USC Marshall
School of Business

DSO 560 – Text Analytics & Natural Language Processing

Instructor: Yu Chen

Midterm Exam

Due Tuesday, April 21st, 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**

- **PICK 4 OF THE 5 SECTIONS TO COMPLETE.**

- **IF YOU DO ALL 5 SECTIONS, I WILL GRADE YOUR BEST 4 SECTIONS.**

SHOW ALL WORK TO RECEIVE CREDIT

1. **Classification & Model Evaluation (4 pts, recommended 20 minutes):**

An NLP startup has developed an initial model that is able to predict the star rating **customers will leave for a product on a fashion website** based solely on the initial comment text that a user writes in forums.

*Consider a rating with 2.0 or less stars to be **negative class**, and a rating with 4.0 and more stars to be **positive class**.*

	Actual Product Rating	**Model A Predicted Sentiment**	**Model B Predicted Sentiment**
1.	1	Positive	Positive
2.	2	Negative	Positive
3.	1	Negative	Negative
4.	1.5	Negative	Negative
5.	4	Negative	Positive
6.	1	Negative	Negative
7.	1	Negative	Negative
8.	4	Positive	Positive
9.	5	Positive	Positive
10.	4.5	Positive	Positive

a. **Which Model (A or B) has the higher metric?**
   i. Accuracy (0.25 pts)
   ii. Precision (0.25 pts)
   iii. F1 Score (0.25 pts)

b. Compute the following probabilities:
   i. P(Model A prediction = Model B prediction | actual_rating = "negative") (0.25pts)

c. Construct the confusion matrix for Model A's test results. Make sure to label whether the columns/rows are actual or predicted results (**1 pt**).

d. The market research team has determined that incorrectly classifying a review as negative when it was actually positive results in an expected financial loss of **$500**. Conversely, incorrectly classifying a review as positive when it was negative incurs only the cost of advertising ($**50**) to a customer who is likely uninterested in the product. **Explain what metric you'd optimize for, and why. Then, explain which model you'd select for this business case, and why. (1pts)**

2. **Naïve Bayes (4 pts, recommended 20 minutes)**

You are a data scientist at HumanFlow, an HR AI startup implementing a **Naïve Bayes NLP model to parse resumes.**

You've collected a small dataset, represented below of employees who have been classified as **strong** or **poor hires** for a data science role:

Strong Hire:

1. SQL, Python, Leadership
2. Python, Looker, SQL
3. AWS, MS Excel, Structured Query Language, Tableau, Python
4. Amazon Web Services (AWS), R

Poor Hire:

1. Stata, Microsoft Excel, SQL
2. Python, Looker, Tableau
3. Amazon Web Services, Leadership
4. R, Stata, Tableau
5. Stata, Leadership
6. Excel, Python

Note – feel free to group together any terms you think are identical in semantic meaning.

A. **List any preprocessing steps** you might take prior to feeding the above corpus into a machine learning model **and explain why they are necessary** (1 pts)

B. Perform any preprocessing and grouping you deem appropriate, and calculate the following probabilities, assuming a Naïve Bayes classifier (2 pts):

- **i.** The **prior** for a **strong hire (0.5pts)**
- **ii.** The **likelihood** for the skillset "SQL, AWS, "Python" given a Strong Hire **(0.5pts)**
- **iii.** The **evidence in Bayes Rule for** "SQL, AWS, "Python" **(0.5pts)**
- iv. The **posterior** for a **strong hire** given a new candidate with the skillset "SQL, AWS, Python" **(0.5pts)**

C. A colleague on your team says that if she sees "SQL" on a candidate's resume, it means she is more likely to see "Python" also on the same resume. Is she correct? Prove why or why not (1 pts).

3. **Vectorization and Similarity (4 pts, recommended 20 minutes)**

You work as a data analyst for a **large fast food restaurant chain**. Your company has conducted several consumer research surveys, with consumers filling in open-response questions about their favorite menu items. This is what 3 customers wrote:

Customer A: nugget shake
Customer B: burgers fries
Customer C. hamburgers nuggets shakes burger

$$\mathbf{TF} = n(t,d)$$
$$\mathbf{IDF} = 1 + \frac{N}{df(t)+1}$$

a. Generate **TF-IDF document vectors** (you may write them as a matrix or table). Calculate IDF for each of the words, then term frequency (TF) for each of document – word combinations (**1pt**)

b. A new user query is submitted via your company's website: **burgers shakes.** Assuming **TF-IDF vectorized** documents and **Euclidean Distance**, does this new query fit more with the search behavior of **Customer B** or **Customer C**?

c. Assume now that a colleague has trained the following 3-dimension **word2vec** word embeddings on the open-ended survey responses. The results are below.

Nugget	-2	0.0	1.0
Shake	-2	0.5	0
Burger	2	-1	2
Fries	3	-2	2.5
Smoothie	-1	0	0.5
Salad	0	2	0
Fruit	0.5	2.5	0.5

From these embeddings, what is one combination (group) of food products the fast food chain should considering **packaging as part of a combo meal**? Explain how you made this determination. (**1 pt**)

d. Using **cosine similarity** and the word2vec embeddings trained by your colleague, is **nugget** more similar to **smoothie** or to **shake**? (**1 pt**)

4. N-Gram Language Models (4 pts, recommended 20 minutes)

Given the following documents:

1. *Classes are today.*
2. *I am late today.*
3. *He went to class.*
4. *They were late to class.*
5. *I went home after class.*

A. Perform lemmatization and stopword removal. For this exercise, consider the words **"the"**, **"at"**, and **"to"** as stopwords. Then construct the transition matrix for this corpus. **Note: Your team has decided on a strategy to deal with out of vocabulary word transitions – they'll set the minimum transition probability to 0.01, instead of 0.** (**2 pts**)

B. Using a **bigram language model,** what is the probability of seeing the sentence ***He went home?*** (**1pt**)

C. Using a **unigram (1-gram) language model,** what is the probability of seeing the sentence ***They went to class***? (**1pt**)

5. **True/False (4 pts, recommended 20 minutes)**

For each 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. The regex **\bThor\b** will result in a higher false positive rate than **Thor** when searching for references to the movie character Thor.

B. In a Hidden Markov Model's **emission matrix**, assuming rows are **observed states** and columns are **hidden states**, the sum of each row should equal 1.

C. A document that has original text **cat litter** and another document that has the text **litter cat** will have identical vectors when using word count vectorization, TF-IDF vectorization, and bag-of-words word2vec vectorization.

D. Two documents:

   a. **cat cat dog dog love love**
   b. **cat dog love**

   Would show a **cosine distance** > 0.

E. There are 3 capture groups in the following regex:
(?:Mr\.|Miss)\s(\w)\s(?P<last_name>\w)

F. **UTF-8** and **ASCII** use the same **Unicode codepoint** for the character "a".

G. If a model's F1 score is 1, it is guaranteed to have 0 false positives and 0 false negatives.

H. If you have a word2vec neural network, with **V** total unique words in your entire vocabulary, and are trying to train word embeddings of size **M** dimensions, the output of the **softmax layer** of word2vec is of shape **M x 1**.