Software Citation project

-----------------------------------

3/20/12

Method:

Journals were chosen from those represented in the 2010 version of the ISI Web of Science database. Using the “Journal Citation Reports” tool (<http://admin-apps.webofknowledge.com/JCR/JCR?SID=3C8h5B7Cn3jJmF5f9Hp>), the journals were collected by choosing the following 18 biology-related subject categories and sorting by impact factor:

BIOCHEMISTRY & MOLECULAR BIOLOGY; BIOLOGY; BIOTECHNOLOGY & APPLIED MICROBIOLOGY; CELL BIOLOGY; DEVELOPMENTAL BIOLOGY; ENTOMOLOGY; EVOLUTIONARY BIOLOGY; GENETICS & HEREDITY; MARINE & FRESHWATER BIOLOGY; MATHEMATICAL & COMPUTATIONAL BIOLOGY; MICROBIOLOGY; MULTIDISCIPLINARY SCIENCES; MYCOLOGY; ORNITHOLOGY; PARASITOLOGY; PLANT SCIENCES; REPRODUCTIVE BIOLOGY; ZOOLOGY

From that list, created three different lists: top 10, next 100 (11-100), and the rest (101-1455).

Sampled 90 issues from each of those 3 strata randomly using journalSampler.scala:

* journalSampler randomly shuffles each list, then takes the first 90 and creates a new list
* I then manually go through that new list of 90 and take the first 30 issues listed that appear to be from journals that publish original research, as opposed to review journals, since software will most likely not be mentioned at all in review articles
  + Review journals generally could be identified by having “review,” “trends in...” or “current opinion in...” in the title
  + Also, “Progress in Lipid Research” and “Human Reproduction Update” are reviews
  + “Briefings in Bioinformatics”
* At the end of this process I have three lists of 30 papers from each of the different strata (90 papers total)

Then manually retrieved each issue from the journal website according to the year, week number, and journal name. (Manually retrieved the dates for that year’s week number from this website: <http://www.epochconverter.com/date-and-time/weeknumbers-by-year.php?year=2006>)

* When an issue was dated during or after the chosen week, we chose the issue that came out prior to that week.
* In cases where that issue did not have any research articles (came across this for 2003-44-SCIENCE -- only reports and news), I went to the issue immediately following and pulled a research article from that issue.
  + 2007-11-GENOME\_RES -- no research papers in the chosen issue, nor in the following issue, so pulled an article from the issue after that (May 2007)
* Didn’t pull these issues because I cannot access these journals (UT does not appear to have a subscription)
  + 2002,20,ANTIOXID REDOX SIGN - strata 2
  + 2009,36,ANTIOXID REDOX SIGN - strata 2
  + 2005,44,AUST J BOT - strata 3
* Did not pull this issue for another reason:
  + 2005,27,J PLANT REGIST - strata 3, left this out because it looked like the articles are all “registrations” of different varieties of plants, not actual research articles.
* Couldn’t pull the issue because it was earlier than the earliest issue listed online. Instead, pulled the earliest issue listed online **[UPDATE -- Because this would bias the sample toward the first issue of the journal, instead deleted these 12 first issues and took the next 12 issues off of their respectives lists of 90 randomly chosen issues]**:
  + ~~2000,47,PLOS PATHOG - strata 2 - retrieved Sept. 2005 instead~~
  + ~~2002,50,CELL METAB - strata 2 - instead Sept. 2005~~
  + ~~2004,2,MOL SYST BIOL - strata 2 - instead Mar. 2005~~
  + ~~2008,19,J MOL CELL BIOL - strata 2 - instead Oct. 2009~~
  + ~~2003,39,CANDOLLEA - strata 3 - instead 2004~~
  + ~~2001,42,MATH MODEL NAT PHENO - strata 3 - instead 2006 Vol 1~~
  + ~~2004,39,IRAN J PARASITOL - strata 3 - instead 2006 Vol 1~~
  + ~~2003,35,RNA BIOL - strata 3 - instead 2004 Vol 1~~
  + ~~2001,38,EXCLI J - strata 3 - instead 2002~~
  + ~~2002,30,IRAN J ARTHROPOD-BOR - strata 3 - instead 2007 Vol 1~~
  + ~~2005,15,GENES NUTR - strata 3 - instead Mar. 2006~~
  + Also, 2001-26-CELL\_STEM\_CELL - strata 2 - deleted and went to next issue on list of random 90
* Pulled these, but some notes of interest:
  + 2006,27,FASEB J - There are no “Research Articles” in FASEB, only: [Research Communications](http://www.fasebj.org/site/misc/rc.xhtml), [Hypotheses](http://www.fasebj.org/site/misc/hyp.xhtml), [Reviews](http://www.fasebj.org/site/misc/rev.xhtml), and [Life Sciences Forums](http://www.fasebj.org/site/misc/life.xhtml). Since “Research Communications” seemed to be Research Article-like, I used a Research Communication paper.
  + 2007,48,UNDERSEA HYPERBAR M - the site that archives this journal -- <http://archive.rubicon-foundation.org/xmlui/> -- appears to only have a search function that pulls all the articles regardless of journal issue onto a results page. I ended up pulling the articles from the journal via Web of Science by sorting by date (in this case, all the articles from Nov/Dec 2007). It is unclear if the sort shown on the Web of Science results page reflects the order of the articles in the journal, but probably unlikely. Still, it probably shouldn’t matter as the article itself was randomly chosen.

