Semester Report

CSE 485 Capstone

Team

Text Geolocators

Team Members

Adam McNabb

Christopher Silvia

Jack Workman

Jang Won

Samantha Juntiff

Weston Neal

Sponsor

Ryan Raub, Global Institute of Sustainability, ASU

Description:

A Natural Language Processing (NLP) algorithm that parses text documents, returns location words found, and generates heat maps for analysis.

2. Table of Contents

3. Executive Summary

In today's connected world, scientists and research institutions perform studies and experiments all over the globe. Their findings are reported at international conferences and in scientific journals. When researchers read and analyze these reports, it is often easy to become lost in the "what" and "why"; however, it is just as important to consider "the where".

Our Text-Geolocator service exists to expose the significance of "the where" in a non-intrusive manner to improve the existing research analysis system. In many areas of study, researchers try to find trends within data to better comprehend findings. Thanks to our Text-Geolocator service, researchers at the school of Sustainability at Arizona State University can now investigate possible trends within the produced location data from our output.

The way it works is simple: a user uploads a document to the web service, our NLP algorithm extracts and analyzes the document's text, and then returns the most probable locations that are mentioned in the document. From there, the user is free to use the data in another report or to compare and analyze for location trends.

We make use of an existing industry standard for representing location data called Geo-JSON. This, along with the (in-progress) heat map display, presents a familiar, intuitive user interface that promotes easy interpretation of the location data and avoids unnecessary confusion. We want our users to be able to retrieve their information quickly and painfree.

We foresee our service becoming a useful tool within the research industry. Almost every day, we hear how location data from smartphones can be used to predict the owner's behavior. So our team proposes a question: why can't the same type of science be applied to researchers and the subjects that they study?

4. Introduction

4.1 Purpose

The purpose of this report is to outline our teams technical accomplishments during the first capstone semester. It is a detailed, professional, carefully-prepared documentation of our project, its development, technical details, and preliminary results. First, we present an overall project description, including scope, user overview, and project plan. Next, we describe our development approach taken, our design, the tools we used, and our preliminary results. Finally, we conclude the semester report, summarize our successes and lessons learned, and provide a future outlook to semester two.

4.2 Project description

Our project sponsors are looking for the locations mentioned in research papers in order to give them meaningful context, which up until now has been a manual process. In this project, our team is tasked with creating a system to automatically determine the most probable locations that are mentioned in a research paper. Our project sponsors will use this geographic information as part of research project management and for many other applications.

The project is a Software as a Service project. Our team created an API that can be accessed by other systems. It is a simple web based interface for feeding in text and returning a JSON object representing the list of most probable locations that the text refers to.

In the first semester of this project, our team focused on creating a Natural Language Processing algorithm and API to determine words and word phrases that are locations. During the second semester, our team will create heat maps with our location data and improve the algorithm with probabilities and deeper discovery abilities.

5. Scope

5.1 Original definition

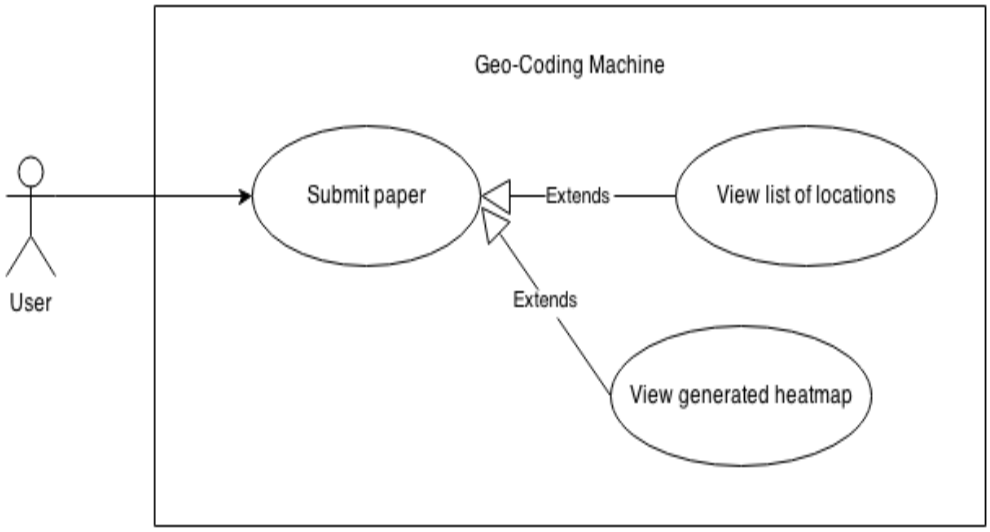
The initial scope of this project was to produce a Software as a Service project containing a Natural Language Processing Algorithm and API that takes formal research text documents in as input and produces a list of most probable locations mentioned in the papers. The algorithm was required to produce consistent, repeatable results, to have the ability to be run across different platforms.

