CS 589 Fall 2020
Information Retrieval Evaluation
Retrieval Feedback

Instructor: Susan Liu
TA: Huihui Liu

Stevens Institute of Technology
Information retrieval evaluation

• Last lecture: basic ingredients for building a document search engine

• You graduate and join Bing
Information retrieval evaluation

• How to know
  • If your search engine has outperformed another search engine
  • If your search engine performance has improved compared to last quarter?

Beat Amazon!
Metrics for a good search engine

- Return what the users are looking for
- Return results fast
- Users likes to come back

- Relevance, CTR
- Latency
- Retention rate
Rank-based measurements

• Binary relevance
  • Precision@K
  • Mean average precision (MAP)
  • Mean reciprocal rank (MRR)

• Multiple levels of relevance
  • Normalized discounted cumulative gain (NDCG)
Precision of retrieved documents

- Fraction of retrieved docs that are relevant

\[
\text{precision} = \frac{\#\text{relevant and retrieved}}{\#\text{retrieved}}
\]

- Fraction of relevant documents that are retrieved

\[
\text{recall} = \frac{\#\text{relevant and retrieved}}{\#\text{relevant}}
\]
Rank-based measurements

- Binary relevance
  - Precision@K
  - Mean average precision (MAP)
  - Mean reciprocal rank (MRR)

- Multiple levels of relevance
  - Normalized discounted cumulative gain (NDCG)
Precision-recall curve

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>1/1</td>
</tr>
<tr>
<td>+</td>
<td>2/2</td>
</tr>
<tr>
<td>-</td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>3/5</td>
</tr>
<tr>
<td>-</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>4/8</td>
</tr>
<tr>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

\[
\frac{1/1 + 2/2 + 3/5 + 4/8}{4}
\]

Slides from UIUC CS598
Mean average precision

- Consider rank position of each *relevant* doc
  - $K_1, K_2, \ldots K_R$

- Compute Precision@K for each $K_1, K_2, \ldots K_R$

- Average precision = average of $P@K$

- Ex: has AvgPrec of

$$\frac{1}{3} \cdot \left( \frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right) \approx 0.76$$

- MAP is Average Precision across multiple queries/rankings

*Slides from Stanford CS276*
### Average precision

<table>
<thead>
<tr>
<th>Ranking #1</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.17</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>0.17</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.33</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>0.67</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ranking #2</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>0.17</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>0.17</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>0.33</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td></td>
</tr>
</tbody>
</table>

**Ranking #1:** \((1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78\)

**Ranking #2:** \((0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52\)
MAP

average precision query 1 = \((1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62\)

average precision query 2 = \((0.5 + 0.4 + 0.43)/3 = 0.44\)

mean average precision = \((0.62 + 0.44)/2 = 0.53\)
Mean reciprocal rank

- Measure the effectiveness of the ranked results
- Assume users are only looking for one relevant document

\[ MRR = \frac{1}{2} \times (1 + \frac{1}{2}) = 0.75 \]
Beyond binary relevance

- Discounted cumulative gain (DCG)
  - Popular measure for evaluating web search and related tasks
  - Information gain-based evaluation
    - For each relevant document, the user has gained some information
    - The higher the relevance, the higher gain
    - The gain is discounted when the relevant document appears in a lower position
Discounted cumulative gain (DCG)

![Diagram showing relevant documents and rankings]

\[
\text{DCG@3 query 1} = \frac{2^2 - 1}{\log_2 2} + \frac{2^1 - 1}{\log_2 4} + \frac{2^2 - 1}{\log_2 5} = 4.79
\]

\[
\text{DCG@3 query 2} = \frac{2^2 - 1}{\log_2 3} + \frac{2^1 - 1}{\log_2 6} + \frac{2^2 - 1}{\log_2 7} = 3.34
\]
Normalized Discounted cumulative gain (nDCG)

\[
IDCG_{@3} \text{ query 1} = \frac{2^2 - 1}{\log_2{2}} + \frac{2^2 - 1}{\log_2{3}} + \frac{2^2 - 1}{\log_2{4}} = 6.39
\]

\[
IDCG_{@3} \text{ query 2} = \frac{2^2 - 1}{\log_2{2}} + \frac{2^2 - 1}{\log_2{3}} + \frac{2^2 - 1}{\log_2{4}} = 6.39
\]

\[
nDCG = \frac{4.79/6.39 + 3.34/6.39}{2} = 0.64
\]
Relevance evaluation methodology

• Offline evaluation:
  • Evaluation based on annotators’ annotation
    • TREC conference
    • Cranfield experiments
    • Pooling
  • Evaluation based on user click through logs

• Online evaluation
  • A/B testing
Text REtrieval Conference (TREC)

- Since 1992, hosted by NIST
  - Number: 794

- Relevance:
  - Title: pet therapy
    - Description:
      How are pets or animals used in therapy for humans and what are the benefits?

