Objective
Implement automatic methods to identify as many relevant documents as possible, while requiring less review effort.

Baseline
• “Seed set” is constructed from the query terms.
• Logistic Regression classification.
• Select the most-likely relevant documents for review.
• Repeat the above process until collecting a sufficient number of relevant documents.

Seed Selection
1. Select Top K documents with the highest BM25 score.
2. Latent semantic indexing and dimension reduction via SVD on features.

Sampling Strategy:
Exploration vs. Exploitation
\[ l_t \in \arg\max_{1 \leq i \leq k} \left\{ \sum_{i=1}^{k} \frac{1}{\left| C_i \right|} \log \frac{\left| C_i \right|}{\left| C_i \cap t \right|} \right\} \]

Graph Strategy:
Run K-Means multiple times and compose un-directed graph based on the times two documents appear in the same cluster.

Jumping Strategy:
Greedy search in one cluster and switch to other cluster when ir-relevant document is found.

Weighed Strategy:
Assign weight for each cluster and decay the weight when encountering ir-relevant document is found.

Feature Engineering
• Unigram & 2-gram TF-IDF value
• The combination of these features.

Reciprocal Rank Fusion (RRF) of Various Features
\[ RRF_{score}(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)} \]

Classifier Selection
- Logistic Regression
- Gradient Boosting
- Decision Tree
- Naive Bayes
- AdaBoost
- RBF SVM
- Linear SVM

Evaluation and Results

<table>
<thead>
<tr>
<th>Methods</th>
<th>tr0</th>
<th>tr1</th>
<th>tr2</th>
<th>tr3</th>
<th>tr4</th>
<th>tr5</th>
<th>tr6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jumping</td>
<td>46</td>
<td>1</td>
<td>2</td>
<td>10</td>
<td>47</td>
<td>49</td>
<td>40</td>
</tr>
<tr>
<td>Weighted</td>
<td>46</td>
<td>0</td>
<td>2</td>
<td>10</td>
<td>47</td>
<td>49</td>
<td>42</td>
</tr>
<tr>
<td>Sampling</td>
<td>45</td>
<td>1</td>
<td>2</td>
<td>14</td>
<td>48</td>
<td>49</td>
<td>46</td>
</tr>
<tr>
<td>Graph</td>
<td>47</td>
<td>2</td>
<td>2</td>
<td>15</td>
<td>45</td>
<td>50</td>
<td>45</td>
</tr>
</tbody>
</table>

At Home
- UWPAH1 (without query expansion) and UWPAH2 (with query expansion).

Sandbox
- UWPAH1 (without query expansion)

Query Expansion
Simple Mixture Model
Estimate feedback query model based on feedback documents and original query:
\[ \hat{\theta}_Q' = (1 - \alpha) \hat{\theta}_Q + \alpha \hat{\theta}_F \]

<table>
<thead>
<tr>
<th>Topic</th>
<th>UWPAH1</th>
<th>UWPAH2</th>
</tr>
</thead>
<tbody>
<tr>
<td>athome100</td>
<td>4019</td>
<td>3968</td>
</tr>
<tr>
<td>athome101</td>
<td>4503</td>
<td>4491</td>
</tr>
<tr>
<td>athome102</td>
<td>1402</td>
<td>1417</td>
</tr>
<tr>
<td>athome103</td>
<td>4307</td>
<td>4305</td>
</tr>
<tr>
<td>athome104</td>
<td>272</td>
<td>291</td>
</tr>
<tr>
<td>athome105</td>
<td>2898</td>
<td>2981</td>
</tr>
<tr>
<td>athome106</td>
<td>12661</td>
<td>12892</td>
</tr>
<tr>
<td>athome107</td>
<td>1914</td>
<td>1892</td>
</tr>
<tr>
<td>athome108</td>
<td>2337</td>
<td>2228</td>
</tr>
<tr>
<td>athome109</td>
<td>2642</td>
<td>2358</td>
</tr>
</tbody>
</table>