CS5624: NLP

Programming Assignment-1

This assignment focuses on the extraction of all the dates mentioned in an input text file and the standardization of them to a unique format. For example, your program should identify all the date expressions (highlighted in bold) in the following pieces of text that give details about India's Independence day:

*Independence Day is celebrated annually on **15 August** as a national holiday in India commemorating the nation's independence from the United Kingdom on **15 August 1947**, the day when the provisions of the 1947 Indian Independence Act, which transferred legislative sovereignty to the Indian Constituent Assembly, came into effect. India retained King George VI as head of state until its transition to a republic, when the nation adopted the Constitution of India on **26 January 1950** (celebrated as Indian Republic Day) and replaced the dominion prefix, Dominion of India, with the enactment of the sovereign law Constitution of India. India attained independence following the Independence Movement noted for largely non-violent resistance and civil disobedience.*

*It was on **15th August 1947** that India was declared independent from British colonialism, and the reins of control were handed over to the leaders of the Country.*

*The Red Fort in Dehli is also an important Independence Day symbol in India as it is where Indian Prime Minister Jawahar Lal Nehru unveiled India's flag on **August 15, 1947**.*

*At midnight between **14/08/1947** and **15/08/1947** free India's first prime minister Pandit Jawaharlal Nehru gave his famous "Tryst with Destiny" speech.*

(Source:
https://en.wikipedia.org/wiki/Independence_Day_(India),https://knowindia.india.gov.in/independence-day-celebration/,
https://www.timeanddate.com/holidays/india/independence-day)

Once a date is identified it should be standardized to **day-month-year** format (note the hyphen). For example, if your program identifies 14/08/1947 as a date then it should standardize it to **14-August-1947**. If the identified date does not have an explicit year

then the default year should be 2022. For example, in the following sentence year is not explicitly associated with *15 August.* In such a case, your program should extract **15-August-2022** as the standardized date.

*Independence Day is celebrated annually on **15 August** as a national holiday in India…*

Similarly, the default values for day and month are 01 and January respectively.

You should write a date_extraction_rollno.py (you should properly document your code). It should take one text file as an input and produce and an XML file as an output (dates_rollno.xml). For the example text given above, your program should produce the following output:

<output>
*Independence Day is celebrated annually on <date std_date=15-August-2022>**15 August** </date> as a national holiday in India commemorating the nation's independence from the United Kingdom on <date std_date=15-August-1947>**15 August 1947** </date>, the day when the provisions of the 1947 Indian Independence Act, which transferred legislative sovereignty to the Indian Constituent Assembly, came into effect. India retained King George VI as head of state until its transition to a republic, when the nation adopted the Constitution of India on <date std_date=26-January-1950> **26 January 1950** </date>(celebrated as Indian Republic Day) and replaced the dominion prefix, Dominion of India, with the enactment of the sovereign law Constitution of India. India attained independence following the Independence Movement noted for largely non-violent resistance and civil disobedience.*

*It was on <date std_date=15-August-1947>**15th August 1947** </date> that India was declared independent from British colonialism, and the reins of control were handed over to the leaders of the Country.*

*The Red Fort in Dehli is also an important Independence Day symbol in India as it is where Indian Prime Minister Jawahar Lal Nehru unveiled India's flag on <date std_date=15-August-1947>**August 15, 1947** </date>.*

*At midnight between <date std_date=14-August-1947>**14/08/1947** </date> and <date std_date=15-August-1947>**15/08/1947** </date> free India's first prime minister Pandit Jawaharlal Nehru gave his famous "Tryst with Destiny" speech.*

</output>