IS 733 - Lab 11

Part I: Distributional Hypothesis
Please use the Distributional Hypothesis Concept to find the solution, show your work step by step

Example 1

1. **A piece of _____ is on the plate.**
2. **Everyone enjoys eating _____.**
3. **You can cut _____ with a knife.**
4. **We make _____ from milk.**

Potential Answers: "cheese," "cake," or "butter"

1. First, read all four sentences and note their contexts something that sits on a plate, that people enjoy eating, that can be cut with a knife, and that is made from milk.
2. Then considering each candidate word in turn. Cake can sit on a plate and be cut, but we bake cake rather than literally make it from milk. Butter can be spread on a plate and cut, but it's churned from cream, not described as "made from milk."
3. Cheese, on the other hand, is commonly placed on a plate, enjoyed as food, cut with a knife, and produced from milk.
4. Therefore, cheese is the only word that fits all contexts.

Example 2

1. **The _____ is parked in the driveway.**
2. **He bought a new _____ for his birthday.**
3. **_____ can drive really fast.**
4. **People often wash their _____ on the weekends.**

Potential Answers: "car," "truck," or "motorcycle"

1. Something parked in a driveway, something someone might give as a birthday present, something that can drive fast, and something people wash on weekends.
2. Then testing "truck", although trucks are parked and sometimes washed, "drive really fast" is less typical for trucks.
3. Later, test "motorcycle", it can be parked, driven fast, and washed, and could be a gift, but motorcycles are less commonly given as birthday presents than cars.
4. Lastly, test "car": cars are routinely parked in driveways, often gifted, commonly associated with driving fast, and washed on weekends. Because car naturally fills all these roles, it is the correct choice.
   Correct choice: car

Example 3

1. **I read an interesting _____ last night.**
2. **Many people enjoy a good _____ before bed.**
3. **_____ often has chapters and a cover.**
4. **You can borrow a _____ from the library.**

Potential Answers: "book," "novel," or "story" – these terms appear in contexts related to reading and library usage.

1. Note the contexts something you read at night, something people enjoy before bed, something that has chapters and a cover, and something you can borrow from a library.
2. Consider “story”, people tell bedtime stories, but “borrow a story” is awkward and stories don’t always have chapters.
3. Then consider “novel”, novels fit most contexts, but the term is more specific than “book.”
4. Consider “book”: you read a book before bed, books have chapters and covers, and you borrow books from libraries.
5. Since book matches every context smoothly, it is the answer.
   Correct choice: Book

Part II: Practice the NLP model to classify data stories

Link to dataset

Show = level 1; Tell = Level 2 and 3

You will be asked to replicate the results from this paper (Figure 6 zero shot results, cross-validation and leave one-plot out); feature extraction steps are as follows:

For each sentence

- we converted it into lowercase
- removed punctuation
- The sentences were then split into tokens, and stop words were removed
- Tokenization and lemmatization
- feature extraction was performed using TF-IDF vectorization

You will apply ML classification models such as logistic regression, Support Vector Machine (SVM), or Naive Bayes (NB). Optionally, you are also welcomed to try other classifiers not used in the paper, such as Random Forest (RF) to see whether it outperformed the existing results.

Bonus: Try additional representations/embedding methods, and compare with the current result

Version	Experiment	Model	Training Data			
			Combined	Zero Shot	One Shot	Two Shot
LLM Generated Stories	Cross Validation	Logistic Regression	0.93	0.85	0.79	0.86
		Naive Bayes	0.92	0.79	0.78	0.82
		SVM	0.94	0.82	0.79	0.83
	Leave One Plot Out	Logistic Regression	0.95	0.94	0.92	0.95
		Naive Bayes	0.95	0.94	0.93	0.95
		SVM	0.95	0.94	0.94	0.95
Student Generated Stories	Train/Validate	Logistic Regression	0.77	0.68	0.69	0.78
		Naive Bayes	0.77	0.67	0.70	0.77
		SVM	0.76	0.67	0.68	0.77

Figure 6: ML model performance (Area Under Curve) in discriminating whether a given narrative sentence is “show” versus “tell”

Summary-

In a standard 5-fold cross-validation, the basic frequency-based classifier had an F1 of around 0.76, so it most commonly labeled the most frequent class. Logistic regression performed slightly better at an F1 of around 0.79, and the random forest model at 0.78. Multinomial Naïve Bayes provided a significant improvement with an F1 score of around 0.83. Linear SVM performed the best with the highest F1 of around 0.86, which also showed it to be best distinguishing "show" and "tell" sentences in cross-validation.

When tested with generalization using the leave-one-plot-out (LOPO) method where a single story is left out at a time performance generally dropped, signifying the difficulty in deploying models to new data. The baseline dummy classifier dropped to about 0.62 accuracy, logistic regression and random forest models both hit about 0.67, and Naïve Bayes about 0.72. Linear SVM again fared the best at about 0.78 accuracy, showing its robustness even with new, unknown narrative plot data.

Overall, all the trained models performed much better than the simple baseline, with linear SVM being the most consistent under both testing methods. Naïve Bayes was a close second and could be beneficial for larger datasets due to more efficient training times. Logistic regression and random forest had modest gains and could have benefits in terms of interpretability or flexibility. Subsequent studies might explore richer feature representations—word embeddings or transformer-based encodings and combinations of methods to achieve even greater than the already substantial performance of the SVM.

Submission:

Part I: solution + steps

Part I: a small paragraph summarizing your finding + GPT statement + Python Notebook