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	Total
Points:	20	20	60	100
Score:				

This assignment has 3 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, please be sure to cite it. Please note that if you are citing a source to bolster your claim then the source should be authoritative (such as our textbook or a book by another credentialed author). Blog posts by strangers are not credible.

1. (20 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. The antonym of a word is more like the word than a word chosen at random
   B. If we could choose *n* to be large enough, an *n*-gram language model would capture a natural language in a both succinct and generalizable manner
   C. Macroaveraging is a superior measure of performance (relative to microaveraging) for datasets in which some classes are substantially more numerous that others
   D. A contraction such as "'d" is an example of a clitic
   E. Laplace smoothing is a time-honored method for dealing with words that have zero counts in a given corpus
   F. The vector space model produces document vectors in which each dimension stands for a word and depends only on how that word occurs in the given corpus
   G. A (dense) distributional representation may be harder to produce than TF-IDF vectors but better captures the context of a word
   H. Descriptive grammar concerns the use of natural language to describe real-world objects
   I. Natural language grammars are ambiguous because we wish to keep them small; if the size of a grammar weren't a consideration, we would be able to eliminate all ambiguity
   J. The Penn Treebank is a pretty good model of how people in United States use English on social media

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 such as *quite* that sometimes intensify and sometimes diminish the meaning of the phrases to which they apply. For example,

   - In *quite a character*, as in *He is quite a character*, *quite* increases the effect of *a character*—because *quite a character* is more of a character than just *a character*.
   - In *quite a few*, as in *Quite a few people attended*, *quite* reduces the effect of *a few*—because *quite a few* is more people than *a few*.

(a) (6 points) Give two such examples, including a word and two phrases illustrating this situation.

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 ∼20 words.

(b) (14 points) Describe an approach for tackling such normal and idiomatic uses by producing separate embeddings for the same word or combining the word in different ways depending on how it is used.

For simplicity, you can assume that all normal uses are unambiguous.

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.

Explain your answer in ∼50–70 words. Be sure to highlight where your approach differs from the original approaches and how your approach addresses the above challenge.

3. Remember to insert brief (∼30–70 words each) explanations for each part.

Consider the Desert Shamrocks, a language community that speaks VeSuObish. VeSuObish is a VSO language in which the default word order is verb–subject–object. VeSuObish has another quirk in that it places articles after nouns. The Wikipedia article on VSO languages https://en.wikipedia.org/wiki/Verb-subject-object_word_order provides interesting background. VSO languages include Arabic, Celtic, and Hebrew—hence the nickname of this language community.

VeSuObish children have a remarkably simple grammar, equivalent to the following productions extracted from the constituency grammar for the language $\mathscr{L}_1$ specified by Jurafsky (Figure 13.1).

S	⟶	NP VP
S	⟶	VP
NP	⟶	Pronoun
NP	⟶	Proper-Noun
NP	⟶	Nominal
NP	⟶	Determiner Nominal
Nominal	⟶	Noun
Nominal	⟶	Nominal Noun
VP	⟶	Verb
VP	⟶	Verb NP

(a) (4 points) Show a constituency parse tree for using the above grammar (in typical English).

The corner shop door man sells drug paraphernalia.

(b) (5 points) Write the equivalent of the following English sentence in VeSuObish.

The corner shop door man sells drug paraphernalia.

(c) (5 points) Provide a constituency grammar for VeSuObish.

(d) (10 points) Convert the above English grammar to Chomsky Normal Form.

(e) (8 points) Show the head word for each constituent in your parse of the following VeSuObish sentence.

Eats lab assistant the apple pie an.

(f) (12 points) Work out the steps of the CYK parser as shown in class for a parse of

Eats lab assistant the apple pie an.

(g) (16 points) Modify the VeSuObish grammar so that we can obtain an interpretation where corner and shop form one nominal compound that modifies another nominal compound formed from door and man.

The corner shop door man sells drug paraphernalia.