Homework 3: Representing Simple Documents

Benjamin Roth, Marina Sedinkina
Symbolische Programmiersprache

Due: Thursday November 16, 2017, 16:00

In this exercise you will:

- Implement a simple document class.
- Get experience using the `unittest` framework.

You can monitor your progress by calling (from the `src` direcory:)

```
python3 -m unittest hw03_documents/test_documents.py
```

Exercise 1: TextDocument class [10 points]

1. Implement the helper method `word_tokenize` that takes a string and returns a list of lower-case tokens. Use nltk for tokenization.
2. Complete the constructor for `TextDocument`. You need to add `word_to_count`, a dictionary that maps every word to the number of its occurrences in this document.
3. Complete the class method `from_file`, that creates a document by reading a file, and calls the constructor with the text read from the file (and the filename as its id).
4. Implement the `__str__` method. It should return a string representation that is at most 25 characters long. If the original text is longer than 25 characters, the last 3 characters of the short string should be *"`...`"*. For example, the document text:
   "`Dr. Strangelove is the U.S. President's advisor.`"
   Should yield the `str` representation:
   "`Dr. Strangelove is the...`"
5. Complete the function `word_overlap` that determines the number of words that occur in both of the documents (`self` and `other_doc`) at the same time. Every word should be considered only once, irrespective of how often it occurs in either document (i.e. we consider word *types*). In other words this should return the size of the intersection of the word sets for both documents.

Using NLTK

If you work on the cip pool computers, nltk should already be installed.
If you use the CIP Pool computers, you may have to download the ressource `'punkt'`:

1. open the Python interactive shell:
   ```
   python3
   ```

2. then execute the following commands:
   ```
   >>> import nltk
   >>> nltk.download('punkt')
   ```

If you use your own computer:

- **Unix (with Python3):**
  ```
  sudo apt-get install python3-pip
  sudo pip3 install -U nltk
  ```
  Test the installation:
  ```
  python3
  >>>import nltk
  ```

- **Windows:** `http://www.nltk.org/install.html`

- If you encounter difficulties, ask fellow students or the tutors.