First Assignment

24 novembre 2022

1 Prompts

The assignment consists in the development, in NLTK, OpenNLP, SketchEngine or GATE/Annie a **Naïve Bayes Classifier** able to detect a single class in one of the corpora available as attachments to the chosen package, by *distinguishing ENGLISH against NON-ENGLISH.* In particular the classifier has to be:

1. Trained on a split subset of the chosen corpus, by either using an existing partition between sample documents for training and for test or by using a random splitter among the available ones;
2. Devised as a pipeline of any chosen format, including the simplest version based on word2vec on a list of words obtained by one of the available lexical resources.

The test of the classifier shall give out the measures of **accuracy, precision, recall on the obtained confusion matrix** and WILL NOT BE EVALUATED ON THE LEVEL OF THE PERFORMANCES. In other terms, when the confusion is produced, then the value of the assignment will be good, independently of the percentage of false positive and negative results.

Deliver a short set of comments on the experience (do not deliver the entire code, but link it on a public repository like GitHub or the GATE Repo). Discuss: size of the corpus, size of the split training and test sets, performance indicators employed and their nature, employability of the classifier as a Probabilistic Language Model.

2 Delivering of the code

The code is in a public repo GitHub:
ChiaraSolito/NLPAssignments/Assignment1. In the repo there are two versions of the code:[1]

- A notebook version, with all the comments from this document directly embedded.
- A python script version, that prints on terminal a formatted version of results.

3 Comments on the experience

3.1 Introduction

The classifier was developed with the nltk package, using its own classifier (`nltk.NaiveBayesClassifier()`) with a simple pipeline, that resembles word2vec:

1. Acquisition of data
2. Cleaning and pre-processing
   (a) Removal of non-alphanumeric characters and words
   (b) Removal of stop-words of all the languages

[1] https://github.com/ChiaraSolito/NLPAssignments/tree/main/Assignment1

3. Tokenization
4. Creation of the Bag of Words
5. Splitting training and testing sets
6. Training the model
7. Testing and Querying the model

3.2 Size of the corpus

The corpus was made mixing 9 pre-existing nltk corpus, from the **Genesis** corpus and the **Universal declaration of human rights** corpus:

- 3 of the corpus are in english (two from the genesis and one from the Udhr)
- The other languages used are: Finnish, French (2 corpus), Portuguese, German and Spanish

The sizes are:

Corpus	**Size**	**Lexical Diversity**
English genesis	44764	0.06230453042623537
English web genesis	44054	0.06033504335588142
Finnish genesis	32520	0.2088560885608856
French genesis	46116	0.0803842484170353
Portuguese genesis	45094	0.08457887967357076
English-latin1 udhr	1781	0.29927007299270075
German udhr	1521	0.3806706114398422
French udhr	1935	0.2930232558139535
Spanish udhr	1763	0.3074305161656268

Total size of the corpus stands at 219548 words, of which 90599 are english.

3.3 Size of the split training and test sets

Common split percentages include:

- Train: 80%, Test: 20%
- Train: 67%, Test: 33%
- Train: 50%, Test: 50%

3.4 Performance indicators

3.5 Employability