Department of Computer Science, Electrical and Space Engineering
Luleå University of Technology

D7041E "Applied artificial intelligence"

IT IS STRICTLY FORBIDDEN TO USE AI GENERATED CODE AND COPY EXISTING CODE FROM THE INTERNET. ALL CASES OF VIOLATION WILL BE REPORTED!

LAB 3: Random high-dimensional representation of structured data

1. Introduction

In this lab we will work with a data representation technique, which is called distributed representation. With distributed representation each feature is represented by a vector from highdimensional random space $H=\{+1;-1\}^d$, where d is typically several hundred or even thousand. We will use this representation to represent feature vectors in data-driven classification problems. In this lab you will see how n-grams could be implemented using HD representation and how to build a simple centroids-based HD classifier for classifying languages of an input texts.

2. The language classification problem

In this lab you will classify the language of an input text. The language recognition will be done for 21 European languages. The list of languages is as follows: Bulgarian, Czech, Danish, German, Greek, English, Estonian, Finnish, French, Hungarian, Italian, Latvian, Lithuanian, Dutch, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish, Swedish.

Task 1.1 Import datasets into Jupyter environment:

1. The training data is based on the Wortschatz Corpora: https://corpora.uni-leipzig.de/en?corpusId=deu_newscrawl-public_2018

2. The test data is based on the Euro Parliament Parallel Corpus: http://www.statmt.org/europarl/
3. Import the data into your Jupyter environment, understand its structure
4. Do necessary pre-processing (removing punctuation, etc.)

2. Encoding of n-grams using d-dimensional {+1,-1} distributed representation

Use encoding procedure for n-grams in {+1;-1} coordinates as described in file: “ArticleEncodingNgramStatistics.pdf” linked in Canvas (see Labs module).

- Use n=3 (tri-grams). Use length of HD vectors d_1=100 and d_2=1000.

Task 2.1 Constructing high-dimensional centroids

An HD vector for a particular input text of certain language is computed by adding all the n-gram vectors. Since we consider 21 European languages at the end of the training phase we will have 21 d-dimensional language HD vectors (centroids) stored in an array.

Question: what will be the size of the n-gram input vector in conventional (local) representation?

Question: Identify difficulties of working with conventional representations of n-grams in the machine learning context.

Task 2.2. Classification using hyperdimensional centroids

In the test phase for an unknown text sample first a query vector in the same fashion as you constructed language vectors in the training phase. To determine the language of this text sample compare its query vector to all the (21) language vectors w.r.t **cosine similarity measure**. Present confusion matrix, compute accuracy and F1-score.

Congrats, you have just become familiar with fundamentals of the cutting edge data representation technique called distributed random representations! Well done!