11-741/11-641/11-441: Machine Learning for Text Mining

Introduction

Yiming Yang
Carnegie Mellon University
yiming@cs.cmu.edu

Instructor and Teaching Assistants

Faculty

- Yiming Yang
- yiming@cs.cmu.edu
- GHC 6717

TA's (office hours on the syllabus)

- Yuexin Wu
- Rose Catherine Kanjirathinkal

Outline

- Course Contents Overview
- Administrative Details

Big Data Era

New data are being created at an unprecedented rate.

Every hour we add

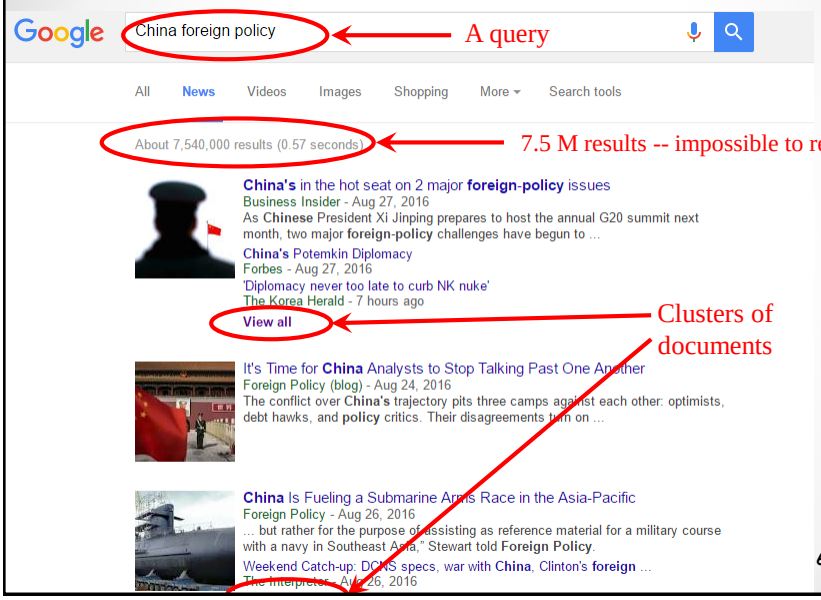
- 1 1,000 articles on Wikipedia.
- 236,000 videos on Youtube.
- 3,400,000 pictures on Flickr.
- 46 million posts on Twitter.

How do we handle such huge and diversified data?

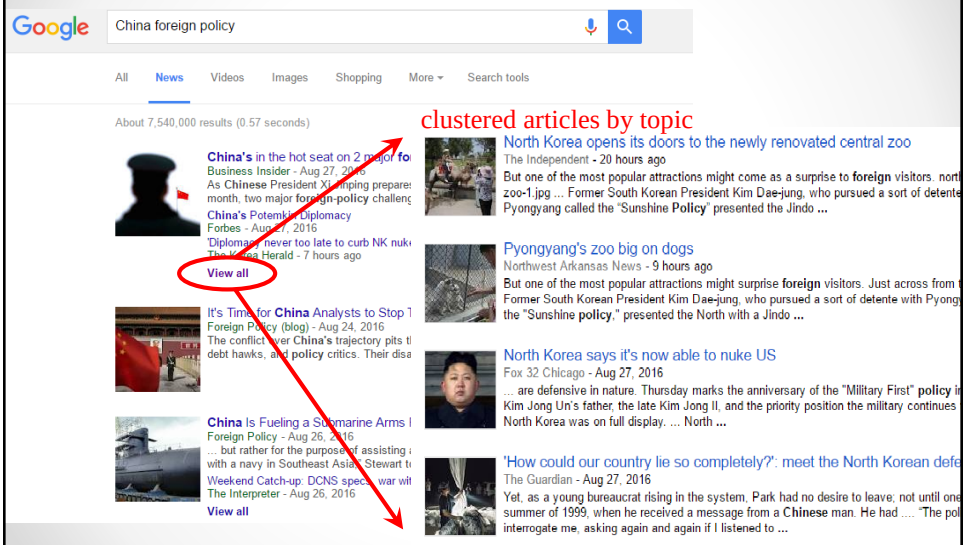
Text Mining Techniques

- **Clustering (of documents, users, shopping items, …)**
  - To discover latent groups, topics, social communities, etc.
- **Classification**
  - To label data using predefined classification taxonomies
- **Authority Detection**
  - To measure the social impacts of web sites, papers, people, etc.
- **Collaborative Filtering (Recommendation)**
  - To predict personal interests based on the tastes of similar users

