The Information Retrieval Project

by Michael Trittin and Andrew McKee

For this project, we developed an application which retrieves the relevancy of certain documents for a given search query and ranks them according to which document is the most relevant. In order to achieve this, our application uses the bm25 algorithm in combination with skip bigrams. Our specific approach implemented the skip bigrams as a dependency of the bm25 algorithm so that for a given phrase, if the entire phrase is found using skip bigrams, the count for the bm25 algorithm uses that. If the skip bigram phrase is found, it is weighted for more heavily than the disparate word version of bm25, which does not account for whether or not the identified words are close together. As a concrete example of this, consider the following:

Document 1: "President Lincoln was assassinated during his presidency. In an unfortunate turn of events, his death took place as he enjoyed his favorite past-time — plays".
Document 2: "Lincoln had always enjoyed theater and went without a security retinue, but after he was assassinated during a performance security was greatly increased."

Search phrase: "Lincoln assassinated"

	Document 1	**Document 2**
Regular BM25 Score	0.0137769	0.0158630
Skip-Bigram BM25 Score	0.027553	0.0
Combined Final Score	0.0413307	0.015863

As you can see, Document 1 is determined to be much more relevant than Document 2, which in the context of the sentences provided would seem to be correct (Document 1 is talking about Lincoln and his assassination, while Document 2 is talking about the security provided to presidents). Our implementation of skip-bigrams ensures that documents that contain the search phrase directly (or near to it) are given priority over documents which happen to have the words in the search phrase but do not contain them in a self-contained sentence.

As far as the assigned phrases went, our program did a *very* good job at identifying the correct documents. There were only a few which it stumbled on, such as "died in office", and I think the reason for this is that there are simply not many occurrences of the words 'died in office' on the relevant presidents' wikipedia pages. On the whole, though, the expected outputs are pretty much always included within the top ten. On the next page we have listed the output for each query.

Search Phrase	Expected Output	Actual Documents Found*
lincoln	Lincoln	Lincoln, Johnson, Buchanan
taft	Taft	Taft, Teddy Roosevelt, Harding
nobel prize	Teddy Roosevelt, Wilson, Carter, Obama	Carter, Obama, Teddy Roosevelt, Wilson
patent	Lincoln	Cleveland, Lincoln, Taft, Washington
Oxford scholar	Clinton	Clinton, Teddy Roosevelt, Madison
war time president	Jefferson, Madison, Monroe, Jackson, Polk, Pierce, Lincoln, Grant, Cleveland, Harrison, McKinley, Wilson, Franklin Roosevelt, Truman, Eisenhower, Lyndon Johnson, Nixon, George H. W. Bush, Clinton, George W. Bush, Obama	Taylor, Polk, Harrison, Fillmore, Taft, Buchanan, Monroe, Johnson, Tyler, Truman
johnson	Andrew Johnson, Lyndon B. Johnson	Andrew Johnson, Lyndon B. Johnson
bush	George H. W. Bush, George W. Bush	George H. W. Bush, George W. Bush
adams	John Adams, John Quincy Adams	John Quincy Adams, John Adams
harrison	William Henry Harrison	William Henry Harrison, Benjamin Harrison
vice President	John Adams, Thomas Jefferson, Martin van Buren, John Tyler, Millard Fillmore, Andrew Johnson, Chester Arthur, Theodore Roosevelt, Calvin Coolidge, Harry Truman, Richard Nixon, Lyndon Johnson, Gerald Ford, George H. W. Bush	Fillmore, Harrison, Tyler, Coolidge, Johnson, Buchanan, Van Buren
died in office	William Henry Harrison, Zachary Taylor, Warren Harding, Franklin Roosevelt	Taylor, Harrison, Fillmore, Polk, Benjamin Harrison, Monroe, Arthur, Hayes, Buchanan, Pierce
assassinated	Abraham Lincoln, James Garfield, William McKinley, John Kennedy	Taylor, Garfield, Kennedy, Lincoln, McKinley
abraham lincoln born	Lincoln	Lincoln
issued emancipation proclamation	Lincoln	Lincoln
author declaration independence	Jefferson	Jefferson, Adams
Founding father	John Adams, Thomas Jefferson, James Madison, George Washington	Madison, Monroe, Adams, Arthur, Washington

* Sorted by score from highest to lowest, so Lincoln, Johnson, Buchanan found Lincoln as the best match, Johnson as the next best, and Buchanan as the third best.

You can see the original output of the above search phrases (including scores) in the "output.txt" file included with our submission.