Contents

About This Guide 5
Revision History 5
Obtaining More Product Information 6
Documentation Comments/Feedback? 6
Technical Support 7
  Contacting Technical Support 7
  Licensing Portal 8
  Customer Care 8
Product Documentation 9

Overview 11

What is Transparent Predictive Coding? 11
  How Does Predictive Coding Work? 11
  What Predictive Coding is Not 11
Using Transparent Predictive Coding for Optimal Performance 12
Types of Transparent Predictive Coding Users 12

Transparent Predictive Coding Best Practices 13

Understanding the Workflow 13
  Items with No Text 13
  Early Case Assessment 13
  Transparent Predictive Coding Phases 13
Best Practices Workflow 14
Folders 15
Workflow Walkthrough 16
Detailed Workflow 17
  Setup 17
  Train and Test 19
  Predict and Finish 20
Alternate Workflows 21
  Prioritizing Linear Review 21
  Using Alternate Sampling Methods 21

Using Transparent Predictive Coding 23

Workflow and Operating Considerations 24
Transparent Predictive Coding Actions 25
Creating a Control Set 27
  Selecting an Controlled Prediction Accuracy Test Control Set Tag 27
  Step 1: Create Control Set (Initial Sample) 29
  Step 2: Enhance Control Set (Additional Sample) 30
  Step 3: Run the Test 32
Training the System 35
  Step 1: Select Predictive Tag 35
  Step 2: Select Items 36
  Step 3: Confirm Training Selections 37

View Job and Training Error Report 38

Predicting 39
  Step 1: Select Predictive Tag 39
  Step 2: Select Items 40

Selecting the Next Training Set 42
  Step 1: Select a Next Training Set Tag 42
  Step 2: Select Items for Next Training Set 43
  Step 3: Copy or Move Items to Folder 44
  Step 4: Confirm the Next Training Set Items 44

Prediction Accuracy Test 45
  Step 1: Select Prediction Test Tag 45
  Step 2: Select Items 46
  Step 3: Confirm Prediction Test Item Selections 47
  Step 4: View the Prediction Test Report 48

Managing Transparent Predictive Coding 49
  Case Settings for Transparent Predictive Coding 49
  Predictive Tag and Training Items Management 49
  Prediction Status 49
    Viewing Predictive Tag Information 50
    Import Training 51
  Using the Review Dashboard 53
  Granting/Restricting Transparent Predictive Coding Access 54

Predictive Analytics with Search and Review 55
  Viewing Prediction Rank Search Results 55
  Reading Prediction Rank Results 56
    Viewing Detail Prediction Rank Results for Attachments and Embeddings 57
  Filtering Prediction Rank Results 58
  Prediction Searches 59
  Viewing Prediction Rankings 60
    Viewing Tagging Prediction Rank: Review Mode 60
    Viewing Attachment Prediction Rank: Review Mode 61
    Viewing Predictive Tagging events: Review Mode 62
    Understanding Predictive Ranking Decisions 62
    Predictive Tag Analysis 63
Appendix: Review Accuracy and Metrics 65

Controlled Prediction Accuracy Test and Prediction Accuracy Test 65

How do these tests work? 65
How do you view what decisions the system will make? 65
Which Prediction Rank Threshold should I choose? 65
How does the Controlled Prediction Accuracy Test differ from the Prediction Accuracy Test? 65

Glossary Terms 67
Predictive Coding User Guide

Welcome to the Clearwell eDiscovery Platform Transparent Predictive Coding User Guide. The Predictive Coding User Guide provides case administrators of the Symantec eDiscovery Platform with details on how to model, prepare, analyze, and review case data using the predictive coding feature in Clearwell.

This section contains the following sections:

• “About This Guide” in the next section
• “Revision History” on page 5
• “Obtaining More Product Information” on page 6
• “Technical Support” on page 7

About This Guide

As a supplement for case administrators, and a companion to the Case Administration Guide, the Transparent Predictive Coding User Guide is intended to guide you through the steps of applying and monitoring the Transparent Predictive coding feature. It describes the best practices on how to train the system to quickly cull through case data for review, and how to review and tag those items as “trained” content.

Note: You must have the Case Administrator role, or have appropriate administrator permissions to perform many of the tasks associated with the Predictive Coding feature. Refer to the “Case Administration Guide” for more information.

Revision History

The following table lists the information that has been revised or added since the initial release of this document. The table also lists the revision date for these changes.

<table>
<thead>
<tr>
<th>Revision Date</th>
<th>New Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sept 2012</td>
<td>• New guide—Documents new predictive coding feature, including import/export modeling workflows, and review, tagging, and filtering capabilities within the Analysis &amp; Review module including:</td>
</tr>
<tr>
<td></td>
<td>• separate tagging of attachments (refer also to the User Guide)</td>
</tr>
<tr>
<td></td>
<td>• search results filtering and review enhancements for attachments</td>
</tr>
<tr>
<td></td>
<td>• searchable ID for attachments and embeddings</td>
</tr>
<tr>
<td></td>
<td>• search report and tag event history search enhancements</td>
</tr>
</tbody>
</table>
Obtaining More Product Information

To obtain more information, refer to:


- **Documentation link** — To obtain the most current online and PDF versions of the documentation, click the **Documentation** link at the bottom of any page in the Symantec eDiscovery Platform.

- **Online help** — Click the **Help** link in the Clearwell user interface to access online help.

Documentation Comments/Feedback?

- Got questions or comments on this guide, or other user documentation? Feel free to contact us at **ClearwellInfoDev@symantec.com**. We appreciate your feedback!
Technical Support

The primary role of the Clearwell Technical Support team is to respond to our customers to help resolve technical issues affecting features or functionality of the Clearwell eDiscovery Platform. This includes configuration, workflow, and internal system issues, but does not include assistance with legal-related or case-specific queries. For additional help with technical issues, refer also to Clearwell’s Knowledge Base content available online at:


For information about Symantec’s support offerings, visit our Web site at:

http://www.symantec.com/business/support/

Contacting Technical Support

Before contacting Technical Support, make sure you have satisfied the system requirements that are listed in your product documentation. Also, you should be at the system on which the problem occurred, in case it is necessary to replicate the problem.

When you contact Technical Support, please have the following information available:

- Product version and patch level
- Service tag number
- Available memory, disk space, and NIC information
- Operating system
- Problem description:
  - Affected case name
  - Screenshots demonstrating behavior
  - Accurate reproduction steps
  - Error messages and log files
  - Troubleshooting that was performed before contacting Symantec
  - Recent software configuration changes and network changes

**Note:** You must have administrative access to the system in order to perform most troubleshooting. Be sure to have this available, as well as a system/network administrator resource in the event access is required.

To contact Clearwell Technical Support, use any of the following methods:

- **Customer Support Portal** — **(Clearwell Products)**
- **Phone** — Toll-Free (North America):
  - 1-877-Clearwell (253-2793)
Licensing Portal

If your Clearwell product requires registration or a license key, access our Licensing Portal:

- **Symantec Licensing Portal**
  - [https://licensing.symantec.com/acctmgmt/index.jsp](https://licensing.symantec.com/acctmgmt/index.jsp)

Customer Care

To contact Symantec Customer Care, use any of the following methods:

- **Customer Care Portal** — (Non-Technical Support) Enterprise Products
  - [http://customer-care.symantec.com](http://customer-care.symantec.com)

- **Phone** — Toll-Free (North America):
  - 1-800-721-3934
  - For other regions: [http://www.symantec.com/business/support/assistance_care.jsp](http://www.symantec.com/business/support/assistance_care.jsp)

Customer Service is available to assist with non-technical questions, such as the following types of issues:

- Questions regarding product licensing or serialization
- Product registration updates, such as address or name changes
- General product information (features, language availability, local dealers)
- Latest information about product updates and upgrades
- Information about upgrade assurance and support contracts
- Information about the Symantec Buying Programs
- Advice about Symantec's technical support options
- Non-technical pre-sales questions
- Issues that are related to manuals
Product Documentation

The table below lists the end-user documentation that is available for the Clearwell eDiscovery Platform product.

### Clearwell eDiscovery Platform Documentation

<table>
<thead>
<tr>
<th>Document</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Installation and Configuration</strong></td>
<td></td>
</tr>
<tr>
<td>Installation Guide</td>
<td>Describes prerequisites, and how to perform a full install of the Clearwell software application</td>
</tr>
<tr>
<td>Upgrade Overview Guide</td>
<td>Provides critical upgrade information, by version, useful prior to upgrading an appliance to the current product release</td>
</tr>
<tr>
<td>Upgrade Guide</td>
<td>Describes prerequisites and upgrade information for the current customers with a previous version of the Clearwell software application</td>
</tr>
<tr>
<td>Utility Node Guide</td>
<td>For customers using utility nodes, describes how to install and configure appliances as utility nodes for use with an existing Clearwell software setup.</td>
</tr>
<tr>
<td>Native Viewer Installation Guide</td>
<td>Describes how to install and configure the Brava Client for native document rendering and redaction for use during analysis and review in Clearwell.</td>
</tr>
<tr>
<td>Distributed Architecture Deployment Guide</td>
<td>Provides installation and configuration information for the Review/Processing Scalability feature (7.x) in a distributed architecture deployment</td>
</tr>
<tr>
<td><strong>Getting Started</strong></td>
<td></td>
</tr>
<tr>
<td>Navigation Reference Card</td>
<td>Provides a mapping of the Clearwell user interface (7.x) compared to 6.x</td>
</tr>
<tr>
<td>Clearwell QuickStart Guide</td>
<td>Describes basic appliance and case configuration</td>
</tr>
<tr>
<td>Reviewer QuickStart Guide</td>
<td>A reviewer’s reference to using the Analysis &amp; Review module in Clearwell</td>
</tr>
<tr>
<td><strong>User and Administration</strong></td>
<td></td>
</tr>
<tr>
<td>Legal Hold Guide</td>
<td>Describes how to set up and configure a Clearwell appliance for Legal Holds, and use the Legal Hold module as an administrator in Clearwell</td>
</tr>
<tr>
<td>Identification and Collection Guide</td>
<td>Describes how to prepare and collect data for processing, using the Identification and Collection module</td>
</tr>
<tr>
<td>Case Administration Guide</td>
<td>Describes case setup, processing, and management, plus pre-processing navigation, tips, and recommendations. Includes processing exceptions reference and associated reports, plus file handling information for multiple languages, and supported file types and file type mapping.</td>
</tr>
<tr>
<td>System Administration Guide</td>
<td>Includes system backup, restore, and support features, configuration, and anti-virus scanning guidelines for use with Clearwell.</td>
</tr>
<tr>
<td>Load File Import Guide</td>
<td>Describes how to import load file sources into Clearwell</td>
</tr>
<tr>
<td>User Guide</td>
<td>Describes how to perform searches, analysis, and review, including detailed information and syntax examples for performing advanced searches</td>
</tr>
<tr>
<td>Export and Production Guide</td>
<td>Describes how to use, produce, and troubleshoot exports</td>
</tr>
<tr>
<td>Transparent Predictive Coding Guide</td>
<td>Describes how to use the Predictive Coding feature in Clearwell to train the system to predict results from control set data and tag settings.</td>
</tr>
</tbody>
</table>
**Clearwell eDiscovery Platform Documentation**

<table>
<thead>
<tr>
<th>Document</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reference and Support</strong></td>
<td></td>
</tr>
<tr>
<td>Collection</td>
<td>A quick reference card of how to collect data in Clearwell</td>
</tr>
<tr>
<td>OnSite Collection</td>
<td>A quick reference for performing OnSite collection tasks</td>
</tr>
<tr>
<td>Review and Redaction</td>
<td>Reviewer's reference card of all redaction functions</td>
</tr>
<tr>
<td>Keyboard Shortcuts</td>
<td>A quick reference card listing all supported shortcuts</td>
</tr>
<tr>
<td>Production</td>
<td>Administrator’s reference card for production exports</td>
</tr>
<tr>
<td>User Rights Management</td>
<td>A quick reference card for managing user accounts</td>
</tr>
<tr>
<td><strong>Online Help</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Includes all the above documentation to enable search across all files. To access this information from within the Clearwell user interface, click Help.</td>
</tr>
<tr>
<td><strong>Release</strong></td>
<td></td>
</tr>
<tr>
<td>Release Notes</td>
<td>Provides latest updated information for the current product release</td>
</tr>
</tbody>
</table>

Overview

Starting with release 7.1.2, Symantec adds Transparent Predictive Coding to its technology-assisted review offerings. Transparent Predictive Coding helps organizations defensibly reduce the time and cost of document review. Transparent Predictive Coding enables legal teams to automatically classify documents according to case issues such as responsiveness or privilege, often requiring only a small fraction of the case to be reviewed manually. The net result is a more defensible automated document review process at significantly reduced cost.

This chapter contains the following topics:

- “What is Transparent Predictive Coding?” on page 11
  - “How Does Predictive Coding Work?” on page 11
  - “What Predictive Coding is Not” on page 11
- “Using Transparent Predictive Coding for Optimal Performance” on page 12
- “Types of Transparent Predictive Coding Users” on page 12

What is Transparent Predictive Coding?

Over the last several decades there have been significant advances in the area of computing called machine learning. The goal of this research is to improve the ability of software to learn from inputs in the environment and use this information to make decisions. This machine learning technology can also be used to help make the review process more cost effective and accurate. For eDiscovery, this is commonly referred to as predictive coding (other terms are suggestive coding, computer-categorized document review and predictive categorization).

Symantec’s Transparent Predictive Coding technology provides visibility into the prediction process, enabling more informed decisions and facilitating greater review accuracy. This solution provides a workflow that adapts to the unique requirements of each case, allowing reviewers to begin using predictive coding immediately and achieve optimal results. Each step in the workflow is documented by comprehensive reporting to help demonstrate the integrity of review to the court.

How Does Predictive Coding Work?

Predictive coding works by having software interact with human reviewers in order to learn the review criteria in the case. As reviewers tag documents in a sample set, the software learns the criteria for assessing documents and can generate accurate estimates of which tags should be applied to the remaining documents. As a result, fewer documents need to be reviewed and reviewers can tag documents more efficiently thereby reducing review costs.

What Predictive Coding is Not

Predictive coding technology is often confused with other types of technology assisted review tools like concept search and clustering. However, unlike those technologies which automatically extract relationships and patterns from documents without human intervention, predictive coding requires human input from the outset. Predictive coding involves training and fine tuning the computer to accurate classify documents as, for example, responsive and not responsive. This requirement is because predictive coding goes a step further than other types of tools to make it easier to search, group, or organize documents.
Using Transparent Predictive Coding for Optimal Performance

Transparent predictive coding is best suited for cases and situations where there is a large amount of documents to review. This means that small review efforts (under 100,000 documents or less than 20 GB) might be better suited for other review strategies and workflows.

The system works best on extracted text documents. Currently, non-text information such as images that cannot be OCR processed, sound, and video are not optimal candidates for transparent predictive coding.

Types of Transparent Predictive Coding Users

In Transparent Predictive Coding, the user type controls what operations users can perform on the Transparent Predictive Coding model and system. The various tasks and actions in the Transparent Predictive Coding workflow are divided between two classes or types of user. Each type of user has different responsibilities and access to different levels of Transparent Predictive Coding features and tasks.

- **Case Administrators and Case Managers** — have full control over the Transparent Predictive Coding system. They can create, modify, monitor and manage the end-to-end Transparent Predictive Coding workflow and model. Additionally, they can set access profiles for reviewers and other key case team members that enable them to perform a subset of tasks and views.

- **Reviewers, Case Users, and Key Review Team Members** — can search on predictions, view, sort, and filter prediction ranks and perform tasks related to review analysis. Reviewers in this class are restricted from making system wide changes to the Transparent Predictive Coding model and workflow.

Below is a summary of Transparent Predictive Coding tasks and features:

<table>
<thead>
<tr>
<th>Actions</th>
<th>Case Administrators</th>
<th>Case Managers</th>
<th>Reviewers/Case Users</th>
<th>Key Review Team Members</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Controlled Prediction Accuracy Test</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Train</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Next Set</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Predict</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Prediction Test</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transparent Predictive Coding Management</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Prediction Status</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Export Predictive Tag</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Access Profile Changes</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Review Dashboard</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>• Prediction Ranks</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Search</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Prediction Ranks</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Understanding the Workflow

Transparent Predictive Coding utilizes a loosely-coupled workflow to ensure full flexibility in its usage. It is uniquely designed to complement existing review workflows. The following information should be taken into account before identifying the set of documents for which Transparent Predictive Coding is to be applied.

Items with No Text

The text content of items is used for both training the system and assigning Prediction Ranks. In order to achieve the highest accuracy with Transparent Predictive Coding, items that have no text content should be separated for manual review and not put through the Transparent Predictive Coding workflow. In addition, any items with processing errors should also be separated for manual review.

Early Case Assessment

Transparent Predictive Coding has powerful analytics that enable effective Early Case Assessment. These analytics can accurately identify and separate items which are determined to be unrelated to the case.

Transparent Predictive Coding Phases

Review workflows typically involve using multiple tags across various tag sets. To achieve the highest accuracy across multiple tags in the most efficient way possible, the training and testing of individual tags should be phased as much as possible.
Best Practices Workflow

The following diagram describes the best practices workflow for identifying all of the positive items in a large set and utilizing the most powerful features in Transparent Predictive Coding. This workflow leverages Clearwell's statistically valid sampling and ensures the richest possible measures of accuracy.
Folders

The following folders are recommended to be created in order to aid the management of this workflow.

Table 1: Folder Setup

<table>
<thead>
<tr>
<th>Folder</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive Coding</td>
<td>Parent folder containing the documents involved in the Transparent Predictive Coding workflow.</td>
</tr>
<tr>
<td>Control Set</td>
<td>Subfolder of Predictive Coding containing the documents selected by the Controlled Prediction Accuracy Test action for use in evaluating the accuracy of the Predictive Tag.</td>
</tr>
<tr>
<td>Initial Sample</td>
<td>Subfolder of Control Set containing the documents selected by Step 1 of the Controlled Prediction Accuracy Test action.</td>
</tr>
<tr>
<td>Additional Sample</td>
<td>Subfolder of the Control Set containing the documents selected by Step 2 of the Controlled Prediction Accuracy Test action.</td>
</tr>
<tr>
<td>Population</td>
<td>Subfolder of Predictive Coding containing unreviewed documents in the Transparent Predictive Coding workflow.</td>
</tr>
<tr>
<td>Training Sets</td>
<td>Subfolder of Predictive Coding containing the documents to be used by the Train action for training the system.</td>
</tr>
<tr>
<td>Training Set 1</td>
<td>Subfolder of Training Sets containing a user selected set of documents to be used for the initial training of the system. The contents of this folder should be carefully selected from the Population by using Clearwell's search features to locate between 100-1,000 items for initial review. For best results, this set of items should have about 50% positive items and 50% negative items.</td>
</tr>
<tr>
<td>Training Set 2 - N</td>
<td>Subfolders of Training Sets containing the documents selected by the Next Training Set action to be used for iterative training of the system.</td>
</tr>
</tbody>
</table>
Workflow Walkthrough

Reference

The following is a reference for the recommended workflow.