5.2 Change of scope and reason for change

There has been no change in our scope.

6. User Overview

6.1 Use case diagram



6.2 Description of actors

The actors who will use this service are scientists, geologists, and any other group interested in knowing what locations have been referenced from a source document, such as a scientific article or any file containing text. These actors are people interested in discovering possible trends from location data, and who have a desire to visualize these trends by means of a heat map.

6.3 User Stories

Biologist in the field:
A biologist is researching why tree frogs are disappearing in a certain section of the rain forest. This biologist has the notion that global warming in this part of the rainforest is a factor. He has the hypothesis that the one degree increase of heat year round has caused the eggs of this type of tree frog to die from being too hot. He documents his findings in a report for other scientists to analyze. Another researcher wants to investigate his findings by generating heatmaps of the gps locations listed in the biologists report. The researcher uploads the scientists documents to geocoders on a monthly basis. The geocoders program analyzes the documents which then generates a heat map. The researcher does this every month for each document provided by the scientist, comparing all heat maps, and generating conclusions based on the locations that the tree frogs are disappearing.

Sustainability researcher:
A sustainability researcher travels the world collecting pollution data in 25 capital cities across the globe. He writes several papers based on his findings, which includes location-result data. He then compares his data with other research papers written on the same topic. The School of Sustainability at ASU desires to know more about pollution in the 25 cities. Researchers from the Sustainability school gather the research papers from the field and upload the papers to the Text-Geolocator Service. The service returns a list of the most probable locations mentioned in

all papers. It then generates a heat map of all the locations, allowing the researchers at ASU to have a better understanding of where pollution is most prevalent from the data collected.

7. Project plan

7.1 First semester

Planned:

Our team planned on developing our own Natural Language Processing (NLP) Algorithm and API that takes formal research text documents in as input and outputs a list of most probable locations mentioned in the papers in GeoJSON format.

Actual:

Instead of developing our own NLP algorithm, we integrated an existing, trained NLP module called the Stanford NER Tagger into our application. Once integrated, we used the Stanford NER Tagger to retrieve a list of locations from a text document submitted through the web service that we built. The end result is a series of GeoJSON objects (one for each location referenced in the document) returned to the user.

7.2 Second Semester

In the second semester, our team will focus on refining the GeoJSON points acquired from the web service. We will design rigorous test cases to ensure the accuracy of our application as well as add more data points to increase the usefulness of our application. Our team will also begin generating heatmaps of the generated locations.

8. Development Approach

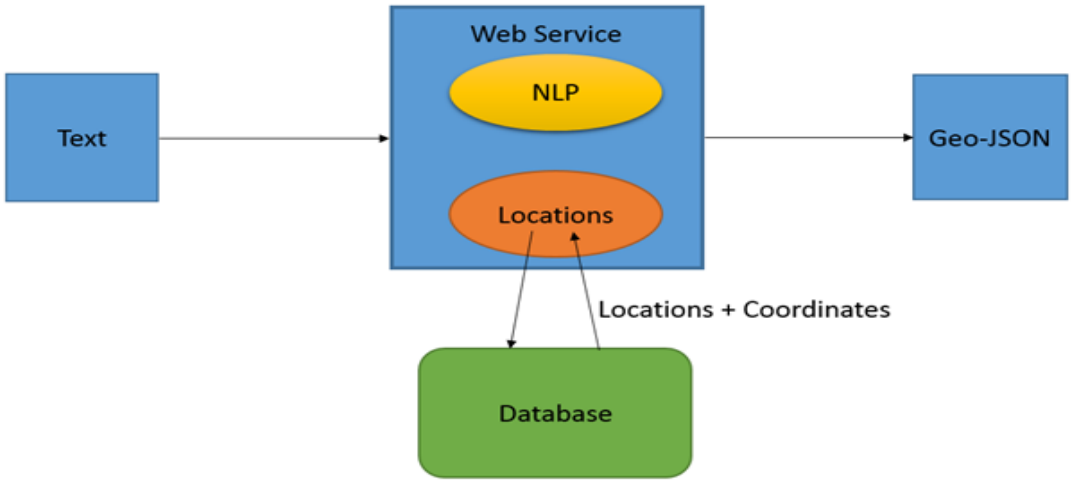
We used a divide-and-conquer approach for the development of our application. After meeting with our sponsors at the start of the semester, we identified four clear tasks:

1. Implementation of the Docker environment (including the database)
2. Setting up a web framework for the web service
3. Researching Geo-JSON
4. Researching location extraction via NLP

Because NLP was such a broad and unknown subject for all members of the team, the decision was made to postpone that task until the first three were taken care of and a universal development environment and version control system were set in place. This left the first three tasks needing to be completed quickly; to complete tasks 1-3, we split into three teams of two. As the semester continued, we retained the three teams and continued to assign new tasks to each pair. During our weekly meetings, each sub-team reported progress made and issues encountered. Our team found that working in smaller teams was an effective way to complete the project requirements.

9. Design Overview and Design Decisions

The following is a visual representation of the high-level design of our application:



Here are explanations of each element in the diagram:

- “Text” - the input of the application, a text document.
- “Web service” - the part of the application presented to the user.
- “NLP” - the natural language processing module used to process the user’s text document.
- “Locations” - the list of locations referenced within the text document and retrieved by the NLP module.
- “Database” - our database containing a massive amount of location data used for acquiring location coordinates.
- “Geo-JSON” - the output of the application, plain-text JSON adhering to the Geo-JSON industry specification.

When designing our application, the main decision was deciding between incorporating a third party NLP module into our application or designing our own. We understood that with our limited understanding of natural language processing applications, we could not create a completely new algorithm reliable enough to return locations of scholarly journals and papers accurately and efficiently. Therefore, a third party module was decided upon as it would be more

advanced and save development time. This saved-time allowed us to focus on the implementation of the algorithm according to the requirements of this project.

10. Technology and tools used

The technology and tools used throughout this project are listed below:

- Docker (https://www.docker.com/) - "An open platform for distributed applications for developers and sysadmins". We used this to easily share and transport our development environment, and to mimic the production environment.
- Fig (http://www.fig.sh/) - a wrapper for Docker used to simplify managing our docker instances.
- Virtualbox (https://www.virtualbox.org/) - A virtual machine managing the application used by our team to develop our project.
- Ubuntu (http://www.ubuntu.com/) - Our server operating system.
- Python (https://www.python.org/) - Our primary programming language.
- Flask (http://flask.pocoo.org/) - A microframework for Python web applications
- Flask-SqlAlchemy (https://pythonhosted.org/Flask-SQLAlchemy/index.html/) -The toolkit extension of flask that offers database access through Python SQL ORM
- Stanford Named Entity Recognizer (NER) Tagger (http://nlp.stanford.edu/software/CRF-NER.shtml) - Our third party NLP module used to identify locations referenced within the user's text document.
- PostgreSQL (http://www.postgresql.org/) - Our database solution.

11. Preliminary results

The results of our work is a basic NLP algorithm that retrieves a list of locations from a text document submitted through the web service that we built. The end result is a series of GeoJSON objects (one for each location referenced in the document) returned to the user. Our sponsors are pleased with our results thus far. Our team has completed all project deliverables required for this semester: a project timeline, development of a design, a plan of action, Natural Language Processing research, implementation of an initial version of our algorithm, and production of a web service.

Here is a link to our service in action:

https://www.youtube.com/watch?v=1xibzI78eKE

12. Problems and risks

Throughout the course of the semester, our team encountered several problems. The initial challenge was making sure that we understood all of the project requirements. We wanted to ensure that communication was clear with our sponsors so that we would be able to deliver everything that they requested. The second challenge was familiarizing ourselves with Natural Language Processing. This involved doing extensive research into existing algorithms and approaches to solve the problem. Another challenge that we faced was testing our completed initial algorithm. We noticed that the "weights" feature of our results (each location returned has a weight associated with it based on how many times that location is mentioned) was returning incorrect values. After running unit tests, we were able to fix the errors in the code to make it work. Another challenge was being able to pull coordinate data from the database. Only a few team members were familiar with databases and how they work, so trying to figure out how to make our NLP algorithm find the right data in a fast/efficient way was challenging. After some research and hard work, we were able to overcome the problem with very few lines of code change. Our final challenge this semester was getting the project to run on a server provided by

our sponsors. Initially, every time we attempted to run the project on the server, it would crash. After exploring different solutions we determined that not enough memory was allocated, which was causing our application to crash. We have now resolved this issue and face no further problems.

