University of the Witwatersrand
School of Computer Science and Applied Mathematics
COMS4054A & COMS7066A: Natural Language Processing/Technology
Lab 1 (Word2Vec)

1 Instructions

For this lab you will be implementing the skip-gram version of Word2Vec discussed in the lecture (mapping from the focal word to context words). You may work in groups of two or three people. To complete the base version of the lab you must:

1. Read in a paragraph of text from a text file from one of the Harry Potter books provided ("HP1.txt" for example is the first Harry Potter book). You **do not** need to use all of the dataset. Only as much as you feel you need to complete the lab. These text files are also not completely cleaned or homogeneous so keep an eye out for that.
2. Split the text by whitespace and remove punctuation so that only the words are left.
3. Extract the unique words from the text and collect them in an array (in order of the word's first appearance in the text).
4. Map each word to a unique 1-hot representation where the dictionary size is the number of words in the original text.
5. For every word in the original paragraph (not unique word - so the same word can be used as input twice) create a dataset where the words are inputs and the corresponding 2-word context (on either side) is the labels. One context word is used as a label at a time.
6. Train a neural network with a linear hidden layer (you may use PyTorch, Jax or Tensorflow/Keras) on this data.
7. Implement an inference function which receives a 1-hot vector and provides the corresponding embedding.

Hint: I strongly recommend you initialize your networks with small random (gaussian) weights (small being around less than 0.1 roughly) and a large learning rate. This will put your networks in the so called "feature learning regime" which is obviously what we want when learning embeddings.

In addition to the above you will be expected to contribute to the class discussion where we will talk about an experiment or topic which we collectively explore throughout the lab. You will be marked within your group for the discussion but any member of the group may be asked to answer. For this lab we will be discussing the broad topic of hyper-parameters, design decisions and training methodology. Some example sub-topics may be:

1. What is the effect of the embedding size on the semantics or interpretability of the learned embeddings?
2. What is the effect of initializing with larger network weights (as opposed to small weights like in the hint above)?

3. What is the effect of using a wider context window on the learned embeddings?

4. Which metrics are better for analysing the learned embeddings?

Essentially any meaningful deviation from the base implementation which may lead to some insight into the working of the model and the topic of embeddings. You will need to explore one such sub-topics. Please ensure that you have added your group to the google sheet and you are welcome to propose your own sub-topic on this sheet. Anyone with a blank sub-topic by the sub-topic deadline (we will use discord to coordinate this) will be allocated one (there's no penalty for not suggesting your own one). Suggested sub-topics are not final and I may still change your sub-topic but I will do my best to not be prescriptive.

You will also be expected to write a one-page (double column) report on the topic you covered (a second page may be used for figures, tables, your contribution statement and references). Please use the IEEE format for this. Please see the rubric below for more details on what should be included in the write-up. This write-up should focus on your methodology, results and insight for your particular topic. Naturally there will be overlap in what is written by each group, however, your aim should be to spend as much time discussing decisions or results for your particular setting.

2 Submission

Due Date: 13 August 2024 at 10:00.

For the submission you may work in pairs or groups of 3. You will be required to:

1. Submit your full code implementation (in a single file called "word2vec.py").

2. Submit a one-page report which includes a small statement on the contribution of each person if you did not work alone. This must be completed on a new paragraph.

3. Partake in the class discussion. It is likely that you/your group will be asked to briefly explain what you found so being prepared is recommended. That said it will be fairly informal. To receive full marks for this you will need to contribute to the discussion meaningfully.

The following is an indication of the rubric which will be used to assess your write-up and discussion.

Mode	0% to 20%	20% to 40%	40% to 60%	60% to 80%	80% to 100%
Write-up Structure	Adequate use of language and structure.	Fair use of language and structure.	Well written and clear	Good use of language and structure.	Excellent use of language, very well written and structured
Method	No clear description of the architecture, data or approach to training	Some description of high-level details	Fair description on the details of the base model but little elaboration of the approach to answering the sub-topic	Base approach is described in detail and steps are well justified. Adequate discussion on the approach to answering the sub-topic	Excellent description and motivation for all parts of the method including on how the sub-topic was answered
Results	Results are not given or irrelevant	Some results given with inappropriate metrics	General results presented with appropriate metrics	General and sub-topic results presented with appropriate metrics	Thorough results which are appropriate for answering the sub-topic and display the general correctness of the model
Discussion Section	No interpretation of results or incorrect interpretation. Little knowledge of the topic displayed	Some general interpretation of results broadly	Results are interpreted which display some understanding of the mechanics of the model. Displays a basic understanding of the topic	Results are interpreted and contextualized to begin to answer the sub-topic. Clear demonstration of knowledge of the general topic	Results are interpreted and contextualized to answer the sub-topic and display insight into the working of the model or original thought on the concepts
Verbal Assessment (Class Discussion)	No display of knowledge on the topic.	Adequate knowledge of the topic but unable to coherently answer the question (gives random facts or irrelevant information)	Provides a concise and correct but shallow answer	The answer is clear, concise and correct. Demonstrates some deeper understanding of the topic	Demonstrates insight into the topic. Answer is clear, concise and correct. Relates the topic to other material or contextualizes the topic more broadly.

3 On the Use of ChatGPT

These labs do not count much individually for the course, and yet they are crucial to understand the material and prepare for other assessments. Thus, I urge you to take them seriously and engage with the material. I am aware though that there is an incentive to just complete these labs as quickly as possible which would result in the use of generative tools to speed up the process. This is an NLP course though and the use of ChatGPT-like software is at least a learnable experience for us. Thus, you are welcome to use generative models for the written portion of these assessments. However, greater emphasis will now be placed on the factual correctness of what is said and the degree of insight which is shown in the write-ups. If I detect something resembling a clear "hallucination" (a generative model making up facts) you will receive 0. Negative marking will also be implemented for poor writing or formatting and incorrect information or incoherent reasoning. You will also receive less marks for extremely generic facts or information. The strategy then is to use these tools to get started but then add insight in afterwards - particularly insight directed towards the sub-topic that you are investigating. Equally, clever prompt-engineering will likely go a long way here. Appropriate use of both strategies will be rewarded.

Similarly, using generative models to help you code is fine but the usual plagiarism rules for the school remain (it is not tolerated). The class discussion will then be used to check that **all** students are engaging with the course material.