The Documents and Text Collection for Pipeline Pilot brings together the utility of document search, analysis and manipulation with the power of process automation and data integration in Pipeline Pilot. The Doc and Text collection uses natural language processing, advanced statistical modeling, and subject-specific ontologies and thesauri to uncover patterns and extract information from text content. Additionally, it provides extensive capabilities for managing and manipulating documents. The Documents and Text Collection can help you achieve your most challenging document analysis and manipulation objectives by linking together search, analysis and reporting steps into automated routines. Integrate literature mining and text analytics with your existing scientific protocols, and run them interactively or automatically every night.

THE DOCUMENTS AND TEXT COLLECTION, PROVIDES A SUITE OF CAPABILITIES FOR:

Documents and Search
- Search PubMed, US and European Patents, Twitter, Bing, websites, SharePoint documents, local files and enterprise databases and integrate with third-party search engines, to find the documents of highest value and importance
- Search multiple data sources easily using a single query language with phrase, wildcard, fielded and synonym matching, allowing searches to be created and maintained efficiently
- Create searchable databases of key documents for ongoing review during the course of a project
- Create rich, interactive reports to explore and mine your documents and text analyses, including specialist visualizations such as Tag Clouds, highlighted terms in reports and document clustering
- Generate Word reports from templates by automatically adding content to Microsoft Word documents to streamline the creation of internal reports and external regulatory documents, reducing error-prone manual document editing
- Do web crawling to extract content and tabular data from the Web

Text Analytics
- Analyze documents / text to extract key concepts to find correlations in documents and online literature for competitive intelligence and to support primary research findings
- Automate information extraction and summarization
- Use relationships between documents for greater understanding

Chemical Text Mining
- For biology, identify Amino Acids, Proteins, DNA, RNA, Cell Lines, and Cell Types.
- For chemistry, identify chemical names and embedded structures in text and convert these to structures with the included OpenEye name-to-structure converter.
  - Names: systematic names (IUPAC, InChI, and SMILES), formulae, family names, abbreviations, identifiers, CAS numbers, and non-systematic (trivial) names
  - Embedded Structures: BIOVIA Draw, ISIS/Draw, ChemDraw and Rccord molecules
- Create chemically-aware search applications by mapping structures to documents in a structure-searchable text database.
SEARCH: LOCAL, ONLINE & ENTERPRISE DOCUMENTS
The Documents and Text Collection gives you the power to extract knowledge from important online document resources such as PubMed, US and European Patents, Bing, TOXNET, Twitter and Wikipedia (user extendable to other remote text data sources). Search these databases with interactive queries, or mine them with large-scale document retrieval and characterization routines. You can also search and mine internal documents in exactly the same way. The Documents and Text Collection indexes and searches folders that contain PDF, Microsoft Office, HTML, or text files (extendable to other file formats). You can even store the results of online searches in local repositories for speedy retrieval and post processing. Local databases of documents stay current automatically by monitoring the folder contents for the introduction of any new or edited documents.

CHEMICALLY-AWARE SEARCH
The Documents and Text Collection uses a chemically aware search algorithm to identify and convert chemical names to structures and enable researchers to query for information in a more intuitive and effective way with substructure, similarity, SMILES, and IUPAC name searches. Also included is a complete end-user application, "Chemically Aware Search", for indexing and searching collections of documents containing structures. It allows you to identify a set of interesting documents, download the content to a local database and index the documents for text or structure search, all without the need to author protocols or additional configuration.

ANNOTATE SCIENTIFIC RESULTS
When reporting the results of pipelined data analyses, it is often useful to include additional information about the output data points. With the Documents and Text Collection, you can easily add a few steps at the end of any Pipeline Pilot protocol and have each data point serve as a query to search a database of literature. For example, after clustering a set of genes with the Biology Collections, you can annotate each gene with summary information from its top reference in PubMed (and a link to further search results). This kind of enhancement makes for more easily interpretable results. Also, you can create your own custom document templates in Microsoft Word, and use the Documents and Text Collection to surgically fill in scientific data and charts to generate beautifully styled final documents.

IDENTIFY EMERGING TRENDS
The Documents and Text Collection can monitor the scientific literature for topics of interest, and it can even alert you when new concepts are emerging for those topics. The latter is achieved by searching for new articles about your topic of interest and detecting the concept words they contain. The association of each concept with the topic of interest is calculated over time to detect emerging new relationships. This allows you to stay on top of a broader class of topics, and learn about breakthroughs before they become widely known.

MINE PATENT DATABASES
The Documents and Text Collection provides you with the tools necessary to characterize research and intellectual property trends in a field of interest. You can search and process the U.S. patent databases (extendable to other patent databases) for trends reflecting the quantity of patents, application areas, companies engaged, and more. For example, by building a protocol to process patents in the field of fuel cells, you can discover how rapidly this emerging field is growing. You can also see that applications for automobiles have come to dominate the area and that Honda and General Motors are leading innovators.
ABOUT PIPELINE PILOT

BIOVIA Pipeline Pilot is a graphical programming environment that optimizes scientific innovation, increases operational efficiency and reduces costs for both research and IT. Pipeline Pilot automates the scientific analysis of data, enabling users across the enterprise to rapidly explore, visualize and report research results. Pipeline Pilot’s Component Collections contain the “scientific building blocks” that allow researchers, developers, engineers and IT professionals to perform both science-specific and generic data-processing functions. By graphically combining components, you can construct workflows for data retrieval, filtering, analysis and reporting.

To learn more about Pipeline Pilot, go to 3DSBIOVIA.com/pipeline-pilot