Clustering web search results



Clustering web search results (cont'd)

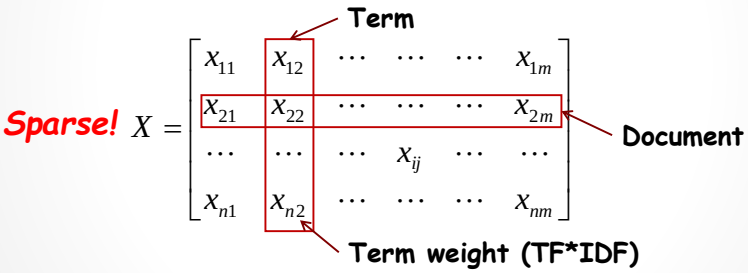


Common Clustering Algorithms

- K-means
- Single-pass
- Bottom-up Hierarchical (Agglomerative)
- Top-down Hierarchical
- Bipartite Reinforcement
- Probabilistic (soft) Clustering
- Spectral Clustering

K-means Clustering of Text Data

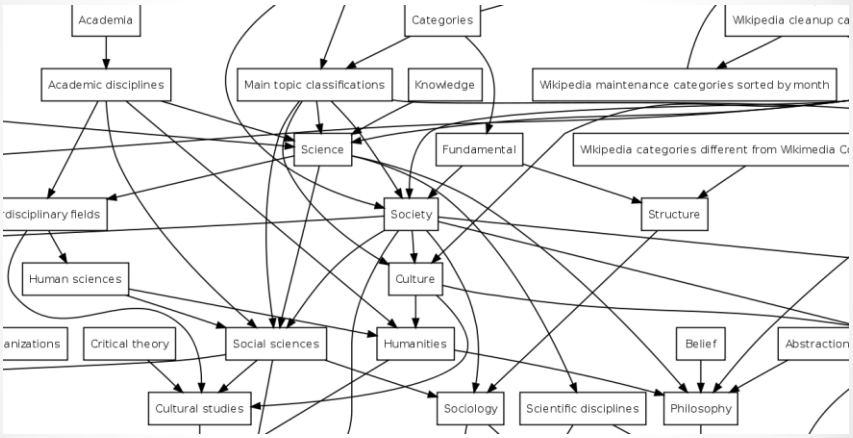
- Input: matrix X (n documents by m unique words)



- Output: K clusters of documents (rows) or words (columns)

