CSE 6240
Web Search and Text Mining
Spring 2020

Instructor: Prof. Srijan Kumar
Teaching Assistants: Roshan Pati, Arindum Roy
Web is a platform for everyone
Web allows...

- **Web enables expression of ideas and social interaction**
- Web is no longer a static library that people passively browse
- **Web is a place where people:**
  - Act as prosumers, i.e., content producers and content consumers
  - Interact with other people:
    - Internet forums, Blogs, Social networks, Twitter, Wikis, Podcasts, Slide sharing, Bookmark sharing, Product reviews, Comments
  - Use services:
    - buy products, stream videos/movies
Web is a…

• Web is a collection of documents
  – E.g., web pages, social media posts

• Web is a network
  – E.g., the hyperlink network of websites, network of people on social networks

• Web is a set of applications
  – E.g., e-commerce platforms, content sharing, streaming services
Web Mining: Opportunities

- Anyone can share and contribute content, express opinions, link to others

- This means: One can data-mine opinions and behaviors of millions of users to gain insights into:
  - Human behavior
  - Marketing analytics
  - Product sentiment
Topics Covered in the Course

• Web is a collection of documents
  – E.g., web pages, social media posts

• Web is a network
  – E.g., the hyperlink network of websites, network of people on social networks

• Web is a set of applications
  – E.g., e-commerce platforms, content sharing, streaming services

Text Mining and Information Retrieval

Network Science

Recommender Systems and Social Media
Unique Value of Textual Web Data

- Useful to many big data applications
- Especially useful for mining knowledge about people’s behavior, attitude, and opinions
- Directly express knowledge about our world: Small text data are also useful!

Data ➔ Information ➔ Knowledge

- This course’s outcome: Learn the basics of processing textual data
Textual Web Data is Prevalent

Topics:
- People
- Events
- Products
- Services, ...

Sources:
- Blogs
- Microblogs
- Forums
- Reviews, ...

- 45M reviews
- 53M blogs
- 1307M posts
- 65M msgs/day
- 115M users
- 10M groups
Applications: Real-time Citizen Journalism

- **Citizen journalism** provides more valuable information than newswire services

- **Challenge:**
  - Many redundant posts, users have to wade through hundreds of posts to locate useful information

- **Goal:**
  - Mine this data in real-time and produce well organized summaries
Applications: Reputation management

• **Consumer Brand Analytics**
  – What are people saying about our brand?

• **Marketing Communications**
  – Significant spending on marketing, advertising:
    Companies trying to position their products
  – Brand analytics helps to determine whether such campaigns are effective

• **Product reviews**
  – Automatically mine product reviews for information on product features, new requests, ...
    • Easy to use, Light weight, Sturdy, Good price, …
Networks are Ubiquitous

Social networks

Economic networks

Communication networks

Information networks: Web & citations

Internet

Networks of neurons
Two Types of Networks

• **Networks (also known as Natural Graphs):**
  – Society is a collection of 7+ billion individuals
  – Communication systems link electronic devices
  – Interactions between genes/proteins regulate life

• **Information Graphs:**
  – Information/knowledge are organized and linked
  – Scene graphs: how objects in a scene relate
  – Similarity networks: take data, connect similar points
Networks: Knowledge Discovery

- **Universal language for describing complex data**
  - Networks from science, nature, and technology are more similar than one would expect

- **Shared vocabulary between fields**
  - Computer Science, Social Science, Physics, Economics, Statistics, Biology

- **Data availability & computational challenges**
  - Web/mobile, bio, health, and medical

- **Impact!**
  - Social networking, Drug design, AI reasoning

- **This course’s outcome:** Learn how to process large scale networks to discover knowledge
Ways to Analyze Networks

- **Predict the type/color of a given node**
  - Node classification
- **Predict whether two nodes are linked**
  - Link prediction
- **Identify densely linked clusters of nodes**
  - Community detection
- **Measure similarity of two nodes/networks**
  - Network similarity
Information and Social Media/Networks

- Google
- Cisco
- Facebook
- Amazon
- Pinterest
Social Media: Polarization on Twitter

- Retweet networks:
  Polarized (left), Unpolarized (right)

Social Media: Misinformation

- **Q:** Is a given Wikipedia article a hoax?
  - Real articles link more coherently:

![Diagram showing real and hoax articles with different connectivity patterns]

