Project Ideas and Datasets

CS345a: Data Mining
Jure Leskovec and Anand Rajaraman
Stanford University

MapReduce & Amazon EC2 session

- Friday 5:30 at Gates B12 5:30-7:30pm
- You will learn and get hands on experience on:
  - Login to Amazon EC2 and request a cluster
  - Run Hadoop MapReduce jobs
  - Use Aster nCluster software
- Amazon have us $12k of computing time
- Each students has about $200 worth of computing time

Projects: intro

- Ideally teams of 2 students (1 (3) is also ok)
- Project:
  - Discovers interesting relationships within a significant amount of data
  - Have some original idea that extends/builds on what we learned in class
    - Extend/Improve/Speed-up some existing algorithm
    - Define a new problem and solve it

Project proposal (1)

- Answer the following questions:
  - What is the problem you are solving?
  - What data will you use (where will you get it)?
  - How will you do it?
    - Which algorithms/techniques you plan to use?
    - Be as specific as you can!
  - Who will you evaluate, measure success?
  - What do you expect to submit at the end of the quarter?

Project proposal (2)

- Due on midnight Feb 1 2010
- Email the PDF to cs345a-win0910-staff@lists.stanford.edu
- TAs will assign group numbers
- Name your file: <group#>_proposal.pdf

Project ideas and datasets

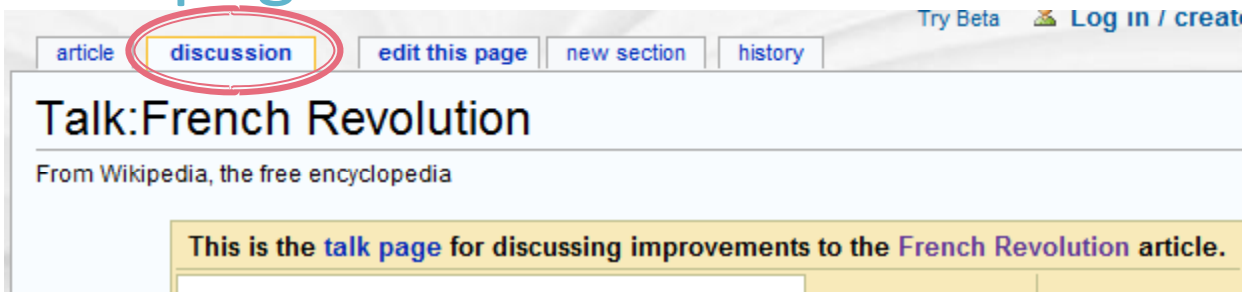
- Wikipedia
- IM buddy graph
- Yahoo Altavista web graph
- Stanford WebBase
- Twitter Data
- Blogs and news data
- Netflix
- Restaurant reviews
- Yahoo Music Ratings

Wikipedia (1)

- Complete edit history of Wikipedia until January 2008
- For every single edit the complete snapshot of the article is saved
- Each **page** has a **talk** page:

Wikipedia (2)