After all the pdfs (and the supplementary info in pdf, word, xls, and eps) had been pulled, I uploaded them to Dedoose (for coding) and Mendeley (to retrieve the references and also to look at the docs?). Since the pdfs were not showing up correctly on Dedoose, I exported the text to plain text files and uploaded those.

* 2000-26-NUCLEIC\_ACIDS\_RES - The pdf for this article was secured by Oxford Journals, so it wouldn’t allow for export. I instead copied the free full text from the website, pasted it into a txt file, and saved it in UTF-8 format for upload into Dedoose.
* 2009-51-MOL\_THER - pdf file was “bad” according to Acrobat, so copied text from website and pasted in txt file. Pdf is still readable by humans.
* The supplementary info often were tables or figures that did not translate well on Dedoose. When a pdf did not show up, I initially tried exporting it in a .doc or a .txt or other text file that would make it more readable. Ultimately, though, most were pretty illegible once uploaded to Dedoose anyway, so most of the time if the supplementary info was in a word doc or a pdf, I kept them in that format. If a document was in an image format (.tif, .gif, .eps, which I think is an image format), I would print to pdf so that it could be uploaded to Dedoose (Dedoose does not appear to have the capability to upload image files). Excel spreadsheets were converted to txt files or printed to pdf.
* 2009-31-NATURE-supp1 - could not create a txt file from this pdf or convert it in any other way. Dedoose would not let me upload the pdf as is (it would freeze), and downloading a new pdf from the Nature site did not fix the problem.

------------------------------------

3/9/12

Issues:

* Does this list of included subject categories seem reasonable? Should I include or exclude any of the categories? **[Include the categories identified by James, marked below:**



* Are the strata (top 10, middle 20, bottom 50) reasonable? **[top 10, the next 100, and the rest - no gaps in sample: justification is that we are stratifying by importance and we are intentionally oversampling journals by importance]**
* Occasionally my script spits out 31 instead of 30 papers. Need to find my bug. **[Use scala.util.random “shuffle” method to shuffle and then take first 30; also, no need for script to determine the article that is being sampled -- once I find the journal issue, I will go to it manually, then use a Random number generator (bounded by the number of articles that appear in that issue) to pick the article.]**
* CAT appears to require that individual txt files are coded at the document level unless they are predelimited with a ==--endcodeableunit--== delimiter if there are multiple txt files or a hard return if there is only one txt file (<http://cat.ucsur.pitt.edu/CAT_quickstart_final.pdf).> **[We are using Dedooce now instead.]**

Method:

Created a list of biology journals from the 2010 version of the ISI Web of Science database (via the “Journal Citation Reports” tool: <http://admin-apps.webofknowledge.com/JCR/JCR?SID=3C8h5B7Cn3jJmF5f9Hp>). The journals were collected by choosing the following subject categories and sorting by impact factor:

* Biochemistry and Molecular Biology
* Biology
* Biotechnology and Applied Microbiology
* Cell Biology
* Developmental Biology
* Evolutionary Biology
* Marine and Freshwater Biology
* Mathematics and Computational Biology
* Microbiology
* Multidisciplinary Sciences (includes Nature, Science, and PNAS)
* Reproductive Biology

I considered but did not include the following subject categories that seem somewhat related to biology:

* Anatomy and Morphology
* Andrology (male menopause, infertility)
* Biochemical Research Methods
* Biodiversity Conservation
* Biophysics
* Cardiac and Cardiovascular Systems
* Cell and Tissue Engineering
* Ecology
* Endocrinology and Metabolism
* Engineering, Biomedical
* Entomology **[INCLUDE]**
* Gastroenterology and Hepatology
* Genetics and Heredity **[INCLUDE]**
* Hematology
* Immunology
* Infectious Diseases
* Materials Science, Biomaterials
* Mycology (study of fungi) **[INCLUDE AS OF 3/14 EMAIL]**
* Neurosciences
* Ornithology **[INCLUDE AS OF 3/14 EMAIL]**
* Parasitology **[INCLUDE AS OF 3/14 EMAIL]**
* Physiology
* Plant Sciences **[INCLUDE]**
* Virology
* Zoology **[INCLUDE]**

Once the categories had been chosen a list of 947 journals was generated, I downloaded a txt file with the names and stats of the journals ranked 1-10 (according to impact factor), a file with journals ranked 464-483 (representing the middle 20 of the list), and a file with journals ranked 896-947 (representing the bottom 50 of the list -- the numbering of this last file looks slightly off because the list was numbered according to rank, which could be shared across multiple journals).

I opened each of those files in Excel, deleted all columns except for the first column showing the journal name, and saved those new files as “JournalNames\_1-10”, “JournalNames\_mid20\_464-483” and “JournalNames\_bottom50\_896-947.”

**[CG NOTE AS OF 6/28/13: During our 3/9/12 meeting, and as shown in 3/9/12 Issues list above, we redefined the strata from “top ten”, “mid 20,” and “bottom 50” to “top ten,” “next 100,” and “the rest.” As a result, the three files used for sampling changed to JournalNames\_1-10, JournalNames\_11-110, and JournalNames\_111-1455.]**

I then wrote a script called journalSampler.scala that took a “JournalNames” file as input and output a list of 30 articles randomly sampled from the file.

Update:

One article chosen from the initial list (SCI CHINA SER A 2003, 49) was actually a logic/mathematics article and was removed from the set. To replace it, we ran a new query for a random set of articles from the 111-1455 strata and chose the first article, 2004-25-J\_THERM\_BIOL to replace it.