Table 2: Workflow Reference Guide

<table>
<thead>
<tr>
<th>Step</th>
<th>Sub-Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setup</td>
<td>1. Setup initial folder structure</td>
</tr>
<tr>
<td></td>
<td>2. Use Controlled Prediction Accuracy Test to create the Initial Sample for the Control Set</td>
</tr>
<tr>
<td></td>
<td>3. Review the Initial Sample to calculate the yield</td>
</tr>
<tr>
<td></td>
<td>4. Use Controlled Prediction Accuracy Test to create the Additional Sample for the Control Set</td>
</tr>
<tr>
<td></td>
<td>5. Review the Additional Sample to finish the Control Set</td>
</tr>
<tr>
<td>Iterate</td>
<td>1. Select items from the Population to become Training Set 1</td>
</tr>
<tr>
<td></td>
<td>2. Review all unreviewed items in Training Sets</td>
</tr>
<tr>
<td></td>
<td>3. Use Train to Train on all items in Training Sets</td>
</tr>
<tr>
<td></td>
<td>4. Use Controlled Prediction Accuracy Test to Predict and Test on all items in Control Set</td>
</tr>
<tr>
<td></td>
<td>5. If results are unsatisfactory, use Next Training Set to have the system select the next training items</td>
</tr>
<tr>
<td></td>
<td>6. Repeat Steps 2-5 until results are satisfactory</td>
</tr>
<tr>
<td>Finish</td>
<td>1. Use Predict to Predict on the Population</td>
</tr>
<tr>
<td></td>
<td>2. Use the Prediction Rank Threshold from the last Test to move positive items out of Transparent Predictive Coding</td>
</tr>
</tbody>
</table>
Detailed Workflow

The following sections describe the individual steps in this workflow. For details on each of the Transparent Predictive Coding actions, see “Transparent Predictive Coding Actions” on page 25.

Setup

The purpose of this section is to help you set up a set of folders for use in Transparent Predictive Coding.

**SETUP STEPS**

**A. Setup initial folder structure**

1. Create the Predictive Coding, Control Set, Initial Sample, Additional Sample, Population, Training Sets, and Training Set 1 folder (see Folder Structure for details)
2. Place all items to go through Transparent Predictive Coding into the Population folder
3. Select items for placement into the Training Set 1 folder
   
   A. Navigate to the Population folder
   
   B. Use Clearwell's search features to locate between 100-1,000 items. See “Training Set 1” on page 15 for more details.
   
   C. Use the Folder action to Move these items to the Training Set 1 folder

**B. Use Controlled Prediction Accuracy Test to create the Initial Sample for the Control Set**

1. Open the Controlled Prediction Accuracy Test action
2. Select the Tag to put through Transparent Predictive Coding
3. Under Step 1, specify the Population folder as the Population Source
4. Specify the Confidence level
5. Specify the Margin of Error
6. Specify the Initial Sample as the Destination Folder
7. Finish and generate the Initial Sample

**C. Review the Initial Sample to calculate the yields**

1. Navigate to the Initial Sample folder
2. Review all items in the folder
D. Use Controlled Prediction Accuracy Test to create the Additional Sample for the Control Set

1. Open the Controlled Prediction Accuracy Test action
2. Under Step 2, check the box “The initial sample has been reviewed.”
3. Specify the Target F-Measure
4. Specify the Target F-Measure Range
5. Specify the Additional Sample as the Destination Folder
6. Finish and generate the Additional Sample

E. Review the Additional Sample to finish the Control Set

1. Navigate to the Additional Sample folder
2. Review all items in the folder
Train and Test
The purpose of this step is to achieve the highest quality Predictive Tag.

<table>
<thead>
<tr>
<th>TRAIN &amp; TEST STEPS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A. Review all unreviewed items in Training Set</strong></td>
</tr>
<tr>
<td>1. Navigate to the Training Set 1 folder (where Training Set 1 is the current iteration)</td>
</tr>
<tr>
<td>2. Review all items in the folder</td>
</tr>
<tr>
<td><strong>B. Use Train action on all items in Training Set</strong></td>
</tr>
<tr>
<td>1. Open the Train action</td>
</tr>
<tr>
<td>2. Select the Predictive Tag that is going through Transparent Predictive Coding</td>
</tr>
<tr>
<td>3. Select &quot;All items in search results&quot; and ensure that &quot;Include previously trained items&quot; is checked to select all items in Training Sets for training</td>
</tr>
<tr>
<td>4. Train on all of the training</td>
</tr>
<tr>
<td><strong>C. Use Controlled Prediction Accuracy Test to Predict and Test on all items in Control Set</strong></td>
</tr>
<tr>
<td>1. Open the Controlled Prediction Accuracy Test action</td>
</tr>
<tr>
<td>2. Under Step 3, ensure that “The final sample has been reviewed.” is checked</td>
</tr>
<tr>
<td>3. Finish and run the Controlled Prediction Accuracy Test</td>
</tr>
<tr>
<td>4. When the Controlled Prediction Accuracy Test job is complete, open the Test Report and evaluate the results</td>
</tr>
<tr>
<td><strong>D. If results are unsatisfactory, use Next Training Set to have the system select the next training items</strong></td>
</tr>
<tr>
<td>1. Open the Controlled Prediction Accuracy Test action</td>
</tr>
<tr>
<td>2. Under Step 2, check the box “The initial sample has been reviewed.”</td>
</tr>
<tr>
<td>3. Specify the Target F-Measure</td>
</tr>
<tr>
<td>4. Specify the Target F-Measure Range</td>
</tr>
<tr>
<td>5. Specify the Additional Sample as the Destination Folder</td>
</tr>
<tr>
<td>6. Finish and generate the Additional Sample</td>
</tr>
<tr>
<td><strong>E. Repeat Steps B through E until results are satisfactory</strong></td>
</tr>
</tbody>
</table>
Predict and Finish

The purpose of this step is to identify positive items to move out of Transparent Predictive Coding.

**PREDICT & FINISH STEPS**

**A. Use Predict to Predict on the Population**

1. Navigate to the Population folder
2. Open the Predict action
3. Select the Predictive Tag that is going through Transparent Predictive Coding
4. Select “All items in search results” to assign Prediction Ranks to all items in the Population
5. Finish and Predict on the entire Population

**B. Use the Prediction Rank Threshold from the last Test to move positive items out of Transparent Predictive Coding**

1. From the last Test Report, get the Best Prediction Rank Threshold
2. Navigate to the Predictive Coding folder
3. Run an Advanced Search for all items at and above the Best Prediction Rank Threshold
4. Use either the Tag or Folder action to mark these items as predicted to be positive
Alternate Workflows

Clearwell’s Transparent Predictive Coding is a highly flexible feature that be integrated into a number of alternate workflows. The following sections describe two potential workflows that use elements of the Transparent Predictive Coding features.

Prioritizing Linear Review

To prioritize a linear review, simply select a Training Set to review, use the Train action on the items to Train the system, and then use the Predict action on all of the items in your case. All items will now have Prediction Ranks, which you may sort on in descending order for prioritization.

Using Alternate Sampling Methods

To utilize alternate sampling methods for your workflow, simply create a sample folder manually with the Folder action. Follow the same workflow as described in the Best Practices Workflow, but instead of using the Controlled Prediction Accuracy Test action, use the Predict action on your sample folder and the Prediction Accuracy Test action on your sample folder to generate a test report.

Note: The Prediction Accuracy Test generates results on the set of items fed into the action. Unlike the Controlled Prediction Accuracy Test, it does not project any results onto the Population.
Using Transparent Predictive Coding

This section takes you through the steps for building and tuning a Transparent Predictive Coding review system. You create, simulate, and build the Transparent Predictive Coding review intelligence with input from your expert reviewers and the Transparent Predictive Coding system. It is an iterative process.

The goal is to have the Transparent Predictive Coding system learn review criteria for classifying items as responsive, privileged, or any other issue code through input from expert reviews. Essentially, you are training and teaching the system through review control and training sets. Over time, the Transparent Predictive Coding system becomes statistically stable, learning enough to correctly predict and rank training sets of documents, progressing to a point where it can be leveraged against your entire item population, and enabling you to achieve high levels of review accuracy at significantly reduced time and cost.

