Project Modules

RSS Aggregator and News Article Downloader

- Periodically download news articles from a list of RSS files.

STATUS: Not started

News Article Downloader – Algorithms Connector

- Performs keyword extraction, topic discovery and classification, and category classification on the downloaded articles.

STATUS: Not Started

Keyword Extraction

- Extract keywords from a news article.

STATUS: Completed

Category Classification

- Identify the categories a news article is in. The categories are Business, Politics, Sports, Entertainment, Health, Technology, and world regions as defined by the United Nations.

STATUS: Algorithm is completed, but need to gather more training data.

Topic Discovery and Classification

- Identify if a news article has the same topic as a previously seen article or if it contains a new topic.

STATUS: Completed

Named Entity Recognition

- Identify the people, places, times, and dates in a news article.

STATUS: Not Started

Web Interface

- Provide an easy to use and friendly interface for navigating the news articles and exploring the relations between categories, topics, and named entities.

STATUS: Not Started

Project Plan

1 RSS Aggregator and News Article Downloader

- This part is important to the project and should be completed soon.
- The algorithms are simple and should be completed easily given time.

RECOMMENDATION: Assign to Teng or Hisazumi
COMPLETION TIME: 1-2 months for coding and debugging

2 News Article Downloader – Algorithms Connector

- This is an extension to the RSS Aggregator and News Article Downloader program.

RECOMMENDATION: RSS Aggregator and News Article Downloader programmer and David
COMPLETION TIME: 1 month for coding and debugging

3 Named Entity Recognition

- Involves finding an implementation that can run on windows and creating and interface to call the program/algorithm from within C#.

RECOMMENDATION: David
COMPLETION TIME: 1 month for coding and debugging

4 Database Design and Implementation

- Involves designing the database that will be used to communicate between the web interface and the algorithms.

RECOMMENDATION: Teng, Hisazumi, and David
COMPLETION TIME: 1 week

5 Server Setup

- Selecting, purchasing, installing, and configuring server(s) for the project.
- I would recommend 1 server for the interface and 1 server for the algorithms and database.

RECOMMENDATION: Teng or Hisazumi
COMPLETION TIME: 1 week for installation and configuration

6 Web Interface

- Provide an easy to use and friendly interface for navigating the news articles and exploring the relations between categories, topics, and named entities.
- Should use UTF-8

RECOMMENDATION: Teng or Hisazumi
COMPLETION TIME: 2-3 months for design and implementation. (perhaps longer depending on the desires of the customer.)

7 Algorithm Refinement

- Testing and updating of the basic algorithms to improve accuracy and performance.

RECOMMENDATION: David
COMPLETION TIME: Until project ends

8 Expansion to Japanese

- Currently the implementations of the algorithms in C# work with English and Japanese.
- Chasen is used for Japanese morphological analysis.
- Expansion to Japanese would require new training data for category classification and a named entity recognizer for Japanese.

RECOMMENDATION: Hisazumi and David
COMPLETION TIME: 2-3 months for full testing and analysis

9 Expansion to Chinese

- The algorithms work on any language that has a morphological analyzer.
- For Chinese, a new plugin for the algorithms would be needed that implements Chinese morphological analysis.
- A Chinese named entity recognizer is also needed.

RECOMMENDATION: Teng and David
COMPLETION TIME: 5-6 months for full testing and analysis

**Completion time means the time to complete if that is the only part of the project being done. It could possibly take longer. I believe for all the parts listed within this document it would take about 1 year to complete and have a functioning web site.