FoxMIS CONNECT | INNOVATE

Fox School of Business
TEMPLE UNIVERSITY

Python and Web Data Extraction: *Introduction*

Alvin Zuyin Zheng
zheng@temple.edu
http://community.mis.temple.edu/zuyinzheng/

Outline

- Overview
- Text Representation
- The Natural Language Toolkit (NLTK)
- Tutorial 3: Computing TF and TF-IDF

Natural Language Processing (NLP)

- Natural language:
  - Language that is used for everyday communication by humans

- Natural Language Processing (NLP):
  - Any kind of computer manipulation of natural language.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*.

Tools

- Text representation
  - Tokenization
  - Stop words removal
  - Stemming
  - Simple summarization
    - Frequency
    - TF-IDF

Outline

- Overview
- Text Representation
  - Overview
  - The Natural Language Toolkit (NLTK)
- Tutorial 3: Computing TF and TF-IDF

Text Representation: A Sample Text

The raw text format is not convenient for any statistical analysis

```
Google is a global technology leader focused
on improving the ways people connect with
information. We aspire to build products and
provide services that improve the lives of
billions of people globally.
```

Tokenization

- Tokenization: splitting text into words and sentences

- The “bag of words” representation
  - Each document is a “bag”
  - The “bag” contains word tokens
  - Word order is ignored

Stopwords Removal

- Stopwords:
  - Typically function words: a, an, and, as, for, in, of, the, to
  - Are usually discarded from a text representation
  - `Google global technology leader focused improving ways people connect information`

Stemming

- A common root may have multiple variants
  - Accounting, accountant, accountants
  - Manage, management, managing, manager
- **Stemming** is the process of reducing words to their word "stem"
  - Accounting, accountant, accountants => account
  - Manage, management, managing, manager => manag
- May not always be used

Term frequency

- Term frequency (tf)
  - How often a word occurs in the document

- Vector Space Model
  - Each document in the corpus is represented by a vector in the word space

$$d_i = \{tf_{i1}, \dots, tf_{ij}, tf_{iM}\}$$

  - tf_{ij} represents the term frequency of word j in doc i
  - M is the number of unique words in the corpus

tf-idf Model

- The tf-idf model further considers the distinctive power of words (i.e., IDF)

$$d_i = \{tf_{i1} * idf_1, \dots, tf_{ij} * idf_j, tf_{iM} * idf_M\}$$

  - tf_{ij} represents the term frequency of word j in doc i. The log scale $\log(1 + tf_{ij})$ is often used in practice
  - idf_j represents the inverse document frequency of word j. The log scale is $\log\left(\frac{N}{df_j}\right)$ is often used in practice

tf-idf versus tf

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

$$\mathrm{tf}(\text{example}, d_2) = 3 \qquad \mathrm{idf}(\text{example}, D) = \log \frac{2}{1} \approx 0.3010$$

$$\mathrm{tfidf}(\text{example}, d_2) = \mathrm{tf}(\text{example}, d_2) \times \mathrm{idf}(\text{example}, D) = 3 \times 0.3010 \approx 0.9030$$

Outline

- Overview
- Text Representation
  - Overview
  - The Natural Language Toolkit (NLTK)
- Text Mining Tools
- Tutorial 3: Computing TF and TF-IDF

Installing NLTK package

- The Natural Language Toolkit (NLTK) provides:
  - A set of tools for the common NLP processes

- Use **`pip`** in your **command line interface** to install

```
pip install nltk
```

NLTK Modules

Task	NLTK modules	Functionality
Accessing corpora	nltk.corpus	standardized interfaces to corpora and lexicons
String processing	nltk.tokenize, nltk.stem	tokenizers, sentence tokenizers, stemmers
Collocation discovery	nltk.collocations	t-test, chi-squared, point-wise mutual information
Part-of-speech tagging	nltk.tag	n-gram, backoff, Brill, HMM, TnT
Classification	nltk.classify, nltk.cluster	decision tree, maximum entropy, naive Bayes, EM, k-means
Chunking	nltk.chunk	regular expression, n-gram, named-entity
Parsing	nltk.parse	chart, feature-based, unification, probabilistic, dependency
Semantic interpretation	nltk.sem, nltk.inference	lambda calculus, first-order logic, model checking
Evaluation metrics	nltk.metrics	precision, recall, agreement coefficients
Probability and estimation	nltk.probability	frequency distributions, smoothed probability distributions
Applications	nltk.app, nltk.chat	graphical concordancer, parsers, WordNet browser, chatbots
Linguistic fieldwork	nltk.toolbox	manipulate data in SIL Toolbox format

Outline

- Overview
- Text Representation
  - Overview
  - The Natural Language Toolkit (NLTK)
- Text Mining Tools
- Tutorial 3: Computing TF and TF-IDF

Tutorial 3: Computing TF and TF-IDF

- Download the 5tfidf.py and put it in the same folder with previous files
- Run the script.
- You will find two new files: tf.csv and tfidf.csv

Other Resources

- Natural Language Processing with Python (for Python 2)