Refer to the following topics in this section:

- “Workflow and Operating Considerations” on page 24
- “Transparent Predictive Coding Actions” on page 25
- “Creating a Control Set” on page 27
  - “Selecting an Controlled Prediction Accuracy Test Control Set Tag” on page 27
  - “Step 1: Create Control Set (Initial Sample)” on page 29
  - “Step 2: Enhance Control Set (Additional Sample)” on page 30
  - “Step 3: Run the Test” on page 32
- “Training the System” on page 35
  - “Step 1: Select Predictive Tag” on page 35
  - “Step 2: Select Items” on page 36
  - “Step 3: Confirm Training Selections” on page 37
  - “View Job and Training Error Report” on page 38
- “Predicting” on page 39
  - “Step 1: Select Predictive Tag” on page 39
  - “Step 2: Select Items” on page 40
- “Selecting the Next Training Set” on page 42
  - “Step 1: Select a Next Training Set Tag” on page 42
  - “Step 2: Select Items for Next Training Set” on page 43
  - “Step 3: Copy or Move Items to Folder” on page 44
  - “Step 4: Confirm the Next Training Set Items” on page 44
- “Prediction Accuracy Test” on page 45
  - “Step 1: Select Prediction Test Tag” on page 45
  - “Step 2: Select Items” on page 46
  - “Step 3: Confirm Prediction Test Item Selections” on page 47
  - “Step 4: View the Prediction Test Report” on page 48
Workflow and Operating Considerations

The system’s flexibility allows you to follow different workflows depending on your review and case requirements. For example, you may perform only some of the Transparent Predictive Coding steps, or perform the steps by varying the sequence, depending on the case review requirements and workflow. As an iterative process, some steps or stages in the system may be repeated a number of times to refine the approach after a better understanding of the data emerges. For specific workflow details, see “Transparent Predictive Coding Best Practices” on page 13.

**Note:** All actions, predictions, and tags operate on the item-level.
Transparent Predictive Coding Actions

The Transparent Predictive Coding workflow combines the efforts of your expert reviewers with the following system actions:

Table 3:  Transparent Predictive Coding Actions

<table>
<thead>
<tr>
<th>Icon</th>
<th>Action</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>📈</td>
<td>Controlled Prediction Test</td>
<td>Setup and run a statistically valid test of prediction accuracy.</td>
</tr>
<tr>
<td>📈</td>
<td>Train</td>
<td>Supply reviewed items to system for training</td>
</tr>
<tr>
<td>🗂️</td>
<td>Next Training Set</td>
<td>Get recommended set of items to review to increase accuracy.</td>
</tr>
<tr>
<td>🌪️</td>
<td>Predict</td>
<td>Apply prediction ranks to items based on system knowledge.</td>
</tr>
<tr>
<td>📈</td>
<td>Prediction Test</td>
<td>Run independent testing on any set of items.</td>
</tr>
</tbody>
</table>
Summary chart of the Transparent Predictive Coding action workflow.

**Table 4: High-Level Summary of Action Workflow**

<table>
<thead>
<tr>
<th>Actions</th>
<th>Method</th>
<th>Inputs</th>
<th>Usage Guidelines - Dependencies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Controlled Prediction Accuracy Test</strong></td>
<td>Switch between human review and Transparent Predictive Coding calculator and guided steps 1-3</td>
<td>Population Single tag</td>
<td>Prerequisites: Manually create folder hierarchy and store population data</td>
</tr>
<tr>
<td><strong>Goal:</strong> Create an accurate control set</td>
<td>May need to repeat steps until you reach desired control set accuracy.</td>
<td></td>
<td>Caution: Do not add or remove removing items from the control set. If you do, you must re-create and re-review the control set.</td>
</tr>
<tr>
<td><strong>Train</strong></td>
<td>Follow the Training steps</td>
<td>Reviewed and tagged training set</td>
<td>Prerequisites: Training items must be viewable within current search.</td>
</tr>
<tr>
<td><strong>Goal:</strong> Identify a sufficient</td>
<td></td>
<td></td>
<td>Caution: Training conflict errors should be resolved. If not, it can potentially lower the prediction accuracy.</td>
</tr>
<tr>
<td><strong>Predict</strong></td>
<td>Follow the Predict steps</td>
<td>Control set Population</td>
<td>Caution: Do not submit training sets to the Predict action.</td>
</tr>
<tr>
<td><strong>Goal:</strong> Arrive at an acceptable prediction ranking</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Next Training Set</strong></td>
<td>Follow the Next Training Set steps</td>
<td>Population Predictive tag</td>
<td>Prerequisites: Population items must be viewable within current search or in a designated folder.</td>
</tr>
<tr>
<td><strong>Goal:</strong> Fetch the next set of items for train</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Prediction Accuracy Test</strong></td>
<td>Follow the Prediction Accuracy Test steps</td>
<td>Control Set or Items that have completed the Predict action.</td>
<td>Prerequisites: Items must be viewable within current search</td>
</tr>
<tr>
<td><strong>Goal:</strong> Quickly test current learning state against random set of reviewed items.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Creating a Control Set

The action Controlled Prediction Accuracy Test produces a highly reliable sample which is then used to test prediction rank accuracy. The goal of Controlled Prediction Accuracy Test is to create an acceptable initial sample (control set) that can be further leveraged against training sets. The workflow involves switching between the work product from your expert reviewers and the Transparent Predictive Coding calculator in order to evaluate and finalize the quality of the control set.

**Important:** Any time you add or remove any items from the control set it causes the system to generate an error and you must re-create and re-review the sample in Transparent Predictive Coding. You can save time and avoid repeating these steps if you first properly identify a set of items for review which comply with the scope of the case.

**Prerequisites:** A folder hierarchy, set up manually, to organize data. For help on creating folders, see "Folders" on page 15

**Inputs:** Population data exists and resides in the folder structure
A single tag

**User:** Case Administrator and Case Managers

Selecting an Controlled Prediction Accuracy Test Control Set Tag

The first task is to select a predictive tag that you want to train the system with.

**To select an Controlled Prediction Accuracy Test control set tag**

1. From the Documents Screen under the Analysis & Review module, click *Action* and select *Controlled Prediction Accuracy Test* to start training the system.

   The Controlled Prediction Accuracy Test menu displays.
2. From the Tag section at the top of the Controlled Prediction Accuracy Test screen, click **Select**.

The **Select Tag** screen opens.

3. Select the tag for the control set.

**Note:** Each tag should have its own control set. This is because each tag is associated with a different accuracy and should be measured and validated separately.
Step 1: Create Control Set (Initial Sample)

Create the first statistically valid sample using the measurements below. Return to this screen after reviewing this sample.

1. Specify the following Create Sample field information:

Table 5: Controlled Prediction Test - Step 1: Create Sample

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Source</td>
<td>Select the source that contains only the set of items that you want to test against or sample from.</td>
</tr>
<tr>
<td>Population Size</td>
<td>Automatically displays the number of items in the Population Source.</td>
</tr>
<tr>
<td>Confidence</td>
<td>Specify a confidence level which is a statistical measure of certainty expressed as a percentage. Confidence level indicates the likelihood that the actual number is within the margin of error. For instance, if the confidence level is 95% with a margin of error of ±3%, then there is a 5% chance the actual value lies outside the margin of error. The default is 95%.</td>
</tr>
<tr>
<td>Margin of Error</td>
<td>Specify the desired margin of error. Margin of error is a statistic expressing the amount of random sampling error related to an estimate. For example, consider a situation where a random sampling of voters are polled to reach an estimate that 60% of voters intend to vote for candidate A in the next presidential election. If the margin of error for the estimate is ±3%, then somewhere between 57-63 percent of the voters are estimated to vote for candidate A. Importantly, all things being equal, a larger sample size will produce a smaller margin of error.</td>
</tr>
<tr>
<td>Initial Sample Size</td>
<td>Automatically displays the minimum number of items needed to achieve the desired confidence level and margin of error.</td>
</tr>
<tr>
<td>Destination Folder</td>
<td>Specify the folder where the control sample of documents will be moved or copied to. <strong>Important:</strong> The folder you specify must be empty.</td>
</tr>
</tbody>
</table>
2. After filling in all the create sample fields, click **Create Sample** to launch the sample creation job.

![Controlled Prediction Accuracy Test](image)

**Note:** The actual folder size may be larger than the sample size. This is because the items sampled may be part of a larger document families.

**Step 2: Enhance Control Set (Additional Sample)**

Enhance the quality of your statistically valid sample using the additional measurements listed in this section. Return to this screen after reviewing your sample.

1. Check that **The initial sample has been reviewed** by your expert reviewers.
2. Specify the following field information

**Table 6: Controlled Prediction Test- Step 2: Enhance Control Set (Additional Sample)**

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
</table>
| Yield          | Automatically displays the computed yield from your review team’s sample. Yield is a measure based on the amount of items tagged with your Predictive Tag which is the percentage of documents in the population which are responsive. A low yield often causes the margin of error to increase, so additional sampling may be required. **Sample Errors to Avoid:**
|                | • 0%—The sample was probably not reviewed.                      |
|                | • 100%—The sample size contains only positive items and the tag is not suitable for use by the system. |

| Desired F-Measure | Measures the accuracy of a test sample. The values must be between 1 and 100 and the default is 80. **Note:** A higher Desired F-Measure results in a smaller Additional Sample Size. |
Desired F-Measure Range: The range of acceptable F-Measure values. The default is 10. For example, if the F-Measure is 80 and the Desired F-Measure Range is 10, then the results will be between 75 and 85. 

Note: A higher Desired F-Measure Range results in a smaller Additional Sample Size. The Desired F-Measure is only guaranteed if the Desired F-Measure is reached.

Additional Sample Size: Automatically displays the statistically valid sample size required based on Yield, Desired F-Measure, and Desired F-Measure Range calculations.

Destination Folder: Specify a folder for the sample.

Important: The folder you specify must be empty.
3. After supplying all the required Enhance Sample Accuracy fields, including a destination folder (which must be empty), click **Add to the Sample**.

![Controlled Prediction Accuracy Test](image)

**Note:** The actual folder size may be larger than the sample size. This is because the items sampled may be part of a larger document families.

**Step 3: Run the Test**

During this step, the system compares the human tagging decisions against its prediction ranks and provides a report. The report gives a ranking of correct and incorrect predictions. The rankings in the report give you an idea of what to expect if you were to bulk tag the remaining items.

