

ĐẠI HỌC KHOA HỌC TỰ NHIÊN

MÔN HỌC: KHOA HỌC DỮ LIỆU

BÁO CÁO ĐỒ ÁN CUỐI KỲ

---

# Word2Vec

---

*Giảng viên*

Nguyễn Ngọc Đức

*Sinh viên*

20424008 - Dương Mạnh Cường



Ngày 9 tháng 2 năm 2022

# Mục lục

<b>1</b>	<b>Section level 1</b>	<b>1</b>
1.1	Section level 2 . . . . .	1
1.1.1	Section level 3 . . . . .	2
1.1.1.1	Section level 4 . . . . .	2

## 1 Section level 1

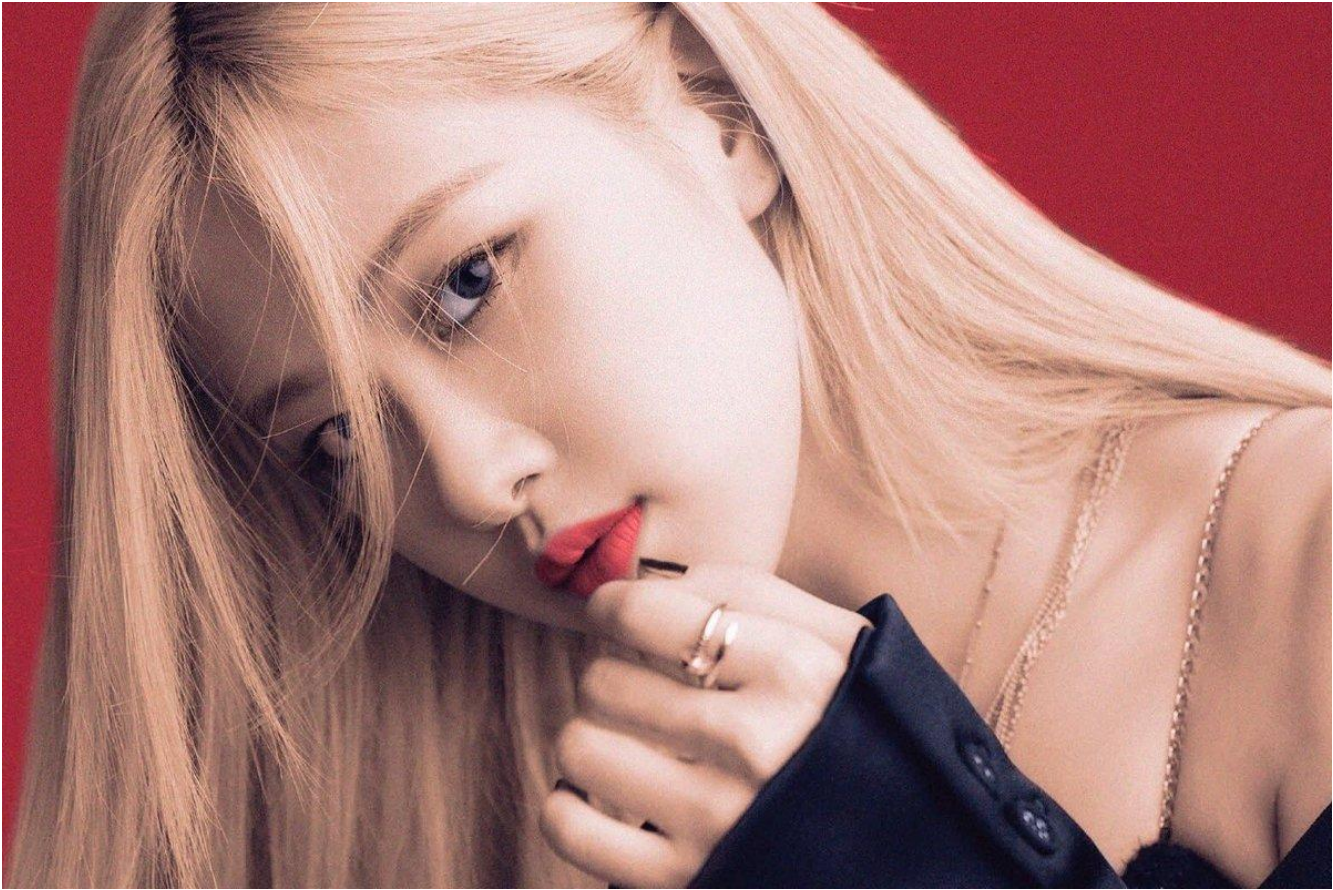
First paragraph under section level 1.

Second paragraph: [https://this\\_is\\_your\\_url](https://this_is_your_url).

### 1.1 Section level 2

First paragraph under section level 2.

**Bước 1:** Title for image



Hình 1: This is caption for image.

This is left left indent [This is href tag](#).

This is title for code block.

```
1 # terminal máy master
2
3 ./start-all.sh
```

### 1.1.1 Section level 3

- `main.ipynb`: some text.
- `main_hadoop.ipynb`: some text.
- `data/cruise_ship_info.csv`: some text.

#### 1.1.1.1 Section level 4

This is text under section level 4.

Trong dataset này sẽ có các feature:

- Age: tuổi của tàu (tính đến năm 2013).
- passengers: số hàng khách ( $\times 100$  người).

Import các thư viện cần thiết.

```
1 from pyspark.sql.session import SparkSession
2 from pyspark.sql.functions import corr
3 from pyspark.ml.feature import StringIndexer, VectorAssembler
4 from pyspark.ml.regression import LinearRegression
5 from pyspark import SparkContext
6
7 import os
8 import findspark
```

#### Nhận xét

- Tập dữ liệu `./data/cruise_ship_info.csv` có tổng cộng 158 quan sát.