|  |  |  |
| --- | --- | --- |
|  | **ASSIGNMENT 02** |  |
|  | Marks: 05 |  |

**NAME: \_\_\_\_Nameer Iqbal\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**CLASS:\_\_\_\_\_BS AI (6A)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**REG. NO. \_\_\_(02-136212-016)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**COURSE:\_\_\_\_Natural Language Processing\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**COURSE CODE:\_CSC-441\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Marks Obtained: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Theoretical Assignment**

Read Carefully:

* The deadline for this assignment is *before* or *on* **03/June/24**

**WARNING**: This is an individual assignment; you must solve it by yourself. Any form of plagiarism will result in receiving a zero on the assignment.

**WARNING**: Late submissions will not be accepted. Any assignment submitted after the cutoff time will receive zero.

* You have to **answer** and submit in **SOFTCOPY** as well as **HARDCOPY** of the given draft. Submit SOFTCOPY on your LMS and HARDCOPY to your CR (CR will submit all assignments to Faculty on the submission date). Both Hardcopy and Softcopy are mandatory.

**QUESTION (Do as Directed) ( 2.5 + 2.5 Marks)**

**Question #1: (CLO 3, PLO 2, BT: C4)**

Develop a comprehensive dataset consisting of a minimum of 500 records in a category or domain of your choice, such as user reviews, social media comments, medical prescriptions, or excerpts from literature. Once you have gathered this dataset do the following

1. ***Identify*** and apply applicable range of Natural Language Processing (NLP) tools and techniques to preprocess and analyze the data. These might include tokenization, stop word removal, stemming, lemmatization, and Information Extraction, among other.
2. ***Select*** an appropriate machine learning algorithm and build a predictive model based on the prepared data. Document each step of your process, from data collection and preprocessing to model selection and evaluation, ensuring a clear and professional presentation of your methodology and results

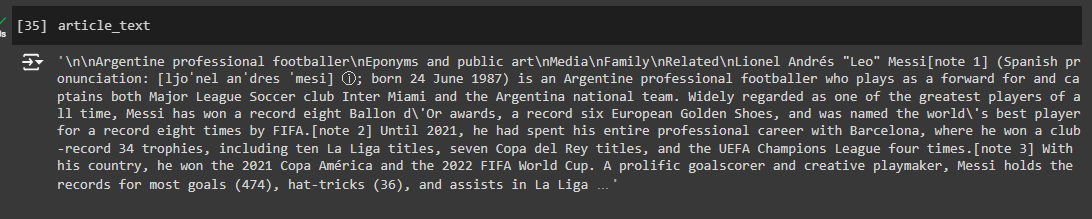
**Solution:**

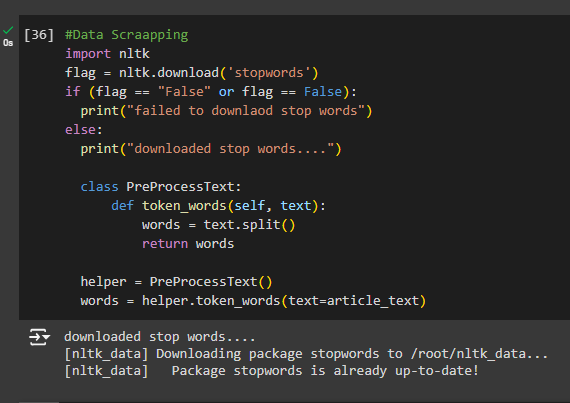
The dataset usedfor this problem is a wikipedia webpage.

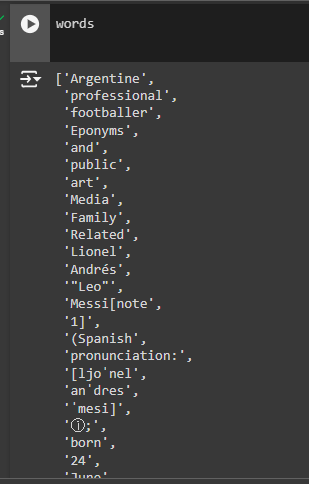
<https://en.wikipedia.org/wiki/Lionel_Messi>

