Assignment 5
Python, NLTK, Regular Expressions

Pick your favourite task(s)!

1 Task 1: Create Natural Language Processing Pipelines

Come up with some research questions of your field. Identify which information has to be extracted from text data to support the analysis. Which of the tools that we discussed in the lecture would be helpful? What are their limits?

2 Task 2: Write your own News Generator!

On the following web page you find an instruction how to generate a political news generator using NLTK:

http://www.gilestthomas.com/?p=279

Probably this will never be required from you, but it is a good way getting used to Python / NLTK and maybe - fun! I prepared a code snippet to read the articles. Make sure you understand what each line does when writing using the code. If the web page is not reachable, you can find it on our course web page as pdf.

2.1 Step 1

Install the NLTK. On the project homepage, you find instructions how to install the data that comes with it (nltk_data).

After installing and downloading the data¹, locate the data folder on your hard disc. It will probably be something like C:\nltk_data. For the exercise, I suggest you use a part of the Reuter’s corpus, which can be found in

nltk_data\corpora\reuters\test

2.2 Step 2

Here is the beginning of your script. Adjust the path to the data according to the data on your hard disc. Instead of using the backslashes between folder names (on Windows “\”, on Linux “/”) use os.sep

¹http://www.nltk.org/data
When it works, you can follow the instructions on the web page. Running the code might take a while, as it is 3000 files or more2!

```python
# -*- coding: utf-8 -*-
import os
import nltk

# read in all files of a directory:
# specify the path to the folder containing text files

# Windows (HardDriveName is probably C)
path = "HardDriveName:"+os.sep+os.sep+"nlkt_data"+os.sep+"corpora"+os.sep+"reuters"+os.sep+"test"+os.sep

# On Linux, it could be:
# path = os.sep+"home"+os.sep+"czirn"+os.sep+"bin"+os.sep+
   "nlkt_data"+os.sep+"corpora"+os.sep+
   "reuters"+os.sep+"test"+os.sep

# list all file names
listing = os.listdir(path)

# empty list in which we will collect articles
articleContents = []

# iterate over all fileNames, open them and save the content
for myFileName in listing:
    print "processing file: " + myFileName
    print "path to file: " + path + myFileName
    currentFile = open(path+myFileName,"r")
    articleContents.append(currentFile.read())
    currentFile.close()

# see what we collected
for article in articleContents:
    print article
```

2.3 Step 3

Follow the instructions on [http://www.gilesthomas.com/?p=279](http://www.gilesthomas.com/?p=279) and add the lines to your script.

Attention: you have to change one in the instructions on the web page if you use the node snippet3. Instead of

```python
content_text = join([article.content for article in articles])
```

2If you receive error messages mentioning a memory error, it might be too many files for your computer. Copy only a few of the files to a new folder and use the new folder.

3In our code, article is just a string. In this guy's code, article is an object with the property content
you slightly change the code to

```python
content_text = ' '.join(article for article in articleContents)
```

In the following line, you have to replace `words_to_generate` with a number.

```python
content = content_model.generate(words_to_generate, starting_words)
```

If you like, try other corpora for your generator. You find many of them in the nltk_data folder. You just have to change the path in the script.

3 Task 3: Analyze Syrian Censorship Logs

If all other assignments are peanuts to you, read the following article, download the Syrian logfile data and try to discover which key words were filtered! ;) For sure, it is an interesting data set.

http://www.heise.de/newsticker/meldung/Auch-Syrien ueberwacht-das-Internet-1355081.html