Assignment 3
Language & Computers | LINS313

Due: Monday August 1, by 11:59pm

The homework is due to Canvas by 11:59pm on its due date. Typeset is preferable, but if you do handwrite it and scan it, please ensure that the writing is clearly legible. If you don't have access to a scanner, there are many free scanner apps for smartphones which will give you a pretty good quality image.

A perfect solution is worth 50 points. **Make sure you show your work**; this will account for the majority of your grade. If any of these instructions or questions do not make sense to you, reach out to me right away.

Problem 1 (10 points) Two paragraphs are given below, one for each of two authors, John Beadle and John Haldane, from texts about traveling in North America. Your job is to determine who wrote the questioned paragraph based on properties of the known paragraph, providing explicit justification for your choices, similarly to what we did in class and which you can see in the first few pages of the Forensic Linguistics slides (no references in your textbook for this task).

Text 2, Author: John Beadle

> From Corsicana the train on the Texas Central Railroad carried me early straight south, leaving the valley of the Trinity and bearing across the high country to the Brazos. Not one acre in ten of this region is under fence. All the rest is common pasture, though most of it belongs to private owners, and is for sale at two to six dollars per acre. The region is high and gently undulating, about one-fifth in timber, the rest fertile prairie. My next stopping place was Houston, which I thought, at first view, the most beautiful place in Texas. There had been a twenty-four-hours' rain, and at 9 A. M

the sun shone out clear: the orange groves, magnolias, and shade trees looked their richest green and Houston presented to the newly arrived Northerner a most enchanting appearance. That city, the original capital of Texas, is at the head of Buffalo Bayou, a long projection of Galveston Bay.

Text 2, Author: John Haldane

What about Chicago? My dear good friend, what can I say about that ninth wonder of the world which people everywhere do not know already, and yet, something fresh may be squeezed out of it at a few points. I found that it had a population of fully 2,000,000, and was so immense in area, that once in it, you required some time to get out of it, even by rail, or by swift tram cars, two of which ran me one evening nearly twenty miles into the suburbs at a speed of thirty miles an hour when possible. The country round about is flat and uninteresting, but nevertheless, Chicago is a wonderful city throughout, and of unique interest amongst the great cities of the globe. The streets were as crowded as those of London, and the shops quite as handsome, but I could not admire such an array of gigantic commercial buildings, hundreds of feet in height, so much alike, so slabby in form, and flat roofed in most cases.

Text 3, Questioned Paragraph

One week sufficed to conclude my business in Oregon, but before leaving a few general notes are in order. Portland is on the west bank of the Willamette, twelve miles above its mouth and near the head of tide-water. But the Columbia often rises so as to cause backwater, giving the Willamette a variation of thirty-two feet. Ocean steamers load at the wharf, and the place has direct water communication with all the ports of the world, the chief exports being wheat, limber, beef and salmon. All the older portion of the city is very beautifully improved; elegant residences abound, with many evidences of taste and wealth. The location is picturesque. The Cascade Range is only occasionally visible, but Mount Hood rears its snowy summit sixty miles eastward, and looks as if it were just out of town. Mount Saint Helens is sometimes in good view, though eighty miles to the north-

east. All the hills around the city are covered with heavy timber, and in town every street is double lined with shade trees.

For this problem, please provide three quantified measures of authorship style (e.g., average sentence length) and three non-quantified observations of similarity (e.g., content or particular expressions). For each measure, you should give its value for each text and you should explicitly state why it supports your determination of who wrote the questioned paragraph. I'm interested in the evidence and reasoning you use based on that evidence, not whether you get the right author.

i. Quantified measures

Provide three quantified measures (be sure to provide quantities for all three paragraphs). Here is a suggested format for writing the values. Don't forget that you also need to explain how it supports your analysis.

Description of measure	Value for Beadle	Value for Haldane	Value for Questioned

ii. Non-quantified observations

Provide three non-quantified measures (state explicitly how each one relates to each paragraph, and how it supports your analysis).

