Student's affirmation: I certify that I have neither taken help in completing this assignment nor helped anyone else with this assignment. I have never discussed this assignment with anyone other than the instructor and TA.

(Signature) _____

Mandatory: Affirmation and signature on the first page; name on every page; submission as PDF.
If you make assumptions about any problem, state them, but be prepared to justify why they were necessary.

Problem	1	2	3	4	Total
Points:	24	24	22	30	100
Score:					

This assignment has 4 problems, for a total of 100 points.

Throughout, I prefer that you think afresh but if you come across a source on which you base your answer, be sure to cite it with *specific page and line numbers*. An acceptable source would be authoritative (such as our textbook or a book by another credentialed author, or a peer-reviewed article). Blog posts by strangers are not credible.

1. (24 points) Mark the following statements true or false. Provide a short explanation of about 10–20 words. You can and should provide a source where appropriate, including the *specific page* and *line numbers*.

   A. We can use intermediate results from n-fold testing to compute statistics such as are needed for statistical hypothesis testing

   B. Newly introduced words such as Linux are an example of a closed class of words changing for English

   C. English sentences have been redefined after the rise of the WWW since constructions like this *I. Mean. It. Now.* have become common

   D. A clitic is an old-fashioned contraction such as *shan't*

   E. Since language is concerned with words, considering n-grams of the characters appearing in a written word don't help deal with out of vocabulary words

   F. We can be certain that if one native speaker of Hausa says something, another native speaker of the same dialect of Hausa will understand them

   G. Pragmatics is the idea that NLP can be used in industry to solve practical problems

   H. If someone writes *winned* instead of *won* we may be able to process them as the same after lemmatization unless the lemmatizer rejects unknown words

   I. The antonym of a word is more like a synonym than a word chosen at random

   J. The vector-space model yields dense vectors that are easy to incorporate as features in machine learning

   K. The Penn Treebank emphasizes sentences that don't match language usage on social media

   L. Descriptive grammar tells us how natural language should be used to describe objects and events

2. Consider a challenge for a distributional representation such as the word embeddings we discussed in class.

   Suppose our embeddings are trained for individual words and the baseline approach to compute the embedding for a phrase is to take the average of the embeddings of the words in that phrase.

   Let us consider words and phrases that carry meaning of sentiment. For example,

   - Words such as *pretty*, *fun*, *solid*, and *happy* are positive
   - Words such as *sad* and *disappointed* are negative
   - Phrases such as *pretty bad* are negative and possibly more so than *bad* alone

- Phrases such as *pretty solid* are positive but less so than *solid* alone

(a) (6 points) Give another such example of three words and two phrases formed of those words, such that the sentiments associated with the words and phrases indicate a nontrivial relationship, especially the strengthening-weakening dichotomy of the above example.

I would prefer English but it is fine to give an example from a language other than English. In that case please document the individual words (original plus phonetic writing plus meanings and links to some online dictionary). I may check your example with someone.

Explain your answer in about 20 words.

(b) (18 points) Describe an approach for either training word embeddings differently or computing embeddings for phrases differently than in the literature.

You could describe your approach in pseudocode or as equations that describe the model. You should follow the original CBOW or Skipgram models or as those approaches are described by Jurafsky and Martin or by Goldberg and Levy.

Be sure to highlight where your approach differs from the original approaches and how your approach addresses the above challenge.

3. The following grammar is extracted from Jurafsky's $\mathscr{L}_1$ miniature grammar (Figure 18.8, 2025-01-12 draft).

An Adverbial Phrase may be a series of one or more Adverbs (separated by conjunctions) as in "She worked silently and methodically." An Adverbial Phrase may be a Prepositional Phrase as "She worked silently and methodically in the office."

S $\longrightarrow$ NP VP	NP $\longrightarrow$ Pronoun	VP $\longrightarrow$ Verb NP PP
VP $\longrightarrow$ Verb NP	NP $\longrightarrow$ Proper-Noun	PP $\longrightarrow$ Preposition NP

(a) (8 points) Enhance this grammar to allow Adverbial Phrases before and after any VP. Explain briefly.

(b) (6 points) Convert your grammar to Chomsky Normal Form. Explain briefly.

(c) (4 points) Show a parse tree for this sentence using your grammar:

Jane silently and methodically dug up carrots in October.

(d) (4 points) Identify the head word for each constituent in your parse tree.

4. Consider a language community, the Scalars, whose speakers follow a language called Scalese. Scalese is just like $\mathscr{L}_1$ as described by Jurafsky (Figure 18.8) with this modification:

- A Scalar always states modifiers redundantly with a scalar inserted.
- The modifiers may be nouns or adjectives.
- A sentence begins with a part that would be well-formed in $\mathscr{L}_1$ but each occurrence of a modifier (or adjective in a copula construction) is echoed with a "yes," the modifier, then a number, then "in," then its attribute.
- Example Scalese utterances are as below. Here the attributes are "flavor," "size," and "time."

  We like coffee icecream, yes, three coffee in flavor.

  This mug is big, yes, four big in size.

  I like coffee icecream, yes, two coffee in flavor, and big cookies, yes, five big in size.

  Was the flight from Houston early, yes, eight early in time?

(a) (24 points) State a context-free grammar that modifies $\mathscr{L}_1$ to support the Scalese dialect. Highlight the productions you would delete and those you would insert to convert the original grammar. You don't need to handle chains of modifiers. All the commas are inserted by your productions.

(b) (6 points) Is your context-free grammar an adequate solution to Scalese? If so, explain how. If not, explain what simplifying assumption is necessary for this solution.