Prove It! (PI)

1 PI goals

Building on a corpus of over 50,000 digitized books, this task investigates the application of focused and semantic retrieval approaches to a collection of digitized books.

2 PI user scenario

The scenario underlying this task is that of a user searching for specific information in a library of books that can provide evidence to *confirm or refute a given factual statement*. Users expect to be pointed directly at book pages that can help them to confirm or refute the claim of the topic. Users are assumed to view the ranked list of retrieved book pages starting from the top of the list and moving down, examining each result. No browsing is considered (only the returned book pages are viewed by users).

3 PI task description

The system needs to do the following:

Search for book pages which are estimated to contain information that can confirm or refute a factual statement expressed in the topic. Before each identified page is returned, the system must verify that the containing book is appropriate (see below) for confirming or refuting the statement. Systems need to return a ranked list of up 1,000 book pages (given by bookID-pageNumber) for which both conditions (page content and book appropriateness) hold.

For a book to be appropriate, its topicality, genre or language is such that a user would trust it in confirming / refuting the statement. For example, a well-known biography on Charles Darwin could be trusted in confirming / refuting Darwin's place and date of birth, while a polemic essay of a novice student about theology could not.

For each result, the system should indicate whether it provides confirming or refuting evidence regarding the claim.

Participants may submit up to 6 runs.

4 PI corpus

The PI task builds on a collection of 50,239 out-of-copyright books, digitized by Microsoft. The corpus contains books of different genre, including history books, biographies, literary studies, religious texts and teachings, reference works, encyclopedias, essays, proceedings, novels, and poetry. 50,099 of the books also come with an associated MAchine-Readable Cataloging (MARC) record, which contains publication (author, title, etc.) and classification information. Each book in the corpus is identified by a 16 character bookID, which is also the name of the directory that contains the book's OCR file, e.g., A1CD363253B0F403.

The OCR text of the books has been converted from the original DjVu format to an XML format referred to as BookML, developed by Microsoft Development Center Serbia. BookML provides additional structure information, including a set of labels (as attributes) and additional marker elements for more complex structures, like table of contents. For example, the first label attribute in the XML extract below signals the start of a new chapter on page 1 (label="PT_CHAPTER"). Other semantic units include headers (SEC_HEADER), footers (SEC_FOOTER), back-of-book index (SEC_INDEX), table of contents (SEC_TOC). Marker elements provide detailed markup, e.g., indicating entry titles (TOC_TITLE) or page numbers (TOC_CH_PN) in a table of contents. The basic XML structure of a typical book in BookML is a sequence of pages containing nested structures of regions, sections, lines, and words, most of them with associated coordinate information, defining the position of a bounding rectangle.

The full corpus, totalling around 400GB, is available on USB HDDs – please contact Gabriella Kazai at a-gabkaz@microsoft.com. A reduced version (13GB compressed) is available via download from http://www.booksearch.org.uk/. The reduced version was generated by removing the word tags and propagating their content into the parent (i.e., line) elements.

5 PI topics

There are 83 topics for this task, each containing a factual statement, see example below.

```
<topic id=2010006 ct_nt=11>
<fact>On November 2, 1917, British prime minister Lord Balfour, in a letter to Lord Rotschild, declared the
British Government's intention of assisting the Jewish people to establish a national home for the jews in
Palestine.</fact>
<query>Balfour declaration national home</query>
<subject>national home for jews in Palestine</subject>
<wikiurl>http://en.wikipedia.org/wiki/Balfour_Declaration_of_1917</wikiurl>
<narrative>
<task> I am working on a school assignment about the establishment of the state of Israel as the fulfillment of
the Zionist aspirations. I was told that the Balfour declaration was an important milestone, and I want the fact
confirmed. </task>
<infneed> All statements in books are relevant that either prove or reject the fact. I need confirmation of the
exact date, and preferably also the text of the declaration. </infneed>
</narrative>
</topic>
```

Topics are available from the download area at http://www.booksearch.org.uk/.

6 PI training data

We have page-level judgements of the degree to which these confirm or refute 30 of the 83 topics, which can be used for training. The judgements provide a graded confirm/refute value for topic/page, and can be downloaded from the download area at http://www.booksearch.org.uk/.

7 PI evaluation

This is a high-precision task, with graded judgements, thus the official measure will be nDCG@10.

8 Submission format

Submissions for the Prove It task should conform to the trec_eval format, where a line has the format:
<query-number><”confirms”|”refutes”|”Q0”><book-page><rank><rsv><run-id>

9 Submission Deadline

The deadline for run submission is **May 19, 2013.** Details on how to submit will be announced closer to the deadline.