Text Mining and Analytics

Session 1: **Introduction**

Instructor: Behrooz Mansouri
Spring 2024, University of Southern Maine

Welcome Note!

Welcome to the exciting world of text mining and analytics!

Thrilled to have you on board as we embark on a journey to explore the vast and valuable insights hidden within the realm of textual data

Text mining, also known as text analytics, is a powerful approach that involves extracting meaningful patterns, information, and knowledge from unstructured text data

In a world where an enormous amount of information is generated through texts, such as emails, social media, articles, and more, the ability to analyze and derive insights from this data is becoming increasingly crucial

Welcome Note!

Welcome to the exciting world of text mining and analytics!

Thrilled t ore the vast and

Text mini ch that involves ge from unstructu

ChatGPT Generated!

In a world where an enormous amount of information is generated through texts, such as emails, social media, articles, and more, the ability to analyze and derive insights from this data is becoming increasingly crucial

What is Text Mining and Analytics?

Text Mining≈Text Analytics

Turn text data into high-quality information and then actionable knowledge

Related to text retrieval

- Can be preprocessor for text mining
- Needed for knowledge

What is Data?

- Data

  Raw representation, symbols/signs

- Information

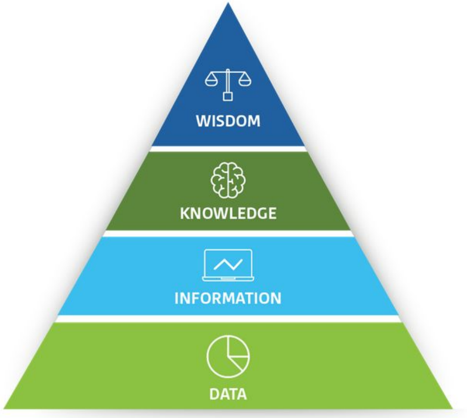
  Processed data that is useful to answer questions (meaning of data in context) – Linked data

- Knowledge

  Organized information that can be acted upon

- Wisdom

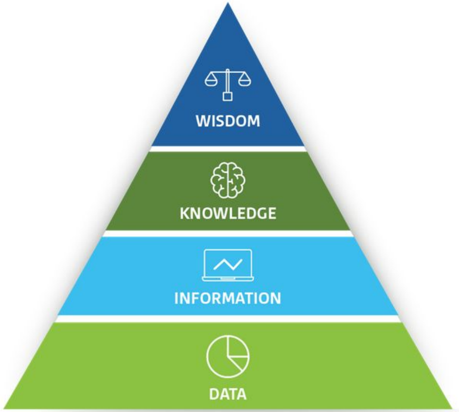
  Applied knowledge



DIKW Pyramid

What is Data?

- Data
  - – 100%, 85%, 70% …
- Information
  - – Course grade: 100%, 85%, 70%, …
- Knowledge
  - – If you have a grade above 93%, you will get A
- Wisdom
  - – Spend enough time on the course to get A!



DIKW Pyramid

What is Text Mining?

IBM: Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights

Amazon: Text analysis is the process of using computer systems to read and understand human-written text for business insights

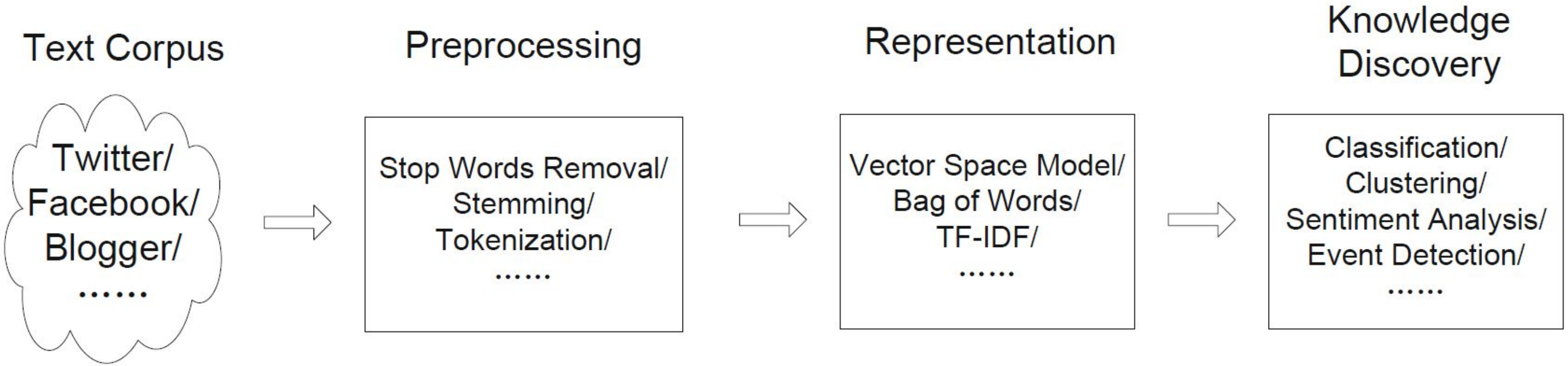
- Text analysis software can independently classify, sort, and extract information from text to identify patterns, relationships, sentiments, and other actionable knowledge

“Text Mining”

Marti Hearst (2003)

- Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources
- A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation
- In search, the user is typically looking for something that is already known and has been written by someone else
- Further Reading: Untangling Text Data Mining

