USC Marshall
School of Business

DSO 560 – Text Analytics & Natural Language Processing

Instructor: Yu Chen

Homework 4

Due Sunday, April 19th, 2020 at 11:59pm PST

Instructions:

- **WRITE ALL ANSWERS ON SEPARATE SHEETS OF PAPER**
- **SCAN EACH PAGE (AS A PDF OR IMAGE) AND SEND TO ME VIA SLACK**
- **PICK 3 OF THE 4 SECTIONS TO COMPLETE**

SHOW ALL WORK TO RECEIVE CREDIT

1. **Classification & Model Evaluation (recommended 15 minutes):**

An online retail startup is using text from descriptions of products on the internet to predict if the clothing item is for female or male customers. The model has so far been tested out on 15 pieces of clothing. **Consider women to be the positive class.**

	Actual Clothing Item	**Predicted Outcome**
1.	Women	Women
2.	Men	Men
3.	Men	Men
4.	Men	Men
5.	Men	Men
6.	Women	Men
7.	Men	Men
8.	Men	Men
9.	Men	Women
10.	Women	Women
11.	Women	Women
12.	Men	Men
13.	Women	Women
14.	Women	Men
15.	Men	Men

a. **Compute the following:**
   i. Accuracy (0.25 pts)
   ii. Precision (0.25 pts)
   iii. Recall (0.25 pts)
   iv. F1 Score (0.25 pts)

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

c. In the context of this model, explain what is the difference between **P(gender_actual = "women")** versus **P(gender_predicted = "women" | gender_actual = "women")**? **(1pt)**

d. If the goal of the model is to make sure that **all women's clothing items are actually captured by the model as women's clothing**, which of the metrics above (accuracy, precision, recall, F1 Score) should be prioritized? Explain why. (**1pt**)

2. **Naïve Bayes (recommended 15 minutes)**

Your marketing firm is managing the social media marketing campaign for a theatrical release called **Shazam!** Each comment has already been tagged with either **Intent to Buy** (indicating a willingness to see the film) or **No Intent to Buy** (no interest in seeing the film)

Intent to Buy Tickets:

1. Love this movie. Can't wait!
2. I want to see this movie so bad.
3. This movie looks amazing.

No Intent to Buy Tickets:

1. Looks bad.
2. Hard pass to see this movie.
3. So boring!

Stopwords to remove: *to, this*

You **do not need to perform stemming or lemmatization, and can disregard punctuation / case-sensitivity** (ie. *Can't = can't*).

A. Calculate the following probabilities:

i. .25 pts - P(x = "so" | y = Intent to Buy)
ii. .25 pts - P(x = "see")
iii. .25 pts - P(x_i = "see", x_j = "movie") - ***probability of seeing see and movie in the same document***
iv. .25 pts - P(y = No Intent | x = "bad")

B. Are the words "love" and "movie" independent in terms of sentence occurrence? Prove why or why not (1 pts).

3. **Vectorization and Similarity (recommended 15 minutes)**

Your company has **an inventory of products tagged with keywords** and a search feature that uses vectorization and similarity to match user queries with desired products based upon these keywords. Assume that part of your company's inventory is described below:

Product A: trendy jeans **Product B:** old blue jeans old **Product C.** old blue trendy red wool jeans

To avoid repetitive computations, you may use the following term frequency (TF) and inverse document frequency (IDF) functions (these are simply the same functions from the textbook, but without the log or square root):

TF = n (t,d) ie. *number of times term* ***t*** *appears in document* ***d***

$$\textbf{IDF} = 1 + \frac{N}{df(t)+1}$$

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: **old jeans.** Assuming **TF-IDF vectorized** documents and **cosine similarity**, would your backend search engine recommend **Product B** or **Product C?** (**1 pt**)

c. Assume now that we have received the following **word2vec** word embeddings:

Trendy	-1	0.0	1.0	2.0	-2
Old	-1	1	1	1	0
Blue	2	0	1	-1	0
Red	3	0	1	-2	0.5
Wool	1	2	-2	0	2
Jeans	2	3	-3	0	2.5

Using a BOW (bag-of-words) approach with **word2vec**, compute the document vector representation for Product A.

Hint: word2vec provides embeddings (vectors) for each word. To create the vector for the document, average the word vectors of the document together.

4. N-Gram Language Models (recommended 15 minutes)

Given the following documents:

1. I love going to the store.
2. He loves working at the restaurant.
3. The store is closed today.
4. Today I am working.
5. He is going to the restaurant.

A. Perform lemmatization and stopword removal. For this exercise, only the word "the" is a stopword. Then construct the transition matrix for this corpus.

B. Using a bigram language model, what is the probability of seeing the sentence ***I love working***?

C. Before you can compare the likelihood of the sentence *I love working* with the sentence *Today I am working at the restaurant* appearing in natural language, what must you do and why?