**Ch14: Text Data for Trading – Sentiment Analysis**

Ch14에서 ch16까지 자연어처리를 다룸. Text data의 경우 구조화되지 않아서 ML알고리즘이 관련 정보를 추출하기가 더 까다롭기 때문에 다른 형태의 data보다 전처리가 더 필요함. 요지는 정보손실을 최소화하면서 text를 numerical format으로 전환해서 ML의 input으로 사용할 수 있게끔 하는 것. 이 챕터에서는 토크나이징과 그것을 이용한 news classification과 sentiment analysis를 다룸.

<Parsing and tokenizing text data>

From text to tokens – the NLP pipeline

NLP pipeline with spaCy and textacy

parsing, tokenizing, and annotating a sentence

NLP with TextBlob

TextBlob is a Python library that provides a simple API for common NLP tasks and builds on the Natural Language Toolkit (NLTK) and the Pattern web mining libraries.

Counting tokens – the document-term matrix

In this section, we first introduce how the bag-of-words model converts text data into a numeric vector space representations.

<https://wikidocs.net/22650>

cosine similarity

countvectorizer

<NLP for Trading>

The naïve Bayes classifier

The conditional independence assumption

pandas factorize

<https://runebook.dev/ko/docs/pandas/reference/api/pandas.factorize>

csr (compressed sparse matrix)

<https://gaussian37.github.io/math-la-sparse_matrix/>