Traditional Framework for Text Analytics



Text Corpus

Twitter/
Facebook/
Blogger/
……

Preprocessing

Stop Words Removal/
Stemming/
Tokenization/
……

Representation

Vector Space Model/
Bag of Words/
TF-IDF/
……

Knowledge Discovery

Classification/
Clustering/
Sentiment Analysis/
Event Detection/
……

Text vs. Non-Text Data

Text, usually generated by humans; which can be subjective (based on perspectives)

Real World → **Sensor** → **Data**

Weather → Thermometer → 10^{O}C, 15^{O}C

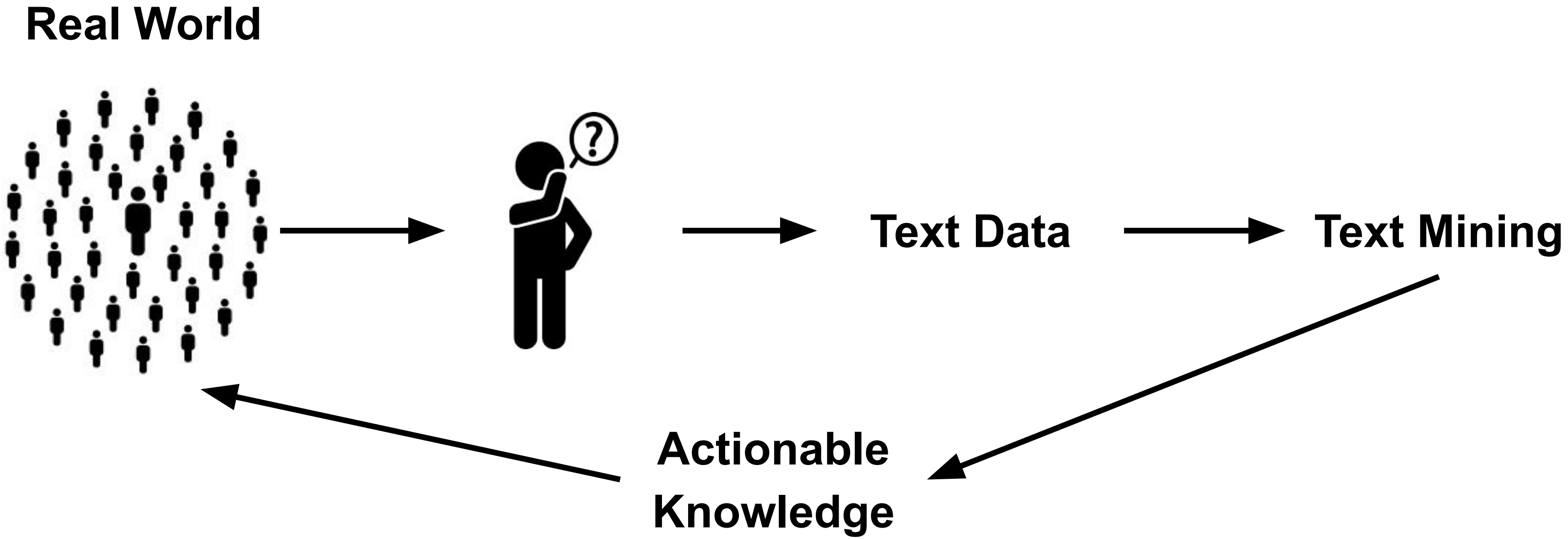
Location → Geo Sensor → 40^{O}S, 110^{O}W

Text vs. Non-Text Data

Text, usually generated by humans; which can be subjective (based on perspectives)

Real World	→	**Sensor**	→	**Data**
Weather	→	Thermometer	→	10^{O}C, 15^{O}C
Location	→	Geo Sensor	→	40^{O}S, 110^{O}W
Perceive	→	Human Sensor	→	Text data