- Different tracks:
  - Web
  - Question answering
  - Microblog

<narr> Narrative:
Relevant documents must include details of how pet- or animal-assisted therapy is or has been used. Relevant details include information about pet therapy programs, descriptions of the circumstances in which pet therapy is used, the benefits of this type of therapy, the degree of success of this therapy, and any laws or regulations governing it.

</top>
The Cranfield experiment (1958)

- Imagine you need to help users search for literatures in a digital library, how would you design such a system?

```
query = "subject = AI & subject = bioinformatics"
```

system 1: the Boolean retrieval system
The Cranfield experiment (1958)

- Imagine you need to help users search for literatures in a digital library, how would you design such a system?

Document-term matrix

<table>
<thead>
<tr>
<th></th>
<th>intelligence</th>
<th>book</th>
<th>the</th>
<th>cat</th>
<th>artificial</th>
<th>dog</th>
<th>business</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Doc2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>query</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

query = “artificial intelligence”  bags of words representation

system 2: indexing documents by lists of words
The Cranfield experiment (1958)

• Basic ingredients
  • A corpus of documents (1.4k paper abstracts)
  • A set of 225 queries and their information needs
  • Binary relevance judgment for each (q, d) pair
  • Reuse the relevance judgments for each (q, d) pair

query = “best phone”, time = 2012, relevance = 1

query = “best phone”, time = 2022, relevance = 0
Scalability problem in human annotation

- TREC contains $225 \times 1.4k = 315k$ (query, documents) pairs

- How to annotate so many pairs?

- Pooling strategy
  - For each system, first run the system to get top 100 results
  - Annotate the union of all such documents
Evaluation based on user click through logs

- TREC style relevance judgment
  - Explicit relevance judgment
  - Difficult to achieve large scalability
  - Relevance is fixed

- Relevance judgment using user clicks
  - Implicit relevance judgment
  - Effortless relevance judgment at a large scale
  - Relevance is fixed, (assume relevance judgment stays the same upon reranking)
Evaluation based on user click through logs

- Click logs for “CIKM”
Evaluation based on user click through logs

- System logs the users engagement behaviors:
  - Time stamp
  - Session id
  - Query id, query content
  - Items viewed by the user (in sequential order)
  - Whether each item has been clicked by the user
  - User’s demographic information, search/click history, location, device
  - Dwell time, browsing time for each document
  - Eye tracking information
Evaluation based on user click through logs

- Click logs are stored in large tables
- Using SQL to extract a subset of query logs

<table>
<thead>
<tr>
<th>Session Id</th>
<th>Timestamp</th>
<th>Action</th>
<th>Action details</th>
</tr>
</thead>
<tbody>
<tr>
<td>123457</td>
<td>1388494920</td>
<td>search</td>
<td>Query = ‘flawless’</td>
</tr>
<tr>
<td>123457</td>
<td>1388494980</td>
<td>click</td>
<td>Page Id = ‘755’</td>
</tr>
<tr>
<td>123457</td>
<td>1388495060</td>
<td>reformulation</td>
<td>Query = ‘flawless beyonce’ =&gt; Reformulation = ‘beyonce’</td>
</tr>
<tr>
<td>123457</td>
<td>1388495115</td>
<td>click</td>
<td>Page Id = ‘170’</td>
</tr>
<tr>
<td>123458</td>
<td>1388495415</td>
<td>search</td>
<td>Query = ‘cikm conference’</td>
</tr>
<tr>
<td>123456</td>
<td>1388361661</td>
<td>reformulation</td>
<td>Query = ‘cikm conference’ =&gt; Reformulation = ‘2014’</td>
</tr>
<tr>
<td>123456</td>
<td>1388361720</td>
<td>click</td>
<td>Page Id = “45”</td>
</tr>
</tbody>
</table>
Online evaluation methodology

- Assumption made by offline evaluation
  - After reranking, relevance judgment stays the same
  - Which is not true…

- Relevance judgment is dynamic, subject to user bias
  - Bias based on positions
  - Preference shifting over time, location
  - Satisficing bias
Position bias [Craswell 08]