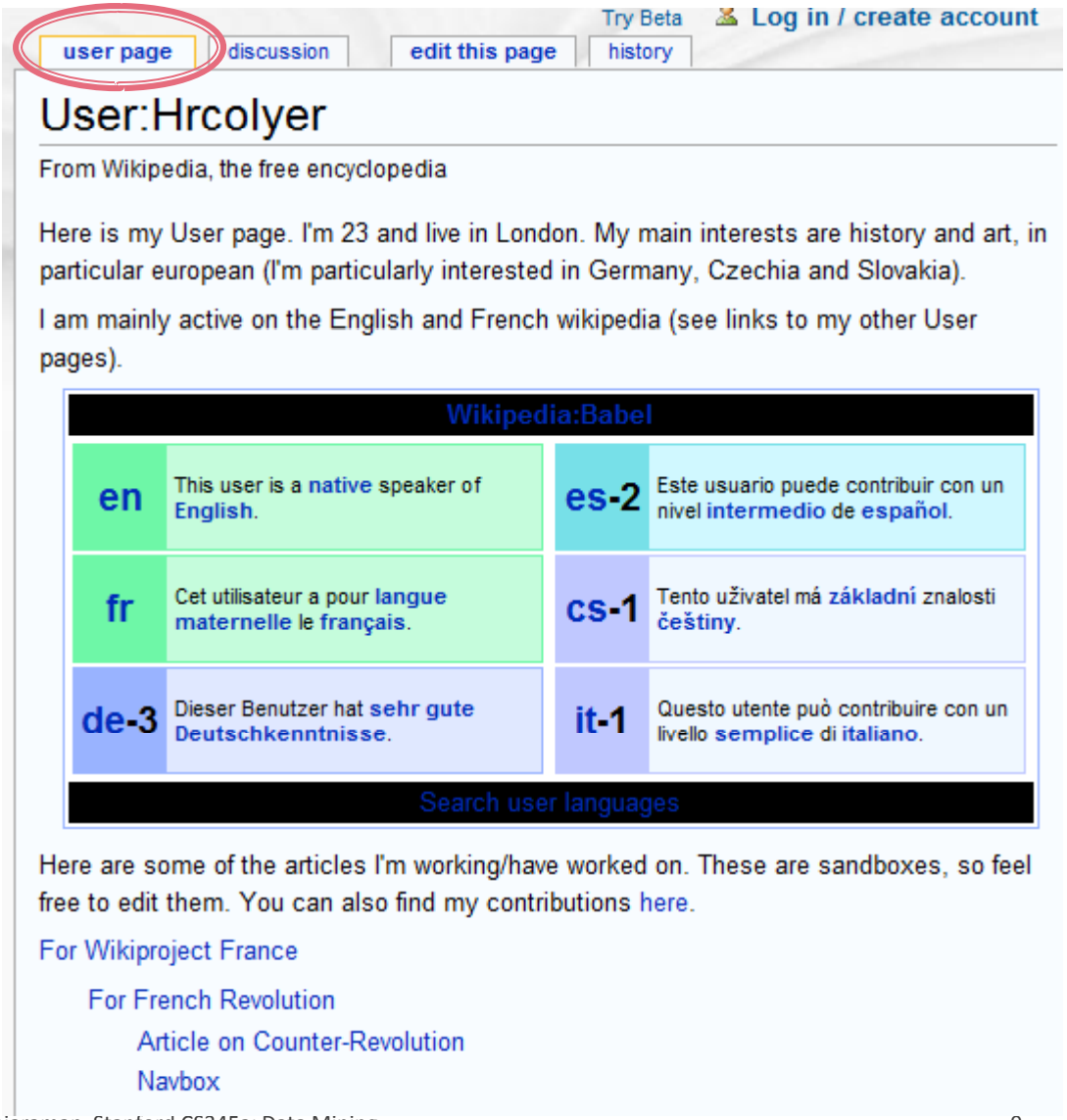
- Talk page:



- Editors discuss things like:

Wikipedia (3): User pages

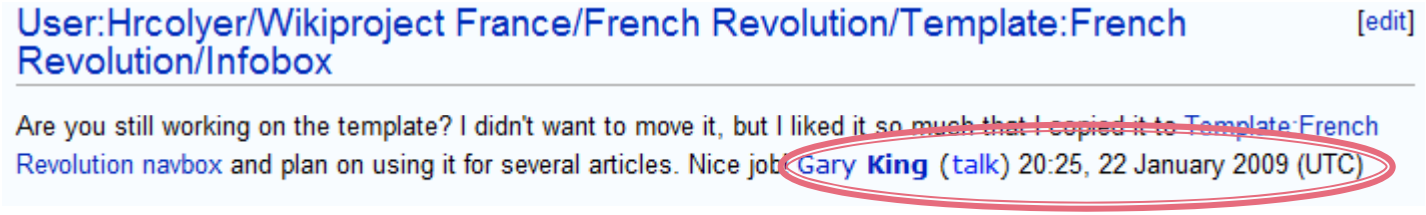
- Every registered use has a page:



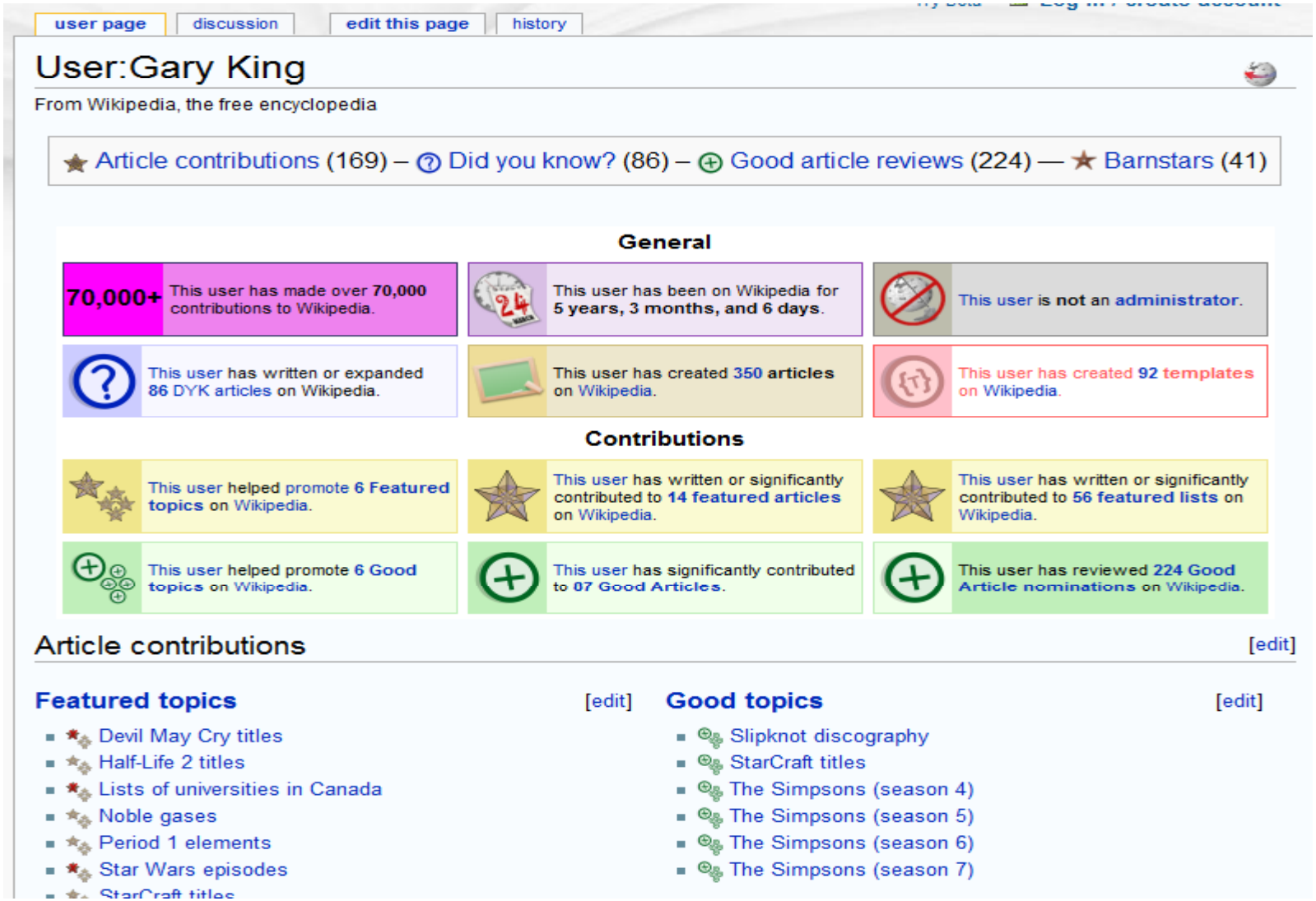
Wikipedia (4): User talk pages

- Every user's page has a talk page:

- Users discuss things:



Wikipedia (5): User pages



Wikipedia: Data format

```
<page>
    <title>Anarchism</title>
    <id>12</id>
    <revision>
      <id>18201</id>
      <timestamp>2002-02-25T15:00:22Z</timestamp>
      <contributor>
        <ip>Conversion script</ip>
      </contributor>
      <minor />
      <comment>Automated conversion</comment>
      <text xml:space="preserve">''Anarchism'' is the political
        theory that advocates the abolition of all forms of
        government.
        ...
      </text>
    </revision>
    <revision>
      <id>19746</id>
      <timestamp>2002-02-25T15:43:11Z</timestamp>
      <contributor>
        <ip>140.232.153.45</ip>
      </contributor>
      <comment>*</comment>
      <text xml:space="preserve">''Anarchism''‘ is the political
        theory that advocates the abolition of all forms of government.
        ...
```