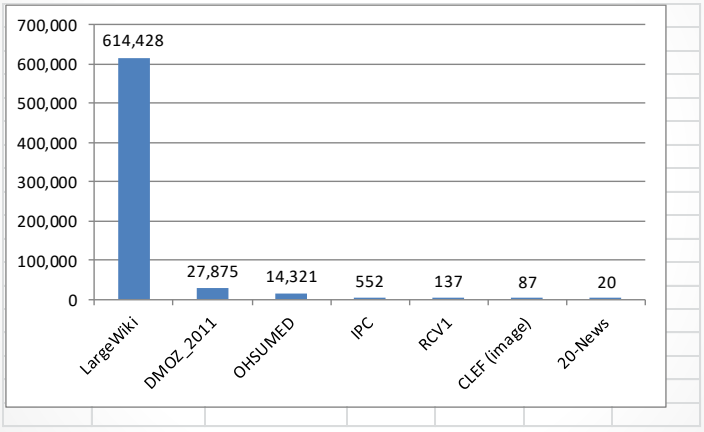
Text Classification

Wikipedia categories form a graph

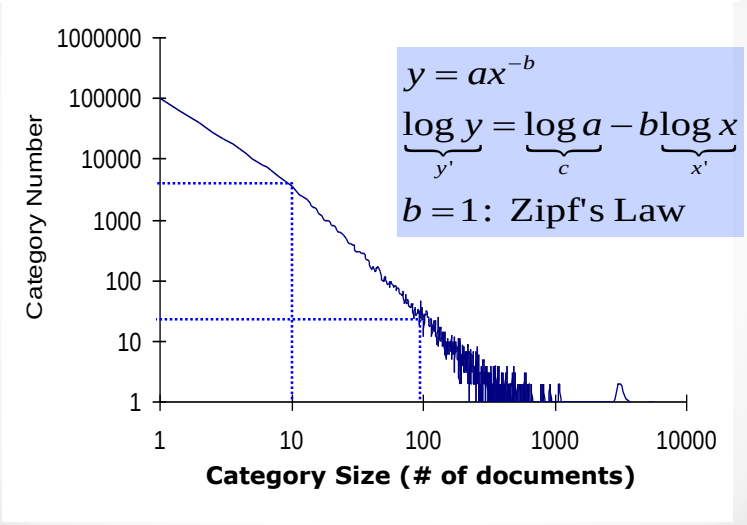


4.45 million articles in English Wikipedia, ~500,000 categories

Large-scale Classification Challenge



The Power Law Phenomenon
(Skewed distribution of categories over documents)



Many Classification Methods

- Rule-based Expert Systems (Hayes, 1990)
- Regression (linear, polynomial, logistic, …) (Furh'91; Yang'92)
- Nearest Neighbor methods (Creecy'92; Yang'94)
- Naïve Bayesian probabilistic methods (Lewis'92)
- Decision trees, symbolic rule induction (Apte'94)
- Neural networks, logistic regression (Wiener'95)
- Error Correcting Output Coding (Kong & Dietterich'95)
- Rocchio-style (Lewis et al., 1996)
- Support Vector Machines (Joachims'98)
- Boosting or bagging (Schapire'98)
- Hierarchical Language Modeling (McCallum'98)
- First Order Inductive Learning (Slattery'99)
- …

Statistical Classification: A Toy Example

(Graph in ESL by Hastie et al.)

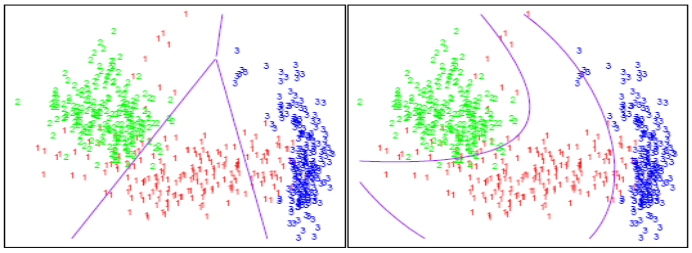
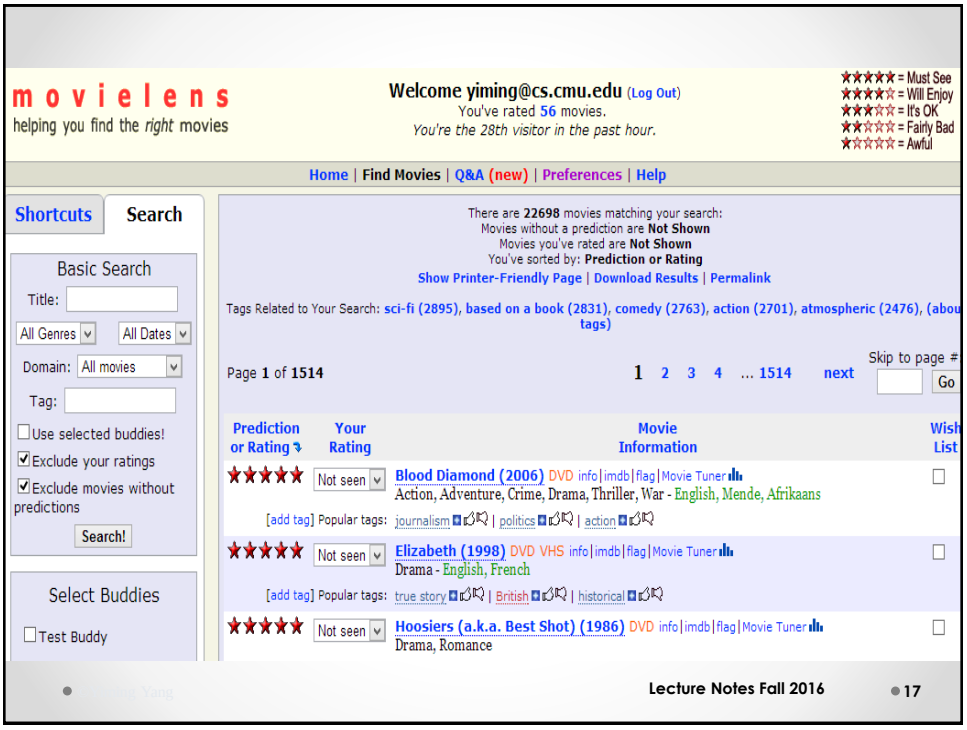


