G. BAG OF WORDS

WORD COUNT REPRESENTATION FOR MEASURING SIMILARITY

(Java IO and String process)

Time Limited: 3 seconds

The most popular model to represent a document, where we simply ignore the order of words that are presented in the document, is called the bag of words model. The reason of the name is that it is like we taking a bag and throwing all the words from that document into the bag, shaking it up, and the new document we've created with the words all jumbled up has exactly the same representation as the original document where the words were ordered[1].

In word count representation for measuring similarity approaches, we need to extract all words in a given file and to compute the appearance frequency of them. It is a basic foundation step in machine learning when we study about text processing. In extraction process, all of single words in a document must be obtained and delimiters must be removed after we based on them to separate the words. The content of file include text, and delimiter characters. The following line list out all delimiter characters:

“ ,.#?!'()|/”

We assume that the content of given text includes English characters only (no Vietnamese characters or other special characters from the others languages are included in given text file).

All character must be converted to lower- case format before make statistic process.

The output format of each word is: word(frequency_value)

And must be in order of alphabet (from A to Z)

[1] Mô hình phổ biến nhất để đại diện cho một tài liệu là việc sử dụng một túi chứa các từ, nơi chúng ta đơn giản bỏ qua thứ tự các từ có trong tài liệu. Và lý do nó được gọi là mô hình túi từ là do chúng ta nghĩ đến việc lấy một túi, ném tất cả các từ từ tài liệu đó vào túi, lắc nó lên. Và tài liệu mới mà chúng ta đã tạo ra đó với tất cả những từ lộn xộn có chính xác đại diện giống như mô tả như tài liệu gốc nơi các từ được đặt theo một trật tự.

Example 1:

Input	Output
Hanoi dang vao thu. Hello ACM ICPC contest, ACM ICPC contest is wonderful.	acm(2) contest(2) dang(1) hanoi(1) hello(1) icpc(2) is(1) thu(1) vao(1) wonderful(1)