|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Kingdom of Saudi Arabia**  **Ministry of Education**  **University of Jeddah**  **College of Computer Science and Engineering**  **Department of Computer Science and Artificial Intelligence** | Logo, company name  Description automatically generated | **المملكة العربية السعودية**  **وزارة التعليم**  **جامعة جدّة**  **كلية علوم وهندسة الحاسب**  **قسم علوم الحاسب والذكاء الاصطناعي** | |  |  |

|  |
| --- |
| **Lab 2** |
| **CCAI323 Machine learning** |
| **First Semester 2023/2024**   |  |  | | --- | --- | | **Lab Date/Time: xxx**  **Lab assignment submission Date/Time: xxxx** |  | | **Student Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  **Student ID: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | | |

|  |  |
| --- | --- |
| **Instructor Name** | **Section** |
|  |  |

**Instructions**:

The lab assignments must be submitted before the allocated Date/Time.

The lab assignments must by uploaded to blackboard.

Plagiarism will be punished according to university rules.

|  |  |
| --- | --- |
| **PLO/CLO** | **SO** |
|  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Max Score** | **Student Score** |
| **PLO S2 / CLO 2 / SO 2** | **Questions 1** | **8** |  |
| **Total** | |  |  |

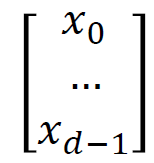
**Lab2 overview content:**

|  |  |  |  |
| --- | --- | --- | --- |
| Title | Description | Jupter notbook | Lab Questions |
| Part A: review of basic | | | |
| What is the Data in Unsupervised Machine Learning? | Demo on text processing and creating word clouds | SklearnText.ipynb | Question1 |
| Sklearn Demo | Examples of generating datasets using sklearn | sklearnDatasets.ipynb | none |

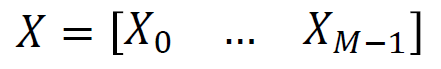
**What is data in unsupervised learning:**

The goal of Unsupervised ML is to find patterns in data, Question: what is the data?

**Data and Datasets:**

* Typically, a data point is a vector in d dimensions.

Data point often called feature vector as each entry represents a feature

* Let **𝑋0, 𝑋1, …, 𝑋𝑀−1** denote the M data points, then dataset is represented as a matrix of dimensions d rows and M columns.

Throughout the course we will call X the dataset or the feature matrix.

**Example: Customer Segmentation**

Data point consists of features of customer.

Example: 4 features

* age = 27
* # of credit cards = 4
* salary = 60,000
* Text

  Description automatically generated# of purchases = 10
* Data point represented as:

Combine feature vectors for multiple customers to create feature matrix.

**Example: Natural Language Processing**

How do we convert text into a feature matrix?

Simple approach is word count.

* Create dictionary of all words (case insensitive)
* Count number of times each word appears in each document.

Consider 3 messages:

“Call me soon”, “CALL to win”, “Pick me up soon”

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| WORDS |  | Feature Matrix | | |
| call |  | 1 | 1 | 0 |
| me |  | 1 | 0 | 1 |
| pick |  | 0 | 0 | 1 |
| soon |  | 1 | 0 | 1 |
| to |  | 0 | 1 | 0 |
| up |  | 0 | 0 | 1 |
| win |  | 0 | 1 | 0 |

Term Frequency Inverse Document Frequency ( Tfidf ) approach:

* Term frequency: number of times word appears in document
* Inverse document frequency: inverse of number of documents in which word appears
* Tfidf is term frequency multiplied by inverse document frequency (with scaling)
* Logic: if word appears in many documents, then its importance/weighting is lowered