13. Summary of tasks completed by each team member this semester

Samantha Juntiff

Work on team presentation

Summary and introduction, created basic design of all slides

Work on report

Reread for mistakes and wrote conclusion

Work on product

Made GeoJson object, created weights, helped teammates merge code

Work on PID

Read over, created the overall diagram of project.

Work on team management (meeting minutes, etc.)

Took notes at times, went to most meetings - if not had an excuse.

Adam McNabb

Work on team presentation

Created the results slide and video setup layout for the presentation. Presented results slide.

Work on report

Setup layout of paper wrote the initial rough draft of sections 1, 2, 4, 5, 6 proofread and formatted paper for consistency.

Work on product

Work with Jack and helped researching NLP and coming up with an example how it will work.

Work on PID

Wrote majority of it, read over it created one use case.

Work on team management (meeting minutes, etc.)

Took most meeting minutes which was shared in a google document to then be uploaded to the website.

Weston Neal

Work on team presentation

Project Technology and Design Decisions

Work on report

Checked the report for accuracy and errors in content and grammar.

Work on product

Worked on infrastructure of the product including setting up the Docker environment and the database.

Work on PID

Checked the document for errors and provided feedback to improve quality

Work on team management (meeting minutes, etc.)

Assisted plans to meet the needs, challenges, and other logistics currently facing the team

Christopher Silvia

Work on team presentation

Created “Lessons Learned” slide. I also added graphics and made edits to other teammates slides.

Work on report

I wrote the description, table of contents, introduction, scope, description of actors, preliminary results, and problems and risks sections. I also proofread and edited all other sections of the report to make them sound consistent.

Work on product

I worked with Samantha to develop the GeoJSON object output and implement the weight feature. I also researched other NLP algorithms and attempted to implement another during the design phase.

Work on PID

I wrote the introduction, table of contents, assumptions, a user story, the non functional requirements, and the first semester plan. I also proofread the entire document and edited all other sections to make them sound consistent.

Work on team management (meeting minutes, etc.)

I reminded the team of certain tasks we assigned, and took notes during meetings when Adam was not present.

Jang Won

Work on team presentation

Worked on Planned, Scope, and Tech pages; Editing

Work on report

Made edits as well as adding more to different sections

Work on product

Helped work on backend and db implementation

Work on PID

Editing and overview of document

Work on team management (meeting minutes, etc.)

Helped with ideas and technologies used

Jack Workman

Work on team presentation

Created “Design and Design Decisions” slide

Work on report

Wrote Executive Summary, Project Plan, Design Overview and Design Decisions, Development Approach

Work on product

Integrated Stanford NER Tagger and assisted with application design

Work on PID

Drew Use Case Diagram

Work on team management (meeting minutes, etc.)

Assisted with separating of tasks and responsibilities throughout semester and drew pictures on the whiteboard.

14. Conclusion

14.1 Success of the project so far

So far our team has met all requirements and deliverables set by the sponsors for semester one. Our team produced a project timeline, developed a design plan, researched Natural Language Processing, implemented an initial version of our algorithm, and produced a web service. Our output currently generates a list of probable locations in the text in GeoJSON format, exactly what our sponsors requested of us thus far.

14.2 Lessons learned

Our team learned that communication is key. We quickly realized that it was very important to contact and to keep the sponsor's interest in mind. At the start of the project we made sure that we understood the overall goal of the sponsors. To enforce this, we met with our sponsor once a week and then gradually in the middle of the semester spread the meetings out every two weeks.

Another lesson that the team learned was how to evaluate each other's strengths and weaknesses. The team chose to maximize the limited time each of us had to work on the project by dividing up parts based on our strengths and experiences. This lesson relied on our communication skills. Once we had established a set schedule, it became easy to divide the parts and present an overall product to the sponsors.

A final lesson that the team learned was how to adapt to fast changes and how to think creatively. The division of labor that the team agreed upon forced us to make rapid and concise decisions on the design of our product. The team usually ended up working two weeks on an individual problem. This required all of the teammates to brainstorm and come up with solutions. Most of the problems the team encountered were foreign and new; therefore, creativity played a large part in our problem solving technique. The team had to pick up a tool and learn how to use it and incorporate it effectively. Furthermore, if the tool did not work as expected, the team had to overcome the problem with a creative solution.

14.3 Future work

For Spring 2015, the team plans on using the current implementation to build a heatmap in order to make location data analysis easy to visualize for our customers. Our team is already brainstorming different tools and languages that we can use to build heat maps. In addition to the heat maps, our team will supply test cases and refine the algorithm to make it faster and more efficient. Our ultimate goal is to give the sponsors a fast, interactive, learning-capable algorithm that can easily be adapted in the future.