The Cycle



Real World

Text Data

Text Mining

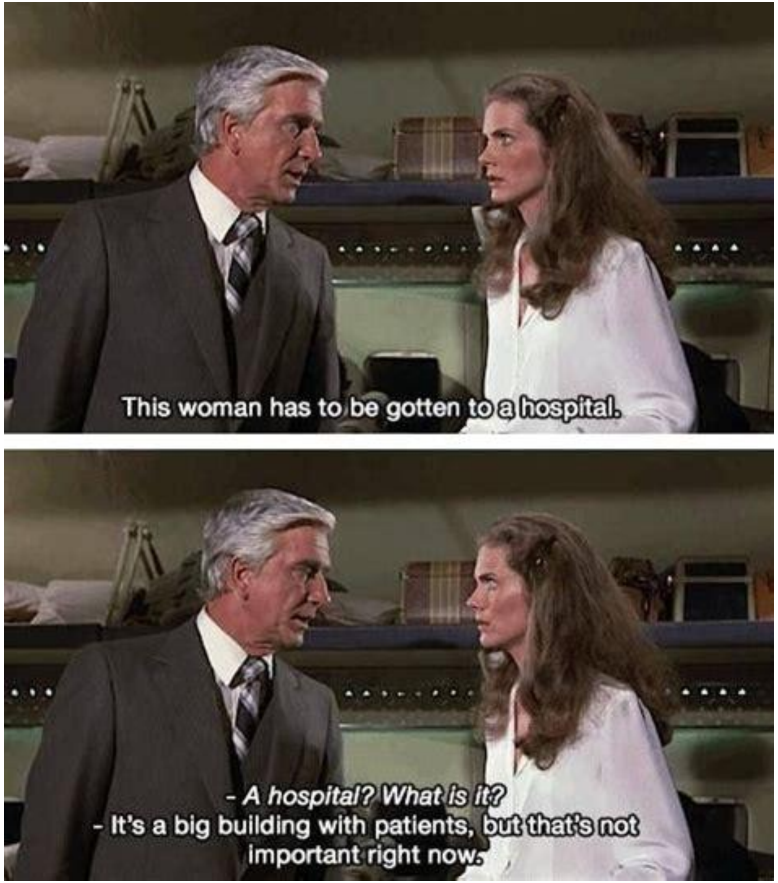
Actionable Knowledge

That shouldn’t be hard!

Ambiguity

Ambiguity: the capability of being understood in two or more possible senses or ways

Models and algorithms in this course are ways to resolve or disambiguate these ambiguities


This woman has to be gotten to a hospital.
- A hospital? What is it?
- It's a big building with patients, but that's not important right now.

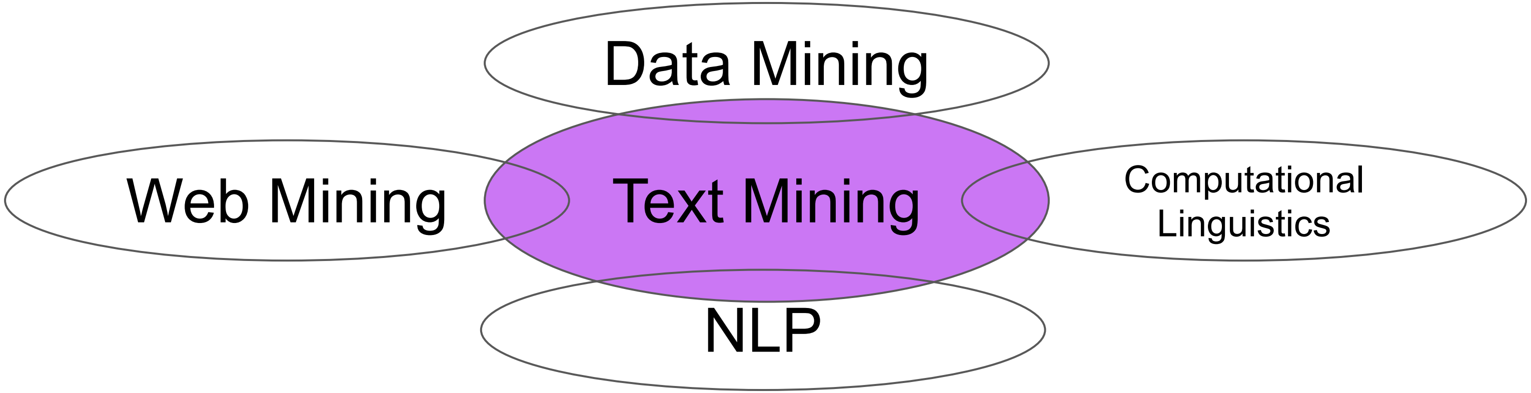
Text Mining vs. Others

Data mining: in text mining, patterns are extracted out of textual data

Web mining: the web sources are usually structured (semi-structured)

Computational linguistics: computational methods used just as in scientific disciplines like computational biology, for linguistics

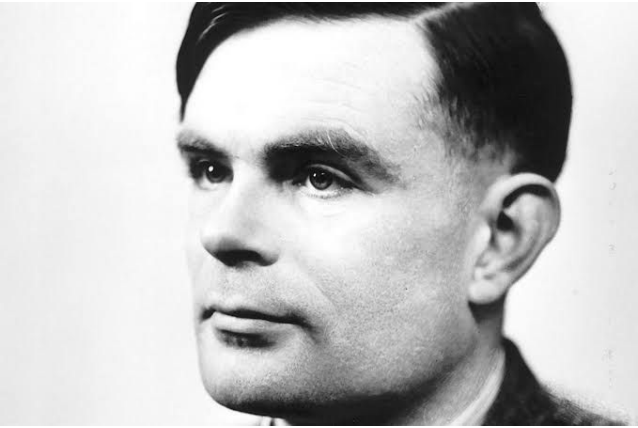
Natural language processing: focused on the design and analysis of computational algorithms and representations for processing natural human language



History of Text Mining & Analytics

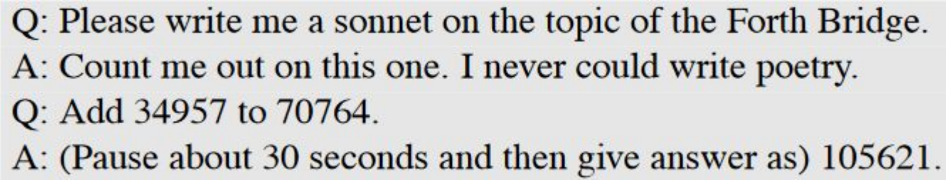
Turing Test (Imitation Game)

“Computing Machinery and Intelligence”
Mind, Vol. 59, No. 236, pp. 433-460, 1950



I propose to consider the question
"Can machines think?”...
We can only see a short distance ahead, but we can see plenty there that needs to be done

In Turing’s game, there are three participants: two people and a computer.
One of the people is a contestant who plays the role of an interrogator. To win, the interrogator must determine which of the other two participants is the machine by asking a series of questions via a teletype. The task of the machine is to fool the interrogator into believing it is a person by responding as a person would to the interrogator’s questions. The task of the second human participant is to convince the interrogator that the other participant is the machine and that she is human.



