**Natural Language Processing for PDF/TIFF/Image Documents  
SYNTAX MODULE  
High Precision Natural Language Processing  
Technical Specification, v1.4**

# 1 Words

## 1.1 Words Overview

The words NLP preprocessor contains the following primary classes, and their relationships:

* Words - This is the base class for the representation of a Natural Language Processing (NLP) preprocessed list of words. The constructor optionally takes as parameters the raw text to tokenize and flags for NLP preprocessing the text.

words = Words(“some text”, flags …)  
  
The constructor calls the private methods \_split(), \_stem() and \_stopwords().

* Word – A single NLP preprocessed word (token).
* Vocabulary – A performance optimized python dictionary for word classification and lemmatizing.

**Vocabulary**

**Words (Base Class)**

**Word**

**Word**

**Word**

**List**

**Fig. 1a High Level view of Words Class Object Relationships**

## 1.2 Words Properties

The words class contains the following properties:

* text – The raw text (non-preprocessed).
* bare – A tokenized list of words from the raw text.
* words – The list of NLP preprocessed words.
* punct – Flag to keep/classify (True) or remove (False) punctuation.
* stem – Value indicating which stemmer to use:
  + builtin
  + NLTK porter, snowball, Lancaster or WordLemmatizer
* Flag to annotate(True) or not annotate (False) NLP preprocessed tokens with parts of speech using NLTK pos\_tag().
* number – Flag to keep/classify (True) or remove (False) numerical numbers. Ex.
  + 1 / 4.5 / 1,000 / three
* unit – Flag to keep/classify (True) or remove (False) units of measurement. Ex
  + Inches / ft / cm
* quantifier – Flag to keep/classify (True) size specifying words or to remove (False). Ex.
  + all / any / more
* preposition – Flag to keep/classify (True) preposition words or to remove (False). Ex.
  + to / from / above
* conjunction – Flag to keep/classify (True) conjunction words or to remove (False). Ex.
  + and / or / but
* article – Flag to keep/classify (True) article words or to remove (False). Ex.
  + a / an / the
* demonstrative – Flag to keep/classify (True) article words or to remove (False). Ex.
  + this / that / these
* question – Flag to keep/classify (True)question words or to remove (False). Ex.
  + who / what / how
* pronoun – Flag to keep/classify (True) pronoun words or to remove (False). Ex.
  + he / she / them
* date – Flag to keep/classify/reduce (True) dates or to remove (False). Ex
  + Jan. 1, 2000 / 01/01/2000 / 2000-01-01
* dob – Flag to keep/classify/reduce (True) date of births or to remove (False). Ex.
  + DOB: Jan 1, 2000 / date of birth is 01-02-2012
* ssn – Flag to keep/classify (True) social security numbers or to remove (False). Ex.
  + 123-84-1234 / 123 84 1234
* telephone – Flag to keep/classify (True) telephone numbers or to remove (False). Ex.
  + (360) 123-1234 / +13601231234
* name – Flag to keep/classify (True) telephone numbers or to remove (False). Ex.
  + Albert Einstein / Donald J. Trump
* address - Flag to keep/classify/reduce street address or to remove (False). Ex.
  + 124 NE 34th Cir, Home Town, AZ, 99123
* gender – Flag to keep/classify (True) gender specifying words or remove (False). Ex.
  + male / man / gal / mom
* sentiment – Flag to keep/classify/reduce (True) sentiment word sequences or remove (False). Ex.
  + not bad / disgusting
* stopwords – Flag to keep (True) all stop words or remove (False). If True, it supersedes all other flags. If False, other flags may be used to include specific categories. The stop words are a superset of the Porter list.

The words class contains the following static variables:

* DECIMAL – The decimal point (US Standard: period , EU: comma)
* THOUSANDS – The thousandths unit separator (US Standard comma , EU period)

## 1.3 Words Overridden Operators

The following operators have their implementations (inherited from the base Object) class overridden:

* len() – The \_\_len\_\_() method is overridden to return the number of NLP tokenized words.
* += The \_\_iadd\_\_() method is overridden to add words to the sequenced word list (append).

