NLP with Python for Machine Learning Essential

<https://www.linkedin.com/learning/nlp-with-python-for-machine-learning-essential-training>

Table of contents

[0 - Setup 2](#_Toc17818769)

[1 - NLP Basics 3](#_Toc17818770)

[2 - Supplemental Data Cleaning 4](#_Toc17818771)

[3 - Vectorizing Raw Data 5](#_Toc17818772)

[4 - Feature Engineering 7](#_Toc17818773)

[5 - Building Machine Learning Classifiers 10](#_Toc17818774)

# 0 - Setup

* Anaconde Installation for Python 3.6
  + Python 3.6
  + Jupyter Notebook
* NLTK (Natural Language Toolkit)
  + pip install -U nltk
* Import NLTK
  + Open Jupyter Notebook and execute:

|  |
| --- |
| import nltk  nltk.download() |

* + Click "All Packages"

**NLTK Function Overview**

Show all functions of the package nltk:

|  |
| --- |
| dir(nltk) |

Some functions to look at:

* pos\_tag (paras speach tagging)
* tokenize

# 1 - NLP Basics

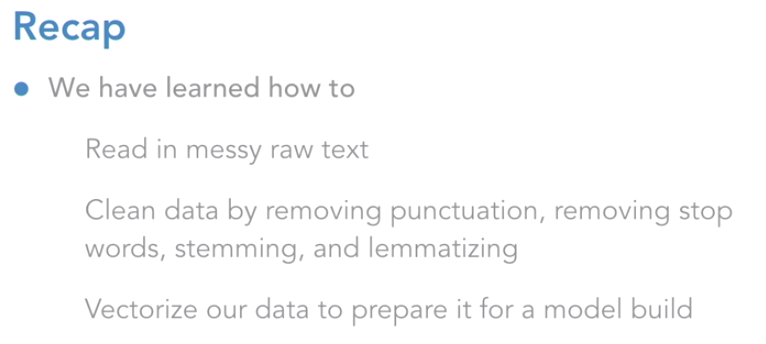
|  |  |
| --- | --- |
|  |  |
|  |  |
|  | |

# 2 - Supplemental Data Cleaning

|  |  |
| --- | --- |
|  |  |
| Stemmer are correct in most cases,  but not always: |  |
|  |  |
|  | Stemming is faster than Lemmatizing.  If Lemmatizing is too slow, try with stemming. |
|  | |

# 3 - Vectorizing Raw Data

|  |  |
| --- | --- |
|  | Document Term Matrix   * Each text is represented as row * Each word of all texts is a column * The number is the count of the word in the corresponding text * The last colum (label) represents the expected resut |
|  | |
|  | The model learns which word correlates with which label.  In reality this are much more columns than in the simplified example below: |
|  | |
| The sample above is called "Count vectorization".  There are different types:   * Count vectorization * N-grams * Term frequency – inverse document frequency (TF-IDF) * Further types, but the one above are the most popular | |
|  | |
|  | |
|  | |



# 4 - Feature Engineering

|  |
| --- |
|  |
|  |
|  |

|  |  |
| --- | --- |
| **Evaluate created features** | |
| Length of text (body\_len) in relation to ham/spam    🡪 This feature provides additional value to the model | |
| Punctuaition of text in relation to ham/spam    🡪 This feature provides NO additional value to the model | |
| **Plot the new features** | |
|  |  |
|  |  |
|  | |
|  | |
|  |  |

# 5 - Building Machine Learning Classifiers

|  |  |
| --- | --- |
|  |  |