COMP90024 Cluster and Cloud Computing Project

* Group Members
* Jiyu Chen 908066

# Abstract

Twitter is a popular social media platform containing large amount of texture data. Aurin provide series of datasets developed and contributed by Australia’s leading researchers. In this project, we will leverage the NECTAR facility to create a four instances cluster environment and mining interesting geoinformation by summarizing tweets from eight cities around Australia and combine them with open sourced Aurin data. We will discuss the system structure, cluster design, tweet crawler, tweet sentiment analyser, sets of Aurin data we have leveraged, views of our data and a demo of the system user interface.

Content

[Introduction 3](#_Toc513331525)

[System Design 4](#_Toc513331526)

[Cluster architecture 4](#_Toc513331527)

[System architecture 4](#_Toc513331528)

[(I guess we have a level structured execution system? ) 4](#_Toc513331529)

[Data Processor 4](#_Toc513331530)

[Hybrid crawler for tweets 4](#_Toc513331531)

[Embedded Sentiment Analyser 4](#_Toc513331532)

[Topic parsing 4](#_Toc513331533)

[Aurin collector and parser 4](#_Toc513331534)

[Format of processed data 4](#_Toc513331535)

[CouchDB as Database 4](#_Toc513331536)

[CouchDB in cluster 4](#_Toc513331537)

[Duplication prevention 4](#_Toc513331538)

[Scenario Study 4](#_Toc513331539)

[System UI and web Implementation 4](#_Toc513331540)

[Reference 5](#_Toc513331541)

# Introduction

Nectar stands for National eResearch Collaboration Tools and Resources project. It gives us fix number of computational resources and allowing researchers to create a cluster in their needs with high flexibility for system architecture design and management. We create an one-master-three-slave structure cluster and making use of CouchDB to control passing messages between nodes including message storing, duplication prevention, resource backup, location transparency, synchronizing. We also implement error handling mechanism and parallel computing to enhance the fault tolerance ability of our system. We have also tested the scalability of the system with different number of instances and result in good performances. Twitter data and Aurin is used in our study. For tweets, the twitter api allows us to search for tweets in the past 7 days. The retrieved tweets are in twitter json format containing the user information, tweets, timestamp, geo tag and so on. We will discuss how we utilize the information in later section. One problem for twitter search api is although it’s fast in retrieving many numbers of data in a short period, a access time limit will force it to stop. The stream api instead can solve the limitation well but relatively slower than search api. Therefore, we implemented a hybrid crawler leveraging both search and stream api. <add intro of aurin n aurin harvester>. In our tweet crawler, we designed a embedded machine learning sentiment analyser for classifying whether the tweet is sentimentally positive or negative. We also designed a baseline for helping us getting a preview of our task. We tested our sentiment analyser on nltk twitter sample [add citation] which result in 98.41% average f1-score and on sent140 corpus [add citation] which result in 66.7% average f1-score. <topic> <hashtag> <map/reduce> <web> <boto>

# System Design

## Cluster architecture

<add something>

## System architecture

## (I guess we have a level structured execution system? )

<add something>

# Data Processor

## Hybrid crawler for tweets

<add something>

## Embedded Sentiment Analyser

<add something>

## Topic parsing

<add something>

## Aurin collector and parser

<add something>

## Format of processed data

<add something>

# CouchDB as Database

## CouchDB in cluster

<add something related to sharding, structure, etc..>

## Duplication prevention

# Scenario Study

<add something>

# System UI and Web Implementation

<add something>

# Reference

[1]

[2]