**Important:** You must run Train on the selected tag before running the Controlled Prediction Accuracy Test.

**To run the Controlled Prediction Accuracy Test on the sample set**

1. Verify that the **Total Control Set Size** displays. This should be a cumulative count of items in both sample folders from Step1: Create Sample and Step 2: Enhance Sample Accuracy.

2. If your expert reviewers have approved the control set, then check **The Control Set has been reviewed** option.
3. Click **Run Test** to apply predictions to all of the items in the control set.

![Step 3: Run the Test](image)

Once you click “Run Test”, a job starts which you can monitor from the Jobs window. The system generates a report with prediction results. When ready, you can click **View Report** to view the results.

4. Click **View Report** to open and view the prediction results.

### Table 7: Controlled Prediction Test - Report Summary

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>populationSize</td>
<td>The total number of items.</td>
</tr>
<tr>
<td>confidence</td>
<td>Input value for Step 1.</td>
</tr>
<tr>
<td>marginOfError</td>
<td>Input value for Step 1.</td>
</tr>
<tr>
<td>yield</td>
<td>Calculated value for Step 2.</td>
</tr>
<tr>
<td>targetFmeasure</td>
<td>Input value from Step 2.</td>
</tr>
<tr>
<td>targetFmeasureRange</td>
<td>Input value for Step 2.</td>
</tr>
<tr>
<td>sampleSize</td>
<td>The actual item count in both sample folders.</td>
</tr>
<tr>
<td>TotalPositives</td>
<td>The number of items that were tagged with your Select Tag option.</td>
</tr>
<tr>
<td>Note:</td>
<td>This number only relates to the sample.</td>
</tr>
<tr>
<td>bestThreshold</td>
<td>An index into the report details section.</td>
</tr>
<tr>
<td></td>
<td>For example, if the value is set to 70, locate that value under the Threshold column in the Report Details (next section).</td>
</tr>
</tbody>
</table>

### Table 8: Controlled Prediction Test - Report Details

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold</td>
<td>Prediction rank threshold</td>
</tr>
<tr>
<td>numDocsAboveThreshold</td>
<td>Number of items with a prediction rank above the prediction rank threshold.</td>
</tr>
<tr>
<td>numTruePositives</td>
<td>Number of items above the prediction rank threshold that are tagged with the predictive tag.</td>
</tr>
</tbody>
</table>
Table 8: Controlled Prediction Test - Report Details

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>numFalsePositives</td>
<td>Number of items above the prediction rank threshold that are not tagged with the predictive tag.</td>
</tr>
<tr>
<td>numFalseNegatives</td>
<td>Number of items below the prediction rank threshold that are tagged with the predictive tag.</td>
</tr>
<tr>
<td>calculatedRecall</td>
<td>Percentage of all positives in sample above the threshold.</td>
</tr>
<tr>
<td>minRecall</td>
<td>Statistically calculated minimum value of recall when extrapolated to the population.</td>
</tr>
<tr>
<td>maxRecall</td>
<td>Statistically calculated maximum value of recall when extrapolated to the population.</td>
</tr>
<tr>
<td>calculatedPrecision</td>
<td>Percentage of all items above threshold that are positive.</td>
</tr>
<tr>
<td>minPrecision</td>
<td>Statistically calculated minimum value of precision when extrapolated to the population.</td>
</tr>
<tr>
<td>maxPrecision</td>
<td>Statistically calculated maximum value of precision when extrapolated to the population.</td>
</tr>
<tr>
<td>calculated Fmeasure</td>
<td>Harmonic mean of calculated precision and recall.</td>
</tr>
<tr>
<td>minFmeasure</td>
<td>Statistically calculated minimum value of F-measure when extrapolated to the population.</td>
</tr>
<tr>
<td>maxFmeasure</td>
<td>Statistically calculated maximum value of F-measure when extrapolated to the population.</td>
</tr>
</tbody>
</table>

For report details, see "Controlled Prediction Accuracy Test and Prediction Accuracy Test" on page 65.
Training the System

The system uses machine learning to predict tag decisions for individual tags. Selecting the action Train teaches the system to recognize patterns in your reviewer’s tagging decisions.

This example uses a subset of training items from the Training Set 1 folder and the Responsive tag to build the Transparent Predictive Coding system.

### Important:
Training clears all the previous prediction ranks.

### Restrictions:
- Maximum number of tags you can train on is 20.
- Maximum number of items you can train on is 10,000.

### Inputs:
The reviewed and tagged training set items.

**Note:** Training items must be viewable within current search.

### User:
Case Administrator and Case Managers

---

#### Step 1: Select Predictive Tag

The first step of the Train process is to select a tag to train the system. The system learns to recognize patterns in the items for a given tag and then predicts how likely other documents are to be tagged. For example, if you train on Responsive, you can predict how likely the other items will also be Responsive.

**To select Predictive Tag**

1. From the Documents Screen under the Analysis & Review module, click **Action** and select **Train** to start training the system.

2. Select a tag for the system to learn.

![Image of Train window](image)

**Note:** If this is not the first training iteration, then the previously trained item count is shown.

3. When you have finished selecting the tag, click **Next**.
Step 2: Select Items

Once you have selected the predict tag, the Select Items screen displays, requesting items to train on.

The system learns on an individual item basis. When you specify the items to train on, it is important to select all items that have been reviewed and not just the ones that have the tag that is being learned. For example, if you are training on the Responsive tag, you also want to include everything that has been reviewed and is Not Responsive.

To select Items

1. Specify the following information, depending on the type of training:

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>All the items in the search</td>
<td>Includes every email message, loose file, attachment, and embedding even if it did not match the search.</td>
</tr>
<tr>
<td>Only items that match the search and filter criteria</td>
<td>Only includes the items which match the search criteria and are tagged. A blank search will cause every item to match.</td>
</tr>
<tr>
<td>Include previously trained items</td>
<td>Training involves building a learning model from all of the training items provided. It is not incremental, it re-learns each time. In most situations select this option unless you want to start over.</td>
</tr>
<tr>
<td>Include conflicts in training</td>
<td>Tagging decisions that mark duplicate items differently are identified by the system as conflicts. For example, if “attach.doc” is tagged Responsive, and another email message also has “attach.doc” but it was tagged Not Responsive, then the items are found to be in conflict. Including items that are marked as conflicts without resolving the underlying reason for the conflict can affect and potentially lower the prediction accuracy. Important: Symantec recommends resolving conflicts and then re-training the system to increase prediction accuracy. If conflicts are not included, the system ignores all conflicts. A conflict report is available for download once the training job completes. See “View Job and Training Error Report” on page 38.</td>
</tr>
</tbody>
</table>
2. After you have finished selecting items, click **Next**.

**Step 3: Confirm Training Selections**

The confirmation window shows a summary and total of the current selected items and previously trained items. If you determine that this training set performs well, you can move on to the next training set.

**Note:** An error displays if the system does not have enough items or has too many similar items selected. Go back and correct the composition of items in order to continue and submit the training job.

**To confirm training items**

- Confirm the total items to train on and click **Finish** to submit the training job.
**View Job and Training Error Report**

A conflict is any contradictory set of data or tagging decisions. The conflict report is generated as part of the training job. It describes the conflict in terms of the training set and the Transparent Predictive Coding system.

**To view the Train results and Error Report**

- The Jobs window displays the status of the train job displays. Under Actions, click the link to view details for the errors or conflicts that occurred during this phase.

- The following sample Train Error Report provides a description of possible errors.

<table>
<thead>
<tr>
<th>Conflict Type</th>
<th>Resolution</th>
<th>Document IDs</th>
<th>Description</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similar Features</td>
<td>DROP</td>
<td>0.7.11.18514-000001</td>
<td>If a group of items with similar features tagged inconsistently then it is considered as conflicting tag decision.</td>
<td>The entire group of items is dropped if there is a conflict. If you specified the ignore conflicts option in the Train action, you can ignore this error.</td>
</tr>
<tr>
<td>No Features</td>
<td>DROP</td>
<td>0.7.11.11240</td>
<td>If the system could not extract any features from the item.</td>
<td>These items are always dropped from training. Please resolve these errors before continuing on.</td>
</tr>
<tr>
<td>Excluded items</td>
<td>DROP</td>
<td>0.7.11.110956</td>
<td>Certain type of items like “EXE” and “DLL”</td>
<td>These items are always dropped from training. Please resolve these errors before continuing on.</td>
</tr>
</tbody>
</table>
Predicting

The Predict action computes a ranking value or score that equates to how likely an item is to be tagged with a selected tag. This computation examines the meta-data and content of prediction items and determines the set of training items with similar meta-data and content that are present in the system. The ranking value range is from 1 to 100 with 1 being least likely and 100 being most likely to be tagged.

Note: When the predict job completes, re-run your search to see additional features, including advanced search options, filters, and ranks for each document.

---

**Important:** Do not apply Predict to items that were used for training.

**Inputs:**
- Any set of items viewable in current search.
- Predictive tags

**User:** Case Administrator and Case Managers

---

**Step 1: Select Predictive Tag**

**To select predictive tag**

1. Select a tag from the list.

2. Click **Next** to advance to Select Items.
Step 2: Select Items

In order to provide flexibility on the set of items to be predicted, the product does not automatically apply ranks to every item. Instead, you specify on the Select Items screen the items for prediction. You should select items for which you want to see prediction ranks but have not been previously selected to train the system (such as the next test set, next review set, or the entire item population).

