Problem 1c: Extra Credit

The `words` parameter of the function `__getitem__()` is treated as a tuple containing 'the' and 'of' when it is called with multiple words, like `embeddings["the", "of"]`. Since the internal implementation of this function loops over `words`, it processes each word correctly. However, `words` is a string 'the' when the function is called with a single word, such as `embeddings["the"]`. This causes the function to mistakenly iterate over each character instead of treating "the" as a single token.

Problem 4a: Syntactic vs. Semantic Relation Types

The accuracies are in percentage, computed as a micro-accuracy, and rounded to three significant figures of precision.

Q1-Q3

EMBEDDING SPACE	SEMANTIC	SYNTACTIC	OVERALL
GloVe 50	40.0	27.6	33.2
GloVe 100	44.5	27.8	35.5
GloVe 200	31.7	21.7	26.2

Q4

The semantic accuracy of the Skip-gram model outperforms that of all the GloVe results, whereas CBOW shows the lowest semantic accuracy (15.5%). In contrast, syntactic accuracy remains relatively stable across different GloVe settings, with Skip-gram and CBOW performing better than GloVes. Among all models, Skip-gram achieves the highest overall accuracy, while GloVe 200 performs the worst. The remaining models (CBOW, GloVe 50, and GloVe 100) show similar overall accuracy at around 35.0%.

Regarding dimensionality effects, semantic accuracy fluctuates with dimensionality, as GloVe 100 has the highest semantic accuracy (44.5%), while GloVe 200 underperforms (31.7%). However, syntactic accuracy remains fairly consistent (GloVe 50, GloVe 100) or slightly decreases (GloVe 100, GloVe 200) as the dimension increases, indicating that increasing dimensionality does not significantly improve syntactic analogy performance.

Problem 4b: Effect of Lenience

EMBEDDING SPACE	SEMANTIC	SYNTACTIC	OVERALL
GloVe 50	56.6	53.7	55.0
GloVe 100	66.5	65.9	66.2
GloVe 200	70.5	67.2	68.7

Introducing lenience improved the accuracy of all GloVe models. Among them, GloVe 200 achieved the highest performance in semantic, syntactic, and overall accuracy, followed by GloVe 100 and GloVe 50. Unlike previous results, where dimensionality had less effect, higher-dimensional embeddings now demonstrate a clear performance advantage.

Problem 4c: Qualitative Evaluation

ANALOGY QUESTION	GOLD ANSWER	GLOVE 50	GLOVE 100	GLOVE 200
france : paris :: italy : *x*	rome	rome	rome	rome
france : paris :: japan : *x*	tokyo	tokyo	tokyo	tokyo
france : paris :: florida : *x*	tallahassee	miami	miami	miami
big : bigger :: small : *x*	smaller	larger	larger	smaller
big : bigger :: cold : *x*	colder	warmer	cooler	colder
big : bigger :: quick : *x*	quicker	quicker	quicker	quicker

Q) How do the different embedding spaces compare to one another?

All GloVe models produce the same result on the semantic analogy test (France: paris) regardless of the embedding size. However, larger embedding sizes tend to perform better on syntactic analogy test (big: bigger). Specifically, as the embedding size increases, the number of correct answers improves, reaching 1, 2, and 3 for embedding sizes 50, 100, and 200, respectively.

Q) How do they compare to the results reported by Mikolov et al.?

As seen from Table 8 in Mikolov et al., 2013, the Skip-gram model outperforms the GloVe models on semantic analogy tests. In particular, it successfully finds the answer for the test "france : paris :: florida : *x*", whereas all GloVe models fail. On the other hand, while the Skip-gram model incorrectly computes "larger" for the test "big : bigger :: small : *x*," GloVe 200 accurately finds the answer, which is "smaller."