

---

# Concept Mapping

R&D Report

By

**Ankith M S** and **Sandip Ghoshal**,

Under the guidance of **Prof. Ganesh Ramakrishnan**

*Department of Computer Science*

*Indian Institute of Technology, Bombay*

---



# Contents

## 1. Introduction

## 2. Training Data generation

- 2.1. Pdf to Text conversion
  - 2.1.1. Tools used
  - 2.1.2. Problem with book corpus
- 2.2. Sentencification
- 2.3. Finding entity mentions
- 2.4. Preprocessing of Knowledge Base
  - 2.4.1. Lemmatizing
  - 2.4.2. Clustering of relations
    - 2.4.2.1. Manual
    - 2.4.2.2. Word2Vec
    - 2.4.2.3. WordNet

## 3. Multi-R Training

- 3.1. Preprocessing
  - 3.1.1. Generate KB
  - 3.1.2. Generate Match File
- 3.2. Named entity linking
  - 3.2.1. YAGO
- 3.3. Exploring features
  - 3.3.1. Dependency parse tree
  - 3.3.2. Word Embeddings
    - 3.3.2.1. Word2Vec

## 4. Multi-r Testing

- 4.1. Test Data generation
  - 4.1.1. Manual
  - 4.1.2. Google Scraper
- 4.2. Results Evaluation
  - 4.2.1. Manual
  - 4.2.2. Evaluator UI

## 5. Information Retrieval

- 5.1. TAAT Algorithms
- 5.2. DAAT Algorithms
- 5.3. Solr-lucene API

**6. Conclusion**

**7. Acknowledgments**

**8. References**