GeoD - A Geo-location Based Topic Model Visualization Tool

Author: Xindi Kang
Director: Dan Baciu
Collaborator: Yichen Li, Junqing Sun

Keywords:
Information Visualization, Maps, Topic Models, Wikification.

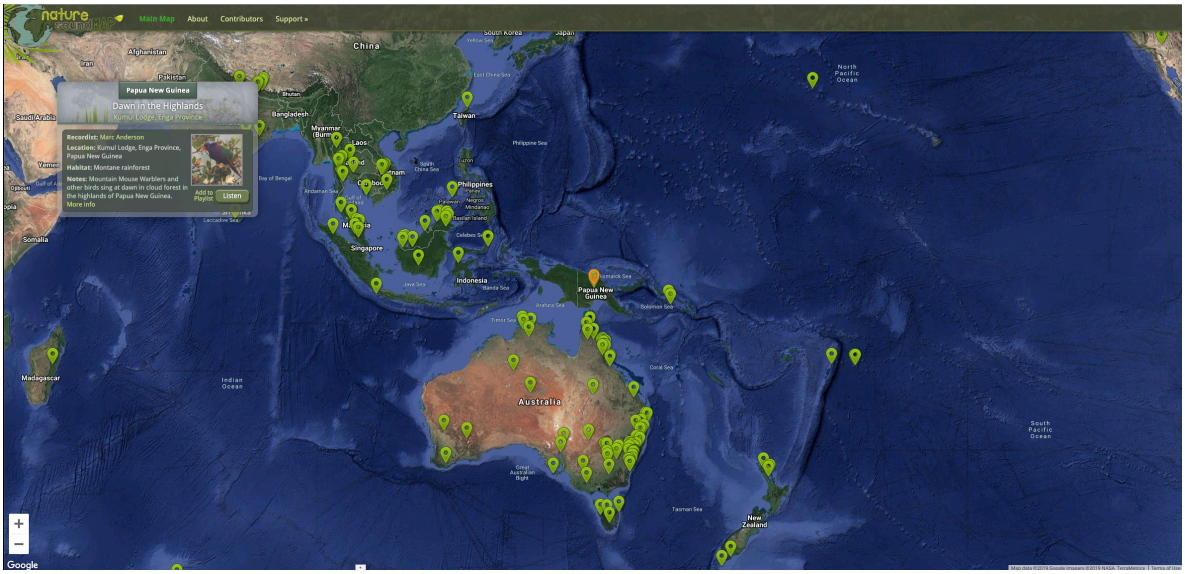
Tools used:
Illinois Wikifier 2013,

Abstract:

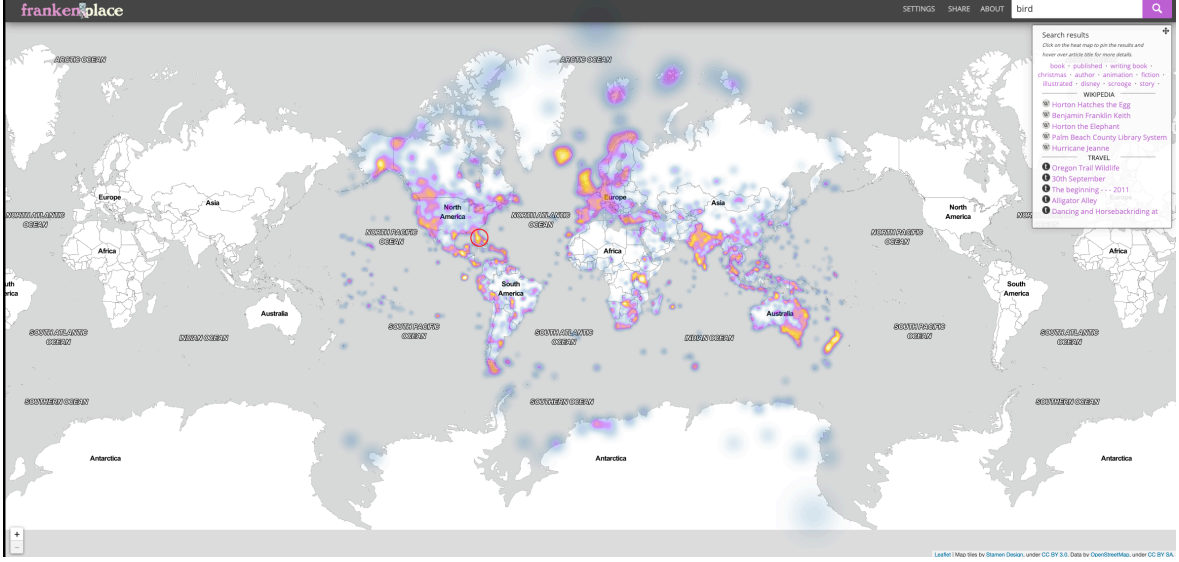
Mapping is one of the most established and prevalent form of information visualizations. While maps can represent geo-location information for the purpose of exploring the landscape, they can also be used as a tool to deduce meta-information about a certain subject. GeoD is a web-based visualization tool designed for exploring the geo-location of topic models, which are clusters of words that represent topic information of a large corpus. In the current version, users are able to see locations mentioned in the entire corpus and in each topic as well.

