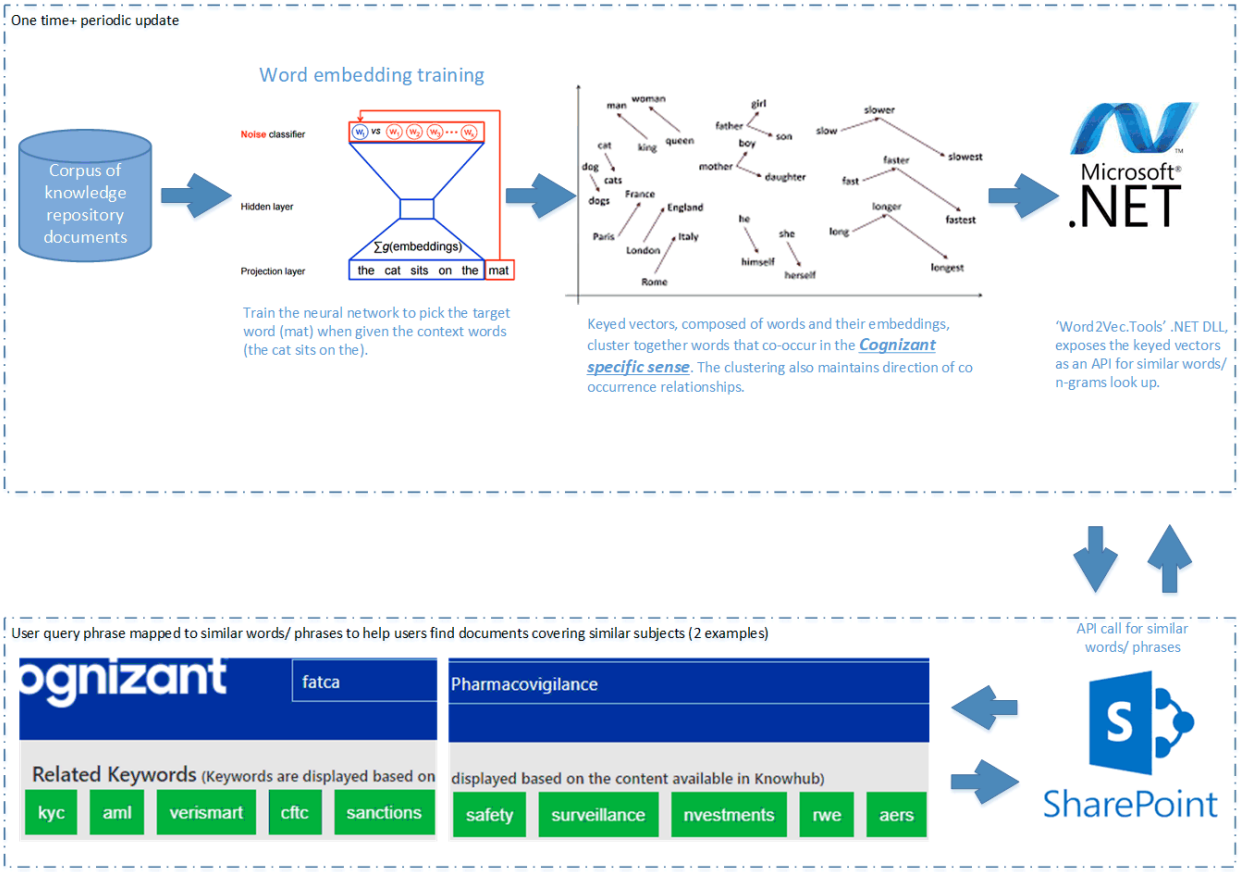
Semantic search for artifacts

Introduction:

Reuse of assets, like ideas and their implementations, is critical. Through careful improvements to such reusable assets over time, their quality and impact rise up to the level of being a competitive advantage. In the best of cases, such assets can provide non-linear/ licensing revenue opportunities.

Discoverability of reusable assets is one of the challenges faced. This is so, despite corporate document repositories containing documents that describe the reusable asset. The same concepts are expressed differently by different team members and by the same people at different times. While the repositories support tagging to help discoverability, the consistency of such expert tagging is doubtful.

Solution - Semantic similarity search:



Microsoft's powerful SharePoint is the document repository at Cognizant. We extended its impressive search capabilities using the following approach.

