An Infrastructure to Store and Analyse Seismic Data as Suffix Trees

MSc Data Analytics - Project Proposal

Tom Taylor

April 2016

Contents

1 Abstract

The purpose of this project is to develop an infrastructure and tool set for converting raw seismic time series data in to a searchable string using SAX (**S**ymbolic **A**ggregate appro**X**imation) and then to store this data as a suffix tree for fast searching and analysis. An interface will then be developed to enable the searching of these suffix trees and provide visualisation of the data. The code should be re-usable in a project to receive live streaming data from many stations.

2 Background

Seismic waves are recorded as movement over three axis: vertical (hereafter referred to as **z**) alongside horizontal in terms of north-south(**n**) and east-west(**e**).

SAX (**S**ymbolic **A**ggregate appro**X**imation) (Lin et al., 2007)

Some words about Suffix Trees and the related infrastructure at Birkbeck. (Chino et al., 2011)

3 Objectives

1. Interpretation of raw data from seismic stations and storage as time-series data in a time-series database such as OpenTSDB for easy access during conversion to SAX and for later rendering during interactions with the data.

2. Conversion of the data to SAX and storage in a suffix tree.

3. An interface for searching and viewing the raw data.

4 Approach

As laid out in the objectives, the development will produce three separate components.

4.1 Interpretation and Storage as time-series data

The raw data received from the stations is accessed as files in the SAC (Seismic Analysis Code) binary format. These will be read using the ObsPy python library which can read

both the headers and return the raw data as a Python object.

This data will be batch processed and imported in to a time series database. Initially OpenTSDB is being considered for performant reasons but there may be scope for this to change depending on how difficult it is to implement this on the Universities hardware. Other options could be InfluxDB or Graphite.

4.2 Conversion of data to SAX and building of the Suffix Tree

4.3 Development of an interface

The interface to query and view the data will be web based as to ensure maximum compatibility with clients. Many open source graph renderers are in existence such as Grafana and Cubism.js and components of both are likely to be used for viewing the raw seismic data alongside the SAX data.

5 Plan

As both ObsPy and the existing infrastructure for Suffix Tree storage are written in Python, this seems the logical choice. Specifically Python 3.5 will be used for all components.

Development will take the approach of Test Driven Development (TDD) where unit tests will be written at a class level before the classes are coded.

References

Chino, D. Y. T., F. A. da Louza, A. J. M. Traina, C. D. de Aguiar Ciferri, and C. T. Jnior
2011. Time series indexing taking advantage of the generalized suffix tree.

Lin, J., E. Keogh, L. Wei, and S. Lonardi
2007. Experiencing sax: a novel symbolic representation of time series.