Assignment: Data Types and Data Representation

Sergei Rogov

All the files belonging to this assignment can be found in my public repository:

https://github.com/SergeiRogov/Data\_representation.git

Activity 1

1. Time in terms of AM or PM.

Binary, qualitative, nominal (If we only care if it’s first half of the day or second)

Discrete, qualitative, ordinal (If we measure time down to hh:mm:ss)

(b) Brightness as measured by a light meter.

Continuous, quantitative, ratio (0 is appoint of reference, presumably brightness can’t be negative)

(c) Brightness as measured by people’s judgments.

Discrete, qualitative, ordinal (if people only have discrete ranking)

(d) Angles as measured in degrees between 0 and 360.

Discrete, quantitative, ratio (If measured only down to angles)

Continuous, quantitative, ratio (If measured continuously)

(e) Bronze, Silver, and Gold medals as awarded at the Olympics.

Discrete, qualitative, ordinal (can rank medals from worst to best)  
  
(f) Height above sea level.

Continuous, quantitative, ratio (0 is appoint of reference. Even though height above sea level might be negative, we can state “the altitude of this place is twice as big as the altitude of the other place”)  
  
(g) Number of patients in a hospital.

Discrete, quantitative, ratio

(h) ISBN numbers for books.

Discrete, qualitative, nominal (Even though it involves some ordering like book edition, as a whole ISBN number is nominal)

(i) Ability to pass light in terms of the following values: opaque, translucent, transparent.

Discrete, qualitative, ordinal

(j) Military rank.

Discrete, qualitative, ordinal

(k) Distance from the center of campus.

Continuous, quantitative, ratio (0 is appoint of reference, distance can’t be negative)

Activity 2

1. Disadvantages of this representation:
2. The tables (matrices) representing texts in this way will be sparse. We can expect a lot of zeros because the majority of words in one text will not be presented in the others.
3. This approach doesn’t take the order of the words into account. Some pieces of information will be lost.

Solution: treat some words not alone, but in pairs or in groups of three to extract more meaning.

1. It does not recognize semantics of the word.

Solution: use special libraries which are “aware” of semantics and can handle it.

1. It does not know if some words are more important than the others.

Solution: analyze the number of word inclusion in a set of documents and calculate some multiplier which scales the “importance” of a certain words.

1. I implemented this method of text representation in two ways: with and without scikit-learn library involvement.

Here are results of this method performed on a small set of sentences:

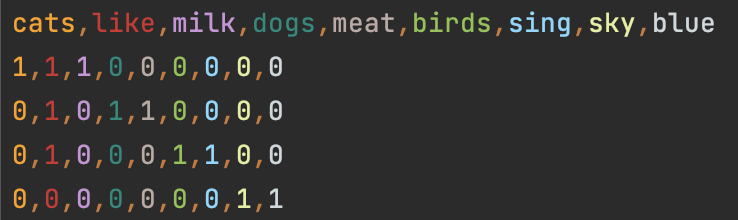
Doc1: Cats like milk.

Doc2: Dogs like meat.

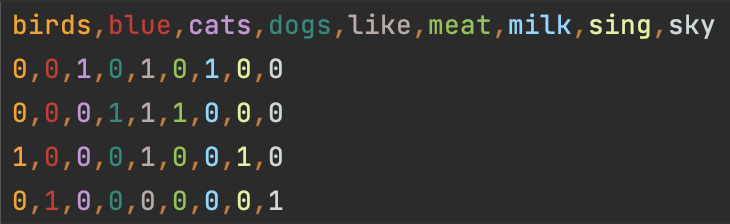
Doc3: Birds like to sing.

Doc4: The sky is blue.

No scikit (order – as words appear in files):

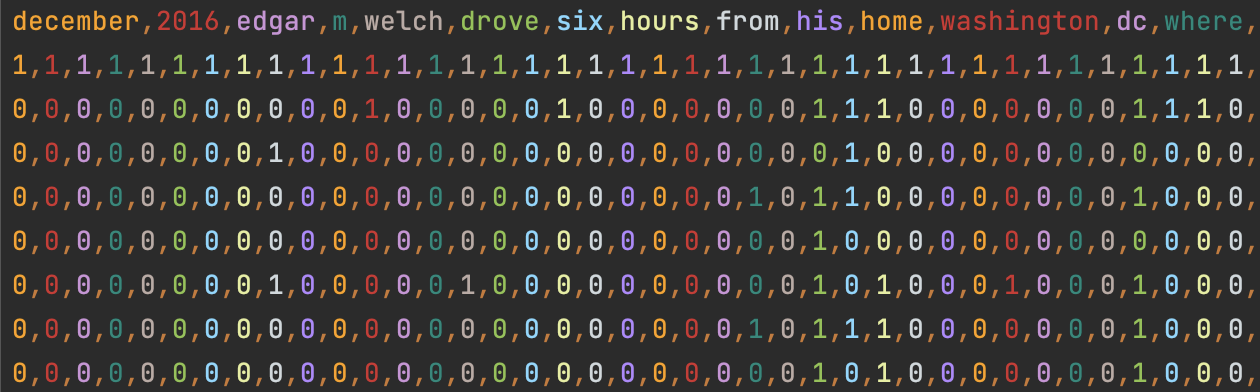


With scikit (presented in alphabetic order):



The result on a bigger set of text files (first columns of the result):

No scikit:



With scikit:

