Proposal of Project for Web Search Engine

Team Members

- Heng Lin, hl2521@nyu.edu
- Liang Niu, ln932@nyu.edu

Problems

Which project we plan to work on

We plan to work on **Project #10: Analyzing text evolution in wikipedia**.

Wikipedia is a collection of several million articles written in a collaborative process by hundreds of thousands of users. Wikipedia stores the entire edit history of all the articles, and it would be very interesting to understand more about text evolution in wikipedia and learn more about the authors in wikipedia.

What exact problems we plan to address

- How text is changed over time of a single entry
- Analyze the changes which happen continuously or in bursts
- Whether incorrect info is removed quickly or not
- Analyze the authors of Wikipedia behaviors including their contribution, areas they are expert in and info of geography
- Create rankings of all entries or the entries in their portal

what exact data sets we plan to use

- English entries of Wikipedia

What approaches we plan to get the data set

- Enwiki dump of 20161101. Wikipedia offers free copies of all available content. But Wikipedia have rate limited downloaders and cap the number of per-ip connections to 2.
- Wikipedia API. The MediaWiki action API is a web service that provides convenient access to wiki features, data, and meta-data over HTTP.
- A web crawler to crawl entries if needed.

What we will implement

- For a single entry, when you type the key word, the program will display its editing history in well-formed visualization showing the timeline of modification.
- As we collect many entries info, we can create a ranking in many dimensions such as the ranking of most frequent modification entries and the ranking of author contribution.
- Find out where contributors are from (using IP info).
- Find out some interesting results such as influential people from wikipedia

Plan

1. First of all, we plan to do some research on rules of wikipedia like the editting process, wikipedia editor's behavior, how editor robot works, how do a change be accepted and do some research on previous work, see if there exists some tools or interesting idea and maybe some resourses.
2. Figure out how to use wikipedia API and crawler to get the dataset, and figure out which protals of entries to analyse, then download the data.
3. Try to analyse a small part of dataset and conclude what result we should get by hand analysing.
4. Build our core program to analyse data, figure out what kind of data structure used to store data when analysing it

    1. Analysis of a single entry
    2. Analysis of contributors
    3. Analysis of all entries in a portal

5. Build a website to show our work we have done, make interface be friendly to human and visualize all the result on it.
6. Show our work to others and then evaluate our work and write the project paper.
7. If still have time left, find some more interesting facts of Wikipedia.

How you will run experiments and evaluate success.

1. Speed: Do a bunch of queries on the website and evaluate the speed(only for single eentry, because ranking cannot be real-time)
2. Show our work to others, collect opinions.
3. Compare our work to existing researches and projects.

Tools and environments

- Python and Python data analysing libraries

- HTML and CSS and Javascript, to visualize our work
- A database management system, probably PostgreSQL or MySQL, to store data, and also used to build website
- A server, to hold service for website. And if needed, we may perform calculation on it
- Slack for communication

Previous work

most relative ones:

- On the Evolution of Wikipedia
- Influential People from Wikipedia
- Structuring Wiki Revision History
- Analyzing and visualizing the semantic coverage of Wikipedia and its authors

also:

- A content-driven reputation system for the wikipedia
- Using text animated transitions to support navigation in document histories

Work Division

Together:

1. Build core analysing program.
2. Evaluate the work.

Heng Lin:

1. Use wikipedia API to fetch data.

Liang Niu:

1. Build the website.