Q: Please write me a sonnet on the topic of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764.
A: (Pause about 30 seconds and then give answer as) 105621.

1950 – 1970

Mid 1950’s – Mid 1960’s: Birth of computational linguistics

- At first, people thought it is easy! Researchers predicted that “machine translation” can be solved in 3 years or so
- Mostly hand-coded rules / linguistic-oriented approaches
- The 3-year project continued for 10 years, but still no good result, despite the significant amount of expenditure

Mid 1960’s – Mid 1970’s: A Dark Era

- After the initial hype, a dark era follows
- People started believing that machine translation is impossible, and most abandoned research for computational linguistics

1970 – 2000

1970’s and early 1980’s – Slow Revival of NLP

- □Some research activities revived, but the emphasis is still on **linguistically oriented**, working on small toy problems with weak empirical evaluation

Late 1980’s and 1990’s – Statistical Revolution!

- □By this time, the computing power increased substantially
- □Data-driven, statistical approaches with simple representation win over complex hand-coded linguistic rules
- “Whenever I fire a linguist, our machine translation performance improves.” (Frederick Jelinek, 1988)

2000’s – Statistics Powered by Linguistic Insights

- □With more sophistication with the statistical models, richer linguistic representation starts finding a new value
- Earliest workshops on text mining:
  - ICML-99 Workshop on Machine Learning in Text Data Analysis
  - KDD-2000: Workshop on Text Mining

Recent Years

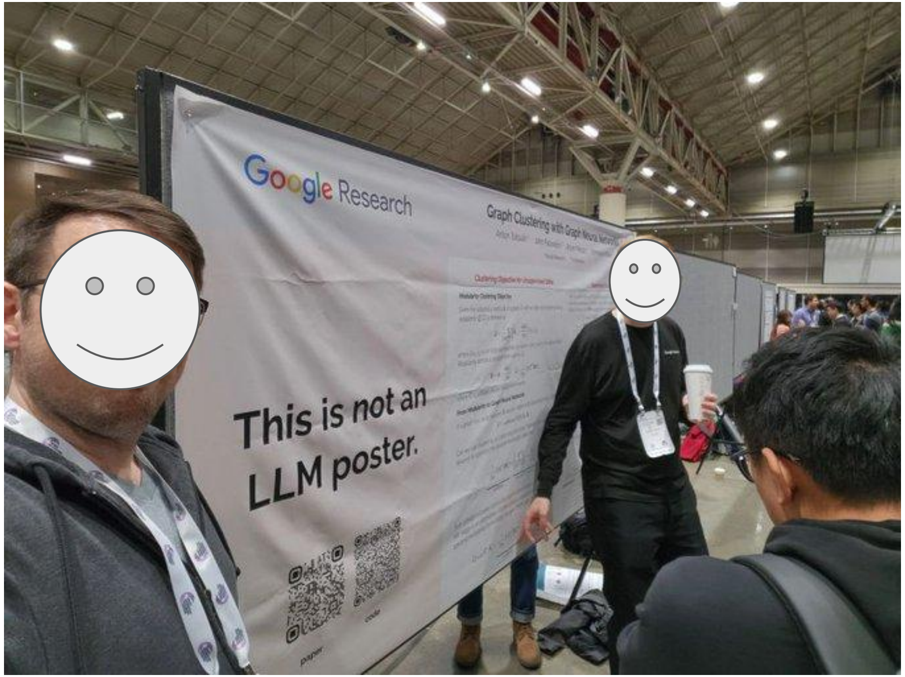
2010’s – Emergence of embedding model and deep neural networks

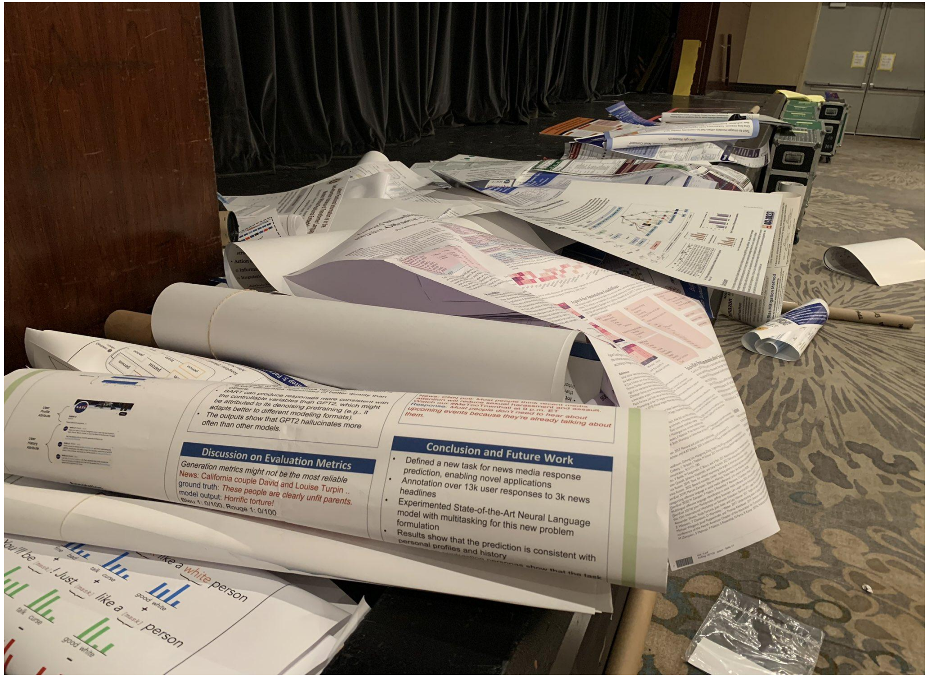
- □Several embedding models for text using neural networks and deep neural networks were proposed including Word2Vec, Glove, fastText, Elmo, BERT, COLBERT, GTP[1-3.5]
- New techniques brought attention to more complex tasks

