Table of contents

List of Figures viii
Abbreviation ix
Abstract x

Chapter 1
Introduction 1
1.1 Understanding Natural Language Processing 1
1.2 Information Retrieval 2
1.2.1 Tokenization and Sentence Segmentation
1.2.2 POS tagging 3
1.2.3 Entity Detection using POS Tags 4
1.2.4 Query Processing 5

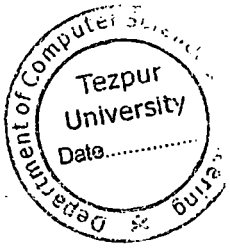
Chapter 2
Getting Deeper into Information Retrieval 7
2.1 Major challenges 7
2.2 Types of IR 7
2.2.1 Boolean Retrieval 7
2.2.2 Ranked Retrieval 8
2.3 Tools and Concepts 9
2.3.1 Vector-Space Model 9
2.3.2 Precision and Recall 9
2.3.3 Stop Words 10
2.3.4 Stemming 10
2.3.5 Compound and Statistical Parsing 10
2.3.6 Compound Splitting 11
2.3.7 Chunking and Shallow Parsing 11
2.3.8 Co-reference Resolution 12

Chapter 3
Implementation 13
3.1 Steps involved 13
3.1.1 Crawling 13
3.1.1 Construction of corpus 13
3.1.3 POS Tagging for query 14
3.1.4 Retrieving relevant text 14
3.1.5 Synonym handling and Morphological Analysis 15
3.1.6 Indexing 16
3.1.7 Displaying relevant results 17
3.2 Architecture 17
3.3 Snapshot of Result 18

Chapter 4
Conclusion and Future prospects 21

Table of contents

References 22
Appendix
A. Installing NLTK, Python and PHP 23
B. Configuring NLTK with required Corpus 23



List of figures

1.1 Architecture of Information Retrieval.....3
1.2 Showing Different types of Tagging.....5
1.3 Effect of query length.....6
2.1Using Set Theory to show Recall.....9
2.2Using Set Theory to show Precision.....10
2.3Chunking.....11
3.1 A stop list of 25 semantically non-selective words which are common in Reuters-RCV1.....14
3.2 Index-time vs Query-time Synonym Handling.....16
3.3 General Architecture of iFAQ.....17
3.4 Snapshots of information retrieval from corpus i).....18
3.5 Snapshots of information retrieval from corpus ii).....19
3.6 Snapshots of information retrieval from HTML Table.....20

Abbreviation

BRILL	Brill tagger
CC	Coordinating Junction
CD	Cardinal Number
CoNLL	Conference on Natural Language Learning
DT	Determinant
ENGTOWL	ENGTOWL tagger
HMM	Hidden Markov Model
IN	in
IR	Information Retrieval
JJ	Adjective
NLP	Natural Language and Processing
NLTK	Natural Language Tool Kit
NN	Noun singular
NNS	Noun plural
NNP	Pronoun singular
PHP	Hypertext
POS	Parts-of-Speech
VBD	Verb, past tense
VBN	Verb, past participle