## 1.4 Words Private Methods

The words class contains of following private methods:

* \_split() – This method performs the first phase of NLP preprocessing of the raw text into a sequenced list of words (bare processing mode).
  + Contractions are expanded (e.g., can’t => can not).
  + Newlines, carriage returns and tabs removed.
  + Duplicated whitespace is removed.
  + Text is split into words and punctuation.
  + Punctuation is removed (except in numerical and date representations when property number and/or date and/or dob is True).
* \_preprocess() - This method performs the second of NLP preprocessing of the ‘bare’ tokenized words.
  + Identify acronyms.
  + Identify proper names.
  + Words are lowercased.
* \_stem() – This method performs the third phase of NLP preprocessing of the tokenized words by removing word endings and reducing word to its root stem (e.g., rider -> ride).
  + Remove plural endings (e.g., flies -> fly).
  + Remove past tense endings (e.g., baked -> bake).
  + Remove present participle endings (e.g., eating -> eat).
  + Remove verb to noun and comparative endings (e.g., rider -> ride, taller-> tall).
  + Remove noun to verb endings (e.g., flatten -> flat).
  + Remove adjective to adverb endings (e.g, costly -> cost).
  + Remove superlative endings (e.g., greatest -> great).
* \_nltkStem() – This method uses the open source NLTK stemmer methods to perform the third phase of NLP preprocessing of the tokenized words, as an alternative to the internal stemmer (i.e. stem)). The Porter, Snowball, Lancaster and WordNetLemmatizer are selectable.
* \_stopwords() – This method performs the fourth phase of NLP preprocessing of the tokenized words by removing/keeping stop words.
  + Remove word (including infinity) and numeric representations of numbers, unless property number is True, then all numbers are retained.
    - If retained, EU decimal and thousandths unit separators converted to US standard.
    - +/- signs preserved.
    - Thousandths unit separator removed.
    - Hex numbers (starting with 0x prefix) are converted to integer value.
    - Text represented numbers (e.g., ten) are converted to integer value.
    - Text represented numeric ordering (e.g., 1st) are converted to integer value.
    - Fractions are converted to floating point value.
  + Remove units of measurement, unless property unit is True.
    - US Standard and Metric, including abbreviations, are recognized.
    - US and EU spelling of metric units are recognized.
  + Remove dates, unless property date is True.
  + Remove date of birth, unless property dob is True.
  + Remove USA social security numbers, unless property ssn is True, where the SSN number is converted to a 9 digit value.
  + Remove USA/CA telephone numbers, unless property telephone is True, where the telephone number is converted to a 10 digit number.
  + Remove USA/CA addresses, unless property address is True, where addresses are converted to the USPO addressing standard.
  + Remove gender indicating words (e.g., man) – inclusive of transgender, unless property telephone is True.
  + Remove proper names and titles (e.g., Dr.), unless property name is True.
  + Remove quantifier indicating words (e.g., all, any), unless property quantifier is True.
  + Remove prepositions (e.g., above, under), unless property preposition is True.
  + Remove conjunctions (e.g., and, or), unless property conjunction is True.
  + Remove articles (e.g., a, an), unless property article is True.
  + Remove demonstratives (e.g., this, that), unless property demonstrative is True.
  + Remove pronouns (e.g., his, her), unless property pronoun is True.
  + Remove question words (e.g., what, why), unless property question is True.
  + Remove common high frequency words i.e., Porter List).
  + Remove sentiment sequence (e.g., good, bad), unless sentiment property is True.
    - Sequence (e.g., not bad) replaced with “positive” or “negative”.
  + Remove punctuation and symbols, unless punct property is True.
* \_isdate() – This method is a support method for \_stopwords(). It will recognize date strings and convert them to ISO 8601 format. The following formats are recognized:
  + MM/DD/YY and MM/DD/YYYY
  + MM-DD-YY and MM-DD-YYYY
  + YYYY-MM-DD (ISO 8601)
  + Month Day, Year (e.g., January 6, 2016)
  + Abbr-Month Day, Year (e.g., Jan 6, 2016 and Jan. 6, 2016)

If the preceding word is birth or DOB, then the date will be tagged as a date of birth (vs. date).