- Position bias
  - Higher position receives more attention
  - The same item gets lower click in lower position
Satisficing bias

$400, 20G \quad \text{vs} \quad $500, 30G

click probability = 0.3

click probability = 0.4

click probability = 0.5

$550, 20G
Online evaluation methodology

- Evaluation by actually having the system deployed and observe user response
  - Less scalable
  - A/B testing

Query: [support vector machines]

<table>
<thead>
<tr>
<th>Ranking A</th>
<th>Ranking B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kernel machines</td>
<td>Kernel machines</td>
</tr>
<tr>
<td>SVM-light</td>
<td>SVMs</td>
</tr>
<tr>
<td>Lucent SVM demo</td>
<td>Intro to SVMs</td>
</tr>
<tr>
<td>Royal Holl. SVM</td>
<td>Archives of SVM</td>
</tr>
<tr>
<td>SVM software</td>
<td>SVM-light</td>
</tr>
<tr>
<td>SVM tutorial</td>
<td>SVM software</td>
</tr>
</tbody>
</table>
Interleaving

Kernel machines
Kernel machines
SVMs
SVM-light
Intro to SVMs
Lucent SVM demo
Archives of SVM
Royal Holl. SVM
SVM-light

Kernel machines
SVMs
SVM-light
Intro to SVMs
Lucent SVM demo
Archives of SVM
Royal Holl. SVM
SVM-light

remove dup

A clicks = 3, B clicks = 1
Online evaluation methodology

- Bing has an existing ranking algorithm A
  - Testing algorithm B is better than A
    - Strategy 1: Running A of 1 month, running B for the next month
    - Strategy 2: Running A 50% of the time, B 50% of the time

- Disadvantage with Strategy 1 and 2:
  - If B fails, it will hurts user experience from the B group

- Running B 5% of the time, running A 95% of the time
Statistical significance testing

- How sure can you be that an observed difference doesn’t simply result from the particular queries you chose?

<table>
<thead>
<tr>
<th>Query</th>
<th>System A</th>
<th>System B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.20</td>
<td>0.40</td>
</tr>
<tr>
<td>2</td>
<td>0.21</td>
<td>0.41</td>
</tr>
<tr>
<td>3</td>
<td>0.22</td>
<td>0.42</td>
</tr>
<tr>
<td>4</td>
<td>0.19</td>
<td>0.39</td>
</tr>
<tr>
<td>5</td>
<td>0.17</td>
<td>0.37</td>
</tr>
<tr>
<td>6</td>
<td>0.20</td>
<td>0.40</td>
</tr>
<tr>
<td>7</td>
<td>0.21</td>
<td>0.41</td>
</tr>
</tbody>
</table>

| Average | 0.20 | 0.40 |

<table>
<thead>
<tr>
<th>Query</th>
<th>System A</th>
<th>System B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.02</td>
<td>0.76</td>
</tr>
<tr>
<td>2</td>
<td>0.39</td>
<td>0.07</td>
</tr>
<tr>
<td>3</td>
<td>0.16</td>
<td>0.37</td>
</tr>
<tr>
<td>4</td>
<td>0.58</td>
<td>0.21</td>
</tr>
<tr>
<td>5</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>6</td>
<td>0.09</td>
<td>0.91</td>
</tr>
<tr>
<td>7</td>
<td>0.12</td>
<td>0.46</td>
</tr>
</tbody>
</table>

| Average | 0.20 | 0.40 |
## Statistical significance testing

<table>
<thead>
<tr>
<th>Query</th>
<th>System A</th>
<th>System B</th>
<th>Sign Test</th>
<th>Wilcoxon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.02</td>
<td>0.76</td>
<td>+</td>
<td>+0.74</td>
</tr>
<tr>
<td>2</td>
<td>0.39</td>
<td>0.07</td>
<td>-</td>
<td>-0.32</td>
</tr>
<tr>
<td>3</td>
<td>0.16</td>
<td>0.37</td>
<td>+</td>
<td>+0.21</td>
</tr>
<tr>
<td>4</td>
<td>0.58</td>
<td>0.21</td>
<td>-</td>
<td>-0.37</td>
</tr>
<tr>
<td>5</td>
<td>0.04</td>
<td>0.02</td>
<td>-</td>
<td>-0.02</td>
</tr>
<tr>
<td>6</td>
<td>0.09</td>
<td>0.91</td>
<td>+</td>
<td>+0.82</td>
</tr>
<tr>
<td>7</td>
<td>0.12</td>
<td>0.46</td>
<td>-</td>
<td>-0.38</td>
</tr>
<tr>
<td>Average</td>
<td>0.20</td>
<td>0.40</td>
<td>p=1.0</td>
<td>p=0.9375</td>
</tr>
</tbody>
</table>

95% of outcomes

---

Slides from UIUC CS598
Retrieval feedback in session search

query = “best phone”

$400, 20G, Nokia

Does the user prefer lower priced phone, or high end phones? Larger storage, better camera?