Wikipedia: Ideas

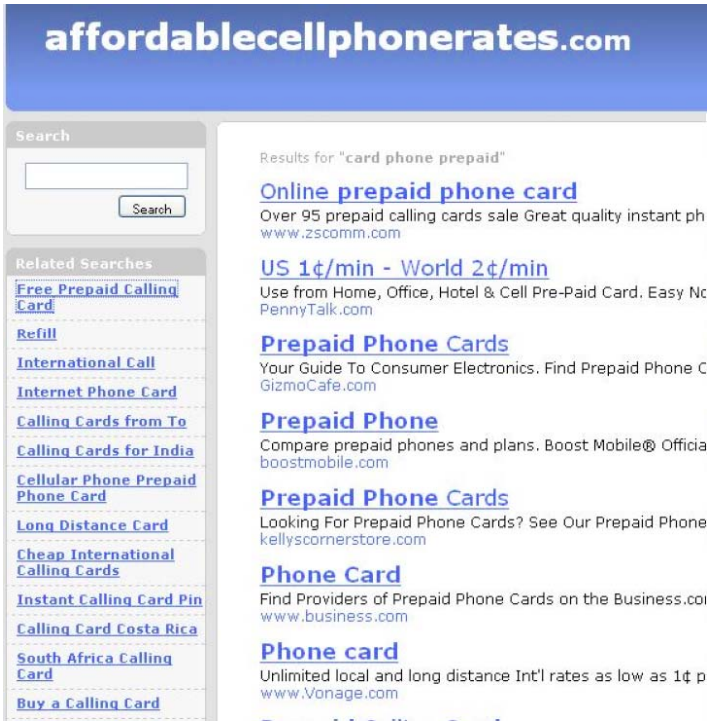
- Complete edit and talk history of Wikipedia:
  - How do articles evolve?
    - Use string edit distance like approach to measure differences between versions of the article
    - Model the evolution of the content
  - Which users make what types of edits?
    - Big vs. small changes, reorganization?
    - Suggest to a which user should edit the page?
  - How do users talk and then edit same pages?
    - Do users first talk and then edit?
    - Is it the other way around?
  - Suggest users which pages to edit

Yahoo Altavista Web Graph

- Altavista web graph from 2002:
  - Nodes are webpages
  - Directed edges are hyperlinks
  - 1.4 billion public webpages
  - Several billion edges
  - For each node we also know the page URL

Altavista: Ideas (1)

- SPAM:
  - Use the web-graph structure to more efficiently extract spam webpages
  - Link farms
  - Spider traps
- Personalized and topic-sensitive PageRank



Altavista: Ideas (2)

- Website structure identification:
  - From the webgraph extract "websites"
  - What are common navigational structures of websites?
    - Cluster website graphs
    - Identify common subgraphs and patterns
  - What are roles pages/links play in the graph:
    - Content pages
    - Navigational pages
    - Index pages
  - Build a summary/map of the website

Stanford webbase

- A collection of focused snapshots of the Web
- Data starts in 2004 and continues till today
  - General crawls
    - start from ~1000 seed webpages
    - Crawl up to ~150,000 pager per site
  - Specialized crawls:
    - Universities
    - US Government
    - Hurricane Katrina (2005) – daily crawls
    - Monthly newspaper crawls

Stanford webbase: Ideas

- Smaller than Altavista but you also have the page content
- Can do topic analysis
- Topic sensitive PageRank
- Study the evolution of websites and webpages

Twitter: Data

- 50 million tweets per month starting June 2009 (6 months)
- Format:

```
T       2009-06-07 02:07:42
U       http://twitter.com/redsoxtweets
W       #redsox Extra Bases: Sox win, 8-1: The Rangers
spoiled Jon Lester's perfecto and his shutout..
http://tinyurl.com/pyhgwy
```

- Two important things:
  - URLs
  - Hash-tags

Twitter: Ideas

- Trending topics: raising, falling
- Inferring links of the who-follows-whom network
- What is the lifecycles of URLs and hash-tags?
- Finding early/influential users?
- Clustering tweets by topic or category
- Sentiment analysis – are people positive/negative about something (a product?)

