**ANALYTICS FOR UNSTRUCTURED DATA**

**MIS 284N Unique 04605/04610 Fall 2021**

**Cohort A: T Th 8:30 – 10:30 a.m.**

**Cohort B: T Th 10:30 a.m. - 12:30 p.m.**

**GSB 3.138**

**Important note about Zoom sessions:** Even though I will teach this course in the classroom (unless directed otherwise by the University), each class will also have its own Zoom session for flexibility. The Zoom links will be announced on Canvas before each class. If you attend a Zoom session, kindly make sure you sign your name at the end of class in the chat window for me to keep track of attendance.

**Instructor:** Professor Anitesh Barua

**Office:** CBA 5.236

**Email:** aniteshb@gmail.com **(strongly preferred),** anitesh.barua@mccombs.utexas.edu

**Office hours (Zoom only):**

**Cohort A: M W 5:00 – 6:15 p.m.**

**Cohort B: M W 6:15 – 7:30 p.m.**

**Zoom link for office hours:** <https://utexas.zoom.us/j/96036960762>

**Teaching Assistant: Ishita Roy** (ishita.roy@utexas.edu)

**TA Office hours: TBA**

**Course Overview**

Massive amounts of user generated content (UGC) have created an unprecedented opportunity for enterprises to engage in real-time interactions with customers, and to enhance brand, customer loyalty, competitiveness, growth and profitability. A large portion of UGC involves unstructured text; thus, while text mining methods have been around for decades, their importance has grown dramatically in recent times. This course is designed to introduce the basics of text mining and text based predictions of business outcomes. The contents of this course will serve as a foundation of subsequent courses such as Social Media Analytics. The pedagogical structure of the course will consist of classroom lectures, use of tools such as Python scripts and packages like SentiStrength (for sentiment analysis), and group assignments involving the categorization and classification of a variety of documents as well as the application of text analytics in solving business problems.

**Learning Objectives**

Students taking this course will develop expertise in the following areas:

1. Text mining methods for document classification, clustering and information retrieval
2. Image analytics
3. Practical analytical and technical skills including Python
4. Techniques for sentiment analysis and predictive models
5. Real world applications of text mining and image analytics

**Course Material**

**Textbook: Natural Language Processing with Python** <https://www.nltk.org/book/> (mainly chapters 5 and 6)

**Articles** (many will be posted on Canvas, those with links have to be downloaded by the students. Additional readings, if used, will also be posted in advance).

* “Reducing Readmissions to Improve Care” (posted on Canvas, easy read, but demonstrates business applications of text analytics)

# “Gaining Business Value from Unstructured Data” (posted on Canvas, easy read, but demonstrates business applications of text analytics)

* “Mine Your Own Business” (posted on Canvas)
* “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews” <http://acl.ldc.upenn.edu/P/P02/P02-1053.pdf>
* [Vader: A parsimonious rule-based model for sentiment analysis of social media text](http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/download/8109/8122)
* “Similarity Measures for Text Document Clustering” by Anna Huang, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.332.4480&rep=rep1&type=pdf>
* “Who is your competitor in the era of the long tail?”

<https://medium.datadriveninvestor.com/who-is-your-competitor-in-the-era-of-the-long-tail-d0ac24fedde8>

* “Tell me what you want: A Recommender system based on customer preferences and product reviews,”

<https://towardsdatascience.com/a-recommender-system-based-on-customer-preferences-and-product-reviews-3575992bb61>

* “Probabilistic Topic Models,” D. Blei, <http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf>
* “An intuitive explanation of convolutional neural networks,” <https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/>

# Introduction to Word Embeddings and Word2Vec

# <https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa>

# “Distributed Representations of Words and Phrases and their Compositionality” <https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>

* “A beginner’s guide to RNN and text generation,”

<https://medium.com/@annikabrundyn1/the-beginners-guide-to-recurrent-neural-networks-and-text-generation-44a70c34067f>

* Generative Adversarial Networks for Text Generation: Parts 1 and 2

<https://becominghuman.ai/generative-adversarial-networks-for-text-generation-part-1-2b886c8cab10>

<https://becominghuman.ai/generative-adversarial-networks-for-text-generation-part-2-rl-1bc18a2b8c60>

**Grading**

Your course grade will be based on the following:

|  |  |  |
| --- | --- | --- |
| Item | Date due | Weight |
| Group assignment 1 | 09/10 | 11% |
| Group assignment 2 | 09/24 | 11% |
| Group Assignment 3 | 10/10 | 11% |
| Group project | 10/07 | 17% |
| Take home exam (individual work) handed out on Canvas, 6:00 p.m. – 9:30 p.m. | 10/12 | 40% |
| Class participation |  | 10% |

**Class participation**

In this course much of the learning is dependent on the accessing the combined knowledge and experience of the group. It is everyone’s job to keep the discussion productive and moving forward. In evaluating your class participation grade, I take the following into consideration:

* useful arguments expressed coherently and succinctly
* good analysis supported by case facts or your own experience
* relevance to previous contributions, i.e. ability to listen and build on what others say
* constructive disagreement
* regard, respect and acknowledgment of others’ contributions
* readiness to contribute to class discussions

**Group assignments**

All assignments in this course are group based. Students will be responsible for creating their own groups. The ideal group size is 5. The group membership should remain unchanged throughout the course, unless there is a truly compelling reason to do so.