$500, 30G, Nokia

session 2

$600, 40G, iphone

observed click
Rocchio feedback

- Feedback for vector-space model
  \[ q_F = \alpha q + \frac{\beta}{|D_r|} \sum_{d_r \in D_r} d_r - \frac{\gamma}{|D_n|} \sum_{d_n \in D_n} d_n \]
  
  \( \alpha \gg \beta \)

- Rocchio’s practical issues
  - Large vocabularies (only consider important words)
  - Robust and effective
  - Requires relevance feedback
Pseudo-relevance feedback

- What if we do not have relevance judgments?
  - Use the top retrieved documents as “pseudo relevance documents”

- Why does pseudo-relevance feedback work?

query = “fish tank”

www.petsmart.com › fish › aquariums ▼

Fish Tanks & Aquariums | PetSmart

125 Items · Shop the latest fish tanks and aquariums at PetSmart to find interesting ways to showcase your favorite fish. Browse large and small tanks, fresh and ... Tanks, Aquariums & Nets · Fish Tanks for Sale: Discount · Fish Aquariums
Relevance feedback in RSJ model

\[ O(\text{rel} = 1|q, d) \overset{\text{rank}}{=} \sum_{w_i=1}^{\text{rank}} \log \frac{\alpha_i(1 - \beta_i)}{\beta_i(1 - \alpha_i)} \]

\[(\text{Robertson \\& Sparck Jones 76})\]

\[ \alpha_i = p(w_i = 1|q, \text{rel} = 1) \]
\[ = \frac{\text{count}(w_i = 1, \text{rel} = 1) + 0.5}{\text{count}(\text{rel} = 1) + 1} \]

Probability for a word to appear in a relevant doc

\[ \beta_i = p(w_i = 0|q, \text{rel} = 0) \]
\[ = \frac{\text{count}(w_i = 0, \text{rel} = 0) + 0.5}{\text{count}(\text{rel} = 0) + 1} \]

Probability for a word to appear in a non-relevant doc
(Pseudo)relevance feedback language model

\[ \text{score}^{JM}(q, d) = \sum_{w_i, w_i \in d, p(w_i | \hat{q})} p(w_i | \hat{q}) \log \left(1 + \frac{(1 - \lambda) \text{count}(w_i, d)}{\lambda p(w_i | C)}\right) \]

\[ p(w_i | q) = \frac{\text{count}(w_i, q)}{|q|} \]

**sparsity**

\[ d \rightarrow \theta_d \rightarrow -D(\theta_q | \theta_d) \]

\[ q \rightarrow \theta_q \rightarrow (1 - \alpha) \theta_q + \alpha \theta^F_q \]

\[ \theta^F_q \rightarrow d_1, d_2, \ldots, d_n \]
Performance of relevance feedback models

