**Natural Language Processing for PDF/TIFF/Image Documents  
Users Guide  
High Precision Natural Language Processing for PDF/TIFF/Image Documents  
Users Guide, v1.5**

# 1. Introduction

The target audience for this users guide are your software developers whom will be integrating the core inner block into your product and/or service. It is not meant to be a complete reference guide or comprehensive tutorial, but a brief get started guide.

To utilize this module you must have installed:

1. This Python module.
2. Python 3.6 or latter
3. Ghostscript ©(open source from Artifex).
4. Tesseract ©(open source from Google).
5. Magick ©(open source from Image Magic).
6. NLTK Toolkit (open source)
7. Unidecode (open source)

# 2. SPLITTER Module

## 2.1 Document Loading

To load a PDF document, TIFF facsimile or image captured document you create a Document (class) object, passing as parameters the path to the PDF/TIFF/image document and a path for storing the split pages/text. Below is a code example.

**from** document **import** Document, Page  
document = Document(“yourdocument.pdf”, “storage\_path”)

## 2.2 Page Splitting

## Upon instantiating a document object, the corresponding PDF document or TIFF facsimile is automatically split into the corresponding PDF or TIFF pages, utilizing Ghostscript (PDF) and Magick (TIFF). Each PDF/TIFF page will be stored separately in the storage path with the following naming convention:

<document basename><pageno>.<suffix> , where <suffix> is either *pdf* or *tif*

The module automatically detects if a PDF document is a digital (text) or scanned PDF (image). For digital documents, the text is extracted directly from the PDF page using Ghostscript and stored separately in the storage path with the following naming convention:

<document basename><pageno>.txt

2.3 OCR  
  
If the document is a scanned PDF, each page image will be extracted using Ghostscript, then OCR using Tesseract to extract the text content from the page image. The page image and corresponding page text are stored separately in the storage path with the following naming convention:

<document basename><pageno>.png  
<document basename><pageno>.txt

If the document is a TIFF facsimile, each page image will be extracted using Magick, then OCR using Tesseract to extract the text content from the page image. The page image and corresponding page text are stored separately in the storage path with the following naming convention:

<document basename><pageno>.tif  
<document basename><pageno>.txt

If the document is an image capture (e.g., JPG), the image is OCR using Tesseract to extract the text content from the page image. The page image and corresponding page text are stored separately in the storage path with the following naming convention:

<document basename><pageno>.<suffix> , where <suffix> is *png* or *jpg*  
<document basename><pageno>.txt

## 2.4 Image Resolution for OCR

The resolution of the image rendered by Ghostscript from a scanned PDF page will affect the OCR quality and processing time. By default the resolution is set to 300. The resolution can be set for a (or all) documents with the static member RESOLUTION of the Document class. This property only affects the rendering of scanned PDF; it does not affect TIFF facsimile or image capture.

# Set the Resolution of Image Extraction of all scanned PDF pages  
Document.RESOLUTION = 150

# Image Extraction and OCR will be done at 150 dpi for all subsequent documents  
document = Document(“scanneddocument.pdf”, “storage\_path”)

## 2.5 Page Access

Each page is represented by a Page (class) object. Access to the page object is obtained from the pages property member of the Document object. The number of pages in the document is returned by the len() builtin operator for the Document class.

document = Document(“yourdocument.pdf”, “storage\_path”)  
   
# Get the number of pages in the PDF document  
npages = len(document)

# Get the page table  
pages = document.pages

# Get the first page  
page1 = pages[0]

# or alternately  
page1 = document[0]

# full path location of the PDF/TIFF or image capture page in storage  
page1\_path = page1.path

## 2.6 Adding Pages

Additional pages can be added to the end of an existing Document object using the += (overridden) operator, where the new page will be fully processed.   
  
 document = Document(“1page.pdf”)

# This will print 1 for 1 page  
print(len(document))

# Create a Page object for an existing PDF page  
new\_page = Page(“page\_to\_add.pdf”)

# Add the page to the end of the document.  
document += new\_page

# This will print 2 showing now that it is a 2 page document.  
print(len(document))

## 2.7 Text Extraction

The raw text for the page is obtained by the text property of the page class. The byte size of the raw text is obtained from the size() method of the page class.

