WikiLink Coloring

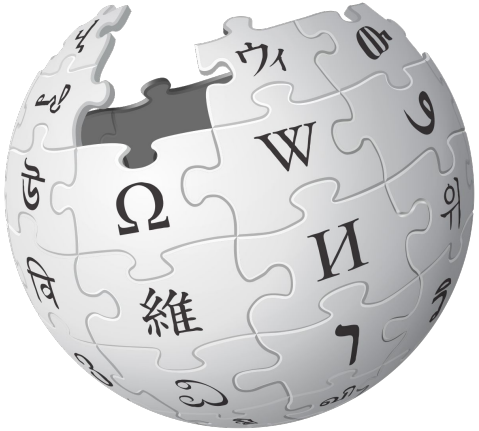
CS 6235: Real-Time Systems
Final Project Presentation

Aman Jain

Georgia Tech

Overview

- Motivation:
  - I am an avid Wikipedia user
  - I want to know which articles to read next
- Combine **two good ideas**:
  - Wikipedia + recommendation engine



Georgia Tech

Original Approach

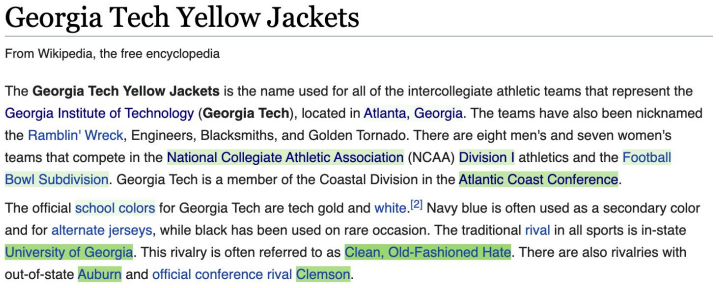
- Articles have **hyperlinks** to other articles
- Highlight these hyperlinks based on “article quality”
- Calculate quality **locally** and **on-demand**

Georgia Tech Yellow Jackets

From Wikipedia, the free encyclopedia

The **Georgia Tech Yellow Jackets** is the name used for all of the intercollegiate athletic teams that represent the Georgia Institute of Technology (**Georgia Tech**), located in Atlanta, Georgia. The teams have also been nicknamed the Ramblin' Wreck, Engineers, Blacksmiths, and Golden Tornado. There are eight men's and seven women's teams that compete in the National Collegiate Athletic Association (NCAA) Division I athletics and the Football Bowl Subdivision. Georgia Tech is a member of the Coastal Division in the Atlantic Coast Conference.

The official school colors for Georgia Tech are tech gold and white.[2] Navy blue is often used as a secondary color and for alternate jerseys, while black has been used on rare occasion. The traditional rival in all sports is in-state University of Georgia. This rivalry is often referred to as Clean, Old-Fashioned Hate. There are also rivalries with out-of-state Auburn and official conference rival Clemson.



Georgia Tech Yellow Jackets

From Wikipedia, the free encyclopedia

The **Georgia Tech Yellow Jackets** is the name used for all of the intercollegiate athletic teams that represent the Georgia Institute of Technology (**Georgia Tech**), located in Atlanta, Georgia. The teams have also been nicknamed the Ramblin' Wreck, Engineers, Blacksmiths, and Golden Tornado. There are eight men's and seven women's teams that compete in the National Collegiate Athletic Association (NCAA) Division I athletics and the Football Bowl Subdivision. Georgia Tech is a member of the Coastal Division in the Atlantic Coast Conference.

The official school colors for Georgia Tech are tech gold and white.[2] Navy blue is often used as a secondary color and for alternate jerseys, while black has been used on rare occasion. The traditional rival in all sports is in-state University of Georgia. This rivalry is often referred to as Clean, Old-Fashioned Hate. There are also rivalries with out-of-state Auburn and official conference rival Clemson.