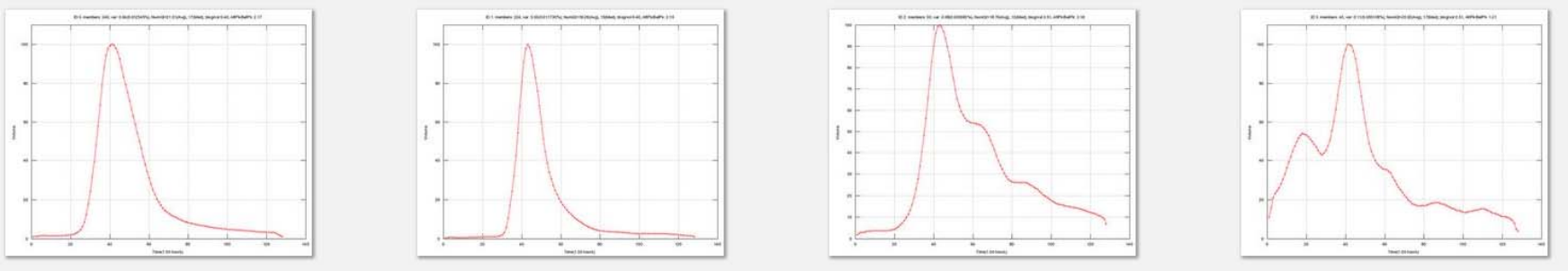
MemeTracker: Data

- More than 1 million newsmedia and blog articles per day since August 2008
- Extract phrases (quotes) and links
- http://memetracker.org
- Format:

```
P       http://cnnpoliticalticker.wordpress.com/2008/08/31/mccain-defends-
palins-experience-level
T       2008-09-01 00:00:13
Q       dangerously unprepared to be president
Q       even more dangerously unprepared
Q       understands the challenges that we face
Q       worked and succeeded
Q       still to this day refuses to acknowledge that the surge has
succeeded
L       http://www.cnn.com
```

MemeTracker: Ideas (1)

- Find all variants (mutations) of the same phrase – cluster phrases based on edit distance and time:
  - lipstick on a pig
  - you can put lipstick on a pig
  - you can put lipstick on a pig but it's still a pig
  - i think they put some lipstick on a pig but it's still a pig
  - putting lipstick on a pig
- Temporal variations of the phrase volume



MemeTracker: Ideas (2)

- Predict the popularity of a phrase over time
- How does information mutate/change over time?
- Which media sites are the most influential? Build a predictive model of site influence
- Which nodes are early mentioners, late comers, summarizers?
- Sentiment analysis – are people positive/negative about something (news, a product)
- Create a model of political bias (liberal vs. conservative)
- What is genuine news, what are genuine phrases and what is spam?

Wikipedia: More ideas

- We also have the Wikipedai webserver logs, i.e., page visit statistics

- How does Wiki page visit statistics correlate with external events, natural disasters?
  - Use Twitter or MemeTracker data to detect those
  - Compare occurrence of phrases and visits to Wikipedia pages

IM Buddy graph: Data

- A large IM buddy graph from March 2005
- 230 million nodes
- 7,340 million undirected edges

- Limitations:
  - Only have the buddy graph with random node ids
  - No communication or edge strength

IM Buddy graph: Ideas

- Find communities, clusters in such a big graph
- Count frequent subgraphs
- Design algorithms to characterize the structure of the network as a whole

Recommendations: Data

- Movie ratings:
  - Netflix prize dataset:
  - http://www.netflixprize.com/
- Yahoo Music ratings:
  - Yahoo Music user ratings of songs with artist, album and genre information
  - 717 million ratings
  - 136,000 songs
  - 1.8 users
- Restaurant reviews

Recommendations: Ideas

- Collaborative filtering:
  - Predict what ratings will user give to particular songs/movies, i.e., which sons will he/she like?
- Supplement the data with additional data sources:
  - Movies -- IMDB
  - Playlists from the web
  - Lyric (text of the song)
- Include taste, temporal component, diversity into the model

Many other ideas/datasets

- Stanford Search Queries
- New York Times articles since 1987
  - Article are manually annotated by subject categories and keywords
  - Entity or relation extraction
  - Extract keywords, predict article category

- Don't feel limited by these
- You can collect the dataset yourself
- And define the project/question yourself