**Hoax article detection performance:**
- Random: 50%
- Human: 66%
- Network: 86%

---

Social Media: Predicting Virality

Information cascade in social networks
Practical Applications of This Course

- **Fraud and Filtering**
  - fraud, trolls/bots/spammers, fake news

- **Recommender Systems**
  - news/literature/movie recommender

- **Categorization**
  - news categorization, help desk email routing, sentiment tagging

- **Topic mining**
  - discovery of topical trends in scientific research
  - discovery of major complaints from customers

- **Prediction and Detection**
  - stock prices from social media posts, voting results
Course Goals

• Provide a **systematic introduction** to text analysis, network analysis, and recommender systems

• Provide an opportunity for students to **explore frontier topics via course projects** (customized toward the interests of students)

• Give students enough training for doing **research** in web mining or **applying** advanced web mining techniques to applications

• Tangible outcomes: **research paper, open source code, and application system**
About CSE6240
Logistics

- Course: Weekly lectures on Monday and Wednesday 3:00pm-4:15 at Boggs B9

- Course website: https://cs.stanford.edu/~srijan/teaching/spring2020/

- Piazza: https://piazza.com/class/k4u6q1g7t672ln
Administrivia

• Office hours:
  – Srijan: 10-11am Wednesday, Coda S1303
  – Roshan (TA): 3-4pm Thursday, Klaus 3rd floor Atrium
  – Arindum (TA): 3-4pm Tuesday, Klaus 3rd floor Atrium

• Piazza as “extended classroom”
  – Post your question on Piazza as soon as you have it
  – Share your expertise by helping answer questions from your peers
  – Initiate discussions of any technical issues related to the course
Prerequisite

• Basic knowledge of **probability and statistics**
• Basic knowledge of **linear algebra**: vectors and matrices
• Knowledge of one or more of the **following areas is a plus**, but not required: Information Retrieval, Machine Learning, Data Mining, Natural Language Processing
• **Programming**
  – Python, Anaconda (miniconda), numpy, scipy, sklearn, pandas
• Contact the instructor if you are not sure
Format and Syllabus

• Weekly two lectures
• **Programming homeworks**: ensure solid mastery of skills of implementation and experimentation
• **Course project**: multiple options, encourage massive collaboration
  – Research Track: In-depth study of a topic ➔ publication/submission
  – Development Track: Implementation of a novel application ➔ useful application
• On [Google docs](#)
Grading Breakdown

• 3 homework assignment: 45%
• 1 course project: 55%
  – Proposal: 5%
  – Milestone report: 20%
  – Final report and poster presentation: 30%
• No midterm or final
Focus of Work

First Day of Instruction

Lectures

Assignments

Project

Jan | Feb | Mar (Spring break) | Apr | Last Day of Instruction

Spring break
Typical Project Pipeline

Big Data → Small Relevant Data → Knowledge → Many Applications

Data retrieval

Clustering

Categorization

Search engines

Filtering

Recommender

Summarization

Data Analysis

Topic mining

Sentiment

Prediction

Applications:

Medical/Health

Security

Education

Business

Social Media
More about Projects I

- **Topics should be related to the WEB:**
  - Information Search
  - Text Analysis
  - Network Science
  - Recommender Systems
  - Social Media

- **Goal:** Get hands-on web mining experience

- **Tangible outcomes:** research paper, open source code, and application system
More about Projects II

- **Empirical analysis** of data to develop a model of behavior
- **Algorithms and models** to make predictions on a dataset
- **Scalable algorithms** for massive datasets
- **Theoretical project** that considers a model/algorithm and derives a rigorous result about it
More about Projects III

• **Topics:**
  – We will release possible topics or you can select your own
  – If you are not sure if your topic lies in the area, come talk to us

• **Collaboration is encouraged**
  – 3 person teams will be most efficient
  – 1 or 2 people teams are fine

• **Proposal is due on Feb 3**
  – Form teams and start thinking about this now!

• **Project milestone check is on Mar 11**

• **Final project is due on Apr 20/22**
More about Projects IV

- **Proposal**: 2 pages
  - Two parts:
    1. Reaction to existing paper/technology
       - Summary
       - Critique/Shortcomings
    2. Proposal
       - How are you improving the existing work?

- **Milestone**: 3-4 pages
- **Final report**: 6-8 pages
- **Final poster**: details after the spring break
Questions?

• Course website: 
  https://cs.stanford.edu/~srijan/teaching/spring2020/

• Piazza: https://piazza.com/class/k4u6q1g7t672ln