Candidate Subtopics:

Resources
- Bing query suggestions and completions
- Google query completions
- Yahoo query completions

Feature Extraction:
- Term frequency based features (DFH, PL2, BM25, etc.)
- Language modeling based features (KL, QLM-JM, SLM-JM, etc.)
- Lexical features (Edit distance, Term overlap, etc.)
- Web hit count based features (NHC, PMI, etc.)

Feature Selection:
Optimization function of Elastic Net:
$$
\text{min}_{\alpha, \beta} \sum_{i=1}^{n} \left( \alpha_i - \beta_i - X_i^T \eta \right)^2 + \lambda \sum_{i=1}^{n} \left( 1 - \frac{1}{2} \beta_i^2 + a||\beta_i||_1 \right)
$$

Relevance Estimation:
Linear ranking model is employed to estimate relevance as follows:
$$
\text{rel}(q, s) = \sum_{k=1}^{N} w_k \cdot f_c(s, q)
$$
- Random forest is utilized to estimate feature importance $w_k$
- $N$ is the number of selected features.

Subtopic Diversification:
Diversifying subtopics by balancing relevance and novelty:
$$
\xi_i = \arg \max_{\forall} \frac{1}{n} \cdot \text{rel}(q, s_i) \cdot (1 - \gamma) \cdot \text{novelty}(s_i, c_i)
$$
$$
\text{novelty}(s_i, c_i) = -\max_{s_i' \neq s_i} \cos(\beta_i, \beta_i')
$$
$\gamma \in [0, 1]$ is a combining parameter.

$n$ indicates the novelty of subtopic $s_i$ given the set $c_i$

Evaluation:

Dataset:
Training: NTCIR-10 INTENT-2 English
Testing: NTCIR-12 IMINE-2 English

Subtopic Mining Subtask:

<table>
<thead>
<tr>
<th>Runs</th>
<th>KDEIM-Q-E-1Q</th>
<th>KDEIM-Q-E-2Q</th>
<th>KDEIM-Q-E-3Q</th>
<th>KDEIM-Q-E-4Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-rec@10</td>
<td>0.7556</td>
<td>0.7556</td>
<td>0.7458</td>
<td>0.7464</td>
</tr>
<tr>
<td>D-nDCG@10</td>
<td>0.6644</td>
<td>0.6644</td>
<td>0.6472</td>
<td>0.5645</td>
</tr>
<tr>
<td>Dir-nDCG@10</td>
<td>0.7100</td>
<td>0.7100</td>
<td>0.6955</td>
<td>0.6565</td>
</tr>
</tbody>
</table>

Query Understanding with Vertical:

<table>
<thead>
<tr>
<th>Runs</th>
<th>V-score</th>
<th>QU-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>KDEIM-Q-E-2Q</td>
<td>0.3014</td>
<td>0.5057</td>
</tr>
<tr>
<td>KDEIM-Q-E-3Q</td>
<td>0.2961</td>
<td>0.4948</td>
</tr>
</tbody>
</table>

Conclusion:
- Proposed a method for mining diversified subtopics
- Proposed a method for vertical selection exploiting word embedding
- Language modelling and query independent features are effective
- Diversification penalize the noisy and redundant subtopics

Future Work:
- Extracting candidate subtopics from other resources
- Top retrieved documents, Wikipedia, Knowledge graph.
- More semantic features for estimating relevance and novelty
- Effectively using word embedding for vertical classification
- Search result diversification using the mined subtopic and vertical