USC Marshall
School of Business

DSO 560 – Text Analytics & Natural Language Processing

Instructor: Yu Chen

Midterm Exam

Due Tuesday, April 20th, 8:00pm PST (80 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 (5 pts, recommended 20 minutes)

1. Question about word2vec (1pt)
2. Question about handling streams of text data (1pt)
3. Question about n-grams (1pt)
4. Question about regex (1pt)
5. Question about Zipf Law (1pt)

Naïve Bayes (3 pts, recommended 15 minutes)

Naïve Bayes Model

- **i.** Calculate the prior (0.5pts)
- **ii.** Calculate the likelihood (1pt)
- **iii.** Calculate the evidence (1 pt)
- **iv.** Calculate the posterior (0.5pts)

Vectorization and Similarity (3 pts, recommended 15 minutes)

a. Either a TF-IDF or Count Vectorization question (**2pt**)
b. Either a Cosine Similarity or Euclidean Distance question **(1pt)**

N-Gram Language Models (3 pts, recommended 15 minutes)

Given the following documents …

A. Constructing a transition matrix (**2 pts**)

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

True/False (3 pts, recommended 15 minutes)

Pick 3 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 TF-IDF
C. Question about Euclidean Distance
D. Question about character encodings
E. Question about character encodings