# YTEX 3.2

## YTEX Overview

YTEX provides the following functionality

- **Semantic Similarity**: YTEX provides a framework for computing the similarity between pairs of concepts; this is integrated with clinical NLP, Data Mining, and Feature Engineering tools.
- **Data Mining**: YTEX provides tools to export UIMA annotations to machine learning packages, including Weka, R, Matlab, SAS, Libsvm, SVMLight, and others.
- **Feature Engineering**: YTEX provides tools that utilize the taxonomical structure of the Unified Medical Language System (UMLS) to improve feature ranking. YTEX implements semantic kernels to integrate semantic similarity measures with machine learning algorithms for document classification.

TODO: YTEX will be added to cTAKES 3.2 (this page and children will wander over to the cTAKES 3.2 documentation when it is ready).

- User's Guide
- YTEX Installation

### Clinical NLP

YTEX provides analysis engines that provide the following functionality:

- Store annotations in a relational database: This allows you to view and extract annotations easily using the structured query language.
- Export annotations to data mining toolkits: YTEX provides tools to declaratively export annotations to WEKA, R, Matlab, and other toolkits /platforms.
- Sentence Splitting: YTEX provides a sentence splitter that does not automatically split sentences at newlines.
- Sectionizer: YTEX identifies document sections using configurable regular expressions.
- Named Entity Recognition: In addition to a dictionary lookup algorithm, YTEX identifies named entities using configurable regular expressions.
- Word Sense Disambiguation (WSD): Named entities may be ambiguous and can be assigned multiple concepts. We've implemented a WSD Annotator that utilizes semantic similarity measures to identify the best concept for a word.
- Negation Detection with Negex: The assertion cTAKES module provides negation detection and much more, but is very heavyweight (the models loaded increase memory and cpu utilization). YTEX provides a lightweight negex-based negation detection module.

For a high-level overview, refer to our paper: Garla, V. et al. The Yale cTAKES extensions for document classification: architecture and application. Journal of the American Medical Informatics Association (2011). doi:10.1136/amiajnl-2011-000093

### Semantic Similarity

YTEX provides a generalizable framework for the computation of semantic similarity measures from any domain ontology; we have tested SNOMED-CT, MeSH, and the UMLS Metathesaurus, and provides programmatic, command line, and a RESTful and XML web services interfaces to users to compute similarity measures. We provide a publicly available web service to compute semantic similarity measures:http://informatics.med.yale.edu/ytex.web/.

For a high-level overview of the semantic similarity methods we've implemented, refer to our paper: Garla, V and Brandt, C. Semantic similarity in the biomedical domain: an evaluation across knowledge sources.

### Data Mining Integration

Data mining tools and statistics packages can access document annotations directly from the database. To support machine learning approaches, YTEX provides tools to extract document features in a bag-of-words representation for use with various data-mining platforms.

### Feature Engineering

For a high-level overview of the feature engineering methods we've developed, refer to our paper: Garla, V and Brandt, C. Ontology-Guided Feature Engineering for Clinical Text Classification (2012) http://dx.doi.org/10.1016/j.jbi.2012.04.010.

We developed novel information-theoretic techniques that utilize the taxonomical structure of the Unified Medical Language System (UMLS) to improve feature ranking, and we developed a semantic similarity measure that projects clinical text into a feature space that improves classification. We improved the performance of the top-ranked machine learning-based system from the I2B2 2008 challenge with these methods.

These feature engineering tools are included in YTEX. We illustrate their usage on the I2B2 2008 Challenge dataset; refer to Feature Engineeringfor a detailed explanation on how to apply these tools

### Text Processing

#### Negation Detection with Negex

YTEX provides the NegexAnnotator, a drop-in replacement for the cTAKES Negation Detection Annotation algorithm. This annotator is based on the Negex source from the Java GeneralNegex package. Advantages over the cTAKES algorithm include:

- long-range negation
- configurable pre- and post- negation triggers

## SegmentRegexAnnotator

To identify document sections in flat files, YTEX includes an annotator that identifies section headings and boundaries based on regular expressions. This relies on a configurable table of regular expressions and segment ids.

## NamedEntityRegexAnnotator

The DictionaryLookup algorithm of cTAKES relies on a lookup table that contains the lexical variants of a named entity. Some concepts have far too many lexical variants to be captured in a lookup table. To address this issue, YTEX includes an annotator that uses regular expressions to identify named entities.

## DBConsumer

YTEX tools stores document annotations in a relational database, thereby simplifying document feature extraction: annotations can be retrieved using the structured query language (SQL).

UIMA annotations are limited in complexity and obey a strict class hierarchy. These restrictions on the structure of UIMA annotations facilitate a high-fidelity relational representation. YTEX currently supports MS SQL Server, Oracle, and MySQL.

## DBAnnotationViewer

YTEX includes a modified UIMA AnnotationViewer to view document annotations directly from the database. In addition to the relational representation; the DBConsumer stores the XML representation of the UIMA CAS of each document in the database. The DBAnnotationViewer simply retrieves the XML CAS for viewing.