DeCAIR Lab Experiment Form

Author(s)	Gheith Abandah		
Author Organization Name(s)	The University of Jordan		
Work Package Number & Title	Work Package 8: Establishing and developing AIR labs		
Activity Number & Title	Activity 8.4: Design of the lab manuals that includes the experiment that will be conducted by the students enrolled in the labs.		
Work Package Leader	Sobhi Abou Shahin, Beirut Arab University		
Due Date of Delivery	30/9/2023	**Project Month**	M33
Submission Date	7/8/2023	**Project Month**	M31

Revision History

Version	Date	Author	Description	Action *	Page(s)
1	7/8/2023	Gheith Abandah	Original (base) document	C	1-3
2	19/4/2024	Gheith Abandah	Update for Spring 2024	U	1-3
3					
4					

(*) Action: C = Creation, I = Insert, U = Update, R = Replace, D = Delete

Disclaimer

This project has been co-funded by the Erasmus+ Programme of the European Union.

Email: DeCAIR@ju.edu.jo

Project Website: http://DeCAIR.ju.edu.jo/

Course Title	Natural Languages Processing
Course Number	0907753
Experiment Number	1
Experiment Name	Text Preprocessing and Word Embeddings
Objectives	The students learn basic skills in natural languages processing (NLP) using Python, NLTK, Gensim, and Keras.
Introduction	This is an introductory experiment in NLP. The student solves two exercises to practice some basic skills in NLP.
Materials	Computer with Python integrated development environment (IDE) software installed (PyCharm is recommended), or Jupyter Notebook (Google Colab is recommended). Dataset files: `GoogleNews-vectors-negative300.bin.gz`
Procedure	**Exercise 1: Text Preprocessing** For this exercise, you will practice text preprocessing techniques. 1. **Tokenization**: Take a sample sentence: "Despite the storm's rapid intensification, the resourceful residents managed to safeguard their community from the potentially devastating impacts." Tokenize this sentence using NLTK's **`word_tokenize`** function. 2. **Lowercasing**: Convert all tokens to lowercase. 3. **Stopword Removal**: Remove stopwords from the list of tokens using NLTK's English stopword list. 4. **Stemming**: Perform stemming on the filtered tokens using NLTK's **`SnowballStemmer`**. **Exercise 2: Word Embeddings** For this exercise, you will practice working with word embeddings using Word2Vec and an Embedding layer in Keras. 1. **Word2Vec**: Load the pre-trained Word2Vec embeddings **`GoogleNews-vectors-negative300.bin`** using Gensim's **`KeyedVectors.load_word2vec_format`**. Print the vector representation for the word 'computer'. Note that you can download the embeddings file from a Jupyter notebook using (warning: the file is about 3.5 GB): `!wget https://figshare.com/ndownloader/files/10798046 -O GoogleNews-vectors-negative300.bin` 2. **Embedding Layer in Keras**: Create a Sequential model in Keras and add an Embedding layer with 10,000 possible tokens and an embedding dimensionality of 32. Print the model summary.

Data Collection	Capture the output of your code for the above two exercises.
Data Analysis	None
Required Reporting	Submit your code and the captured output.
Safety Considerations	Standard safety precautions related to using computer.
References	1. Course slides available on https://www.abandah.com/gheith/ 2. H. Lane, C. Howard, and H. Hapke, Natural Language Processing in Action: Understanding, analyzing, and generating text with Python, Manning, 2019. 3. Aurélien Géron, Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow, O'Reilly, 3rd Edition, 2022.