# Get the page table  
pages = document.pages

# Get the first page  
page1 = pages[0]

# Get the total byte size of the raw text  
 bytes = page1.size()

# Get the raw text for the page  
 text = page1.text

The property scanned is set to True if the text was extracted using OCR; otherwise it is false (i.e., origin was digital text).

# Determine if text extraction was obtained by OCR  
scanned = document.scanned

## 2.8 Asynchronous Processing

To enhance concurrent execution between a main thread and worker activities, the Document class supports asynchronous processing of the document (i.e., Page Splitting, OCR and Text Extraction). Asynchronous processing will occur if the optional parameter ehandler is set when instantiating the Document object. Upon completion of the processing, the ehandler is called, where the document object is passed as a parameter.

**def** done(d):  
 “”” Event Handler for when processing of document is completed “””  
 print(“DONE”, d.document)

# Process the document asynchronously  
document = Document(“yourdocument.pdf”, “storage\_path”, ehandler=done)

## 2.9 NLP Preprocessing of the Text

NLP preprocessing of the text requires the SYNTAX module. The processing of the raw text into NLP sequenced tokens (syntax) is deferred and is executed in a JIT (Just in Time) principle. If installed, the NLP sequenced tokens are access through the words property of the Page class. The first time the property is accessed for a page, the raw text is preprocessed, and then retained in memory for subsequent access.

# Get the page table  
pages = document.pages

# Get the first page  
page1 = pages[0]

# Get the NLP preprocessed text  
words = page1.words

The NLP preprocessed text is stored separately in the storage path with the following naming convention:

<document basename><pageno>.json

## 

## 2.10 NLP Preprocessing Options (Config)

NLP Preprocessing of the text maybe configured for several options when instantiating a document object with the optional config parameter, which consists of a list of one or more predefined options.

document = Document(“yourdocument.pdf”, “storage\_path”, config=[options])

# options:

bare # do bare tokenization  
stem = internal | # use builtin stemmer  
 porter | # use NLTK Porter stemmer  
 snowball | # use NLTK Snowball stemmer  
 lancaster | # use NLTK Lancaster stemmer  
 lemma | # use NLTK WordNet lemmatizer  
 nostem # no stemming

pos # Tag each word with NLTK parts of speech  
roman # Romanize latin-1 character encodings into ASCII

## 2.11 Document Reloading

## Once a Document object has been stored, it can later be retrieved from storage, reconstructing the Page and corresponding Words objects. A document object is first instantiated, and then the load() method is called specifying the document name and corresponding storage path. The document name and storage path are used to identify and locate the corresponding stored pages.

# Instantiate a Document object  
document = Document()

# Reload the document’s pages from storage  
document.load( “mydoc.pdf”, “mystorage” )

# This will reload pages whose filenames in the storage match the sequence:  
# mystorage/mydoc1.json  
# mystorage/mydoc2.json  
# ….

## 2.12 Document and Page Classification

Semantic Classification (e.g., category) of the document and individual pages requires the CLASSIFICATION module. The classification is deferred and is executed in a JIT (Just in Time) principle. If installed, the classification is access through the classification property of the document and page classes, respectively. The first time the property is accessed for a document or page, the NLP sequenced tokens for each page are processed for classification of the content of individual pages and the first page is further processed for the classification of the content of the entire document.

# Get the classification for the document  
document\_classification = document.class

# Get the classification for each page  
**for** page **in** document.pages:  
 classification = page.class

# 3. SYNTAX Module

## 3.1 NLP Processing

The Words (class) object does the NLP preprocessing of the extracted (raw) text. If the extracted text is from a Page object (see SPLITTER), the NLP preprocessing occurs the first time the words property of the Page object is accessed.

**from** document **import** Words, Vocabulary

# Get the first page in the document  
page = document.pages[0]

# Get the raw text from the page as a string  
text = page.text

# Get the NLP processed words (Words class) object from the page as a list.  
words = page.words

# Print the object type of words => <class ‘Document.Words’>  
type(words)

## 3.2 Words Properties

The Words (class) object has four public properties: text , words, bagOfWords and freqDist. The text property is used to access the raw text and the words property is used to access the NLP processed tokens from the raw text.

# Get the NLP processed words (Words class) object from the page as a list.  
words = page.words

# Get the original (raw) text as a string  
 text = words.text

# Get the NLP processed words from the original text as a Python list.  
words = words.words

# Print the object type of words => <class ‘list’>  
type(words)

The bagOfWords and freqDist properties are explained later in the guide.

