# Advanced Machine Learning - Assignment 4 Part 1 Report

**Submitted by:**   
Miryam Horovicz - 313246704, Yaniv Tal - 031431166

## Notebook Overview

Our submitted notebook consists of three sections:

1. **General section** – handles the logistics:
   1. Imports
   2. Mount Google drive
   3. Download data
   4. Preprocess data i.e. arrange full text, prepare data frame
2. **Task part 1 –** Comparison model training by LDA topic modelling
3. **Task part 2** – Comparison model Analysis and Evaluation

## Part 1 Comparison model training

Comparison model training consists of three sections:

1. **Preprocess data** – prepare data for topic modelling:
2. Tokenizing the text
3. Removing stop words
4. Convert the tokenized object into a corpus and dictionary by bag of words method.
5. **Topic model training -** train LDA model applying following architecture

Diagram

Description automatically generated

Trained model results can be shown by words and probabilities in each topic as Shown below

Text

Description automatically generated

1. **Similarity function -** function that get two documents and return similarity measure between them:
2. Apply bag of words for each document
3. Apply the trained LDA model and get representation vector
4. Return cosine similarity between the two vectors.

## Part 2 – Comparison model Analysis and Evaluation

For LDA model analysis and evaluation we used two techniques:

1. **Word Cloud** – Show most frequent words in each topic, bigger font means higher word probability in this topic:

Qr code

Description automatically generated with medium confidence

1. **Interactive Visualization -** interactive visualization in the notebook that shows topics views. User can navigate between the different topics and see the words distribution.

