Getting Ready to Start on Projects

- Next Wednesday: project selection, in class
- Teams of four members
- If you have a project idea, please email a 1-2 slide project overview to Rada by 02/09
- 02/11: Project selection -- identify a project from all the available projects
- More on timeline, requirements, etc. on 02/11
What is on People’s Mind? Geolocation on Yahoo Answers

- Goal: understand differences between questions asked by people from different countries
- Data: Yahoo Answers includes questions and answers divided into topics, and divided into countries. E.g., uk.answers.yahoo.com, sg.answers.yahoo.com, etc.
- Approach:
  - Collect data for at least three different English speaking countries, for several predetermined topics (e.g., Computers; Family; etc.
  - Build a classifier to automatically predict the country of the question
  - Analyze the differences using topic models and linguistic ethnography (tools already available)
- Evaluation:
  - Classification accuracy on test data
Course Webpage Classifier

• Goal: build a system that can automatically identify course webpages

• Data: webpages crawled from the Web, likely from university websites

• Approach:
  – Build a collection of likely-course webpages starting with a large number of university webpages
  – Manually annotate a subset (of at least 750-1000 pages) for whether they are course webpages or not
  – Build a classifier to automatically determine if a webpage is a course webpage or not

• Evaluation:
  – Classification accuracy on manually annotated dataset