To select the items that will be predicted

1. Specify the items for prediction.

Table 11: Predictive Tag - Step 2: Select Items

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>All items in search results</td>
<td>All items in the search results includes every email message, loose file, attachment, and embedding even if it did not match the search. (Recommended)</td>
</tr>
<tr>
<td>Only items that match the search and filter criteria</td>
<td>Only items that match the search and filter criteria will only include the items which match the search criteria. A blank search will cause every item to match</td>
</tr>
</tbody>
</table>
2. Click **Next** to confirm selections.

![Image of Predict window with summary]

**Table 12: Predictive Tag - Item Summary**

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selected items to rank</td>
<td>Includes items in the current search results.</td>
</tr>
<tr>
<td>Previously trained items</td>
<td>Prediction ranks are not applied to documents which have been used for training. These items are not included in the Predict action.</td>
</tr>
<tr>
<td>Previously predictive items</td>
<td>The total number of items to be predicted.</td>
</tr>
</tbody>
</table>

3. Click **Finish** once you validate the item summary.
Selecting the Next Training Set

If the Transparent Predictive Coding system has not learned enough to predict and meet desired recall and precision metrics, the action Next Training Set allows you to generate the next set of items for the system to learn and improve its prediction capabilities. This action allows you to copy or move items within the Transparent Predictive Coding workflow.

**Important:** Typically, items that comprise the Next Training Set have *not* been reviewed.

**Inputs:**
- Predictive tags
- Population either viewable in current search or in a predictive coding specific folder.

**User:**
- Case Administrator and Case Managers

---

**Step 1: Select a Next Training Set Tag**

**To select Next Training Set tag**

1. From the Documents Screen under the Analysis & Review module, click **Action** and select **Next Training Set**. The Next Training Set screen displays.

2. Select the tag for which you want to improve accuracy.

   ![Next Training Set Screen](image)

3. When you have finished selecting the tag, click **Next** to advance to the Items screen.
Step 2: Select Items for Next Training Set

1. Select a batch of items from the population to comprise the next training subset.

![Image of the Next Set window]

**Table 13: Next Training Set - Step 2: Set Items**

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>All items in search results</td>
<td>Includes every email message, loose file, attachment, and embedding even if it did not match the search.</td>
</tr>
<tr>
<td>Only items that match the search and filter criteria</td>
<td>Include the items which match the search criteria. A blank search will cause every item to match.</td>
</tr>
</tbody>
</table>

**Note:** You can also specify a folder for the population. Use the same folder from the Controlled Prediction Accuracy Test if that action was performed.

2. Click **Next** to advance to the Copy/Move screen.
Step 3: Copy or Move Items to Folder

1. Click the first drop-down menu to select whether you want to copy or move items to a folder.

2. Click the second drop-down menu to select the appropriate training folder and click Next.

Step 4: Confirm the Next Training Set Items

1. Verify the information and when ready, click confirm to launch the Next Set job.

Table 14: Next Training Set - Step 4: Confirm Items

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive Tag</td>
<td>Displays the selected predicted tag.</td>
</tr>
<tr>
<td>Selected items to use</td>
<td>These items are used for the Next Set.</td>
</tr>
<tr>
<td>Action</td>
<td>Either:</td>
</tr>
<tr>
<td></td>
<td>• Copy to folder</td>
</tr>
<tr>
<td></td>
<td>• Move to folder</td>
</tr>
<tr>
<td>Subset size</td>
<td>The approximate size of the next training review set.</td>
</tr>
</tbody>
</table>
Prediction Accuracy Test

Prediction Test allows you to quickly test the current learning state of the system against a random set of items that have been reviewed. The role of Prediction test differs from Controlled Prediction Accuracy Test in that it is not as comprehensive and does not utilize or measure against the controlled and statistically robust Controlled Prediction Accuracy Test. For full defensibility and transparency of your work product, you should use Controlled Prediction Accuracy Test.

Prediction Test is flexible and can be very useful if you have devised your own sampling technique. It is not a replacement for the Controlled Prediction Accuracy Test but rather a complimentary testing tool.

---

Important: Items submitted and run through Predict action

Inputs: Items viewable in current search results.

User: Case Administrator and Case Managers

---

Step 1: Select Prediction Test Tag

The system will compare the human tagging decisions against the system’s own prediction ranks and provide a report. For each rank, the report will tell you how many items were predicted correctly and what you can expect if you were to bulk tag the remaining items.

To select Prediction Test tag

1. From the Documents Screen under the Analysis & Review module, click Action and select Prediction Test. The Prediction Test screen displays.

2. Select a tag to validate the accuracy of system predictions.

3. Click Next to advance to the Items screen.
Step 2: Select Items

Prediction Test relies on and works best with a statistically valid sample. Item selections should be completely random, so you may want to select only the items that match your search and filter criteria (depending on how you set up your test sample).

1. **Select the Prediction Test items**

![Image of selection options](image.png)

Table 15: Prediction Test - Step 2: Select Items

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>All items in the search results</td>
<td>Includes every email message, loose file, attachment, and embedding even if it did not match the search.</td>
</tr>
<tr>
<td>Only items that match the search and filter criteria</td>
<td>Only include the items which match the search criteria. A blank search will cause every item to match.</td>
</tr>
</tbody>
</table>

2. **Click Next to advance to confirmation menu.**
Step 3: Confirm Prediction Test Item Selections

1. View the summary of selections. You also have the option of notifying your review team of the summary.

   ![Prediction Test Summary](image)

   **Note:** If there were any issues or errors preventing the Prediction Test, an explanatory warning message displays at the top of the screen.

2. Optional: You can get an email notification when the job completes. From the **Send Email Summary to (Optional)** drop-down menu, select the email recipient.

3. If the summary list is complete, click **Finish**. A Prediction Test job starts and a report is available for pickup in the Jobs window. You can also view the report from the Prediction Status page.
Step 4: View the Prediction Test Report

The following is an explanation and breakdown of the summary Prediction Test report fields.

**Table 16: Prediction Accuracy Test - Report Summary**

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>confidence</td>
<td>Input value for Step 1.</td>
</tr>
<tr>
<td>marginOfError</td>
<td>Input value for Step 1.</td>
</tr>
<tr>
<td>sampleSize</td>
<td>The actual item count in both sample folders.</td>
</tr>
<tr>
<td>TotalPositives</td>
<td>The number of items that were tagged with your Select Tag option. Note: This number only relates to the sample.</td>
</tr>
<tr>
<td>bestThreshold</td>
<td>An index into the report details section. For example, if the value is set to 70, locate that value under the Threshold column in the Report Details (next section).</td>
</tr>
</tbody>
</table>

**Table 17: Prediction Accuracy Test - Report Details**

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold</td>
<td>Prediction rank threshold</td>
</tr>
<tr>
<td>numDocsAboveThreshold</td>
<td>Number of items with a prediction rank above the prediction rank threshold.</td>
</tr>
<tr>
<td>numTruePositives</td>
<td>Number of items above the prediction rank threshold that are tagged with the predictive tag.</td>
</tr>
<tr>
<td>numFalsePositives</td>
<td>Number of items above the prediction rank threshold that are not tagged with the predictive tag.</td>
</tr>
<tr>
<td>numFalseNegatives</td>
<td>Number of items below the prediction rank threshold that are tagged with the predictive tag.</td>
</tr>
<tr>
<td>calculatedRecall</td>
<td>Percentage of all positives in sample above the threshold.</td>
</tr>
<tr>
<td>calculatedPrecision</td>
<td>Percentage of all items above threshold that are positive.</td>
</tr>
<tr>
<td>calculated Fmeasure</td>
<td>Harmonic mean of calculated precision and recall.</td>
</tr>
</tbody>
</table>
Managing Transparent Predictive Coding

This section describes the various management tasks that you can perform to track and monitor the progress and results of Transparent Predictive Coding workflows.

Refer to the following topics in this section:

- “Case Settings for Transparent Predictive Coding” on page 49
- “Predictive Tag and Training Items Management” on page 49
- “Prediction Status” on page 49
- “Using the Review Dashboard” on page 53
- “Granting/Restricting Transparent Predictive Coding Access” on page 54

Case Settings for Transparent Predictive Coding

Starting with 7.1.2, Transparent Predictive Coding is automatically enabled. Clearwell recommends that you do not disable the feature even if you are not using it.

**Note:** If you had previously disabled Transparent Predictive Coding for a case, and you want to enable it, make sure that you enable the feature and rerun post-processing.

For more information on how to enable or disable the feature, refer to the Case Administrator Guide “Enable Predictive Coding”.

Predictive Tag and Training Items Management

There are two types of limits in the Train action. First, the number of predictive tags cannot exceed 20 and secondly, the maximum number of items cannot exceed 10,000. If you have more than 20 predictive tags, you can use the Export Training function for a predictive tag that is not needed and then use Delete Training to allow you to train on another predictive tag.

Prediction Status

The Prediction Status page provides a listing of all the actions and activity performed on predictive tags. You can view current training cycle information and see how various actions and events ran in the Transparent Predictive Coding system and who ran them. There are two types of information displayed:

- **Predictive Tags** — Lists all the available predictive tags
- **Predictive Tag Status** — Record of events on the selected predictive tag. These events include a timestamp, activity type, description, user and a report.
Viewing Predictive Tag Information

To view predictive tags and predictive tag status

1. From the Case Home module, select **Predictive Status** to view predictive tag activity.

   The Predictive Status screen shows the status (right column) and lists the predictive tag (left column).

