USC Marshall
School of Business

DSO 560 – Text Analytics & Natural Language Processing

Instructor: Yu Chen

Final Exam (Practice)

Due Tuesday, May 10th, 8:30pm PST, 90 minutes

No exams will be accepted past 8:35pm PST

Instructions:

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

- **DO ALL SECTIONS.**

ONCE YOU SUBMIT YOUR EXAM, YOU CAN LEAVE CLASS

SHOW ALL WORK TO RECEIVE PARTIAL CREDIT

Short Answer (5 pts, recommended 30 minutes)

Pick 5 of the short answer questions below to answer. Write no more than 2 sentences in your explanation. Each question is 1pt: 0.5pts for the correct answer and 0.5pts for a correct explanation.

Use the following word2vec embedding table for Q1-Q2

Word			
A	**-1.0**	**2.0**	**1.0**
B	**1.0**	**0**	**1.0**
C	**2.0**	**-2.0**	**0**
D	**1.0**	**-1.0**	**0**
NULL/UNKNOWN	**0**	**0**	**0**

1. Which pair of words would have the highest cosine similarity score with each other? Why?

2. Write what the input sequence to a sequential model would look like for the document "A B A D" and sequence length 5.

3. Write a regex with named capture groups that extracts the username and domain (username@domain) from an email address.

4. Provide an example of when an **InvalidContinuationByte** error could occur.

5. What is one way a Naïve Bayes model could deal with a test document containing a token it has not seen before during training?

6. Provide the binary encoding for the character 🥰 in UTF8. Identify which bits are continuation bytes.

The following sections will be condensed, and only 2 of the following sections will appear. You can look through prior years' exams – not much has changed / will change:

Naïve Bayes (3 pts, recommended 15 minutes)
Vectorization and Similarity (3 pts, recommended 15 minutes)
N-Gram Language Models (3 pts, recommended 15 minutes)

True/False (5 pts, recommended 20 minutes)

Pick 5 of the statements below, indicate if it is true or false. **In both cases (true or false), explain your reasoning in a brief sentence. Each question is worth 1pt: 0.5pts for the correct answer, 0.5pts for explanation.**

A. The documents "John loves cats" and "cats love John" would have the same vector if we used TF-IDF or word count vectorization.
B. In a Hidden Markov Model, the word we observe at sequence step 2 is determined by the word we observe at sequence step 1.
C. If you want to make sure you keep human-readable tokens, you should pick lemmatization over stemming.
D. According to Zipf Law, there is a relatively uniform distribution of word counts across all the words in a corpus' vocabulary.
E. Using a longer sequence length for your RNN models will reduce the chance of vanishing/exploding gradients.
F. LSTMs calculate all of their hidden states across all sequences at once.