Dan Cohen's Digital Humanities Blog » Blog Archive » Wikipedia Vs. Encyclopaedia Britannica Keyword Shootout Results

In my post "Wikipedia vs. Encyclopaedia Britannica for Digital Research"[1], I asked you to compare two lists of significant keywords and phrases, derived from matching articles on George H. W. Bush in Wikipedia and the Encyclopaedia Britannica. Which one is a better keyword profile—a data mining list that could be used to find other documents on the first President Bush in a sea of documents—and which list do you think was derived from Wikipedia? The people have spoken and it's time to open the envelope.

Incredibly, as of this writing everyone who has voted has chosen list #2 as being the better of the two, with 79% of the voters believing that this list was extracted from Wikipedia. Well, the majority is half right.

First, a couple of caveats. For some reason Yahoo's Term Extraction service returned more terms for the second article than the first (I'm not sure why, but my experience has been that the service is fickle in this way). In addition, the second article is much shorter than the first, and Yahoo has a maximum character length for documents it will process. I suspect that the first article was truncated on its way to Yahoo's server. Regardless, I agree that the second list is better (though it may have been helped by these factors).

But it may surprise some that list #2 comes from the Encyclopaedia Britannica rather than Wikipedia. There are clearly a lot of Wikipedia true believers out there (including, at times, myself). Despite its flaws, however, I still think Wikipedia will probably do just as well for keyword profiling of documents as the Encyclopaedia Britannica. And qualitative considerations are essentially moot since the Encyclopaedia Britannica has rendered itself useless anyway for data-mining purposes by gating its

content.

This entry was posted on Monday, February 6th, 2006 at 2:18 pm and is filed under APIs[2], Text Mining[3], Wikis[4], Yahoo[5]. You can follow any responses to this entry through the RSS 2.0[6] feed. You can leave a response[7], or trackback[8] from your own site.

References

1. ^ "Wikipedia vs. Encyclopaedia Britannica for Digital Research" (www.dancohen.org)
2. ^ View all posts in APIs (www.dancohen.org)
3. ^ View all posts in Text Mining (www.dancohen.org)
4. ^ View all posts in Wikis (www.dancohen.org)
5. ^ View all posts in Yahoo (www.dancohen.org)
6. ^ RSS 2.0 (www.dancohen.org)
7. ^ leave a response (www.dancohen.org)
8. ^ trackback (www.dancohen.org)