

## DeCAIR Lab Experiment Form

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

You are free to share, copy and redistribute the material in any medium or format, as well as adapt, transform, and build upon the material for any purpose, even commercially, provided that you give appropriate credit to the project and the partnership, and indicate if any changes were made. You may do so in any reasonable manner, but not in any way that suggests the partnership, or the European Commission endorses you or your use. You may not apply legal terms or technological measures that legally restrict others from using the material in the same manner that you did.

Copyright © DeCAIR Consortium, 2021-2024

Email: [DeCAIR@ju.edu.jo](mailto:DeCAIR@ju.edu.jo)

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

The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

<b>Course Title</b>	Natural Languages Processing
<b>Course Number</b>	0907753
<b>Experiment Number</b>	1
<b>Experiment Name</b>	Text Preprocessing and Word Embeddings
<b>Objectives</b>	The students learn basic skills in natural languages processing (NLP) using Python, NLTK, Gensim, and Keras.
<b>Introduction</b>	This is an introductory experiment in NLP. The student solves two exercises to practice some basic skills in NLP.
<b>Materials</b>	Computer with Python integrated development environment (IDE) software installed (PyCharm is recommended), or Jupyter Notebook (Google Colab is recommended).  Dataset files: <code>GoogleNews-vectors-negative300.bin.gz</code>
<b>Procedure</b>	<p><b>Exercise 1: Text Preprocessing</b> For this exercise, you will practice text preprocessing techniques.</p> <ol style="list-style-type: none"> <li><b>Tokenization:</b> Take a sample sentence: "<a href="#">Despite the storm's rapid intensification, the resourceful residents managed to safeguard their community from the potentially devastating impacts.</a>" Tokenize this sentence using NLTK's <code>word_tokenize</code> function.</li> <li><b>Lowercasing:</b> Convert all tokens to lowercase.</li> <li><b>Stopword Removal:</b> Remove stopwords from the list of tokens using NLTK's English stopwords list.</li> <li><b>Stemming:</b> Perform stemming on the filtered tokens using NLTK's <code>SnowballStemmer</code>.</li> </ol> <p><b>Exercise 2: Word Embeddings</b> For this exercise, you will practice working with word embeddings using Word2Vec and an Embedding layer in Keras.</p> <ol style="list-style-type: none"> <li><b>Word2Vec:</b> Load the pre-trained Word2Vec embeddings <code>GoogleNews-vectors-negative300.bin</code> using Gensim's <code>KeyedVectors.load_word2vec_format</code>. Print the vector representation for the word '<a href="#">computer</a>'. Note that you can download the embeddings file from a Jupyter notebook using (warning: the file is about 3.5 GB):  <pre>!wget https://figshare.com/ndownloader/files/10798046 -O GoogleNews-vectors-negative300.bin</pre></li> <li><b>Embedding Layer in Keras:</b> 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.</li> </ol>

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