Github Repo

CSDX 226 – WEB ANALYTICS AND SOCIAL MEDIA MINING

CAT 1 Assignment

ALI IZZATH SHAZIN K

220071601028
BTECH CSE A

TASK 1. **Tokenize the following contents about "Natural Language Tool Kit" into words and sentences using Python libraries (nltk, etc..)**

"Natural Language Toolkit (NLTK) is one of the largest Python libraries for performing various Natural Language Processing tasks. From rudimentary tasks such as text pre-processing to tasks likes vectorized representation of text – NLTK's API has covered everything."

TASK 2. **Remove stop words from the following text using NLTK.**

"Natural Language Toolkit (NLTK) works as a powerful Python library that a wide range of tools for Natural Language Processing (NLP). From fundamental tasks like text pre-processing to more advanced operations such as semantic reasoning, NLTK provides a versatile API that caters to the diverse needs of language-related tasks."

TASK 3. **Convert the text below to lowercase and remove all punctuation.**

“Let’s eat, Grandma!

Grandma, Let’s eat!

Silvia, Are you free tomorrow?

Yes, I’m free on Saturday.”

TASK 4. **Clean the below given text about "Importance of Text Preprocessing in NLP" by removing extra whitespaces and special characters.**

" @@Natural Language Processing (NLP)!!! is a field of AI that focuses on ... enabling computers to understand, interpret, & generate human language.

NLP includes tasks like **tokenization, lemmatization,** && sentiment analysis.

It helps in applications such as chatbots, machine translation, and voice assistants!!!

However, cleaning text—removing extra spaces, punctuations, && special $$$ characters—is crucial.

Without preprocessing, NLP models may not perform accurately !!!

So, can you clean this messy text & make it structured??? "

TASK 5. **Given a blog on "The impact of Automation and AI on Modern Industries", apply stemming and lemmatization and compare the results.**

" The researchers are analyzing various datasets to study the effects of automation.

They observed that automated systems perform tasks more efficiently than humans.

Many industries have been adopting AI-driven solutions to improve productivity.

Running complex algorithms helps in predicting future trends accurately.

Several companies are investing in developing smarter and more adaptive models.

Data scientists continuously refine their models to achieve better performance.

The advancements in technology have transformed the way businesses operate. "

```
Sentences: ['Natural Language Toolkit (NLTK) is one of the largest Python\nlibraries for performing various Natural Lang
uage Processing tasks.', "From rudimentary tasks such as text pre-processing to tasks likes\nvectorized representation o
f text - NLTK's API has covered everything."]

Words: ['Natural', 'Language', 'Toolkit', '(', 'NLTK', ')', 'is', 'one', 'of', 'the', 'largest', 'Python', 'libraries',
'for', 'performing', 'various', 'Natural', 'Language', 'Processing', 'tasks', '.', 'From', 'rudimentary', 'tasks', 'such
', 'as', 'text', 'pre-processing', 'to', 'tasks', 'likes', 'vectorized', 'representation', 'of', 'text', '-', 'NLTK', "'
s", 'API', 'has', 'covered', 'everything', '.']
```

```
Filtered Words (Without Stopwords): ['Natural', 'Language', 'Toolkit', '(', 'NLTK', ')', 'works', 'powerful', 'Python',
'library', 'wide', 'range', 'tools', 'Natural', 'Language', 'Processing', '(', 'NLP', ')', '.', 'fundamental', 'tasks',
'like', 'text', 'pre-processing', 'advanced', 'operations', 'semantic', 'reasoning', ',', 'NLTK', 'provides', 'versatile
', 'API', 'caters', 'diverse', 'needs', 'language-related', 'tasks', '.']
```

```
Lowercase & Punctuation Removed:

lets eat grandma
grandma lets eat
silvia are you free tomorrow
yes im free on saturday
```

```
Cleaned Paragraph:

Natural Language Processing NLP is a field of AI that focuses on enabling computers to understand interpret generate hum
an language NLP includes tasks like tokenization lemmatization sentiment analysis It helps in applications such as chatb
ots machine translation and voice assistants However cleaning textremoving extra spaces punctuations special charactersi
s crucial Without preprocessing NLP models may not perform accurately So can you clean this messy text make it structure
d
```

```
Stemmed Words:         ['the', 'research', 'are', 'analyz', 'variou', 'dataset', 'to', 'studi', 'the', 'effect', 'of', '
autom', '.', 'they', 'observ', 'that', 'autom', 'system', 'perform', 'task', 'more', 'effici', 'than', 'human', '.', 'ma
ni', 'industri', 'have', 'been', 'adopt', 'ai-driven', 'solut', 'to', 'improv', 'product', '.', 'run', 'complex', 'algor
ithm', 'help', 'in', 'predict', 'futur', 'trend', 'accur', '.', 'sever', 'compani', 'are', 'invest', 'in', 'develop', 's
marter', 'and', 'more', 'adapt', 'model', '.', 'data', 'scientist', 'continu', 'refin', 'their', 'model', 'to', 'achiev'
, 'better', 'perform', '.', 'the', 'advanc', 'in', 'technolog', 'have', 'transform', 'the', 'way', 'busi', 'oper', '.']

Lemmatized Words:      ['The', 'researcher', 'are', 'analyzing', 'various', 'datasets', 'to', 'study', 'the', 'effect',
'of', 'automation', '.', 'They', 'observed', 'that', 'automated', 'system', 'perform', 'task', 'more', 'efficiently', 't
han', 'human', '.', 'Many', 'industry', 'have', 'been', 'adopting', 'AI-driven', 'solution', 'to', 'improve', 'productiv
ity', '.', 'Running', 'complex', 'algorithm', 'help', 'in', 'predicting', 'future', 'trend', 'accurately', '.', 'Several
', 'company', 'are', 'investing', 'in', 'developing', 'smarter', 'and', 'more', 'adaptive', 'model', '.', 'Data', 'scien
tist', 'continuously', 'refine', 'their', 'model', 'to', 'achieve', 'better', 'performance', '.', 'The', 'advancement',
'in', 'technology', 'have', 'transformed', 'the', 'way', 'business', 'operate', '.']
```