Figure 4.1: *The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space $X_1, X_2, X_{12}, X_1^2, X_2^2$. Linear inequalities in this space are quadratic inequalities in the original space.*

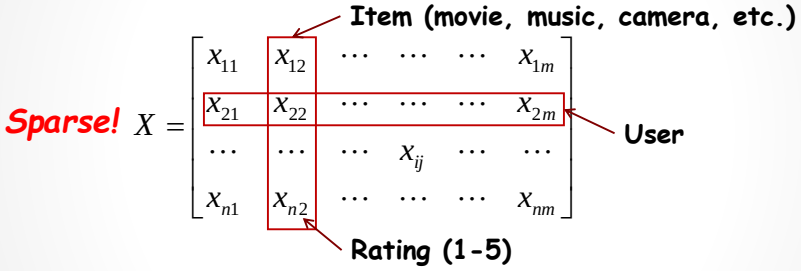
Text Mining Techniques

- **Clustering (of documents, users, shopping items, …)**
  - To discover latent groups, topics, social communities, etc.
- **Classification**
  - To label data using predefined classification taxonomies
- **Authority Detection**
  - To measure the social impacts of web sites, papers, people, etc.
- **Collaborative Filtering**
  - To predict personal interests based on the tastes of similar users



Collaborative Filtering (CF)

- Training Set: matrix X (n users by m items)

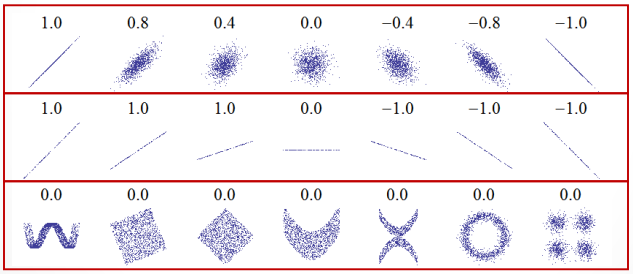


- Input: the profile (a few rated items) of a new user
- Output: the system predicted (ranked) list of items
- System predictions are based on the tests of similar users

CF: How do we measure the similarity of the tastes b/w two users (e.g., x and y)?

Pearson's Correlation Coefficient (PCC)

$$r_{xy} = z_x \cdot z_y \qquad -1 \le r_{xy} \le 1$$



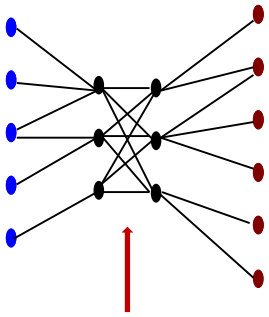
Full Range

Highly Correlated
|r| = 1

Not Correlated
r = 0

(a graph from Wikipedia)

Matrix Factorization: Introducing Hidden Layers



Dimension-reduced Representation of Users and Items

Topic Coverage

http://nyc.lti.cs.cmu.edu/classes/11-741/f16/index.html

- High-dimensional vectors & sparse matrices (1 lecture + HW1)
- Clustering (3 lectures + HW2)
- Link Analysis (2 lectures + HW3)
- Collaborative Filtering (3 lectures + HW4)
- Classification (3 lectures + HW5)
- Learning to Rank (2 lectures + HW6)
- Significance Testing (3 lectures + HW7)
- Others (SVD, Matrix Factorization, SGD, Deep Learning, 4 lectures)
- CPP (Capstone Project Proposal) (12 topics, more details later)