2020’s – Large language model (Soon to be dark era!)

- Llama, Chat-GPT, Bard, Orca, Mistral, Mixtral, …
- Why dark era?

Do an Actual Research!


Google Research
Graph Clustering with Graph Neural Networks
This is not an LLM poster.
paper
code



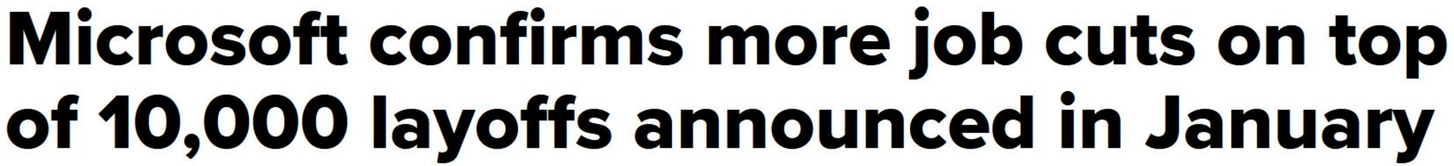
Large Language Models are not the only research topic!

Training with more data and additional transformer layers is not research!

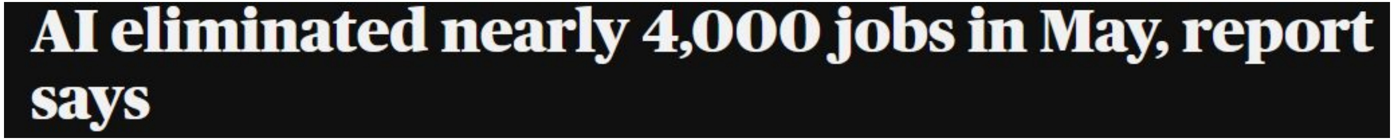
Do an Actual Research!

1. You must do a literature review and be able to link the works and discuss their pros and cons
2. Before developing your proposed model, you should write your hypothesis
   a. What research questions are you trying to answer?
   b. How would you design your experiment to answer that question?
   c. Upon what result will you accept your hypothesis?
3. No matter what results you will get (accept/reject), you need to **analyze your results** on both positive and negative sample and then decide what to do next!

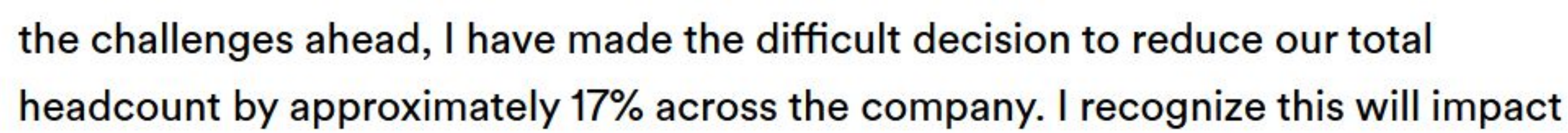
Are we Losing Jobs?



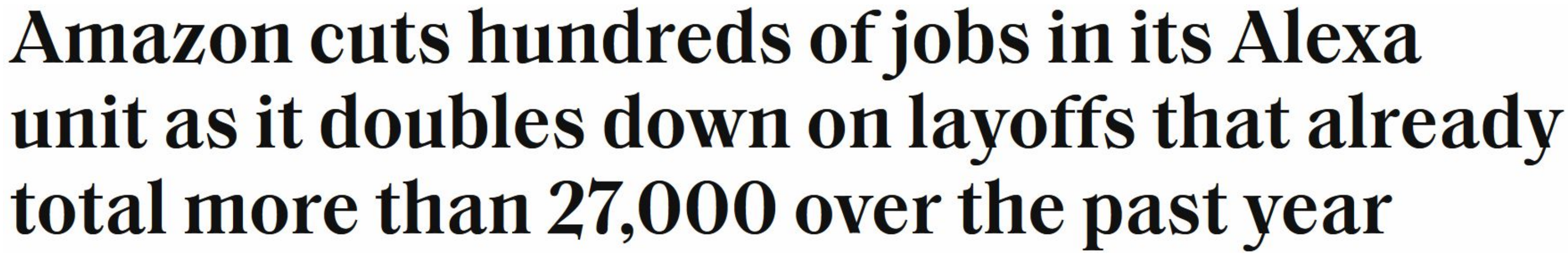
Microsoft confirms more job cuts on top of 10,000 layoffs announced in January



