sparkcognition™
Deep NLP

The Challenge of Unstructured Data

Structured data is data that can easily be represented as a table—think spreadsheets. If your data contains only numbers and categories, existing analytics tools can easily handle it. However, this accounts for only about 20% of data produced by organizations.[1]

The rest is the messier twin, unstructured data: PDFs, books, journals, audio, video, images, notes, analog data, and any other source imaginable. This unstructured data is primarily meant for human use and consumption but hard to analyze at scale, so it's rarely leveraged in aggregate to provide deeper insights. Businesses carefully (or not so carefully) collect this data and file it away, where it sits unused, rarely ever to see the light of day. Despite the current fervor surrounding the power of big data, this is the fate of 80% of all data produced.These numbers aren't likely to go down anytime soon. The total amount of data, both structured and unstructured, is increasing year over year by 39%.[2] IDC and EMC both project that data will grow to 40 zettabytes by 2020, which would be a growth of 50 times in 10 years.[3] To put these numbers further in perspective, linguist Mark Liberman has calculated that all human speech ever spoken in the history of humankind would total to about 42 zettabytes.[4] And 80% of this is unstructured. The amount of unstructured data being produced and stored is beyond human comprehension—and almost none of it is being used. All the while, organizations struggle with bottlenecks in organizational workflows, increasing operational costs, lack of visibility into processes, and the loss of tribal knowledge from the workforce.

Why is this massive resource allowed to lie fallow, even as businesses hook up increasing numbers of sensors to try to fuel data analytics? Because unstructured data cannot be understood and analyzed by most machines, accessing and analyzing unstructured data is a difficult, expensive prospect. Humans can understand this data, of course, but analysis driven entirely by humans doesn't scale to large operations, opens up the risk of human errors, and is a waste of time and resources. Human beings may be phenomenal at understanding language and images, but they're not equipped to handle data on the order of zettabytes.

While machines aren't naturally inclined to understand unstructured data, they can be taught. Natural language processing, or NLP, is a field dedicated to teaching machines to use and understand language in a human-like fashion. Major airlines are already using NLP, paired with machine learning, to extract tribal knowledge hidden in maintenance logs and make it available across the workforce. By leveraging historical data, machine learning is also elevating historically proven troubleshooting techniques to help diagnose and solve problems faster and more effectively.

NATURAL LANGUAGE PROCESSING

The field of NLP has existed in its modern form longer than even the study of artificial intelligence, with work on automatic translation and similar projects dating back to the early 1950s. But it's the recent machine learning boom that has revolutionized the subject, allowing it to flourish in new ways. This is because machine

Unstructured Data

The Lifeblood of Organizations

Process Logs	Event logs, server data, application logs, business process logs, audit logs, CDRs, mobile location ...
Sensor Data	Medical devices, electric meters, cameras, ECUs, engines, HVAC, machinery, high value assets ...
Business Data	Project management, marketing, productivity, CRM, contracts, procurement, HR, expenses ...
Public Records	Government, competitive, regulatory, compliance, health care services, public finance, stock ...
Archives/Storage	Scanned documents, statements, insurance forms, customer information, paper archives, system records ...
Documents	XLAS, PDF, CSV, email, DOCX, PPT, HTML, plain text, XML, JSON ...
Social Media	Twitter, LinkedIn, Facebook, Tumblr, Blog, SlideShare, Youtube, Google+, Instagram, Flickr, RSS, Pinterest ...
Media	Images, videos, audio, Flash, live streams, podcast, webinars ...

learning revolves around writing algorithms that can learn beyond their initial programming, rather than being constrained by the rules coded into them. Rather than trying to hand-code all of the rules of language—a daunting task even if the scientific community agreed on them—programmers feed text into a machine learning program and let it glean the rules for itself, often using probabilistic models to figure out usage in a more fleshed-out, natural way. This also makes improving the model easier. Instead of writing rules of increasing complexity, simply feed the model more text and let it learn how a human might.

NLP technology has enormous implications for businesses and organizations, specifically in how it allows computer programs to understand unstructured data and leverage it for analysis. By automating workflows of unstructured data, NLP can drive and streamline high-value business decisions. This includes minimizing operational costs, reducing the risk of human error, and gaining visibility and insight into processes to drive decision-making.

SAMPLE USE CASES FOR NLP

Here are a few examples of how SparkCognition™ Deep NLP has been incorporated into the workflows of major businesses.

Maintenance Advisory Application

Using deep learning, SparkCognition developed an advisory tablet application for aircraft front-line staff. This application allowed maintenance technicians to conduct machine-to-human dialogue to troubleshoot asset failures and mechanical issues with high accuracy, assess faults and troubleshoot using queries in natural language, and optimize their workflow and deliver relevant documentation with a faster turnaround. This solution lowered the cost of maintenance and improved asset availability for operators by up to 10%.

Financial Document Classification

SparkCognition is enabling digitization and compliance processes for a major bank with 900,000 contracts under management and a daily global transaction volume of $3 trillion. Each transaction requires access to a wide range of document types, many of which are unstructured. It takes roughly 1,000 human touchpoints and 72 hours to reconcile a single transaction. Using machine learning and NLP, SparkCognition is extracting information and classifying financial documents to support compliance, with a goal of increasing accuracy and reducing transactions to only 50 human touchpoints.

To learn more about Deep NLP and how it unlocks unstructured data for organizations, visit https://www.sparkcognition.com/product/deepnlp/.

ABOUT SPARKCOGNITION

SparkCognition's award-winning AI solutions allow organizations to predict future outcomes, optimize processes, and prevent cyberattacks. We partner with the world's industry leaders to analyze, optimize, and learn from data, augment human intelligence, drive profitable growth, and achieve operational excellence. Our patented AI, machine learning, and natural language technologies lead the industry in innovation and accelerate digital transformation. Our solutions allow organizations to solve critical challenges—prevent unexpected downtime, maximize asset performance, optimize prices, and ensure worker safety while avoiding zero-day cyberattacks on essential IT and OT infrastructure.

To learn more about how SparkCognition's AI solutions can unlock the power in your data, visit www.sparkcognition.com.

REFERENCES

[1] https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem/#249dac5a493a

[2] https://www.veritas.com/news-releases/2016-05-18-new-veritas-research-information-governance

[3] https://www.kdnuggets.com/2012/12/idc-digital-universe-2020.html

[4] http://itre.cis.upenn.edu/~myl/languagelog/archives/000087.html

Contact SparkCognition today at info@sparkcognition.com