## 3.3 Vocabulary Dictionary

The words property returns a sequenced Python list of words as a dictionary from the Vocabulary class. Each word in the list is of the dictionary format:

{ ‘word’ : word, # The stemmed version of the word  
 ‘lemma’: word, # The lemma version of the word  
 ‘tag’ : tag # The word classification  
 }

## 3.4 Traversing the NLP Processed Words

The NLP processed words returned from the words property are sequenced in the same order as the original text. All punctuation is removed, and except for detected Acronyms, all remaining words are lowercased. The sequenced list of words may be a subset of the original words, depending on the stopwords properties and may be stemmed, lemma, or replaced.

# Get the NLP processed words from the original text as a Python list.  
words = words.words

# Traverse the sequenced list of NLP processed words **for** word **in** words:  
 text = word.word # original or replaced version of the word  
 tag = word.tag # syntactical classification of the word  
 lemma = word.lemma # The lemma version of the word

## 3.5 Stopwords

The properties which determine which words are removed, stemmed, lemmatized, or replaced are set as keyword parameters in the constructor for the Words class. If no keyword parameters are specified, then all stopwords are removed after being stemmed/lemmatized. The list of stopwords is a superset of the Porter list and additionally includes removing additionally syntactical constructs such as numbers, dates, etc. For a complete list, see the reference manual.

If the keyword parameter stopwords is set to False, then all word removal is disabled, while stemming/lemmatization/reducing are still enabled, along with the removal of punctuation. Note in the example below, while stopwords is disabled, the word jumping is replaced with its stem jump.

# No stopword removal  
words = Words(“The lazy brown fox jumped over the fence.”, stopwords=**False**)  
# words => “the”, “lazy”, “brown”, “fox”, “jump”, “over”, “the”, “fence”

# All stopword removal  
words = Words(“The lazy brown fox jumped over the fence.”, stopwords=**True**)  
# words => “lazy”, “brown”, “fox”, “jump”, “fence”

## 3.6 Bare

When the keyword parameter bare is True, all stopword removal, stemming/lemmatization/reducing ad punctuation removal are disabled.

# Bare Mode  
words = Words(“The lazy brown fox jumped over the fence.”, bare=**False**)  
# words => “the”, “lazy”, “brown”, “fox”, “jumped”, “over”, “the”, “fence”, “.”

## 3.7 Numbers

When the keyword parameter number is True, text and numeric version of numbers are preserved; otherwise they are removed. Numbers which are text based (e.g., one) are converted to their numeric representation (e.g., one => 1). The tag value for numbers is set to Vocabulary.NUMBER.

# keep/replace numbers  
words = Words(“one twenty-one 33.7 1/4”, number=**True**)

print(words.words)  
# will output:  
# [  
# { ‘word’: ‘1’, tag: Vocabulary.NUMBER },  
# { ‘word’: ‘21’, tag: Vocabulary.NUMBER },  
# { ‘word’: ’33.7’, tag: tag: Vocabulary.NUMBER },  
# { ‘word’: ‘0.25’, tag: tag: Vocabulary.NUMBER },  
# ]

If a number is followed by a text representation of a multiplier unit (i.e., million), the number and multiplier unit are replaced by the multiplied value.

words = Words(“two million”, number=True)

print(words.words)  
# will output:  
# [  
# { ‘word’: ‘2000000’, tag: Vocabulary.NUMBER},   
# ]

## 3.8 Unit of Measurement

When the keyword parameter unit is True, US Standard and Metric units of measurement are preserved; otherwise they are removed. Both US and EU spelling of metric units are recognized (e.g., meter/metre, liter/litre). The tag value for units of measurement is set to Vocabulary.UNIT.

# keep/replace unit  
words = Words(“10 liters”, number=**True**, unit=**True**)

print(words.words)   
# will output:  
# [  
# { ‘word’: ‘10’, tag: Vocabulary.NUMBER },   
# { ‘word’: ‘liter’, tag: Vocabulary.UNIT },  
# ]

## 3.9 Standard vs. Metric

When the keyword parameter standard is True, Metric units of measurement are converted to US Standard. When the keyword parameter metric is True, Standard units of measurement are converted to Metric Standard.