AI eliminated nearly 4,000 jobs in May, report says

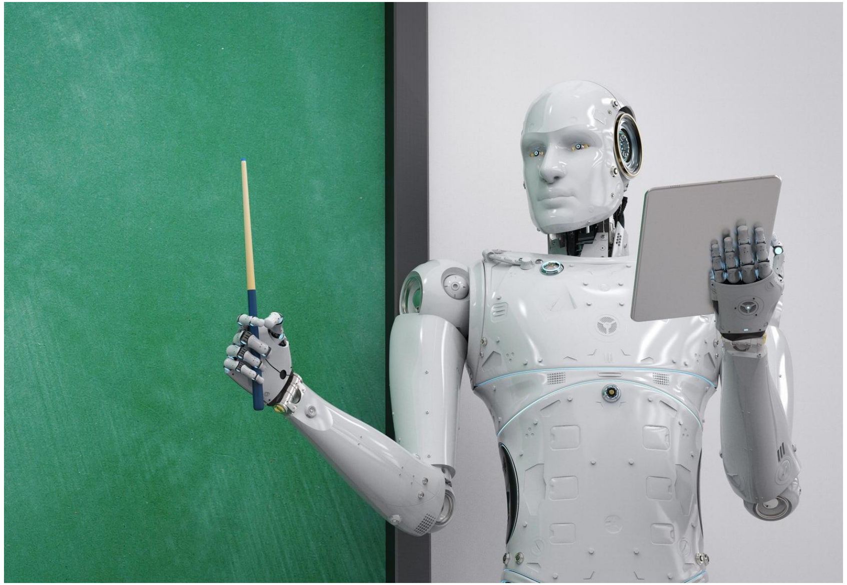


the challenges ahead, I have made the difficult decision to reduce our total headcount by approximately 17% across the company. I recognize this will impact



Amazon cuts hundreds of jobs in its Alexa unit as it doubles down on layoffs that already total more than 27,000 over the past year

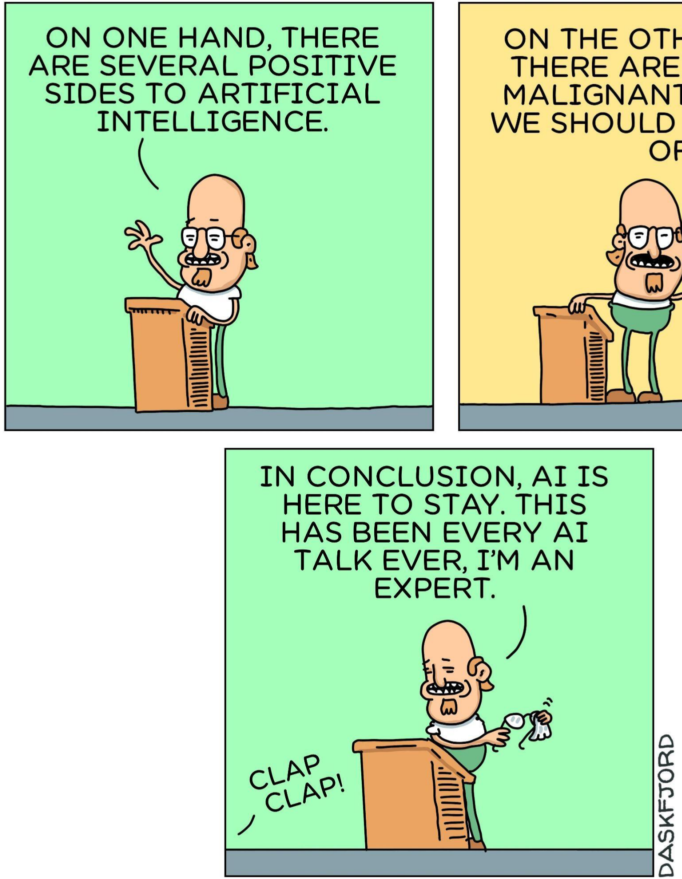
What Jobs?



AI replacing professors?

Do we still have elevator drivers?

Learn how to be a Computer Scientist!


ON ONE HAND, THERE ARE SEVERAL POSITIVE SIDES TO ARTIFICIAL INTELLIGENCE.
ON THE OTHER HAND, THERE ARE SEVERAL MALIGNANT ASPECTS WE SHOULD BE AWARE OF.
IN CONCLUSION, AI IS HERE TO STAY. THIS HAS BEEN EVERY AI TALK EVER, I'M AN EXPERT.
CLAP CLAP!
DASKFJORD

Almost everyone claiming to be an AI expert!

Because of the massive, often quite unintelligible publicity that it gets, artificial intelligence is almost completely misunderstood by individuals outside the field.

