Assignment 1: Authorship Attribution

Classification with Naive Bayes

In this assignment, you will implement the Naive Bayes classification method. First you will train the classifier on the text of two famous authors: Mary Anne Evans (22 November 1819 – 22 December 1880), known by her pen name as George Eliot and Charlotte Brontë (21 April 1816 – 31 March 1855). Then you should evaluate the model on two other texts of George Eliot and Charlotte Brontë. The task is to try and find out how well your model manages to identify the works of each author.

George Eliot

Charlotte Brontë

You should write python codes to solve the tasks provided in the following section. Write answers to the discussion points preferable as comments in your code. Email the code and the answers to Mehdi (mehdi.ghanimifard@gu.se). Submit your answers individually. You are allowed to discuss with your fellow students, but not write code together.

Deadline: March 3

Readings

- Read about Naive Bayes and Python using scikit-learn.
- See also Mehdi's code, which he added in the website.

and also

- Ryan L. Boyd and James W. Pennebaker (2015). Did Shakespeare Write Double Falsehood? Identifying Individuals by Creating Psychological Signatures With Text Analysis. *Psychological Science* 26(5) 570–582
- Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. J. Am. Soc. Inf. Sci. Technol., 60(1), 9-26. doi:10.1002/asi.v60:1
- Kredens, Krzysztof, Coulthard Malcolm (2013). Corpus Linguistics In Authorship Identification. In Paul Foulkes and Peter French (2013). The Oxford Handbook of Language and Law. Oxford: Oxford University Press.

1. READ THE DATA: PREPROSSESING

Download the data here.

Here is an example of a line in the text corpus.

CBronte Jane cbronte_jane.txt take care of him said mr rochester to the latter and keep him at your house till he is quite well i shall ride over in a day or two to see how he gets on...

A line in the file is organized in columns:

0: author: Cbronte= Charlotte Brontë, Eliot (George Eliot)= Mary Anne Evans

1: work: Jane=Jane Eyre, Professor = The Professor (by Brontë) and Adam=Adam Bede, Middlemarch (by Eliot).
2: document identifier: *.txt
3 and on: the document tokens

Here is some Python code to read the entire corpus. This code is also available in the file assignment1.py in the package you unpacked.

```
def read_corpus(corpus_file):
    out = []
    with open(corpus_file) as f:
        for line in f:
            tokens = line.strip().split()
            out.append( (tokens[0], tokens[3:]) )
    return out
```

Note that the there was already some sort of prepossessing of the data, for example words were converted to lowercase, and some punctuation was removed.

2. TRAINING PHASE

Write a Python function that uses a training set of documents to estimate the probabilities in the Naive Bayes model:

$$p(C_k|x)=\frac{p(c_k)\,p(x|C_k)}{p(x)}$$

where $x = (x_1, \ldots, x_n)$

C_k are the classes, i.e., authors.

$p(C_k|x_1, \ldots, x_n)$

Return some data structure containing the probabilities.

3. TEST PHASE

Import the texts that will be used for estimating how well the algorithm performs. Note that the test texts have the same structure as the training texts. Then your trained set to estimate the test data.

4. EVALUATION

For the purposes of this assignment, you are requested to report the accuracy of the texts. That is how well the classifier performs on a new text.

5. OPTIONAL TASKS

1. Exclude the grammatical words and evaluate the classifier again (there are functions that can do that e.g.,)

```
from nltk.corpus import stopwords

stop = set(stopwords.words('english'))

sentence = "this is a test bar sentence"

print [i for i in sentence.lower().split() if i not in stop]

['test', 'bar', 'sentence']
```

2. Evaluate the classifier on a different author and work and explain the findings: for example on Charle Dickens David Copperfield.