# keep/replace unit  
words = Words(“10 liters”, number=**True,** unit=**True** standard=**True**)

print(words.words)   
# will output:  
# [  
# { ‘word’: ‘2.64172’, tag: Vocabulary.NUMBER },   
# { ‘word’: ‘gallon’, tag: Vocabulary.UNIT },  
# ]

## 3.10 Date

When the keyword parameter date is True, USA and ISO standard date representation and text representation of dates are preserved; otherwise they are removed. Dates are converted to the ISO standard and the tag value is set to Vocabulary.DATE.

# keep/replace dates  
words = Words(“Jan 2, 2017 and 01/02/2017”, date=**True**)

print(words.words)   
# will output:  
# [  
# { ‘word’: ‘2017-01-02’, tag: Vocabulary.DATE },   
# { ‘word’: ‘2017-01-02’, tag: Vocabulary.DATE },  
# ]

## 

## 3.11 Date of Birth

When the keyword parameter dob is True, date of births are preserved; otherwise they are removed. Date of births are converted to the ISO standard and the tag value is set to Vocabulary.DOB.

# keep/replace dates  
words = Words(“Date of Birth: Jan. 2 2017 DOB: 01-02-2017”, dob=**True**)

print(words.words)   
# will output:  
# [  
# { ‘word’: ‘2017-01-02’, tag: Vocabulary.DOB },   
# { ‘word’: ‘2017-01-02’, tag: Vocabulary.DOB },  
# ]

If date is set to True without date of birth set to True, date of births will be removed while other dates will be preserved.

## 

## 3.12 Social Security Number

When the keyword parameter ssn is True, USA Social Security numbers are preserved; otherwise they are removed. Social Security numbers are detected from the prefix presence of text sequences indicating a Social Security number will follow, such as SSN, Soc. Sec., Social Security, etc. Social Security numbers are converted to their single 9 digit value and the tag value is set to Vocabulary.SSN.

# keep/replace dates  
words = Words(“SSN: 12-123-1234 Social Security 12 123 1234”, ssn=**True**)

print(words.words)   
# will output:  
# [  
# { ‘word’: ‘121231234’, tag: Vocabulary.SSN },   
# { ‘word’: ‘121231234’, tag: Vocabulary.SSN },  
# ]

## 3.13 Telephone Number

When the keyword parameter telephone is True, USA/CA telephone numbers are preserved; otherwise they are removed. Telephone numbers are detected from the prefix presence of text sequences indicating a telephone number will follow, such Phone:, Mobile Number, etc. Telephone numbers are converted to their single 10 digit value, inclusive of area code, and the tag value is set to one of:

Vocabulary.TELEPHONE,   
Vocabulary.TELEPHONE\_HOME  
Vocabulary.TELEPHONE\_WORK  
Vocabulary.TELEPHONE\_OFFICE  
Vocabulary.TELEPHONE\_FAX

# keep/replace dates  
words = Words(“Phone: (360) 123-1234, Office Number: 360-123-1234”, telephone=**True**)

print(words.words)   
# will output:  
# [  
# { ‘word’: ‘3601231234’, tag: Vocabulary.TELEPHONE },   
# { ‘word’: ‘3601231234’, tag: Vocabulary.TELEPHONE\_WORK},  
# ]

## **3.14 Address**

When the keyword parameter address is True, USA/CA street and postal addresses are preserved; otherwise they are removed. Each component in the address is tagged according to the above street/postal address component type, as follows:

* Postal Box (Vocabulary.POB)
* Street Number (Vocabuary.STREET\_NUM)
* Street Direction (Vocabuary.STREET\_DIR)
* Street Name (Vocabuary.STREET\_NAME)
* Street Type (Vocabuary.STREET\_TYPE)
* Secondary Address (Vocabuary.STREET\_ADDR2)
* City (Vocabulary.CITY)
* State (Vocabulary.STATE)
* Postal (Vocabulary.POSTAL)