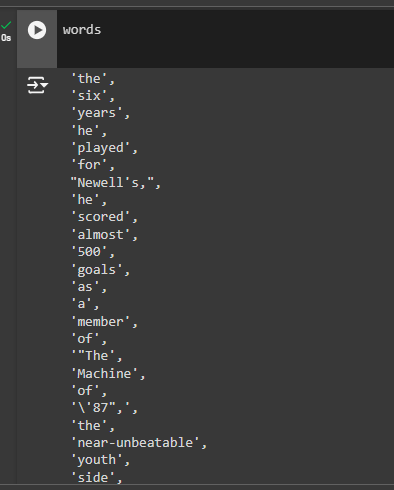
**Data Preprocessing:**





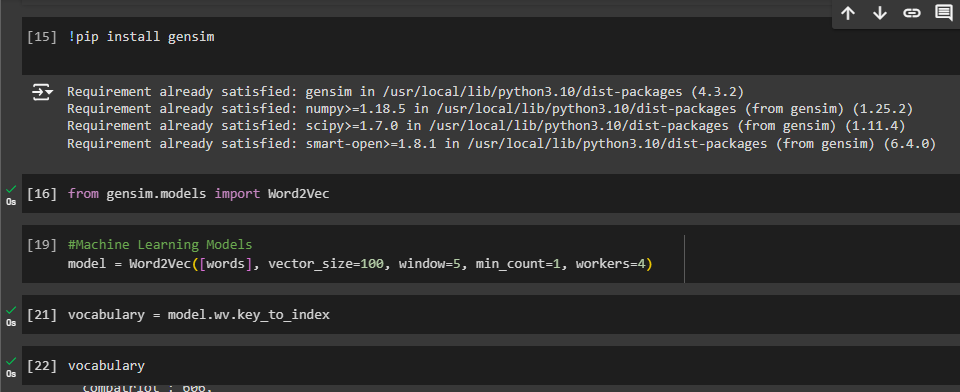


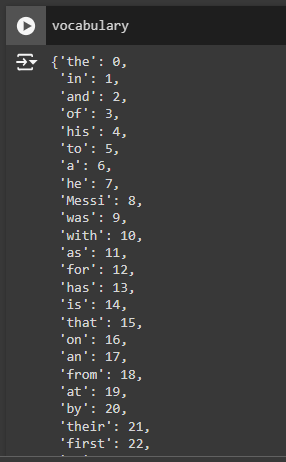


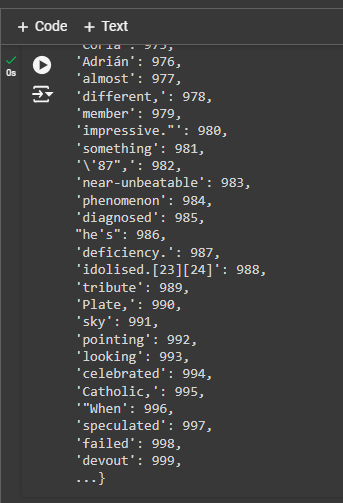


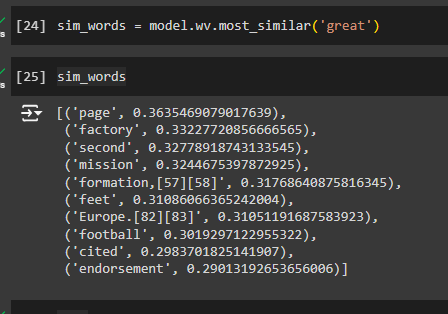
**Machine Learning Models:**

- Word2Vec is used to represent words as vectors (array). These vectors capture the meaning of words by placing similar words close to eachother in the vector space.

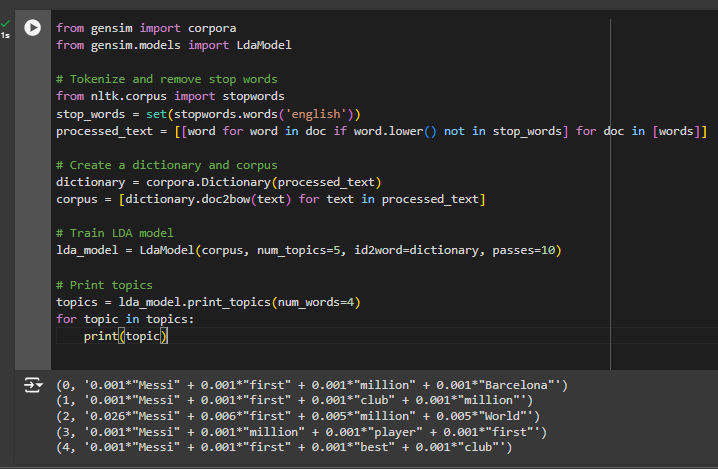








- LDA model is used to find hidden topics within a collection of documents. Each document is viewed as a mixture of topic and each topic is a mixture of words.



**Summary:**

- Data Scraping: Fetches and processes text from a Wikipedia page.

- Text Preprocessing: Tokenizes text and removes stopwords.

- Word2Vec: Creates word embeddings and finds similar words.

- LDA: Identifies topics within the text and lists the main words in each topic.