Georgia Tech

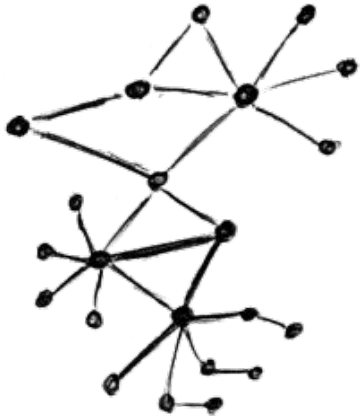
Roadblock & Pivot

- Wikipedia has **rate limiting** on queries
  - Results in HTTP 429 Errors
- Pivot:
  - On-demand querying not possible
  - Information needs to be **preprocessed** and stored for quick lookup

Georgia Tech

Preprocessing

- Wikipedia publishes **monthly dumps** of all data on site
- Goal - generate a **graph** from the dump
  - Nodes: articles
  - Edges: hyperlinks between articles
- **Data is too big** (~100 GB) to handle locally
  - Solution: process dump on COC-ICE



Georgia Tech

Link “Quality” Algorithm

- Looking for links to **similar articles**
- Definitions:
  - **Input Space** of article X: set of all articles that link to X
  - **Output Space** of article X: set of all articles that X links to
- Similar input and output spaces -> articles are similar

$$inputSimilarity = \frac{|inputs(X) \cap inputs(Y)|}{|inputs(X) \cup inputs(Y)|}$$

$$outputSimilarity = \frac{|outputs(X) \cap outputs(Y)|}{|outputs(X) \cup outputs(Y)|}$$

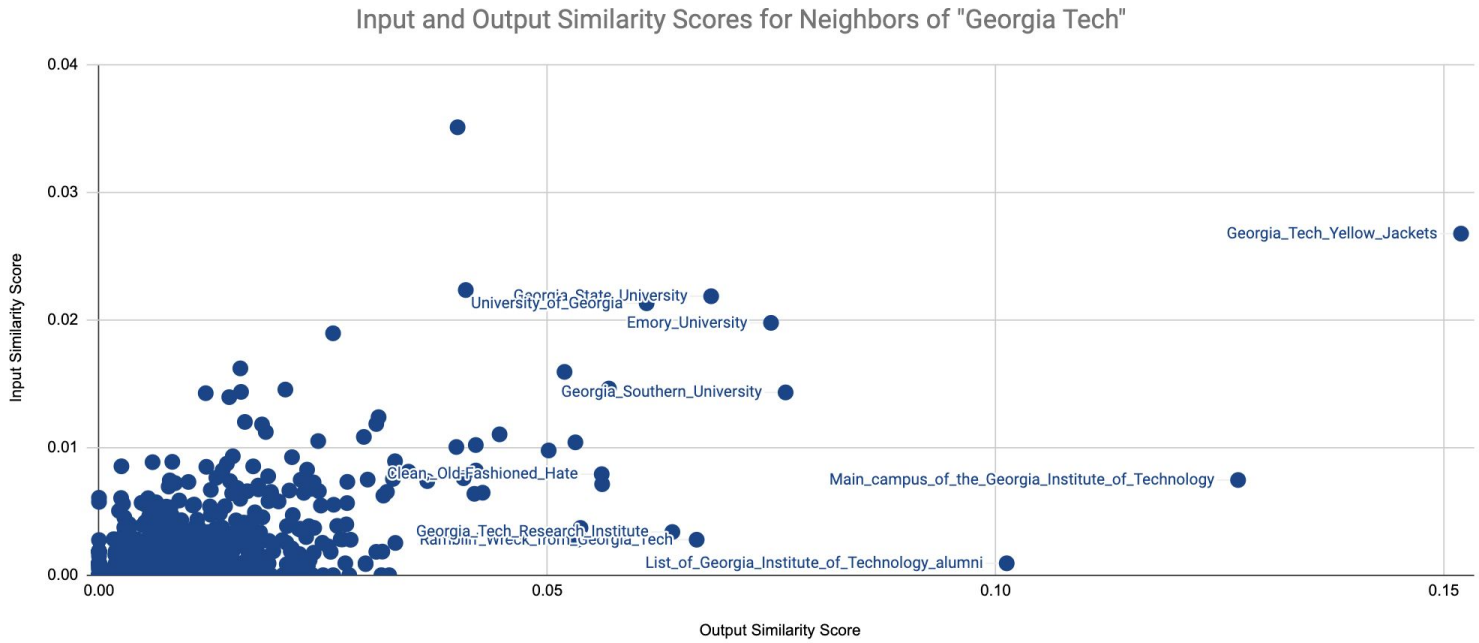
$$similarity = \sqrt{inputSimilarity^2 + outputSimilarity^2}$$

