NATURAL LANGUAGE PROCESSING

Lecturer: Doctor Bui Thanh Hung
Data Science Laboratory
Faculty of Information Technology
Industrial University of Ho Chi Minh city
Email: hung.buithanhcs@gmail.com (buithanhhung@iuh.edu.vn)
Website: https://sites.google.com/site/hungthanhbui1980/

Bài 1: Sử dụng BeautifulSoup và urllib để lấy các tin tự động trên trang https://news.google.com/news/rss

Viết các hàm xử lý để in ra màn hình 3 thông tin của mỗi tin: Tựa tề, Đường link và Ngày phát hành, lưu kết quả vào 1 file theo từng dòng với cấu trúc như trên.

Bài 2: Cho dữ liệu như sau:
Data = [{'Thanh pho': 'Hai Phong', 'Nhiet do': 32}, {'Thanh pho': 'Da Nang', 'Nhiet do': 29}, {'Thanh pho': 'Can Tho', 'Nhiet do': 34}]

1- Hãy chuyển dữ liệu trên thành số sử dụng thư viện DictVectorizer
2- In ra tên các feature

Bài 3: Cho dữ liệu sau:

```
data= [{'word-2': 'con',
        'pos-2': 'DT',
        'word-1': 'mèo',
       'pos-1': 'NN',
        'word+1': 'trên',
       'pos+1': 'PP',    }]
```

1- Hãy chuyển dữ liệu trên thành số sử dụng thư viện
2- In ra tên các feature

Bài 4: Cho dữ liệu sau:

```
dulieu=[
‘Hôm_nay tôi đi_học',
' Hôm_nay tôi đi_học ở trường',
' Hôm_nay tôi nghỉ ở nhà',
 'Hôm_nay tôi có đi_học không?']
```

1- Hãy chuyển dữ liệu trên thành số sử dụng thư viện CountVectorizer
2- In ra tên của các feature
3- Vào một câu bất kỳ, in ra vector giá trị số của câu đó dựa trên dữ liệu ở trên

Bài 5: Cho dữ liệu sau:
Tôi là sinh viên trường Đại học Công nghiệp thành phố Hồ Chí Minh

Chuyển câu trên thành unigram, bigram và trigram sử dụng thư viện có sẵn và in kết quả ra màn hình.

Bài 6: Cho dữ liệu huấn luyện như sau:
<s> tôi là IUH </s>
<s> IUH là tôi </s>
<s> IUH tôi học ở </s>
<s> IUH tôi đã học ở </s>
<s> tôi đã học ở IUH sao </s>
Giả sử chúng ta sử dụng bigram language model dựa trên dữ liệu đã huấn luyện ở trên. .
1. Hãy cho biết từ có thể xuất hiện tiếp theo theo model huấn luyện ở trên trong các câu sau là gì?
(1) <s> IUH . . .
(2) <s> IUH tôi học . . .
(3) <s> IUH tôi là IUH . . .
(4) <s> tôi đã học . . .
2. Trong các câu sau, câu nào cho xác suất cao nhất với mô hình đã huấn luyện ở trên?
(5) <s> IUH tôi đã tôi học </s>
(6) <s> IUH tôi là </s>
(7) <s> tôi đã học IUH tôi là </s>
3. Hãy tính complexity của câu sau:
<s> tôi đã học ở IUH
4. Tính xác suất bigram dựa trên mô hình huấn luyện ở trên cho các từ sau:
P(học |<s>)
P(học | IUH)
P(IUH|<s>)
P(IUH| học)
P(tôi|IUH)
P(tôi|học)
P(học|tôi)

5. Hãy tính xác suất của câu sau:

(8) <s> tôi đã học ở IUH sao

(9) <s> IUH tôi đã học

Bài 7: Cho dữ liệu như ở bài 4:

dulieu=[
'Hôm_nay tôi đi_học',
' Hôm_nay tôi đi_học ở trường',
' Hôm_nay tôi nghỉ ở nhà',
'Hôm_nay tôi có đi_học không?']

Với các định nghĩa như sau về TF-IDF:

Tf means **term-frequency** while tf–idf means term-frequency times **inverse document-frequency**:

tf-idf(t, d) = tf(t, d) * log(N/(df + 1))

- ✓ *tf(t,d) = count of t in d / number of words in d*
- ✓ *df(t) = occurrence of t in documents*
- ✓ *idf(t) = log(N/(df + 1))*
- t — term (word), d — document (set of words)
- n — count of corpus, corpus — the total document set

Kết quả sẽ được chuẩn hóa bằng công thức Euclidean

$$v_{norm} = \frac{v}{||v||_2} = \frac{v}{\sqrt{v_1^2+v_2^2+\cdots+v_n^2}}$$

Hãy tự tính TF-IDF của văn bản đã cho theo các công thức ở trên