Administrative Details

Okay, that's the content … now for the administrative stuff

Support System: Resources

- Textbooks (available at the bookstore)
  - Primary: Introduction to Information Retrieval (IR), Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze, Cambridge University Press. 2008
  - Reference: Pattern Recognition and Machine Learning (ML), Christopher M. Bishop, Springer 2006
- Course materials via URL's
  - http://nyc.lti.cs.cmu.edu/classes/11-741/f16/index.html for homework assignments, lecture notes, copies of papers (when necessary) and data with restricted access to .cmu.edu (or via VPN)
  - https://piazza.com/class/irv5bgvra6b33u
  - http://www.cmu.edu/blackboard/

Course Organization

- ❑ **Lectures**
  - Problems (sub areas), Methods, Algorithms, Evaluations
- ❑ **Hands-on experience (homework)**
  - 5 programming assignments (Clustering, Link Analysis, etc.)
  - 2 problem solving exercises (basic matrix algebra, significance tests)
- ❑ **Weekly reading summaries**
  - Required for 11-741 students only (**but everyone is highly encouraged**)
- ❑ **Mid-term Exam: No**
- ❑ **Final Exam**
  - **Closed-book, no cheating sheets, no arrangement for exam-time exception**
- ❑ **CPP (Capstone Project Proposal)**
  - Team work (presentations + 4-page write up)

Difference among 741, 641 and 441

(More details in the syllabus)

- **11-741**
  1. Count as a PhD-level course (12 units) in LTI
  2. Reading summary per week is required
  3. CPP work/peer-review is required; HW 2-7
- **11-641**
  1. Count as a MS-level course (12 units) but not PhD-level in LTI
  2. Reading summary is not required
  3. CPP work/peer-review is required; HW 1-6
- **11-441**
  1. Count as UG course (9 units)
  2. Reading summary is not required
  3. CPP work is not required, but CPP peer-review is required; HW 1-6

HW	11-741 Weight	11-641 Weight	11-441 Weight
HW0	0%	0%	0%
HW1	0%	5%	10%
HW2	5%	10%	10%
HW3	10%	10%	10%
HW4	12.5%	12.5%	15%
HW5	12.5%	12.5%	15%
HW6	10%	10%	10%
HW7	5%	0%	0%

Grading

- Students in 11-741 (PhD level course) will be graded by 55% on 6 homework assignments (2-7), 17.5% CPP, 17.5% final exam and 10% on reading summaries.
- Students in 11-641 (MS level course) will be graded by 60% on 6 homework assignments (1-6), 20% CPP and 20% final exam. Reading summaries are not required but encouraged.
- Students in 11-441 (UG level course) will be graded by 70% on 6 homework assignments (1-6), 20% final exam, and 10% on peer-reviews of the CPP work (including oral presentations and written reports) by the students in 11-741 and 11-641. CPP work, except the peer-review part, is not required for 11-441 students. Reading summaries are also not required but encouraged.

CPP (Capstone Project Proposal)

1. Students are divided into teams, each focuses on one topic
2. Each student submits 3 candidate topics by TBD
3. Instructor/TA's assign one topic (3 papers) per team
4. Presentation on literature overview + new idea + proposed thorough work
5. Write-up proposal (4 pages)
6. Peer-reviewed evaluations on 4-6 (50%) + by TA's (50%) on 4-6 and review questions/comments
7. The entire team will receive the same evaluation score, so good coordination among team members is important for quality work.

Homework Policies

- All homework must be submitted via Blackboard
  - Due by 11:59 pm of the due date
- Late homework:
  - Deduct 10% for each day late.

Don't fall behind.

Cheating, Copying, Plagiarism, Etc

- You must be the author of everything that you submit for a grade
- Revising or modifying someone else's work does not make you the author
- It is okay to discuss homework with other students, share ideas, experience, and lessons learned
- Turn in the signed form (otherwise you will not be graded)

http://nyc.lti.cs.cmu.edu/classes/11-741/f15/Policy/cheating_policy_form.pdf

Cheating, Copying, Plagiarism, Etc

Penalties

- Usually failure of the course
- Possibly expulsion from the graduate program

If you are having problems meeting your deadlines

… submit the assignment late, or don't submit it at all

- Being late or taking a zero just lowers your grade
- Cheating causes you to fail the course