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 22 tháng 2 năm 2022

Mục lục

1 Section level 1

First paragraph under section level 1.

Second paragraph: 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.

```
# terminal máy master

./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.

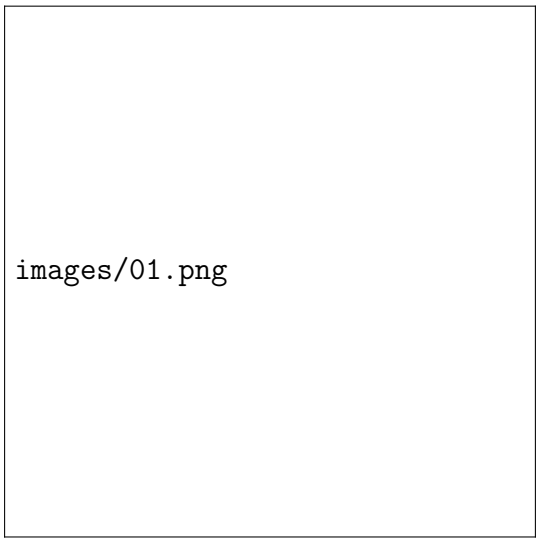
```
from pyspark.sql.session import SparkSession
from pyspark.sql.functions import corr
from pyspark.ml.feature import StringIndexer, VectorAssembler
from pyspark.ml.regression import LinearRegression
from pyspark import SparkContext

import os
import findspark
```

Nhận xét

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

1. **CBOW model**.
2. **Skip-gram model**.



Hình 2: Các từ có ý nghĩa tương đồng nhau nằm gần nhau.