1. Word similarity: Semantically similar words frequently occur in the context of the same word groups (sentences). Extensive documentation for this assertion exists and we refer the reader to Mikolov et al's paper at https://arxiv.org/abs/1301.3781.

2. Cognizant specific word embeddings: While many pre-trained embeddings exist, they are trained on generic documents like news and Wikipedia articles. However, by training a custom embedding for the ~100,000 documents in Cognizant's repository, we were able to exploit domain specific word similarities.
3. Word embedding implementations: There are many word embedding implementations, each based on different techniques. The reader is referred to this white paper for details https://arxiv.org/pdf/1901.09785.pdf. We chose to work with a Word2Vec skip-grams library to generate "keyed vectors", which contains the words and n-grams that occur in a corpus of documents, and their vector representation. This space is rapidly evolving and further evaluation is warranted.
4. .NET Word2Vec DLL: The above "keyed vectors" are exported as a binary file that is then read by the Microsoft Word2Vec.Tools DLL. This DLL exposes an API through which similar words and n-grams can be looked up.
5. SharePoint: When users enter a search terms the documents corresponding directly to the search term are displayed as usual. In addition, a parallel call is made to the Word2Vec API and a ribbon of similar terms is populated, clicking on which updates the documents returned. Users can now look at documents that match the exact search term or documents that match similar search terms.

Results:

As seen in the image above, searching for pharmacovigilance prompts the user to also consider documents covering Real World Evidence (RWE) and Adverse Event Reporting System (AERS). A user, like a Business Development team member, can effectively gather Cognizant's artifacts related to a particular deal she is working on and not leave out important information from her sales pitch.

Such benefits are delivered to Cognizant employees across industry verticals, technology horizontals and different roles (e.g. a technical architect might search for a reusable Continuous Integration/ Continuous Deployment pipeline design without having to rebuild one).

Users have now reported much improved satisfaction levels with the repository as a reuse enabler.

Image sources:

1. Word embedding training: TensorFlow team
2. Word vector similarity https://medium.com/analytics-vidhya/implementing-word2vec-in-tensorflow-44f93cf2665f

About the Authors

Bala Kesavan is a Data Scientist, Predictive Analytics, Delivery Excellence, within Cognizant Digital Systems & Technology line of service**,** specializing in natural language processing. He focuses on tracking this continuously evolving space and applying new approaches to solve new problems and to improve current solutions. Bala has domain experience in manufacturing and banking apart from IT projects in a career spanning 20-plus years. He has a PGDM from MDI, Gurgaon and can be reached at BalakrishnaSaravanan.Kesavan@cognizant.com | https://www.linkedin.com/in/scmguru/.

Akhil Goyal is a Director, Predictive Analytics, Delivery Excellence, within Cognizant's Digitial Systems & Technology line of service, where he leads the design and development of advanced analytics solutions across the company's delivery excellence initiatives such as knowledge management and reuse, risk management and contract management. He has over 20 years of experience in the IT industry, across delivery management, client engagement, risk consulting and data science. He has led multiple programs for global, *Fortune* 500 clients and successfully deployed large organization change initiatives. Akhil received a B. Tech. degree in electrical engineering from Indian Institute of Technology, Delhi. He can be reached at Akhil.Goyal@cognizant.com | www.linkedin.com/in/akhilgoyal1.

Ravishankar Ganesan is Vice President, Delivery Excellence within Cognizant's Digital Systems & Technology line of service, with the responsibility of strengthening pursuit robustness through market intelligence, industry benchmarks, risk assessments, solution and estimation review. He is also responsible for driving improvement in managed services revenue and deploying advanced analytics solutions to solve business problems. Ravi has over 30 years of diverse experience in quality assurance and delivery excellence and is a certified Black Belt and a certified assessor on MBNQA (Malcolm Baldrige National Quality Award). He holds a bachelor's degree in mechanical engineering from Annamalai University, Tamilnadu, India; a postgraduate diploma in statistical quality control and operations research from Indian Statistical Institute; and an MBA from Indira Gandhi National Open University (IGNOU). He can be reached at Ravishankar.Ganesan@Cognizant.com | https://www.linkedin.com/in/ravishankar-ganesan-43309823/.