Problem 2 (10 Points) In the course slides the relative frequency of the words "I" and "the" were calculated for five texts of three authors: Austen, Doyle, and Krugman. These two dimensions were used to calculate the centroids for the three authors using the K-means algorithm. There are many other values that could be used for clustering with K-means; for this problem, you'll work with the relative frequency of "we", "he", and "a". In particular, you are given the measurements for six texts, some of them by Austen and some by Doyle, and your job is to cluster them with k-means.

Here are the documents and their measurements:

Doc ID	x	y	z
dl	2.1	10.9	15.8
d2	4.6	11.3	23.0
d3	1.8	11.4	19.2
d4	1.7	9.2	17.1
d5	2.1	9.5	19.2
d6	7.8	15.4	22.0

You can probably spot the clusters pretty easily just by inspecting the values, but for this problem you need to compute the centroids of each cluster using K-means.

You are given the following two initial centroids:

Centroid ID	x	y	z
cl	1.8	10.0	16.4
c2	2.0	10.6	19.0

Note that in the slides, we used two of the document points as initial centroids: here, they are different points, so there will be non-zero distances from them to all documents.

i. **Distance Computation**
   For every document, compute the distance between it and the two centroids cl and c2. In the slides you saw how to compute the squared distance in two dimensions. For three it is not different: you just need to include the z dimensions in your calculation:

   (1) $$distance(d_i, d_j) = (x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2$$

   Here is a table to help organize your values:

	d1	d2	d3	d4	d5	d6
c1						
c2						

ii. **Group Membership**
Based on the distances, write down the group memberships for each centroid. Then, compute the new centroids based on the group memberships.

iii. **Repeat**
Do this for **two** more iterations (the centroids won't change after that). At each iteration, give the distances, the centroids, and the group memberships.

iv. **Including a new document**
You are given a new document, d7, with the measurements (3.2, 13.3, 24.9) and are told it is written by Doyle. Calculate the distance between d7 and the final centroids you computed for Part 3. (Show your work.) Based on this result, which of the first six documents were likely to have been written by Doyle and and which by Austen?

Problem 3 (10 points) FriendPage has asked you to build a system that monitors and flags hate speech on its platform. Your system reads comments on news articles posted by users classifies comments as hate speech or not. Here are some examples[1]:

"Women have such a tough time being women. They have periods, do the lifting, and always have to ask for directions."	"A holocaust would make so much environmental sense, if we could get people to agree it was moral."

One important feature of the Naive Bayes classifier is modularity. We can obtain conditional probabilities (like P("holocaust" | hate_speech)) from one dataset, and prior probabilities like P(hate_speech)) from another. This is especially helpful in cases like hate speech detection. In order to have a sense of what kind of language people use in hate speech, we will need a dataset that draws disproportionately from this category. However, we know based on externally collected statistics that the prior probability that any given comment contains hate speech is actually fairly low. Using Bayes' Theorem, you will compute the probability that the second comment should be flagged as hate speech, based on the fact that it contains *holocaust* and

[1] These comments were auto-generated by the GPT3 language model, when prompted to say something about women or the Holocaust. Source: https://twitter.com/an_open_mind/status/1284487376312709120

based on the following training data:

Dataset 1: A balanced hate speech dataset consisting in 10,000 comments, 5,000 of which have been flagged by moderators as hate speech. The word "holocaust" is in 1500 of the hate speech tagged comments and 30 of the normal comments.

Dataset 2: An independent study of 1 million comments on the platform which determined that 0.5% of all comments contain hate speech.

In answering the questions below, be sure to use dataset 1 and dataset 2 appropriately for computing the probabilities, as indicated above.

i. What is the probability of hate speech? (Use dataset 2).

ii. What is the probability of "holocaust" given hate speech? (Use dataset 1)

iii. What is the probability of hate speech given "holocaust"? (Show your work)

Your final answer may seem low. Check: is it still higher than the bast rate for hate speech found in dataset 2? If so, then your model is successfully taking into account the evidence (seeing the word "holocaust") in updating its prior probability for hate speech. Given that the word "holocaust" is used in many contexts which do not constitute hate speech, this is desirable behavior. Ultimately, would want to base our final judgment off not just this probability, but off the probability of hate speech given the joint probability of seeing this word together with others, such as "hate" itself.