Even AI's practitioners are somewhat confused about what AI really is (Schank, 1987)

About the Course

Course Description

The course provides a complete overview of the state of the art and research perspective in the field of text mining and analytics, with an introduction to some relevant and correlated problems

The course has two parts:

- In the first half of the semester, the focus is on understanding general solutions to text mining problems
- In the second part, we focus on specific text mining problems and apply techniques we have learned

Topics covered in this course includes:

- Web scraping and data cleaning
- Language modeling
- Neural networks
- Large language models
- Text classification and clustering
- Sentiment analysis
- Semantic parsing
- Summarization
- Machine translation
- Sequence labeling
- Information retrieval and question answering

Course Learning Outcomes

- Describe the fundamental concepts and techniques of text mining
- Analyze the performance of a text mining system by applying the proper evaluation measures
- Design and implement real applications using text mining techniques
- Analyze large volume text data generated from a range of real-world applications

Course Assessment

- 5 assignments helping students to prepare for the course project
- 5 quizzes (bi-weekly), open notes
- Course project broken down into three phases

Assessment	**Value**
Class Activities	5%
Quizzes	15%
Assignments	20%
Project: Phase 1	15%
Project: Phase 2	20%
Project: Phase 3	25%

Course Schedule

Date	Session	Topic	Quiz	Assignments/Project	Due
01/17	1	Introduction			
01/22	2	Web Scraping and Data Cleaning		Assignment 1	
01/24	3	N-gram and Language Models			
01/29	4	Vector Semantics	Quiz 1		Assignment 1
01/31	5	Naive Bayes Classification			
02/05	6	Neural Networks			Project: Team and Topic
02/07	7	Backpropagation and Pytorch		Assignment 2	
02/12	8	Word2Vec	Quiz 2		
02/14	9	Recurrent Neural Networks			
02/21	10	LSTM and Attention			Project Progress Report
02/26	11	Transformers		Assignment 3	Assignment 2
02/28	12	Large Language Models	Quiz 3		
03/04	13	Project Part I			
03/06	14	Project Part I			Assignment 3

Course Textbooks

- Related content and textbooks will be provided by the instructor through course.maine.edu
- Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (Third Edition) by Daniel Jurafsky and James H. Martin
  Draft available online: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf and here
- A Programmer's Guide to Data Mining: http://guidetodatamining.com/
- Mining Text Data: https://link.springer.com/book/10.1007/978-1-4614-3223-4
- Natural Language Processing with Python by Steven Bird, Ewan Klein, and Edward Loper: https://www.nltk.org/book/
- Hugging Face Tutorial: https://huggingface.co/course/chapter1/1

Contacting the Instructor

Instructor: Dr. Behrooz Mansouri
Email: behrooz.mansouri@maine.edu
Website: https://cs.usm.maine.edu/~behrooz.mansouri/
Course Website: https://cs.usm.maine.edu/~behrooz.mansouri/courses/TM2024.html

Student Meeting hours:
Monday 14 – 15, Tuesday 10:00 – 12:00
Room 224 Science Building – Dubyak Center – AIIR Lab

Email me with Subject: **TM-MainMessage**

e.g., TM-Question About Assignment 1

- Emails are responded mostly during the students hours
- Emails after 4 PM are replied to the earliest by the next day
- Emails after Friday noon are replied to earliest by Monday
- Emails should not be longer than two paragraphs – **keep it short and to the point**

Class Policies

- More than 3 absences will lead to elimination from the class – “L” as the grade
- Late submission policy (48 hours): 20% off
  - No submission will be accepted after 48 hours
- No late arrivals
  - late arrival = absence
- No phone, no laptops at no time
  - Except for coding sessions
- Plagiarism will result in F and will be dealt with according to the University of Southern Maine policies
- Students are allowed (and encouraged) to use online resources and existing codes for learning purposes

Contacting Classmates

Slack Channel (for student’s communication only):

https://join.slack.com/t/textmining-talk/shared_invite/zt-2ap4y9k79-k~_BwmFD2BKblQsPjpBDnw

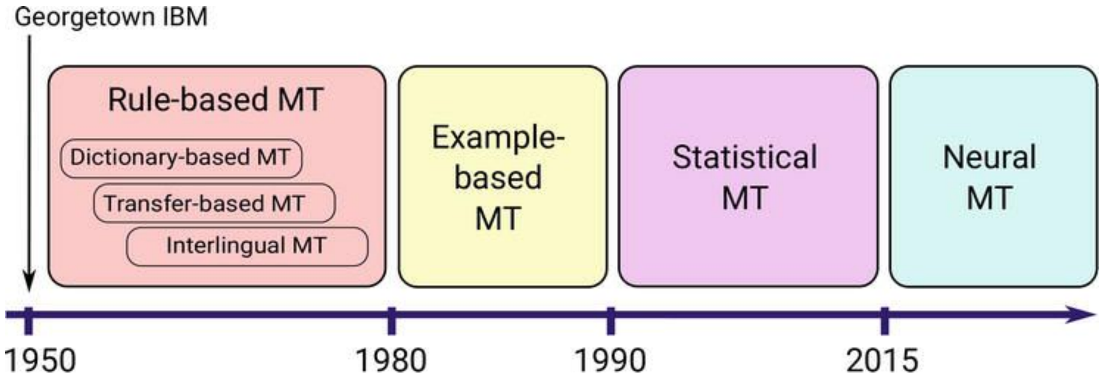
- **Please use the channels accordingly**
- **Inform your classmate on using GPUs on server channel**

Tasks/Applications in Text Mining and Analytics

A few of the Text Mining Tasks

- We will see a set of tasks in text mining that you can consider for your course project!
    - CLEF 2024 Labs
        - Each student is responsible for one unique task
    - You can choose topics of your own, but need instructor’s approval
    - You can consult the instructor for other topics

Machine Translation



Low resource languages can be challenging?

