Name: Drew Walker

Teaching for Public Health

Tf-idf texts as data and

1. **Class and Topic** – Please use the space below to explain the setting for your microteaching

This class will take place within a Texts as Data- or NLP type courses offered to Emory graduate

students. Similar to BSHE 797R TADA or POLS 585: Variable Topics Texts as Data.

Its focus is on learning natural language processing techniques, with an emphasis on its ethical

applications in social and health sciences.

1. **Course description narrative**

This lesson will be one of the first lectures (likely 3rd or 4th) following a series taking students from overview of natural language processing applications in health sciences, through basic data pre-processing tasks (punctuation, stop-word removal, regularization of texts), and word tokenization. This will be the last half of the introductory analysis lectures covering simple text frequency analyses. Following this lecture, students will progress towards regular expressions, lexicon/dictionary matching and sentiment analysis, word embedding techniques, supervised/unsupervised machine learning in text analysis, and use of neural network techniques.

1. **Description of intended (target) learners**

* Graduate students (Masters or PhD, not restricted to public health)
* Basic proficiency in R (data loading, cleaning, transformation, up to Chapter 23 in R for Data Science <https://r4ds.had.co.nz/index.html>
* Expectation for students to have access to a dataset with text data to use for class, if not instructor can provide

1. Pre-class preparation:
   1. Software
      1. Make sure R, RStudio, and Git are installed on computer.
      2. Create GitHub Account
      3. Make sure R is linked with Git
   2. Readings:
      1. Chapter 1 of Eisenstein J. *Introduction to Natural Language Processing*. MIT Press; 2019.
2. **Learning Objectives**
   1. Explain in plain terms the term frequency-inverse document frequency method of analyzing text data.
   2. Describe appropriate applications of tf-idf to multidisciplinary real-world research and analytic solutions.
   3. Understand how to incorporate code examples of tf-idf analysis in R for their own research projects
3. **Instructional Materials**
   1. A/V Setup with projector, laptop HDMI plug-in or internet connectivity
   2. Link to code repo and lesson materials <https://github.com/drew-walkerr/tf_idf_lesson>
4. **Description of teaching Strategies** 
   1. Beginning: Talking about Word Frequencies, stop words, bag of word approach, tf-idf intro
   2. Middle: Mathematical formula, example, class activity, overview of tf-idf-process, results discussion
   3. End: why log (deepening foundation knowledge), examples of applications, further reading (useful for homework)
5. **Evaluation** – Please use the space below to briefly explain how you might assess whether students have met the learning objective. Note: you will not administer your assessment during your microteaching session.
   1. Assessment for this class will be conducted informally during participation in the discipline-specific words activity using two class volunteers to introduce themselves and describe their department or field of study, and then try to come up with words that are more specific to their field than the other student’s.
   2. Appropriate identification of plausible terms will be discussed as a group, taking time to discuss different uses of terms in disciplines. Common words will also be discussed to illustrate the difference between tf and idf components of the equation.
6. **Homework**
   1. Following the lesson, students will apply the tf-idf method either using the instructor’s code or other R packages on their own texts as data corpus. We will expect a write-up including a table of the most frequent terms without stopwords, and the highest tf-idf valued words (top 10 at least). A short paragraph will be included on interpretation of results and uploaded via Rmarkdown file and knitted html file to class Github assignment.