USC Marshall
School of Business

DSO 560 – Text Analytics & Natural Language Processing

Instructor: Yu Chen

Final Exam

Due Thursday, Dec 9thth, 8:00pm PST (90 minutes)

Instructions:

- **WRITE ALL ANSWERS ON SEPARATE SHEETS OF PAPER**
- **SCAN EACH PAGE (AS A PDF OR IMAGE) AND SEND TO ME VIA SLACK**

SHOW ALL WORK TO RECEIVE CREDIT

1.

Short Answer (pick 5 of the 6 questions, 5 pts, recommended 30 minutes)

1. Question about word2vec (1pt)
2. Question about text encoding (1pt)
3. Question about n-grams (1pt)
4. Question about regex (1pt)
5. Question about RNNs (1pt)
6. Question about interpreting attention values (1pt)

Naïve Bayes (3 pts, recommended 15 minutes)

Naïve Bayes Model Calculation: Calculate the posterior probability and determine independence (3pts)

Vectorization and Similarity (3 pts, recommended 10 minutes)

Either a TF-IDF or Count Vectorization question with Euclidean Distance or Cosine Simarility

N-Gram Language Models (3 pts, recommended 10 minutes)

Given the following documents …

A. Constructing a transition matrix after performing text preprocessing (**2 pts**)

B. Calculating probability/perplexity (**1pt**)

True/False (5 pts, recommended 20 minutes)

Pick 5 of the statements below, indicate if it is true or false. If it is false, explain why it is false and provide an example. You may provide an explanation if it is true in case you are wrong and would like to receive partial credit.

A. Question about similarity and count vectorization
B. Question about topic modelling / dimensionality reduction
C. Question about BERT
D. Question about Hidden Markov Models and Part of Speech Tagging
E. Question about regex
F. Question about stemming/lemmatization