Introduction:

Maps are one of the most established form of information visualizations. The origin of mapping, also known as Cartography, dates back to as far as ancient Greece and Rome, when people used geometries and patterns to represent geographical information such as buildings, roads and vegetation. In modern context, while many different ways of visualizing information are available, the representation of geo-location information using geographic coordinate system remains unchanged. There are various visualizations based on geo-location that are developed for the web for research purposes. Here are some examples:



Nature sound map is a visualization tool for exploring nature sounds recorded around the world. http://www.naturesoundmap.com/



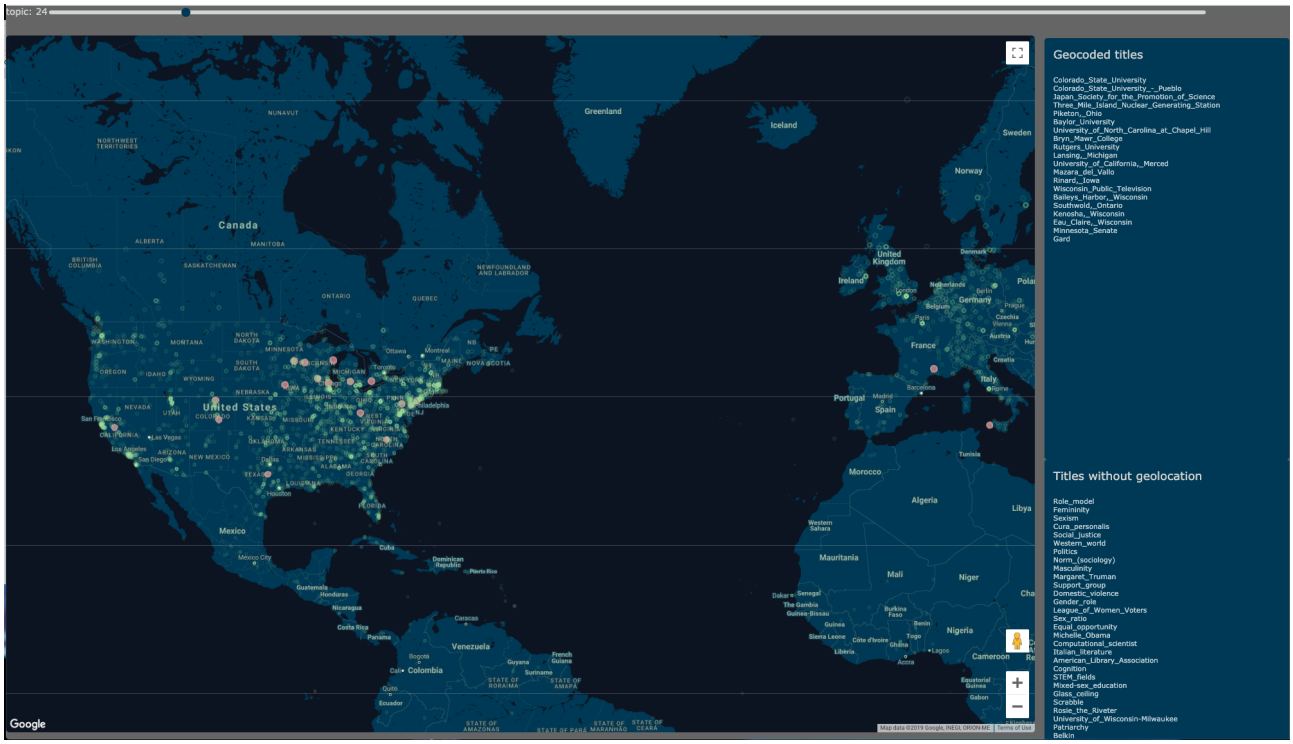
Frankenplace is an interactive thematic map search engine that uses geographic context as a means to discover, organize, and interactively visualize the documents related to a search query. http://www.frankenplace.com/