Georgia Tech

Algorithm Explanation

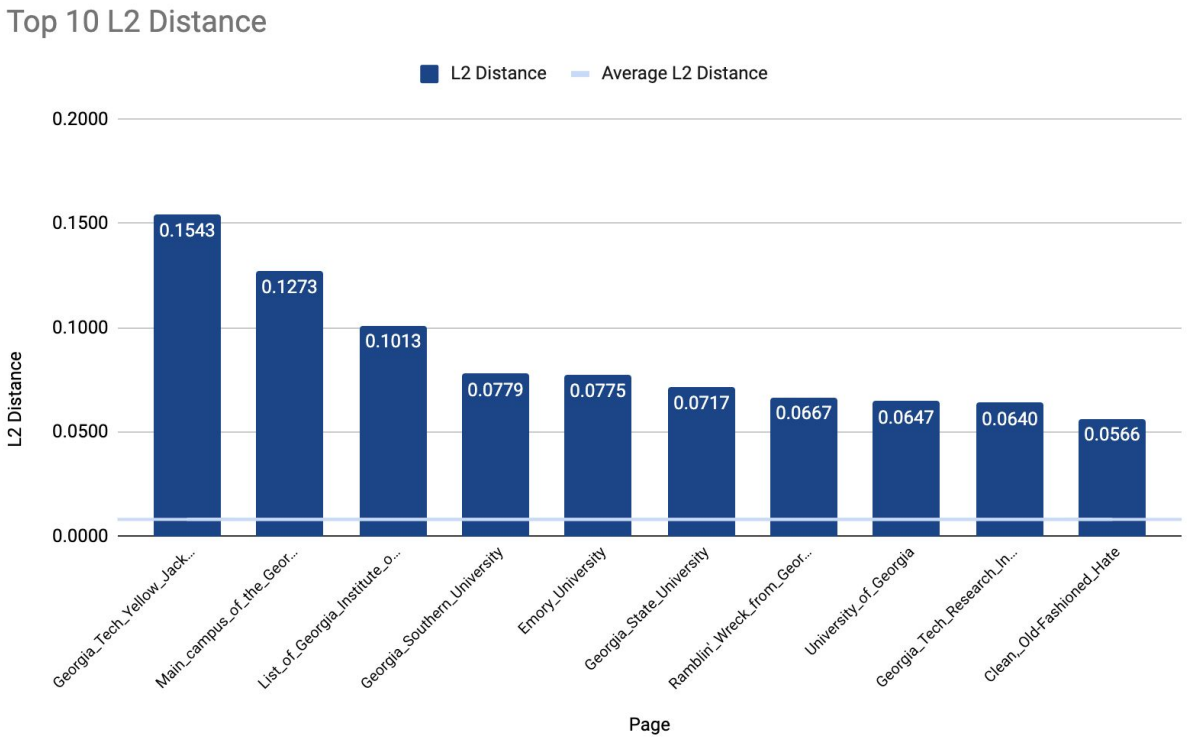
- Intersection gives us the number of similar links in space
- Dividing by union normalizes for articles that have much larger input spaces (e.g. “Atlanta” compared to “Georgia Tech”)
- L2 distance to equally weigh input and output similarity

Example: “Georgia Tech”



Georgia Tech

“Georgia Tech” L2 Distances



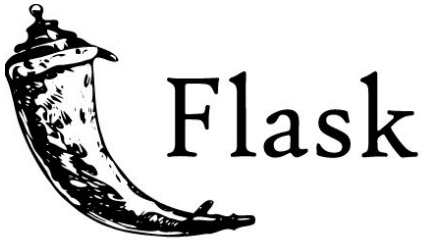
Georgia Tech

Preprocessing Wrapped Up

- We convert the dump to a graph ahead of time
  - Each node has a list of inputs and outputs
- On demand, we calculate similarity using the algorithm specified above

System Architecture

- Storage Solution: Redis
  - Tried Firestore and MongoDB before deciding on Redis
  - **Fast random reads** with less than 5GB of data
- Backend: Python Flask
- Frontend:
  - Chrome Extension written in Javascript



Georgia Tech

Live Demo

en.wikipedia.org/wiki/Georgia_Tech

Georgia Tech

Challenges and Learning Outcomes

- Having to pivot early in the process was stressful but ultimately a good learning experience
  - Writing parallel software to handle large amounts of data
  - Using a supercomputer
- Learning Javascript - I mostly avoided this

Future Work

- Knowledge Obsolescence
  - Dumps only updated monthly: might miss new pages
  - Solution: update redis every time extension is used
- Productionize
  - Everything currently hosted locally, can be hosted on cloud
- User Interface
  - Extensions is bare-bone right now, can be made prettier

Thank You!

github.com/amanj120/wikilink-coloring

Georgia Tech