   ![Predictive Status Screen](image)

   - Controlled Prediction Accuracy Test
   - Train
   - Predict
   - Prediction Test
   - Next Training Set
   - Clear
   - Import
   - Export

2. On the Predictive Tags screen, all of the predictive tags are displayed. For import training information, see "Import Training" on page 51.

3. The Predictive Tag Status column displays the following record information for each entry:

   **Table 18: Predictive Tag Status**

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timestamp</td>
<td>Date and time when the action completed.</td>
</tr>
</tbody>
</table>
   | Activity Type | The following activity types are recorded here:  
   |             | • Controlled Prediction Accuracy Test  
   |             | • Train  
   |             | • Predict  
   |             | • Prediction Test  
   |             | • Next Training Set  
   |             | • Clear  
   |             | • Import  
   |             | • Export  
   | Description | • Text summary of the activity. |
   | User        | Displays the user who ran the action. |
   | Report      | Downloadable reports for the following activity types:  
   |             | • Controlled Prediction Accuracy Test: Test Report  
   |             | • Train: Conflict Report (if applicable)  
   |             | • Prediction Test: Test Report |
4. You can perform the following actions from the Predictive Tag Status screen.

**Table 19: Predictive Tag Status Actions**

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Export History</td>
<td>Export the history report to either a CSV file.</td>
</tr>
<tr>
<td>Clear Predictions</td>
<td>Clear predictions for a given tag.</td>
</tr>
<tr>
<td>Delete Training</td>
<td>Delete training for a predictive tag and clears all history.</td>
</tr>
<tr>
<td>Import Training</td>
<td>Import predictive tag data to the system.</td>
</tr>
<tr>
<td>Export Training</td>
<td>Export predictive tag data to the system.</td>
</tr>
</tbody>
</table>

Import Training allows you to leverage training from other cases. For example, if privilege or another issue is similar across cases you could immediately set aside privileged items. Similarly, you may want to import training in order to utilize Transparent Predictive Coding for more effective culling of spam emails.

During this phase, only the Predictive Tag is imported, and not the documents that were used for training to create it. This means that subsequent training only uses new documents from the current case and you will not see any information displayed in the prediction insight or exemplars since the training documents are not available.

**Important:** When importing training, only tags from cases that you have access to are shown. Similarly, the system only displays tags that were exported from the original case.

**To import training**

1. From the Predictive Tags screen, select **Import Training**.

Import Training screen opens.

![Import Training Screen](image)

**Note:** Import training from file system refers to the operating system's file system and not to an upload feature of the Transparent Predictive Coding platform.

2. Select whether you want to import your training from a case using the drop-down menu or file system and click **Next**.
3. Map the imported training to a tag and click **Next** to advance to the Confirm screen.

For this example, the Responsive tag is selected.

4. On the Confirm screen, verify the selections and click **Finish** to start the Import Training job.
Using the Review Dashboard

The Review Dashboard provides a set of interactive charts and reports that measure prediction accuracy and analyze items by probability score. Case Administrators and reviewers (with appropriate permissions) can view Prediction Rank information in either chart or table format.

To access Prediction Rank information from the Dashboard

1. From Case Home, in the Review Dashboard Chart Options box, select Analysis: Prediction Rank.

2. From the Folders drop-down list, select all documents, all folders, or a specific folder to display data.

3. Select a predictive tag to view from the Models menu. This example shows the predictive tag Responsive.

4. From the Labels drop-down list, select all labels or a specific label to display filter criteria data.

5. Select whether you want unused columns to display in the chart.

6. Click Chart.

   A chart displays showing the number of prediction tag items by filter criteria.

7. To view the displayed information in table format, select the Table view.
Granting/Restricting Transparent Predictive Coding Access

As a Case Administrator or Manager you can allow or restrict the viewing of prediction ranks, actions and management views. For information on granting or restricting Transparent Predictive Coding access, refer to "Administering User Accounts" in the System Administration Guide.
Predictive Analytics with Search and Review

To better understand and interpret the Transparent Predictive Coding results, the Search and Review capabilities display prediction ranking information. You are already familiar with the benefits of Advanced Search and Review on the platform and the prediction ranking additions are logical extensions to assist in review analysis and case strategy. These analytics make sense of the available prediction results using a bar ranking, action status and percentage prediction metric and are designed to inform your next step in the review workflow process.

Refer to the topics in this section:

• “Viewing Prediction Rank Search Results” on page 55
• “Reading Prediction Rank Results” on page 56
• “Filtering Prediction Rank Results” on page 58
• “Prediction Searches” on page 59
• “Viewing Prediction Rankings” on page 60

Viewing Prediction Rank Search Results

Starting with 7.1.2, a Prediction Rank column has been added to the search results page. This new column allows you to select a single tag, sort, and view a bar chart or a status icon of the prediction ranking for all items that have been included in the building of the system. You can examine prediction ranks at the parent item level or drill-down with the Detail view of prediction rankings on any single item.

There is nothing about sorting on prediction score. In particular, it should be noted that sorting the column uses the “family score” and not the score for the email item itself. So when you sort, the bars will not necessarily (and in fact will probably not) appear to be in order. You have to use the hover to determine the reason why a document was sorted a particular way.

Depending on the state of the item in the predictive model, the Prediction Rank column can display different types of information:

• Ranking Bar— Gives tag and prediction rank for the parent item. You can sort on the column to display in descending or ascending order. The sorting is based on the highest Prediction Rank in the entire document family. See “To read Prediction Rank results (Ranking Bar example)” on page 56.

• Training—Indicates the item was used for training.

• Unknown—Indicates no Prediction Rank for the item due to content characteristics that are not known to the system. The unknown state can change with additional training cycles.

• No Indexable Content—Indicates no Prediction Rank because the item has no text content.
Reading Prediction Rank Results

**Information:**

- The Prediction Rank column:
  - Appears only if Transparent Predictive Coding actions have been run and you have View access.
  - Remembers and displays the previously viewed Predictive Tag.
  - Displays other conditions such as Training, Unknown and No Content, when appropriate, instead of the bar chart.

**User:**

Case Administrator, Case Managers and Review Team

**To read Prediction Rank results (Ranking Bar example)**

1. Select the predictive tag from under the edit icon.
2. Hover over the prediction bar to get additional predictive tag information.

**Note:** Sorting on the Prediction rank column uses the document family score and not the score for the email item itself. This means that when you sort, the Prediction rank bars may not be in order. You should use the hover action to determine the reason why a document was sorted a particular way.

This example uses the Responsive tag. The Prediction rank displays a ranking of 43.22% for the parent item and also gives you the highest (96%) and lowest (35%) prediction ranking of the attachment items which were part of the parent-child ranking.
Table 20: Search Prediction Rank Information

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive Tag Name</td>
<td>Name of the tag</td>
</tr>
<tr>
<td>Parent Prediction Rank</td>
<td>Percentage that the parent matches the predictive tag</td>
</tr>
<tr>
<td>Highest Attachment or Embedding Prediction Rank</td>
<td>Highest percentage ranking of all the attachments or embeddings in the family</td>
</tr>
<tr>
<td>Lowest Attachment or Embedding Prediction Rank</td>
<td>Lowest percentage ranking of all the attachments or embeddings in the family</td>
</tr>
</tbody>
</table>

3. Click on the Prediction Rank bar and the View Training Examples menu displays. A list of the items that were considered the most influencing in the tagging prediction for this item displays. For more information, see “Understanding Predictive Ranking Decisions” on page 62.

Viewing Detail Prediction Rank Results for Attachments and Embeddings

While the main search page listing gives you a prediction rank for the parent item, you can view prediction ranks for all the related attachments of an item by selecting the Detail mode for an item.

To view detailed Prediction Rank results for attachments and embeddings

1. Click on Detail for the item you want to view.

   The Detail screen opens.

   ![Detail screen image](image)

   You can see the prediction ranking for each attachment.

2. Click on the prediction ranking bar for the attachment and the View Training Examples menu opens with more detail. See “Understanding Predictive Ranking Decisions” on page 62.
Filtering Prediction Rank Results

You can apply different Prediction Rank filters to gain a better understanding of how the Transparent Prediction Coding model is working and intelligently cull down your review population. The platform gives you a ranking of the prediction tags in the order of most to least relevant. For general filtering information, see “Filtering Search Results” in the User Guide.

Prediction Rank filtering can be valuable in whittling down and prioritizing a subset of items that will ultimately lead to a reduced set of key items for the case. For example, you may want to manage the large item review for responsive documents by assigning review of certain percentage ranges to your reviewers for culling. These filtering ranges can be useful as a First-Pass Review in order to get a quick initial batch of relevant review items for your review team.

To filter on Prediction Rank Results

1. Click on the filter category to sort the items.

Table 21: Filter Search Results: Prediction Rank

<table>
<thead>
<tr>
<th>Filter Name</th>
<th>Usage Guidelines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction Rank</td>
<td>Includes all prediction tags. It also includes 2 subfilters:</td>
</tr>
<tr>
<td>• Predictions subfilter — Lists all the items that have been run through the Predict action.</td>
<td>- 5 predefined ranges of prediction ranks.</td>
</tr>
<tr>
<td></td>
<td>- No Indexable Content — Includes items that could not be assigned a Prediction Rank.</td>
</tr>
<tr>
<td></td>
<td>- Unknown — Includes all items that do not match the training set.</td>
</tr>
<tr>
<td>• Training subfilter — Lists all items that have run through the Train action.</td>
<td>- Training — Includes all items used for training the predictive tag.</td>
</tr>
<tr>
<td></td>
<td>- Not Used — Includes all items that were not used for training due to training errors.</td>
</tr>
</tbody>
</table>

Note: The search training filter separates items into Trained and Not Used whereas Advanced Search Training Document combines both of these elements into Training Document.
Prediction Searches

Find particular prediction rankings and ranges on your items with Prediction search to gain insight into your data. For example, depending on the review requirements of the case, you can use Prediction search to locate all the Responsive items between a certain percentage range and exclude all Privileged items in your search.