* \_isnumber() - This method is a support method for \_stopwords(). It will recognize numerical sequences and convert them to decimal base 10 format. The following formats are recognized
  + Base 10 integer, floating point, exponent, fraction
  + Base 16 hex integers
* \_isSSN() - This method is a support method for \_stopwords(). It will recognize USA Social Security numbers. The following formats are recognized:
  + Prefixed with SSN or Soc. Sec. No. or Social Security Number
  + Number Format:
    - 12-123-1234 / 12 123 1234 / 121231234
* \_isTele() - This method is a support method for \_stopwords(). It will recognize USA/CA Telephone numbers. The following formats are recognized:
  + Prefixed with Tele, Phone, Mobile, Office, etc, optionally followed by Number, Num, No, #
  + Number Format:
    - 1231231234 / 123 123 1234 / 123-123-123 / (360) 123-1234 / …
* \_isAddr() – This method is a support method for \_stopwords(). It will recognize USA/CA postal addresses. The following formats are recognized:
  + [POB[,]] Street-Number [Street-Direction] Street-Name [Street-Type] [Street-Direction] [,] [POB[,]] [Secondary-Address[,]][City[,]State]
  + POB[,] City State
* \_streetnum() – This method supports the \_isAddr() method in recognizing street numbers. The following formats are recognized:
  + [(N|S|W|E)]digits[letter][-][digits|letter
  + Ex. N1300 / 123-33 / 33A / 33
* \_streetdir() – This method supports the \_isAddr() method in recognizing directional phrases. The following formats are recognized:
  + North|South[sp][West|East]
  + N|S[.][w|e][.]
* \_citystate() – This method supports the \_isAddr() method in recognizing city/state references in a postal address. The following formats are recognized:
  + City[,](Full-State|Abbr-State)

USA and Canadian state names are replaced with their ISO 3166-2 codes (e.g., Alabama => ISO3166-2:US-AL).

* \_pob() – This method supports the \_isAddr() method in recognizing Post Office Boxes in street addresses. The following formats are recognized:
  + ( P.O.B | POB | P.O. Box | P.O. | PO )digits
* \_streetaddr2() – This method supports the \_isAddr() method in recognizing secondary address components in street addresses. The following formats are recognized:
  + (Apt|Ste|Rm|Fl|Bldg|Dept)[#](word[-][word)
* \_isGender() – This method supports gender recognition in stopwords. It recognizes phrases:
  + (Sex|Gender)[:] (M|F|Male|Female)
* \_conversion() – This method performs the fifth phase of NLP preprocessing of the tokenized words of converting Standard to Metric (standard=True) and vice-versa (metric=True).
* \_partsofspeech() – This method performs the sixth phase of NLP preprocessing of the tokenized words of tagging words with their parts of speech tag (using NLTK).

## 1.5 Words Public Methods

The words class contains of following public methods:

* There are no public methods.

## APPENDIX I: Updates

*v1.1*

* Refactored Stopword Removal for Parts of Speech, Numbers and Dates
* Added Vocabulary class and Lemmatizing
* Added support for Date of Birth
* Added support for recognizing word version of numbers.
* Fixed handling of Hex numbers

*v1.2*

* Added support for Social Security Numbers
* Added support for Telephone Numbers
* Added support for Proper Names
* Fix not recognizing Acronym if first word
* Refactored tokenization and added bare mode.

*v1.3*

* Added support for Spanish punctuation: ¿¡
* Added support for numeric multipliers (e.g., 10 million).
* Added support for unit of measurements.
* Added conversion of unit of measurements between Standard and Metric.
* Added support for Sex/Gender[:] M/F
* Added support for USA street addresses
* Fix not recognizing single letter abbreviations.
* Fixed not recognizing title (name) proceeded by a comma.
* Fixed not recognizing number followed by unit of measurement when combined (e.g., 2cm).

*v1.4*

* Add NLTK options for stemming, lemmatization and parts of speech.
* Change outputting of US State names to ISO 3166-2 standard

# APPENDIX II: Anticipated Engineering

The following has been identified as enhancement/issues to be addressed in subsequent update:

1. Support for detecting Abbreviations
2. Support for splitting into paragraphs.
3. Support for email addresses
4. Add support for mail stop Secondary Street Address components.
5. Add support for Canadian Street/Postal addresses.
6. Add support for Currency.
7. Add <SOS> and <EOS> annotation.
8. Fix lose next word after street/postal address

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