SML312 — Research Projects in Data Science: Mini Project #3

29/11/22

Jonathan Hanke

Justin H. Cai

Problem 1

(a). The idea of vector embeddings is to encode words into vectors such that they can be read by a computer, and to encode them in such a way that similar words have correspondingly nearby vectors. Word2Vec constructs embeddings from a corpus (set) of words using CBOW or Skip-gram such that similar words have nearby vectors. The nearness of vector embeddings is measured by cosine similarity, which measures the size of the angle between the vector embeddings. Word2Vec creates embeddings that store similarity in semantics and/or syntax within the vector embeddings that can then be used effectively in further natural language processing tasks.

(b). See code.

(c). See code.

Problem 2

(a)-(d). See code.

(e). See code. Topic 1: Simple words/prepositions/pronouns. Topic 2: Nations and governments. Topic 3: Democratic presidential election campaigns.

Problem 3

(a)-(b). See code.

(c). See code. I expect the accuracy on the validation data is more characteristic of model performance on new data because the model has not yet seen the validation data, whereas the model was trained on the training data and thus has seen it before.

(d). See code. The model performance on the test data is mostly in line with my expectations from part (c) because the model performance on the test data, with an accuracy of 0.954, is closer to the model performance on the validation data, 0.959, than to the model performance on the training data, 0.962.

Acknowledgements

Honor Statement

This problem set represents my own work in accordance with University regulations.
— Justin Cai, 29/11/22

Collaborators

None