**Assignment for Unit 5**

Name: Date:

Year and Section:

*Note: For problems with code, create a separate python file for each questions and upload them to a public repository under your GitHub account.*

1. (20 points) Using Wikipedia as the corpus, obtain 5 different topics that will serve as your documents, and create a term-document matrix. You can use the shared code on GitHub as a reference.
   1. Term-document matrix using raw frequency.
   2. Term-document matrix using TF-IDF weights.
2. (10 points) What are the differences between using TF-IDF weights and raw frequency?
3. (10 points) Using cosine similarity, compare two documents and find out which of the documents is most similar.
4. (30 points) Using the same dataset used above, use the word2vec package to create a classifier for dense vectors.
   1. Use Logistic Regression, with the appropriate configuration for the model and dataset.
5. (20 points) What are the differences of using word2vec compared to the tf-idf in terms of:
   1. Vector Space?
   2. Vector Size?
6. (10 points) How do we evaluate the performance of Semantic Models (i.e TF-IDF and Word2Vec)?