You perform a Prediction Search from the Advanced Search page. The Prediction search can be expanded using multiple rows, each of which contains three drop-down boxes and a percentage min and max field.

![Prediction Search Interface](image)

**Table 22: Advanced Search: Predictions**

<table>
<thead>
<tr>
<th>Field/Option</th>
<th>Usage Guidelines</th>
</tr>
</thead>
<tbody>
<tr>
<td>any/not any</td>
<td>Enables positive or negative searches</td>
</tr>
<tr>
<td>with/without</td>
<td></td>
</tr>
<tr>
<td>Tag name</td>
<td>Displays a list of all predictive tags</td>
</tr>
<tr>
<td>Rank Between/</td>
<td>Finds items according to the following rules:</td>
</tr>
<tr>
<td>Unknown/</td>
<td>• <strong>Ranked Between</strong> specifies the range of Prediction Ranks</td>
</tr>
<tr>
<td>No Indexable Content/</td>
<td>• <strong>Unknown</strong> includes all items that do not match the training set.</td>
</tr>
<tr>
<td>Training Document</td>
<td>• <strong>No Indexable Content</strong> includes items that could not be assigned a Prediction Rank</td>
</tr>
<tr>
<td>Min%/Max%</td>
<td>• <strong>Training Document</strong> includes all items used for training the predictive tag.</td>
</tr>
<tr>
<td></td>
<td>Specifies the minimum and maximum prediction ranking</td>
</tr>
</tbody>
</table>
Viewing Prediction Rankings

Transparent Predictive Coding Platform offers complete visibility into an item’s prediction. Not only does the system calculate prediction rankings but it shows why and what the system considered when assigning a particular prediction ranking. Seeing how the system arrived at its results ensures the quality, defensibility, and statistical veracity of those results and validates the training process.

The review platform displays prediction rankings in the tagging panel for both an item and its attachments. In addition, the View Training Examples screen along with the Analyze functionality provide in-depth analysis of why the model applies a tagging decision for a given prediction ranking. This transparency provides assurances into the reliability and enables you to make informed decisions about your review strategy and scope.

Refer to the following Review topics:

- “Viewing Tagging Prediction Rank: Review Mode” on page 60
- “Viewing Attachment Prediction Rank: Review Mode” on page 61
- “Viewing Predictive Tagging events: Review Mode” on page 62
- “Understanding Predictive Ranking Decisions” on page 62

Viewing Tagging Prediction Rank: Review Mode

You can view the prediction rank of every predictive tag in the tagging panel.

To view prediction ranking in the tagging panel

1. In review mode, select the tag and hover over it to see its prediction rank.

2. To open the View Training Examples menu, click on the prediction ranking bar. For more information, see “To view the View Training Examples” on page 62.
Viewing Attachment Prediction Rank: Review Mode

You can view the prediction ranks for all the attachments of the item you are reviewing. In this example, there are 8 attachments with prediction rankings for the Responsive tag and it shows the prediction rank for the third attachment.

To view prediction ranks of attachments

1. In review mode, hover over the prediction bar to get the prediction rank for the attachment.

2. To open the View Training Examples menu, click on the prediction ranking bar for the attachment. See “Understanding Predictive Ranking Decisions” on page 62.
Viewing Predictive Tagging events: Review Mode

Use the Related Items pane to view prediction tag event history.

1. The Related Items view, which is part of the Review mode, displays prediction rankings and other event history.

![Related Items pane](image)

Understanding Predictive Ranking Decisions

The View Training Examples screen shows you how the system derived its prediction ranking. Use the table entries to understand which training items have the highest influence on the tagging prediction for the item. From this screen, you can export the contents to a CSV file or select Analyze to drill down to the Predictive Tag Analysis menu for further analysis and, if necessary, perform actions to retrain the system.

To view the View Training Examples

1. Click on the prediction bar ranking from either the Search or Review screen display of the item.

   The View Training Examples screen displays.

   ![View Training Examples](image)

2. Clicking on an item displays all the entries tagged with the Predictive Tag (for example Responsive) or close matches of those that were under the Features In Common heading.

3. Click Export to copy the contents to a CSV file.

4. Click Analyze to open the Predictive Tag Analysis menu. See “Predictive Tag Analysis” on page 63.
Predictive Tag Analysis

You can glean deeper insight, adjust and correct tagging decisions from the Predictive Tag Analysis screen. The transparency of the system enables you to evaluate its approach by looking at the primary item, positive and negative examples, and the highlighted Features in common elements. This information helps you to determine if further refining is necessary. For example, if you decide that a certain item should be tagged differently by the system, you can retag, retrain and then predict with the new information. You can use the Actions menu for the retraining workflow of: Tag, Train and Predict.

Important: If you do change the tagging, make sure that you immediately incorporate the changes into the system by cycling through the Train and Predict workflow. If you do not retrain the model with the new tagging decisions, the changes have no effect until the system is made aware of the changes during training.

To analyze predictive tag choices

The following example shows the primary item along with negative and positive examples that account for the predictive tagging decision of Responsive. Examining both the negative and positive examples gives you insight into why the system applied what it learned and how useful it is in its current state for predictive tagging.

• From the View Training Examples screen, click Analyze.

The Predictive Tag Analysis Screen displays.

The Primary Item displays first.
This is a positive example.

The heading indicates this is a positive example and lists the associated **Features in common**.

This is a negative example.

The heading indicates it is a negative example and lists the **Features in common**.
Appendix: Review Accuracy and Metrics

Controlled Prediction Accuracy Test and Prediction Accuracy Test

How do these tests work?

The tests compare the system’s predictive tagging decisions against the humans’ tagging decisions and compute the number of True Positives, False Positives, True Negatives, and False Negatives. These metrics factor into the F-Measures and the accuracy of the predictions. The F-Measures calculated by the test can help you determine whether to continue iterating or to move items outside of the Transparent Predictive Coding workflow.

How do you view what decisions the system will make?

To see what decisions the system would make, you should select a Prediction Rank Threshold. At this point, you can see exactly which items would be considered to be positive, (the remainder of the items in the Population would be considered to be negative), by running an advanced search on all items in the Population that have Prediction Ranks at or above the Prediction Rank Threshold.

Which Prediction Rank Threshold should I choose?

For the most balanced results, choose the Prediction Rank Threshold that specified in the test result as the best threshold.” The best Prediction Rank Threshold will coincide with the highest possible F-Measure value calculated by the test, which is available in the table of the test report. The test report also allows you to see how changing the Prediction Rank Threshold affects the F-Measure, Precision, and Recall values. For example, a lower Prediction Rank Threshold may give a higher Recall value with a lower Precision and F-Measure, while a higher Prediction Rank Threshold may give a higher Precision value with a lower Recall and F-Measure.

How does the Controlled Prediction Accuracy Test differ from the Prediction Accuracy Test?

The Controlled Prediction Accuracy Test operates on a system-created Control Set, and contains detailed information about the nature of the Population and the statistics involved with the creation of the Control Set. This means that the generated test results not only contain calculated values of the F-Measure, Precision, and Recall based on the Control Set, but they also contain projected minimum and maximum values for these metrics and are statistically guaranteed for the Population.

The Prediction Accuracy Test, on the other hand, simply operates on a set of items with Prediction Ranks that are specified by the user. Thus, the generated test results only contain calculated values of the F-Measure, Precision, and Recall based on the user-specified items.
Glossary Terms

F-Measure
The harmonic mean between Precision and Recall. As Precision and Recall are often inversely correlated (such as Precision increases and Recall decreases, and vice versa), the F-Measure is useful in determining the balance between the two metrics.

False Negative
This occurs when an item’s Prediction Rank is below the Prediction Rank Threshold but the item has been tagged as positive.

False Positive
This occurs when an item’s Prediction Rank is at or above the Prediction Rank Threshold but the item has not been tagged as positive.

Precision
The ratio of True Positives to the total number of Positive items identified by the system. A higher precision indicates a smaller number of False Positives, while a lower precision indicates a larger number of False Positives.

Prediction Rank
This refers to how closely a given item’s content and metadata match what the system deems to be Positive.

Prediction Rank Threshold
This refers to a Prediction Rank value selected by using a test report. All items that have a Prediction Rank at or above the Prediction Rank Threshold would then be considered as positive, while all items that have a Prediction Rank below the Prediction Rank Threshold would then be considered as negative.

Population
All of the items to which Transparent Predictive Coding is being applied.

Recall
This is the ratio of True Positives to the total number of Positive items identified by the system. A higher recall indicates a smaller number of False Negatives, while a lower precision indicates a larger number of False Negatives.

Training Conflicts
For a given set of items for training, Transparent Predictive Coding identifies tagging decisions that mark similar items differently as conflicts.

Note: Including items in training that are marked as conflicts without resolving the underlying reason for the conflict can affect and potentially lower the prediction accuracy.

True Negative
This occurs when an item’s Prediction Rank is below the Prediction Rank Threshold and the item has not been tagged as positive.

True Positive
This occurs when an item’s Prediction Rank is at or above the Prediction Rank Threshold and the item has been tagged as positive.