

Hierarchical Topic Detection from Text Data

Final Year Project Plan

31st August 2017

Chen Ziao

Supervised By –

Prof. Ke Yi Ping, Kelly

School of Computer Science and Engineering

TABLE OF CONTENTS

[1 Introduction 2](#_Toc491902384)

[1.1 Project Objectives 2](#_Toc491902385)

[1.2 Scope 2](#_Toc491902386)

[2 PRoject breakdown 2](#_Toc491902387)

[2.1 Task to complete 2](#_Toc491902388)

[2.2 Project Schedule 4](#_Toc491902389)

[3 Approach 4](#_Toc491902390)

[4 Risk Management 4](#_Toc491902391)

# Introduction

## Project Objectives

This project aims to study the concept of hierarchical topic detection from text data. An effective algorithm needs to be designed and implemented after comparing different approaches. Web Crawler and website API would be used to crawl large scale of data from the Internet for model training and testing purpose. Natural Language Processing would be used to preprocess the data into model readable format. Training data and validation data would be used in Machine Learning to train and validate the model. Testing data is used to evaluate the models’ performance when comparing different models generated by different algorithms.

## Scope

Through the project, 5 major components need to be achieved.

1. Research on online papers on texting mining, especially text topic detection to understand available methodologies.
2. Collect and preprocess large-scale text data from Internet. Preprocess and store them in appropriate format.
3. Design and implement effective algorithms to detect topics in a hierarchical manner from the text data.
4. Compare and contrast performance of different algorithms. Discuss the benefits and limitations of each method.
5. Visualize the final result and deliver it in the form of final report and presentation

# PRoject breakdown

## Task to complete

|  |  |  |  |
| --- | --- | --- | --- |
| **Task** | **Description** | **Efforts Estimate** | **Dependencies** |
| Project Plan | Complete project plan and understand its objective | 3 weeks | Not Applicable |
| Literature Review | Read published papers related to text mining and topic detection. Learn existing methods and technologies | 6 weeks | Project Plan |
| Data Collection | Collect large scale data from online website for next stage model implementation | 3 weeks | Project Plan |
| Algorithms Implementation | Choose from available algorithms or design own algorithm and implement it using software tools | 6 weeks | Literature Review & Data Collection |
| Algorithms Comparison (Result and Discussion) | Based on results of different algorithms (models), discuss respective pros and cons, and select the best one as final algorithm | 2 weeks | Algorithms Implementation |
| Result Visualization | Visualize the experiment result using visualization tool such as matlibplot in Python and Excel | 1 week | Algorithms Comparison |
| Final Report | Aggregate all the steps completed through the project and compile them into a single report | 2 weeks | Result Visualization |
| Final Presentation | Prepare PowerPoint slides and relevant materials to do a presentation to professors and judges to demonstrate the project achievements | 1 week | Result Visualization |

## Project Schedule





# Approach

This project would use agile methodology. All the steps including literature review, data collection, experimental study and implementation would be broken into small increments which are iterative. Each iteration would last for about 2 weeks and the task in certain iteration would be planned beforehand. After iteration completes, a summary of achievement and next iteration task would be sent to Prof. Ke Yiping Kelly to seek for suggestion so that feedback can be obtained frequently throughout the project and new changes can be adapted quickly.

The main programming language used for algorithm implementation and data collection would be Python. Natural Language Processing and Machine Learning techniques would be utilized for data preprocessing and data modeling. Last but not least, Latex would be used to write the final report.

# Risk Management

| **Risk Description** | **Mitigation/Contingency Plan** | **Criticality (Low/Medium/High)** |
| --- | --- | --- |
| Network connection fails or connection is blocked when crawling data from online resources | Use dynamic IP when crawling or use API provided by the website | Low |
| Crawled data loss | Perform frequent data backup | Medium |
| Scheduled task cannot be completed within the planned timeframe | Plan buffer time for each task. Communicate with Prof. Ke and modify the plan accordingly. Lit | Medium |