Word embeddings

Word embeddings are key elements of large language models. Basically, they are word representations in a vector space. Research on embedding has been active for more than a decade.

You can find various pre-compiled word vectors online.

Let's look at two vector lists with 300 dimensions:

1) Here, you'll find a version published by Meta: https://fasttext.cc/docs/en/english-vectors.html

Download the 'wiki-news-300d-1M.vec.zip' vectors.

We'll call this 'FastText'.

2) Here, you'll find a version published by Stanford: https://nlp.stanford.edu/projects/glove/

Download the 'glove.6B.zip' vectors (we will want to use glove.6B.300d.txt).

We'll call this 'Glove'.

Write the code to load the FastText and the Glove vector lists.

Write the code to compare the word 'Ferrari' and the word 'chevrolet' using cosine similarity on FastText and Glove. Use the lowercase version of these words.

Do the same for the word 'Ferrari' and the word 'banana' using cosine similarity on FastText and Glove. Use the lowercase version of these words.

Your answer should contain the following:

1) The code used to load the vector lists and compare words.
2) the cosine similarities of the word pairs on FastText and Glove as float numbers with 3 decimal precision (e.g., 0.445).
3) A paragraph, in English, explaining your findings.