The Class

- Extensive experience in managing Big Data in projects funded by DARPA
- What is Big Data??
- Digital File Formats
  - Detecting them
  - Extracting data and metadata
  - Language identification and translation
- Classifying Files
- Information Clustering and Similarity
- Understanding Metadata and Evaluating it
- Language Identification and Detection
- Big Data Sets
- Named Entity Recognition (NER)…and more

Class Typical Format

- 1 class per week, so
  - Try not to miss it – if you miss it, make a buddy or two that can help inform you of what's going on
    - If you miss class, step one is to inform Professor and TA
- Format
  - Weekly lectures(s)
  - Presentations (beginning in week 2) and discussion
  - Periodic Breakout groups and report back (in-class participation)
  - Supplemental videos and ancillary material for review

Roadmap / Requirements

- Python, Java
  - Moderate proficiency, more Python than Java
- Package Management Systems
  - PIP, Conda, Setuptools, Maven, Ant/Ivy
- Course Includes
  - Understanding of core theory
  - Intro to Tika as a Framework for Content Detection & Analysis
  - Python, Java programming for Content Detection and Analysis using Tika, ElasticSearch™, Solr, Nutch and Apache Hadoop™
  - Group work and assignments
  - An Individual Presentation on one of the Research Papers assigned
  - Participation

What is participation? Good ideas

- **Each class there will be *in class* group activity**
  - **Don't miss class**
  - **Class is recorded, but lecture slides are not posted online**
    - **Key point: take notes!**
  - **You will participate with your in class group and there will be in-class deliverables**
    - **DEN students will participate remotely**
- **Speaking and asking questions professionally and politely**
  - **Asking questions to understand concepts and help you and your fellow classmates**

What we'll cover

- **Develop sufficient proficiency in theory and application of Data Science including being capable of automatically identifying files, extracting information from them including their text and metadata and language.**
- **Develop sufficient proficiency in Data Science techniques with Large Data sets collected from the Web and other places (Intranet, Science Data Sets, Public Data Sets).**
- **Develop sufficient proficiency in Java and Python to write and execute software that automatically extracts text and metadata from large data sets.**

- **You will get to leverage one or a combination of several Apache software technologies**
  - **Tika, Nutch, Lucene, Solr, Hadoop, HBase, Hive, Cassandra, etc.**
- **You will make a significant contribution to one or more of the above communities**
- **Deliverables**
  - **A report for each assignment**
  - **Source code and final demonstration to me at end of class**
  - **You can do some of the assignments in teams, we'll talk about this later.**

≤ 5 persons, 2wk weeks to finish

Individual Presentation

- You will be assigned one of the required paper readings in the class
- These are full-up research papers, of varying length, on topics in Informatics, Data Science, Machine Learning, Content Detection & Analysis, Search, Big Data, Analytics, Information Retrieval, and other areas
  - You can't read them the night before and do a presentation, not a good idea

- What I'm looking for (~20 minutes of presentation, with ~5 mins questions at the end)
  - You understood the paper
  - Discussion of related work and background
  - Discussion of why should I care about the topic
    - And more importantly why your fellow classmates should care
  - Relation of your paper to the lecture slides I gave on the topic
  - *Simple* summarization and description of the algorithm and/or technology introduced in the paper
  - What were the results/contributions/conclusions of the paper
  - Your evaluation of Pros of the paper
  - Your evaluation of Cons of the paper

- What I'm *NOT* looking for
  - Plagiarism
  - Repetition
  - Cutting/Pasting out of the paper
  - Regurgitation
  - You to follow the EXACT set of bullets that I gave on the prior slide
- You should be looking to be innovative – show the class and me that you really understood what was in the paper
  - Treat it like a conference presentation

Exam

- You will have a single exam given in week 9 of the course per the syllabus
  - Haven't solidified the room as of yet, but will likely be the same as the class
  - Exam will not be comprehensive (will only cover the material through the first 8 weeks of the course)
- Will include both an in-class question/answer component with multiple parts and additional material

write about the topic 相关的简答

A Few Updates

- Reminder about the book we will be using
  - Chris A. Mattmann, and Jukka Zitting. Tika in Action, 256 pages. New York: Manning Publications, November 2011. ISBN: 9781935182856.
- (Available on Amazon)
- !!Syllabus has changed!!
  - Syllabus assigned readings will now be shown on the date they are due
  - See here: https://sites.google.com/view/dsci-550-spring2022/home
  - Here is some clarity on assignment due dates.
    - Assignment 1 due date: Friday, March 4th, 2022 (day after class, 11:59:59 p.m. PT)
    - Assignment 2 due date: Friday, April 1st, 2022 (day after class, 11:59:59 p.m. PT)
    - Assignment 3 due date: Tuesday, May 10 @ 3:59:59 p.m. PT (note different day of week & time)
    - Also: presentations assigned on the last day of class can be sent as a video on May 10th 3:59:59 p.m. PT.
  - The final exam due dates in May give all of you more time to finish you final assignment and final presentation, respectively, to make up for the extremely busy time you will otherwise have at the end of the semester.