<table>
<thead>
<tr>
<th>S.w.</th>
<th>Metric</th>
<th>MLE</th>
<th>RM3</th>
<th>RM4</th>
<th>DMM</th>
<th>SMM</th>
<th>RMM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Trained on AP1 and Tested on AP2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/</td>
<td>AvgPr</td>
<td>0.220</td>
<td>0.295</td>
<td>0.301</td>
<td>0.290</td>
<td><strong>0.304</strong></td>
<td>0.299</td>
</tr>
<tr>
<td></td>
<td>Pr@10</td>
<td>0.386</td>
<td>0.408</td>
<td>0.418</td>
<td><strong>0.422</strong></td>
<td>0.400</td>
<td>0.398</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>3074</td>
<td>3810</td>
<td>3892</td>
<td>3681</td>
<td><strong>3933</strong></td>
<td>3859</td>
</tr>
<tr>
<td>w/o</td>
<td>AvgPr</td>
<td>0.231</td>
<td>0.312</td>
<td>0.321</td>
<td>0.289</td>
<td><strong>0.324</strong></td>
<td>0.323</td>
</tr>
<tr>
<td></td>
<td>Pr@10</td>
<td>0.398</td>
<td>0.436</td>
<td><strong>0.448</strong></td>
<td>0.424</td>
<td>0.432</td>
<td>0.446</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>3154</td>
<td>3913</td>
<td>3908</td>
<td>3674</td>
<td>3921</td>
<td><strong>3927</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Trained on TREC6 and Tested on TREC78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/</td>
<td>AvgPr</td>
<td>0.217</td>
<td>0.249</td>
<td>0.242</td>
<td>0.235</td>
<td><strong>0.251</strong></td>
<td>0.243</td>
</tr>
<tr>
<td></td>
<td>Pr@10</td>
<td>0.437</td>
<td>0.438</td>
<td>0.426</td>
<td>0.443</td>
<td>0.443</td>
<td><strong>0.451</strong></td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>5114</td>
<td>5805</td>
<td>5739</td>
<td>5476</td>
<td><strong>5821</strong></td>
<td>5625</td>
</tr>
<tr>
<td>w/o</td>
<td>AvgPr</td>
<td>0.217</td>
<td>0.251</td>
<td>0.243</td>
<td>0.235</td>
<td><strong>0.252</strong></td>
<td>0.249</td>
</tr>
<tr>
<td></td>
<td>Pr@10</td>
<td>0.434</td>
<td><strong>0.454</strong></td>
<td>0.446</td>
<td>0.433</td>
<td>0.441</td>
<td>0.443</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>5107</td>
<td>5799</td>
<td>5776</td>
<td>5500</td>
<td><strong>5896</strong></td>
<td>5833</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Well-Tuned on WT2G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/</td>
<td>AvgPr</td>
<td>0.293</td>
<td><strong>0.338</strong></td>
<td>0.319</td>
<td>0.327</td>
<td>0.330</td>
<td>0.309</td>
</tr>
<tr>
<td></td>
<td>Pr@10</td>
<td>0.450</td>
<td><strong>0.500</strong></td>
<td>0.470</td>
<td>0.494</td>
<td>0.496</td>
<td>0.458</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>1830</td>
<td>1822</td>
<td>1806</td>
<td>1843</td>
<td><strong>1856</strong></td>
<td>1811</td>
</tr>
<tr>
<td>w/o</td>
<td>AvgPr</td>
<td>0.306</td>
<td><strong>0.344</strong></td>
<td>0.328</td>
<td>0.326</td>
<td>0.331</td>
<td>0.319</td>
</tr>
<tr>
<td></td>
<td>Pr@10</td>
<td>0.456</td>
<td><strong>0.490</strong></td>
<td><strong>0.490</strong></td>
<td>0.476</td>
<td>0.476</td>
<td>0.482</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>1870</td>
<td>1862</td>
<td>1879</td>
<td>1873</td>
<td><strong>1889</strong></td>
<td>1863</td>
</tr>
</tbody>
</table>
Query expansion

- what is the most common blood type
- what is the most shared video on tiktok
- what is the most expensive car
- what is the most expensive car in the world
- what is the most expensive thing in the world
- what is the most popular game
Query expansion

• Query expansion/reformulation techniques
  • Using manually created synonyms
  • Using automatic derived thesaurus
  • Using query log mining

<table>
<thead>
<tr>
<th>Word</th>
<th>Nearest neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>absolutely</td>
<td>absurd, whatsoever, totally, exactly, nothing</td>
</tr>
<tr>
<td>bottomed</td>
<td>dip, copper, drops, topped, slide, trimmed</td>
</tr>
<tr>
<td>captivating</td>
<td>shimmer, stunningly, superbly, plucky, witty</td>
</tr>
<tr>
<td>doghouse</td>
<td>dog, porch, crawling, beside, downstairs</td>
</tr>
<tr>
<td>makeup</td>
<td>repellent, lotion, glossy, sunscreen, skin, gel</td>
</tr>
<tr>
<td>mediating</td>
<td>reconciliation, negotiate, case, conciliation</td>
</tr>
<tr>
<td>keeping</td>
<td>hoping, bring, wiping, could, some, would</td>
</tr>
<tr>
<td>lithographs</td>
<td>drawings, Picasso, Dali, sculptures, Gauguin</td>
</tr>
<tr>
<td>pathogens</td>
<td>toxins, bacteria, organisms, bacterial, parasite</td>
</tr>
<tr>
<td>senses</td>
<td>grasp, psyche, truly, clumsy, naive, innate</td>
</tr>
</tbody>
</table>