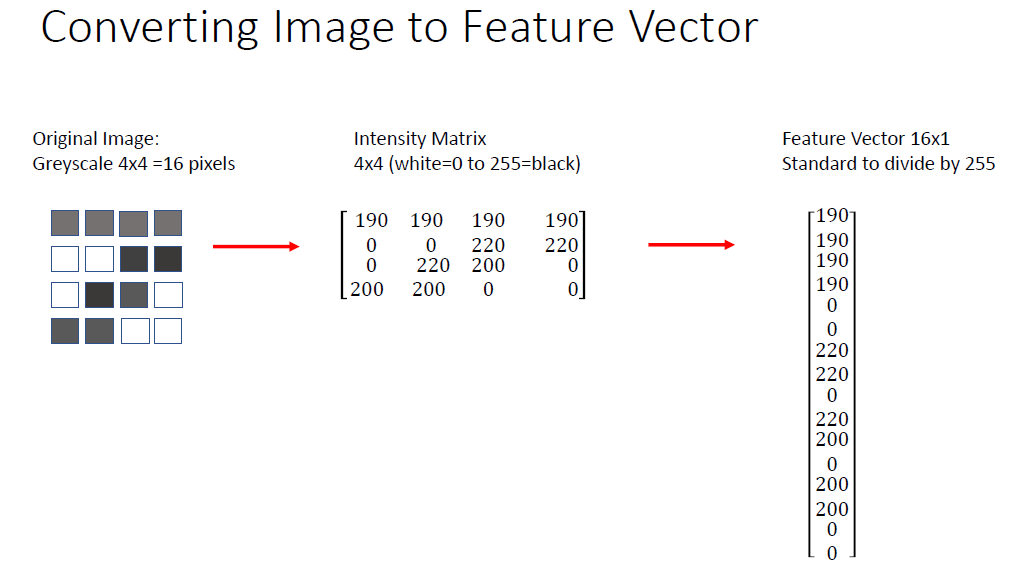
Messages: “Call me soon”, “CALL to win”, “Pick me up soon”

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| WORDS |  | Feature Matrix | | |
| call |  | 0.58 | 0.47 | 0.00 |
| me |  | 0.58 | 0.00 | 0.43 |
| pick |  | 0.00 | 0.00 | 0.56 |
| soon |  | 0.58 | 0.00 | 0.43 |
| to |  | 0.00 | 0.62 | 0.00 |
| up |  | 0.00 | 0.00 | 0.56 |
| win |  | 0.00 | 0.62 | 0.00 |

**Example: Images**

* Images typically are composed of rectangular arrays of pixels
* For black and white images, intensity of greyscale for each pixel is represented by a number between 0 and 255 (0=white, 255=black)
* Feature vector for image is vector of intensities for all pixels
* For colour images, each pixel represented by 3 values intensities of red, blue, and green components for that pixel feature vector in colour case vector will be 3 times longer than in black and white case.

**Converting Images to feature vector:**



**sklearn dataset demo:**

* sklearn is shortened version of name of package scikit learn
* sklearn is a package for machine learning: There is a plenty of functionality in this package, including codes for supervised and unsupervised machine learning.
* Demo shows how to create datasets for testing clustering algorithms
* See following sites for details:

https://scikit learn.org/stable/modules/generated/sklearn.datasets.make\_blobs.html

https://scikit learn.org/stable/modules/generated/sklearn.datasets.make\_moons.html

https://scikit learn.org/stable/modules/generated/sklearn.datasets.make\_circles.html

Please implement the commands to explore example datasets from sklearn that will be used in this course.

**Download Jupyter Notebook for demo:**

* sklearnDatasets.ipynb

**Websites for Data:**

Sklearn Toy Datasets:

* https://scikit learn.org/stable/datasets/toy\_dataset.html
* 7 easy to use datasets

University of California, Irvine Machine Learning Data Repository:

* <https://archive.ics.uci.edu/ml/index.php>
* Contains 100s of freely available machine learning datasets

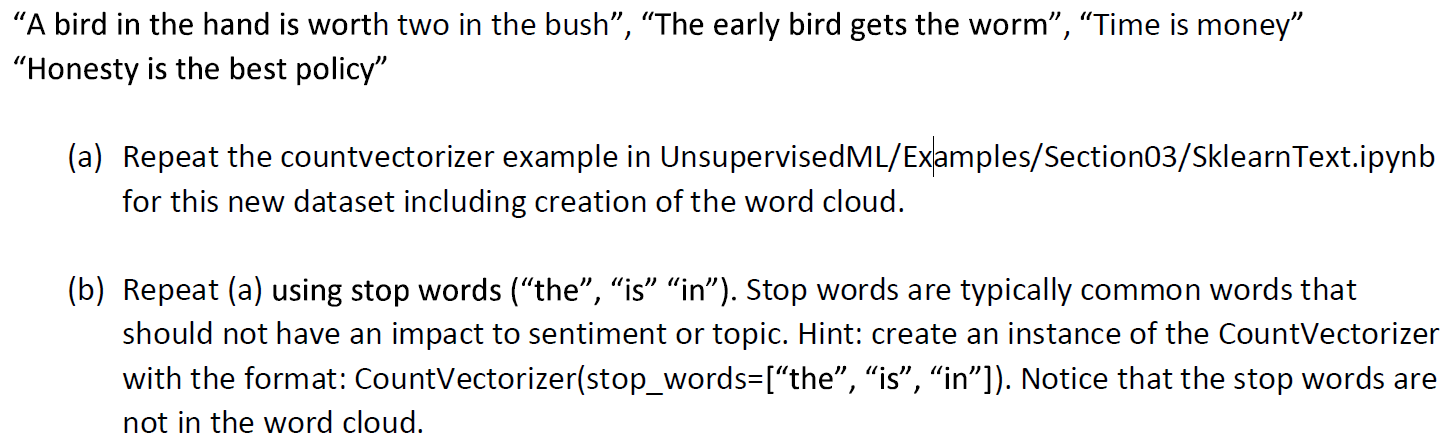
Kaggle:

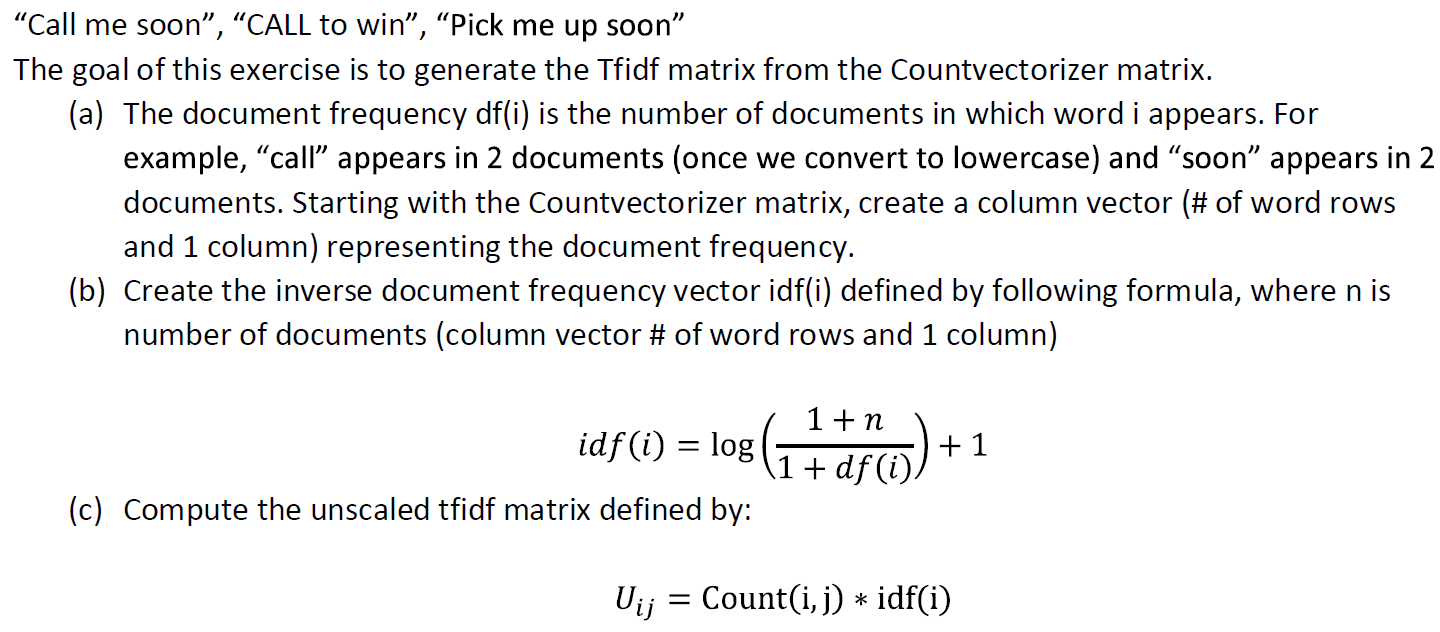
* [www.kaggle.com](http://www.kaggle.com)
* Site for data science competitions (often with prize money) with freely available data
* Can learn from tutorials and notebooks created by others
* You will need to create a free account to access Kaggle resources

**Download Jupyter Notebook for demo:**

SklearnText.ipynb

**Question 1: [PLO K1 / CLO 1 / SO 1] [8 marks]**

**Q1.1: Consider the messages:**

**Q1.2: let’s go back to the example from the demo with documents:**

Text, letter

Description automatically generated