INTRODUCTION OF DATA SCIENCE

Giảng viên: TS. Bùi Thanh Hùng
Bộ môn Khoa học dữ liệu, Khoa Công nghệ thông tin
Đại học Công nghiệp thành phố Hồ Chí Minh
Email: buithanhhung@iuh.edu.vn
Website: https://sites.google.com/site/hungthanhbui1980/

Bài 1:

Cho dữ liệu như file SampleInput.txt có nội dung như sau:
School = Riverdale High
Grade = 1
Student number, Name
0, Phoebe
1, Rachel

Student number, Score
0, 3
1, 7

Grade = 2
Student number, Name
0, Angela
1, Tristan
2, Aurora

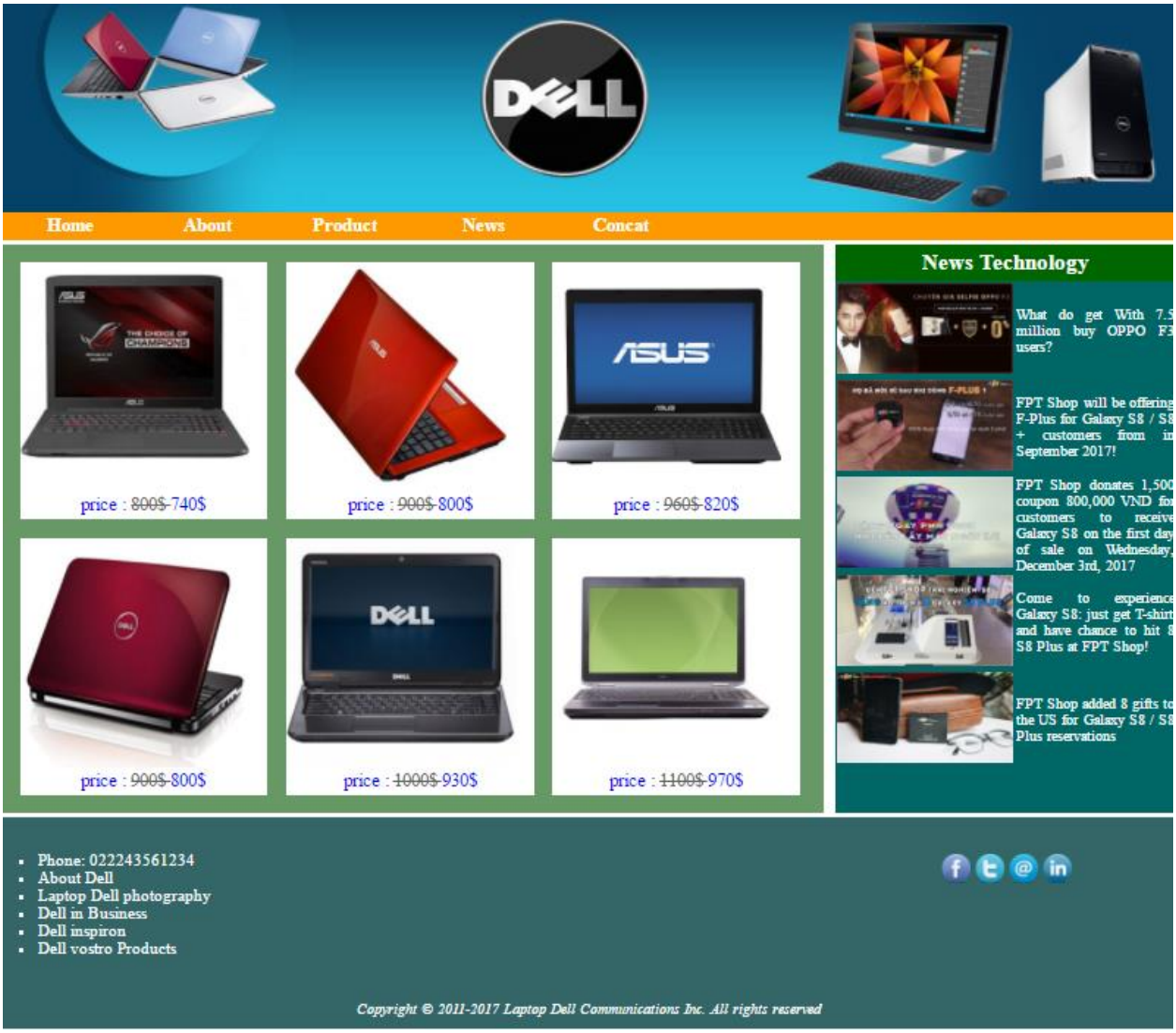
Student number, Score
0, 6
1, 3
2, 9
…..

Hãy viết code sử dụng regular để chuyển dữ liệu trên về dữ liệu có cấu trúc như sau:

School	Grade	Student number	Name	Score
Hogwarts	1	0	Ginny	8
		1	Luna	7
	2	0	Harry	5
		1	Hermione	10
	3	0	Fred	0

		1	George	0
Riverdale High	1	0	Phoebe	3
		1	Rachel	7
	2	0	Angela	6
		1	Tristan	3
		2	Aurora	9

Bài 2:
Sử dụng html tự học từ trang www.w3schools.com xây dựng một website đơn giản với giao diện như sau:



Gợi ý: Hình ảnh có thể lấy các hình khác và sử dụng các tags : div, ul, li

Bài 3:
Cho dữ liệu thô trong file gửi đính kèm.

https://drive.google.com/file/d/1TFRQ6MXXXhIL5NyCbmmQK-KJLNHt1X6g/

Viết 1 hàm chuyển từ dữ liệu thô sang chuỗi các từ theo các bước như sau:
3.1 Loại bỏ các thẻ HTML
3.2 Loại bỏ các non-letters
3.3 Chuyển các từ thành chữ thường và tách thành các từ riêng biệt với giả thuyết là các từ cách nhau bởi khoảng trắng
3.4 Loại bỏ các stop words
3.5 Chuyển các từ về từ gốc: Stem words (sử dụng thư viện SnowballStemmer của nltk)
3.6 Nối các từ thành một chuỗi, các từ cách nhau bởi khoảng trắng và trả về kết quả
3.7 Chuyển thành đặc trưng TF-IDF
3.8 Tính độ đo tương đồng Cosin dựa trên đặc trưng TF-IDF của 2 câu bất kỳ