**Notes for any classroom session**

**Safety and Class Participation/Masks:** We will all need to make some adjustments in order to benefit from in-person classroom interactions in a safe and healthy manner. Our best protections against spreading COVID-19 on campus are masks (defined as cloth face coverings) and staying home if you are showing symptoms. Therefore, for the benefit of everyone, this is means that all students are required to follow these important rules.

* **Every student must wear a cloth face-covering properly in class and in all campus buildings at all times**.
* **Students are encouraged to participate in documented daily symptom screening.** This means that each class day in which on-campus activities occur, students must upload certification from the symptom tracking app and confirm that they completed their symptom screening for that day to Canvas.  Students should not upload the results of that screening, just the certificate that they completed it. If the symptom tracking app recommends that the student isolate rather than coming to class, then students must not return to class until cleared by a medical professional.

**Course Agenda**

|  |  |  |  |
| --- | --- | --- | --- |
| **Class #** | **Topic** | **Reading(s)** | **Assignments, Exam** |
| 1 (08/26) | Introduction to unstructured data analytics: The business context | “Reducing Readmissions to Improve Care” (posted on Canvas) “Gaining Business Value from Unstructured Data” (posted on Canvas) | Install Python (Anaconda), if not done already in other courses on your computer |
| 2 (08/31) | Text analytics basics like tokenization, stemming, feature vector, etc.  Discovering new knowledge and insights from conversations (Part I) | Natural Language Processing with Python, Chapter 5 <https://www.nltk.org/book/ch05.html> (I would strongly suggest you browse through earlier chapters for basics).  “Mine your own business” (posted on Canvas) |  |
| 3 (09/02) | Discovering new knowledge and insights from conversations (Part II)  Discussion of assignment 1 | “Mine your own business” (to be posted on Canvas) |  |
| 4 (09/07) | Zipf’s law, weights on words | Zipf’s law  https://en.wikipedia.org/wiki/Zipf%27s\_law |  |
| 5 (09/09) | Sentiment analysis Part I | “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews” by  Peter D. Turney, to be posted on Canvas. | **Group assignment 1 (Edmunds) due on 09/10 by 11:59 p.m. on Canvas** |
| 6 (09/14) | Sentiment analysis Part II  Document similarity, resonance & crowdsourced recommendation systems | “[Vader: A parsimonious rule-based model for sentiment analysis of social media text](http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/download/8109/8122).” (To be posted on Canvas)  “Similarity Measures for Text Document Clustering” by Anna Huang (to be posted on Canvas)  “Who is your competitor in the era of the long tail?”  <https://medium.datadriveninvestor.com/who-is-your-competitor-in-the-era-of-the-long-tail-d0ac24fedde8> |  |
| 7 (09/16) | Clustering of documents  Topic modeling | “Similarity Measures for Text Document Clustering” by Anna Huang (to be posted on Canvas)  Probabilistic Topic Models, D. Blei, <http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf> (to be posted on Canvas) |  |
| 8 (09/21) | Image analytics | <https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/> |  |
| 9 (09/23) | Advanced text analytics using conditional random fields | Introduction to conditional random fields  <http://blog.echen.me/2012/01/03/introduction-to-conditional-random-fields/> | **Group assignment 2 due on 09/24 by 11:59 p.m. on Canvas** |
| 10 (09/28) | Meeting with groups to discuss final project topic | | |
| 11 (09/30) | Advanced text analytics with word embedding & word2Vec | Introduction to Word Embeddings and Word2Vec<https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa>“Distributed Representations of Words and Phrases and their Compositionality”<https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>An Introductory Guide to Understand how ANNs Conceptualize New Ideas (using Embedding) <https://www.analyticsvidhya.com/blog/2018/04/introductory-guide-understand-how-anns-conceptualize-new-ideas/> |  |
| 12 (10/05) | Generating text with Adversarial Networks | Generating text  <https://medium.com/@annikabrundyn1/the-beginners-guide-to-recurrent-neural-networks-and-text-generation-44a70c34067f>  <https://becominghuman.ai/generative-adversarial-networks-for-text-generation-part-1-2b886c8cab10>  <https://becominghuman.ai/generative-adversarial-networks-for-text-generation-part-2-rl-1bc18a2b8c60> |  |
| 13 (10/07) | **Group Project Presentations** | | **Group assignment 3 due on 10/10 by 11:59 p.m. on Canvas** |
| 14 (10/12) | **Final take home exam (individual effort) 6:00 p.m. - 9:30 p.m. due on Canvas** | | |