Haoming Li, 20426226
Kristina Wong, 76513468
Shengjie Xu, 10616769
Yirui Jiang, 64137163

Report

(used the provided HTML file collection)

1 – mondego :

sdcl.ics.uci.edumondego-group
mondego.ics.uci.edu
www.ics.uci.edu~djp3classes20060330ICS105ResourcesAnteaterIdol.html
www.ics.uci.edu~lopes
mondego.ics.uci.edudatasetsC=M3BO3DA

2- machine learning :

www.ics.uci.edu~pazzaniPublicationsOldPublications.html
www.ics.uci.edu~pazzaniPublicationsAPubs.html
www.ics.uci.edu~qliu1MLcrowdICMLworkshop
www.ics.uci.edu~qliu1MLcrowdICMLworkshopindex.html
www.ics.uci.edu~kibler

3- software engineering :

www.ics.uci.edu~taylorPublications.htm
www.ics.uci.edu~wscacchipublications.html
www.ics.uci.edu~andrepublications.html
www.ics.uci.edu~wscacchi
www.ics.uci.edu~redmilespublications.html

4 - security :

drzaius.ics.uci.edu~swirlimpromptu-0.30apidocsindex-all.html
www.ics.uci.edu~gtspubs.html
drzaius.ics.uci.edu~swirlimpromptu-0.20apidocsindex-all.html
www.ics.uci.educommunitynewsnotes
www.ics.uci.educommunitynewsnotesindex.php

5 - student affairs :

kdd.ics.uci.edudatabasesmoviesdatacasts.html
www.ics.uci.eduugradqa

www.ics.uci.eduugradqaindex.php
www.ics.uci.edugradpolicies
www.ics.uci.edugradpoliciesindex.php

6 - graduate courses :

www.ics.uci.eduugradqa
www.ics.uci.eduugradqaindex.php
www.ics.uci.edugradpolicies
www.ics.uci.edugradpoliciesindex.php
www.ics.uci.eduugraddegreessecondbaccs.php

7- informatics :

drzaius.ics.uci.edumetaclassesinformatics161fall06papers01kling.html
www.ics.uci.edufaculty
www.ics.uci.edufacultyindex.php
www.ics.uci.edufacultyindex.phpdepartment=Informatics
www.ics.uci.edu~kaycoursesprevious.html

8 - REST :

www.ics.uci.edu~fieldingpubsdissertationrestarchstyle.htm
www.ics.uci.edu~fieldingpubsdissertationevaluation.htm
www.ics.uci.edu~fieldingpubsdissertationconclusions.htm
www.ics.uci.edu~fieldingpubsdissertationtop.htm
www.ics.uci.edu~fieldingpubsdissertationintroduction.htm

9 - computer games :

www.ics.uci.edu~danclass267datasetscalgarynews
www.ics.uci.educommunitynewsnotes
www.ics.uci.educommunitynewsnotesindex.php
www.ics.uci.edu~wscacchi
www.ics.uci.educommunitynewsarticlesviewarticleid=77

10 - information retrieval :

www.ics.uci.edu~gbowkerconverge.html
www.ics.uci.edu~kobsaprivacyGerman.htm
www.ics.uci.edu~sharadpubs.html
www-db.ics.uci.edupagesresearchmars
www-db.ics.uci.edupagesresearchmarsindex.shtml

Some explanations about indexing and searching:

Indexing:

I used tf-idf for indexing. For tf, I used to normalize it by the total number of words in each document (tf / number-of-words-in-the-document). However, I found out that by doing this, some very small documents/pages will always have very high ranks (since these documents only have a very small amount of words). Therefore, I tried and used log normalization for tf for the project (1 + log(tf)), and found out that log normalization is better. I also used a kind of wtf, more precisely, the words in the title are weighted twice as much as the words in the body. Moreover, I treat the words with all capitalized letters as special terms (for instance, 'REST' and 'rest' are two terms, but ('Rest'/'rEst' /…) and 'rest' is one term).

Searching:

If the user just wants to search a single term (for instance, 'information'), the rank of documents will only depends on the documents' tf-idf values for that term. However, if the user wants to search multiple terms (for instance, 'information retrieval', 'computer game science'), the rank of a document will not only depend on the sum of the document's tf-idf values for each term.

For example:

	Doc1	Doc2
information	0.1	0.3
retrieval	0.8	0.4

The score of Doc1 for 'information retrieval' will not be 0.1 + 0.8 = 0.9, instead, it will be ((0.1+0.8)/2)/2 + 0.1 = 0.325 (average-tf-idfs + the-lowest-tf-idf). This is because if the users type in 'information retrieval', they want to search the 'information retrieval' as a whole, not by separate terms. For Doc1 above, the document contains far more 'information' than 'retrieval' (for instance, the document is actually talking about how to hide information, and only mentioned retrieval once), which means the document may be less relevant to ''information retrieval'. However, for Doc2, although 0.3+0.4 < 0.1+0.8, Doc2 is likely to be more relevant to 'information retrieval' than Doc1. By using score = (average-tf-idfs + the-lowest-tf-idf) , the score of Doc2 will be ((0.3+0.4)/2)/2 + 0.3 = 0.475 > Doc1 = 0.325. Although it is only a very rough way to score a document, at least it makes more sense.