6,800 living languages
600 with written tradition
100 spoken by 95% of population

Question Answering

IBM-Watson Defeats Humans in "Jeopardy!"

CheckThat!: https://checkthat.gitlab.io/clef2024/

Task 1: Check-Worthiness

- Assess whether a given statement (e.g., a tweet), is worth fact-checking; does it contain a verifiable factual claim?
- Paper: https://ceur-ws.org/Vol-3497/paper-019.pdf

Task 2: Subjectivity

- A binary classification tasks in which systems have to identify whether a text sequence is subjective or objective
- Paper: https://ceur-ws.org/Vol-3497/paper-020.pdf

Other tasks:

- Task 3 Persuasion Techniques
- Task 4 Detecting hero, villain, and victim from memes
- Task 5 Authority Evidence for Rumor Verification
- Task 6 Robustness of Credibility Assessment with Adversarial Examples

SimpleText: http://simpletext-project.com/2023/clef/

Task 1: Content Selection

- Retrieving passages to include in a simplified summary
- For a given topic, find all the relevant passages from a scientific corpus

Task 2: Complexity Spotting

- Identifying and explaining difficult concepts
    - Decide which terms require explanation, and provide explanations

Task 3: Text Simplification

- Simplify scientific text

Paper: https://ceur-ws.org/Vol-3497/paper-239.pdf

Exist: http://nlp.uned.es/exist2024/

sEXism Identification in Social neTworks

- **Task 1:** Sexism Identification in Tweets
  - Decide whether a given tweet contains sexist expressions or behaviors
- **Task 2:** Source Intention in Tweets
  - Categorize the message according to the intention of the author
  - DIRECT, Reported, Judgmental
- **Task 3:** Sexism Categorization in Tweets
  - Predefined categories: Ideological and Inequality, Stereotyping and dominance, …
- **Task 4:** Sexism Identification in Memes
  - Binary classification task consisting on deciding whether a given meme is sexist
- **Task 5:** Sexism Categorization in Memes

ERisk: https://erisk.irlab.org/

Early risk prediction on the Internet

- **Task 1**: Search for symptoms of depression
  - Ranking sentences from a collection of user writings according to their relevance to a depression symptom
- **Task 2:** Early Detection of Signs of Anorexia
- **Task 3**: Measuring the severity of the signs of Eating Disorders
  - Estimating the level of features associated with a diagnosis of eating disorders from a thread of user submissions

QuantumCLEF: https://qclef.dei.unipd.it/

Research in Quantum Computing (QC)

- Development of very powerful devices
- Able to tackle even realistic problems
- Promising a great improvement in terms of performance
  - For some computationally intensive tasks

The objective of QuantumCLEF is to design and develop an evaluation infrastructure for QC algorithms and, in particular, for Quantum Annealing

QuantumCLEF addresses two different tasks involving computationally-intensive problems that are closely related to the Information Access field: Feature Selection and Clustering

Nicola Ferro Slides from FIRE 2023: link

Task of your Choice: Example 1- Math Domain

- Does a math question, needs clarification?

Counting bounded integer solutions to $\sum_i a_i x_i \leq n$

CQ: Do the x_i need to be all positive, or just non-negative?

- Multilingual math search
- Math-word problem

Problem: Nick has \$100, Shea has \$10. How much money they have in total?
Equation: x = 100+10
Solution: \$110

Task of your Choice: Example 2 - Legal Domain

Artificial Intelligence: Historical Context and State of the Art: https://link.springer.com/chapter/10.1007/978-3-031-41264-6_1

Tasks:

- Question Answering and Conversational Systems
- Summarization
- Information Extraction: Keyword or Keyphrase extraction
- Legal case outcome predictions supported on textual evidence

Where to find Tasks and Test Collections?

KDD: Knowledge Discovery and Data Mining https://kdd2024.kdd.org/

CIKM: Conference on Information and Knowledge Management https://www.cikm2024.org/

WWW: The Web Conference https://www2024.thewebconf.org/

TREC: Text Retrieval Conference https://trec.nist.gov/pubs/call2023.html

Fire: Forum for Information Retrieval http://fire.irsi.res.in/fire/2023/home

Where to find Tasks and Test Collections?

EMNLP: Conference on Empirical Methods in Natural Language Processing

ACL: Association for Computational Linguistics

NAACL: Annual Conference of the North American Chapter of the Association for Computational Linguistics

https://www.aclweb.org/portal/content/joint-call-workshops-proposals-eaclaclnaaclemnlp-2024

CoNLL: Conference on Computational Natural Language Learning https://www.conll.org/2023

COLING: International Conference on Computational Linguistics https://lrec-coling-2024.org/

CLEF: Conference and Labs of the Evaluation Forum https://clef2024.imag.fr/

SemEval: Workshop on Semantic Evaluation https://semeval.github.io/SemEval2024/

Next Session

Web Scraping and Data Cleaning

In the next session, we will explore

- Sources for textual data and web scraping
- Tools and techniques for data cleaning

To do:

- Explore CLEF labs
- Join Slack Channels and find teammates!