Topic Modeling:

In the WE1S research group, the work is focused on finding out what everyone says (WE1S) about the humanities as an academic discipline by exploring a large amount of articles with topics concerning the humanities. In this process, one of the research questions would be, which places in the world are being talked about in the articles, and what do people say about these places.

In order to first find out what people talk about in general, topic modeling becomes a very useful tool in the process. Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes. [1] In a typical WE1S topic model, a topic is consisted of a topic number, and a cluster of keywords that belong to this topic.

The Interface:



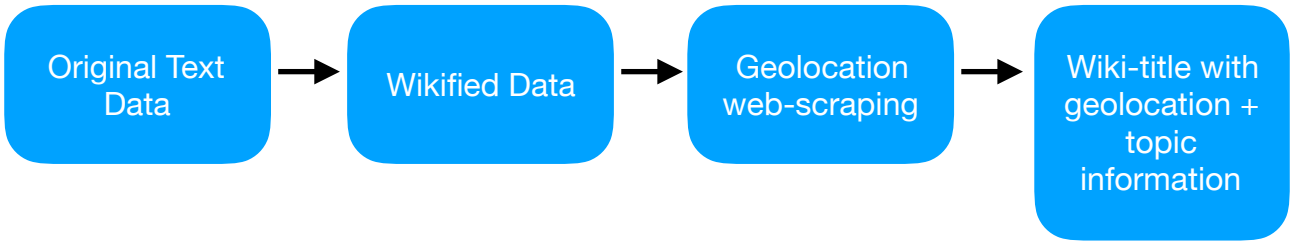
GeoD is a tool designed for visualizing topic information in the geographic domain. It is particularly useful for researching on locations that are discussed by the entire literature corpus and a specific topic. At the current stage of the visualization, users can easily get an overview of the density distribution of all topics by looking at the green dots shown on the world map. Each of the green dots represent a data point of a particular location mentioned in a topic. Hovering over a particular dot shows the name of that location, followed by the topic number. Sliding the slide bar on top of the interface allows the user to select a particular topic number. Upon selecting a particular topic, the relevant geolocations is highlighted in red, and all of the names of the locations are listed on the side bar located on the right of the interface. The keywords without geolocations, but are included in the topic, are shown in the module in the lower right corner.

Data Processing:

The visualization tool uses topic model data generated from the University Wire literature database. The text-data of the literature runs through a Python program which searches the Wikipedia database (Illinois Wikifier 2013) for any existing titles. The process is called Wikification, which is the task of identifying and linking expressions in text to their referent Wikipedia pages.[2]

The output of the Wikification is a .xml file which contains the all the titles found in the input text data, which is then run though a python program to scrape geo-locations from the web. With more processing it was finally turned into a comprehensive .json file with contains all of the wiki-titles with corresponding geolocations.

Workflow of data processing:



Snippet of the final data output:


```
{
  "Topics": [
    {
      "Topic": "1",
      "Wikidata": [
        {
          "WikiID": "57164",
          "Title": "Cajun",
          "Count": "7",
          "coord": null
        },
        {
          "WikiID": "163416",
          "Title": "Hank_Williams",
          "Count": "2",
          "coord": null
        },
        {
          "WikiID": "151916",
          "Title": "T-shirt",
          "Count": "27",
          "coord": null
        },
        {
          "WikiID": "2380366",
          "Title": "Catwalk_(theater)",
          "Count": "7",
          "coord": null
        },
```

The visualization uses GoogleMapsAPI to achieve visualizations of the data points. Original code is available upon request (due to googleMaps license key distribution problem). The visualization is available online at https://www.gettoby.com/p/vs0whs15jsyt.

References:

1. Blei, David (April 2012). "Probabilistic Topic Models". *Communications of the ACM*. **55** (4): 77–84. doi:10.1145/2133806.2133826.

2. Lev Ratinov and Dan Roth and Doug Downey and Mike Anderson, Local and Global Algorithms for Disambiguation to Wikipedia. *ACL* (2011) pp.