CS 1671/2071
Human Language Technologies

Session 8: Project match day, tf-idf, PPMI

Michael Miller Yoder

February 5, 2025

University of Pittsburgh | School of Computing and Information

Course logistics: quiz and homework

- Quiz on Canvas due **this Thu Feb 6**
  - What readings it covers is specified in the description on Canvas
- Homework 2 is due **Feb 20**
  - Build a text classification system to predict deception in a game (Diplomacy)
- Next project milestone: project proposal due Feb 28
  - Stay tuned for more details on that
  - If your data or task is a bit unspecified, book a meeting with Michael next week or later to discuss data

Overview: Project match day

Project match process

Review term-document and term-term matrices

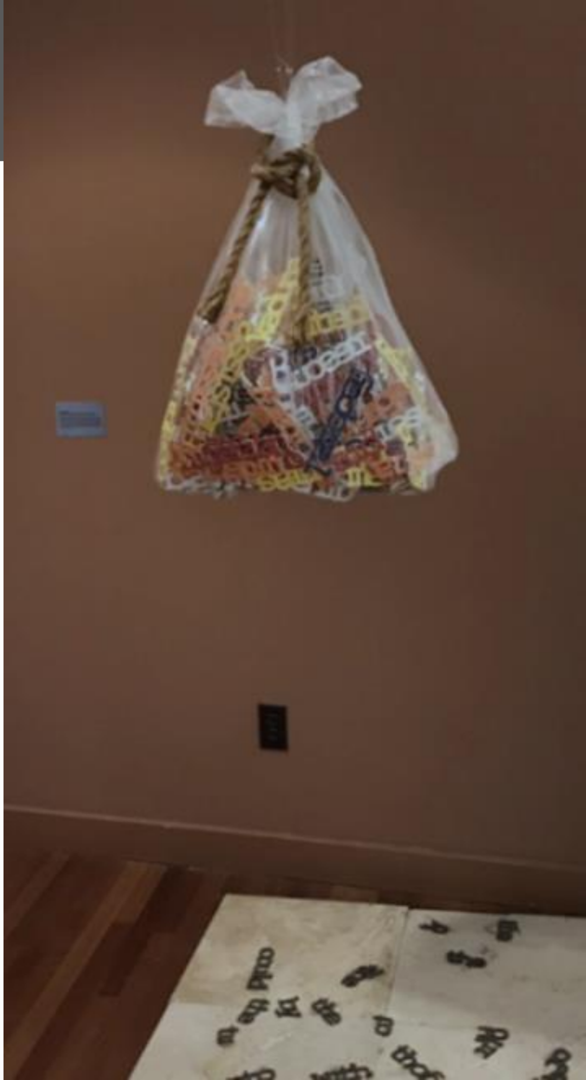
Project match

- Go to the spot in the room that corresponds to the project you are most interested in working on
  - We will likely do this for several rounds
- Goal: groups of 2-4 on projects
  - Groups of 3 or 4 students are ideal

Review activity

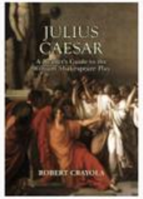
For term-document and term-term matrices:

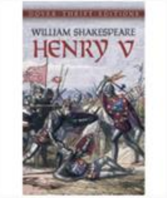
1. What do the dimensions (numbers of rows and columns) correspond to?
2. What does the value in each cell mean?



Term-document matrix

- Each cell is the count of term t in a document d ($tf_{t,d}$).
- Each document is a count vector in $\mathbb{N}^V$, a column below.

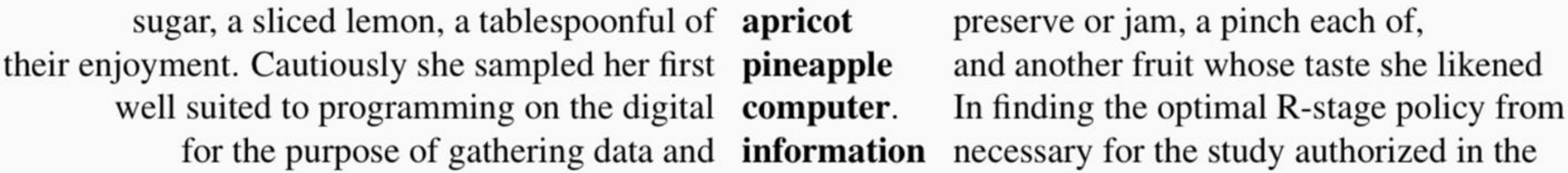




	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	6	117	0	0

Slide credit: David Mortensen

Sample Contexts of ±7 Words



sugar, a sliced lemon, a tablespoonful of	**apricot**	preserve or jam, a pinch each of,
their enjoyment. Cautiously she sampled her first	**pineapple**	and another fruit whose taste she likened
well suited to programming on the digital	**computer**.	In finding the optimal R-stage policy from
for the purpose of gathering data and	**information**	necessary for the study authorized in the

	aardvark	digital	data	pinch	result	sugar ...
⋮						
apricot	0	0	0	1	0	1
pineapple	0	0	0	1	0	1
computer	0	2	1	0	1	0
information	0	1	6	0	4	0
⋮						

Slide credit: David Mortensen