# keep/replace street addresses  
words = Words(“12 S.E. Main Ave, Seattle, WA”, gender=**True**)

print(words.words)   
# will output:  
# [  
# { ‘word’: ‘12’, tag: Vocabulary.STREET\_NUM },   
# { ‘word’: ‘southeast’, tag: Vocabulary.STREET\_DIR },   
# { ‘word’: ‘main’, tag: Vocabulary.STREET\_NAME },   
# { ‘word’: ‘avenue’, tag: Vocabulary.STREET\_TYPE },   
# { ‘word’: ‘seattle’, tag: Vocabulary.CITY },   
# { ‘word’: ‘ISO316-2:US-WA’, tag: Vocabulary.STATE },   
# ]

3.15 Gender  
  
When the keyword parameter gender is True, words indicating gender are preserved; otherwise they are removed. Transgender is inclusive in the recognition. The tag value is set to one of Vocabulary.MALE, Vocabulary.FEMALE or Vocabulary.TRANSGENDER .

# keep/replace gender indicating words  
words = Words(“man uncle mother women tg”, gender=**True**)

print(words.words)   
# will output:  
# [  
# { ‘word’: ‘man’, tag: Vocabulary.MALE },   
# { ‘word’: ‘uncle’, tag: Vocabulary.MALE },   
# { ‘word’: ‘mother’, tag: Vocabulary.FEMALE },   
# { ‘word’: ‘women’, tag: Vocabulary.FEMALE },   
# { ‘word’: ‘transgender’, tag: Vocabulary.TRANSGENDER },  
# ]

## 3.16 Sentiment

When the keyword parameter sentiment is True, word and word phrases indicating sentiment are preserved; otherwise they are removed. Sentiment phrases are reduced to the single primary word indicating the sentiment and the tag value is set to either Vocabulary.POSITIVE or Vocabulary.NEGATIVE.

# keep/replace sentiment indicating phrases  
words = Words(“the food was not good”, sentiment=**True**)

print(words.words)   
# will output:   
# [  
# { ‘word’: ‘food’, tag: Vocabulary.UNTAG },  
# { ‘word’: ‘not’, tag: Vocabulary.NEGATIVE},  
# ]

## 3.17 Parts of Speech

When the keyword parameter pos is True, each tokenized word is further annotated with it’s corresponding NLTK parts of speech tag.

# add parts of speech tagging  
words = Words(“Jim Smith”, pos=**True**)

print(words.words)   
# will output:   
# [  
# { ‘word’: ‘food’, ‘tag’: Vocabulary.UNTAG, ‘pos’: NN },  
# { ‘word’: ‘not’, ‘tag’: Vocabulary.NEGATIVE, ‘pos’: NN },  
# ]

## 3.18 Romanization

When the keyword parameter roman is True, the latin-1 character encoding of each tokenized is converted to ASCII.

# Romanization of latin-1 character encodings  
words = Words(“Québec”, roman=**True**)

print(words.words)   
# will output:   
# [  
# { ‘word’: ‘quebec’, ‘tag’: Vocabulary.UNTAG,   
# ]

## 3.19 Bag of Words and Word Frequency Distribution

The property bagsOfWords returns an unordered dictionary of each occurrence of a unique word in the tokenized sequence, where the word is the dictionary key, and the number of occurrences is the corresponding value.

# Get the Bag of Words representation  
words = Words(“Jack and Jill went up the hill to fetch a pail of water. Jack fell down and broke his crown and Jill came tumbling after.”, stopwords=True)  
  
print(words.bagOfWords)  
# will output:  
# { ‘pail’: 1, ‘the’: 1, ‘a’: 1, ‘water’: 1, ‘fetch’: 1, ‘went’: 1, ‘and’: 2, ‘jack’: 2, ‘jill’: 2,  
# ‘down’: 1, ‘come’: 1, ‘fell’: 1, ‘up’: 1, ‘of’: 1, ‘tumble’: 1, ‘to’: 1, ‘hill’: 1, ‘after’: 1 }

The property freqDist returns a sorted list of tuples, in descending order, of word frequencies (i.e., the number of occurrences of the word in the tokenized sequence.

# Get the Word Frequency Distribution  
words = Words(“Jack and Jill went up the hill to fetch a pail of water. Jack fell down and broke his crown and Jill came tumbling after.”, stopwords=True)  
  
print(words.freqDist)  
# will output:  
# [ (‘jack’, 2), (‘jill’, 2), (‘and’, 2), (‘water’, 1), (‘the’, 1), …. ]

**Proprietary and Confidential Information  
Copyright ©2018, Epipog, All Rights Reserved**