Homework 3
Report

A0153992U, A0000000U, A0000000U
Team ChickenPox

1 Introduction

1.1 We are awesome

Hello world!

2 Feature generation

Alike image recognition, natural language processing is something that our brains are accustomed to due to millions of years of evolution and fine tuning. We are very good at recognizing the important information and context contained in a written passage, but however natural it may seem, it is certainly not simple to imitate. Given a passage, how can we automatically extract the important information, and how do we connect this to formulate a coherent response?

In order to run machine learning classifiers on the given dataset we decided to play with a number of distinct methods for generating natural language features. Most are based on computing simple statistics such as the number of occurrences on the words or collections of words (bigrams). Some more sophisticated ones tried to also take into account the word order. All are described below.

2.1 Bag of words

2.2 Cantor mapping

2.3 Bigrams, Trigrams

3 Regularization and validation

4 Results

5 Improvements

```
for i:=maxint to 0 do
begin
{ do nothing }
end;
Write('Case insensitive ');
Write('Pascal keywords.');
```

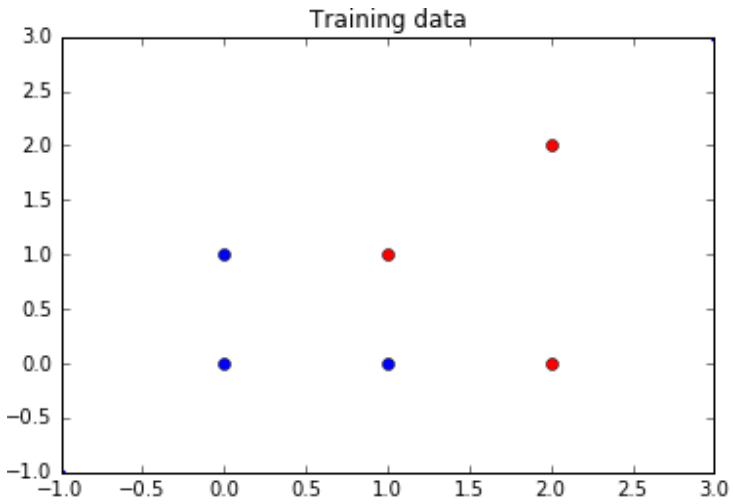

Figure 5.1: Hehehe