Problem 4 (10 points) In a dataset of tweets annotated for polarity, there were 155 tweets annotated as neutral, 167 as positive, and 242 as negative.

I create a subjectivity classifier that identifies tweets as "objective" (neutral) or subjective (either positive or negative). It labels 153 tweets as objective and the rest as subjective. Of the 153, only 112 were actually objective.

i. Fill in the following table based on the data given above

Description of Measure	Classifier: Objective	Classifier: Subjective
Annotation: Objective		
Annotation: Subjective		

ii. Now, answer the following questions about the classifier's ability to identify objective tweets.

- How many false negatives did the classifier produce as a detector of objective tweets?
- What is the precision of the classifier as a detector of objective tweets?
- What is the recall of the classifier as a detector of objective tweets?

iii. Next, answer the following questions about the classifier's ability to identify subjective tweets.

- How many false positives did the classifier produce as a detector of subjective tweets?
- What is the precision of the classifier as a detector of subjective tweets?
- What is the recall of the classifier as a detector of subjective tweets?

Problem 5 (10 points) The goal of a subjectivity classifier is to identify tweets that should be labeled as positive or negative by a polarity classifier. The polarity classifier is what will be used to understand the *overall* positivity or negativity toward health care reform on Twitter, and it will be used to present positive and negative *examples* to policy makers. Recall that probabilistic classifiers provide not only a label, but also a confidence value (the probability), and that probability can be used as a threshold. In class, we used the following decision rule to determine a candidate tweet's label:

$$\hat{c} = \underset{x \in \{subj, obj\}}{\text{argmax}} \; P(\text{subjectivity=c}|\text{Tweet contains 'breakfast'})$$

where $\hat{c}$ is the predicted category, and $\underset{x \in \{subj, obj\}}{\text{argmax}}$ returns whichever category (subj or obj) has the highest probability according to the classifier (i.e., we plug each one in for c and calculate that probability, then take the highest). You can instead choose the label based on a *threshold*, which we'll call θ (theta):

$$\hat{c} = subj \text{ if } P(\text{subjectivity=c}|\text{Tweet contains 'breakfast'}) > \theta$$
$$otherwise, \hat{c} = obj$$

where θ is a threshold value between 0 and 1. If θ = 0.5, then the decision is the same as the argmax above, but if θ =0.8, then we are saying that we will only consider tweets subjective if they have a probability of being subjective greater than .8, and so on. This allows us to use the confidence value to select more or fewer tweets as subjective. (At the extremes: if θ = 0, everything is subjective, and if θ = 1, nothing is.) This is important because when a classifier is more confident, it is typically more correct. So, if we have a higher threshold, it means that precision is higher because each decision is more likely to be right. However, if we have too high a threshold, we might fail to label many instances, so this can damage recall.

This question asks to you to consider whether precision or recall is more important for identifying subjective tweets for different contexts. Another way of thinking about this question is this: is it more important (a) to catch all of the subjective tweets at the risk of including many objective ones and thus have better recall (by using values for θ less than 0.5), or (b) to select fewer tweets that the subjectivity classifier is more confident about as subjective at the risk of not providing many subjective tweets to the polarity classifier and thus have better precision (by using values for θ greater than 0.5)?

With that in mind, and also considering that there may be 100,000 or more tweets per day that are relevant to your topic of interest, say whether you think precision or recall for subjective tweets is more important for the following purposes:

i. A single person monitors the pulse of the people (on Twitter) by tracking the positive/negative tweet ratio over time.

ii. A group of ten researchers must find interesting comments (in specific tweets) to demonstrate specific sentiments as examples in a policy decision.

iii. An automated system will use the positive/negative ratio toward many different companies in order to decide whether to buy or sell shares of those companies' stocks.

iv. Write 4-5 sentences for each (a short paragraph). Think about the effort that goes into various aspects of these tasks, such as how many tweets one person can look at in an hour and what the cost of mistakes about sentiment—including missing important subjective tweets or assigning labels positive and negative to many tweets that are actually objective (in which case they are incorrect by definition).