Workshop 2

Basic Text Processing

Requirements

- Python 3
- nltk library
- nltk modules

Install NLTK

- **Install NLTK Package**

```
!pip install nltk
```

- **Download NLTK modules**

```
import nltk
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
```

Task 1

- Pick your favorite song and use its lyric as the input.
- Pre-process the input.
  - Tokenize
  - Remove stopwords
  - Apply normalization
  - Compare stemming vs lemmatization

Task 1: Output Format

```
Original text:
<text>

Tokens:
[...]

Cleaned:
[...]

Stems vs Lemmas:
| Word | Stem | Lemma |
```

Task 2

- Recall Task 1, which preprocessing step changed the text the most?
- In your opinion, would stemming or lemmatization work better for each of the following NLP applications?
  - search engines
  - machine translation
  - sentiment analysis

Task 3

Write regular expressions for the following languages.

1. the set of all binary strings with odd length
2. the set of all lowercase English alphabetic strings ending with a letter ‘b’
3. the set of all strings from the alphabet {*a*, *b*} such that each *a* is immediately preceded by and immediately followed by a *b*;

Submit your answers in a PDF file.

Task 4 (1/3)

Implement a program that extracts the URLs, e-mail addresses and hashtags from the given text.

- URLs => (urls.txt)
  - A URL is generally following this structure: scheme://hostname/path?query
  - Scheme: The communication protocol used to access the resource (e.g. https:// http:// ftp://).
  - Hostname: The domain name like example.com. This can include subdomains (e.g., blog.example.com).
  - Path: The specific location of the resource on the server, with directories separated by forward slashes (e.g., /software/index.html).
  - Query string: an optional part that provides additional parameters to the resource. It begins with a question mark (?) and consists of key-value pairs separated by ampersands (&) (e.g., ?id=1234&sort=asc).

Task 4 (2/3)

- E-mail address => emails.txt
  - An email address follows the format username@domain.com, where username is a unique local part before the "@" symbol and domain.com is the domain name of the mail server.
  - Username: The part of the email address before the "@" symbol. It can include letters, numbers, and certain special characters like periods (.), hyphens (-), and underscores (_).
  - @ Symbol: This symbol separates the username from the domain name and is a required part of the format.
  - Domain Name: The part after the "@" symbol, which identifies the mail server. It includes the domain name and the top-level domain (TLD). In this task consider only .com, .org, .edu, and .net.

Task 4 (3/3)

- Hashtags (hashtags.txt)
  - A hashtag starts with the pound sign (#) directly before a word or phrase (a combination of letters in uppercase or lowercase, digits, underscores), with no spaces, special characters ($, %, &, @) or punctuation (. , !, ?) in between.

- Submit your python code along with input file and the three text files (with the extracted information).