UNIVERSITY OF NAIROBI

COLLEGE OF BIOLOGICAL AND PHYSICAL SCIENCES

SCHOOL OF COMPUTING AND INFOMATICS

DOCUMENT RETRIEVAL SYSTEM

By

JOYCE MUTUA
REG. NO (P/56/P/7866/2002)

SUPERVISOR: LAWRENCE MUCHEMI

October 2012

Submitted in partial fulfilment of the requirements of the Master of Science in Information Systems
DECLARATION

This project is my original work and has not been submitted in support of a degree or any other award in any university.

Signed ______________ Date ______________

Joyce Mutua

This project has been submitted with my approval as university supervisor.

Signed ______________ Date ______________

Lawrence Muchemi

School of Computing and Informatics,

University of Nairobi
ABSTRACT

To effectively utilise repositories of data, document retrieval systems act as a means of performing the task of sifting through these repositories to extract documents that meet an individual’s information need.

The projects’ document corpus was drawn from 71 Master of Science in Information Systems and Computer Science project abstracts done at the University of Nairobi, School of Computing and Informatics between the years 2006 and 2010. The project utilised the vector space model as its basis for document matching and ranking. Based on the gold standard of relevance, these documents were put in document categories that reflected their content. This provided the basis against which recall and precision measures of the system accuracy was measured.

The system achieved an average precision score of 0.781667 and recall score of 0.833333. Recall scores were mainly affected by the presence of homonyms and homographs while precision scores were affected by synonyms. The study also showed that the presence or absence of a term in a document is what influences the retrieval and ranking of relevant documents in the vector space model, not the size of the document corpus.

Keywords: Vector Space Model, Tfidf score, Document retrieval system, Recall, Precision.