website scraping would add power to this TM scheme – one could get to the end state faster than monitoring Twitter for extended periods of time.

Finally, not all subjects will be amenable to an approach such as TM. In such cases, the onus for technology scouting will remain with SMEs and others. Thus, ultimately, this effort is an experiment to see if such a scheme is realistic and of value. Initial impressions are that the approach has potential, but there will always be a need for SME input to the process to make it meaningful.

Attribution

The idea for this project came from the work of a Booz Allen Science Technology Analytic Intelligence Cell effort that developed a tool capable of using next generation text analytics to analyze unstructured data, with the goal of enabling a team of analysts to overcome information overload and quickly understand a new topic with little a priori knowledge.

- **splicemachine** Splice Machine is a transactional SQL-on-Hadoop database for real-time Big Data applications.
- **pivotal** Pivotal is a Platform-as-a-Service (PaaS) provider. It is a spin-off of VMware and EMC that seeks to compete with Amazon Web Services.
- **gemfire** GemFire is a Pivotal product that provides a real-time distributed data store.
- **cloudera** Cloudera is a software company that provides Apache Hadoop-based software, support and services, and training to business customers.
- **azure** Windows Azure is a cloud computing platform and infrastructure, created by Microsoft, for building, deploying and managing applications and services through a global network of Microsoft-managed datacenters. It provides both PaaS and IaaS services and supports many different programming languages, tools and frameworks, including both Microsoft-specific and third-party software and systems.
- **hdinsight** HDInsight is a Hadoop-based service from Microsoft.
- **hortonworks** Hortonworks is a business computer software company that focuses on the development and support of Apache Hadoop.
- **platfora** Platfora is a Big Data analytics company whose software works in tandem with the open-source software framework Apache Hadoop to assist companies and government organizations with rapid data analysis, data visualization, and sharing.
- **inmemory** In-memory refers to an emerging construct of storing (relatively large amounts of) data in local memory as opposed to on disk.
- **mapr** MapR is a company that provides a Hadoop platform.