Homework 04

Entropy, Natural Language Processing & SQL

Table of Contents

[Introduction 1](#_Toc512341582)

[Part I 1](#_Toc512341583)

[1. Identify the first feature 1](#_Toc512341584)

[2. Entropy of S 3](#_Toc512341585)

[Part II 4](#_Toc512341586)

[1. What is the difference between a Bag Of words Model in NLP and a Word2vec Model, discuss advantages of one over the other? 4](#_Toc512341587)

[2. What is a word vector? What is a word Embedding? On what factors does the word embedding of a word depend (explain it from a NLP perspective)? 4](#_Toc512341588)

[3. What is a corpus in NLP? How is the vocabulary of a model different from the corpus? 4](#_Toc512341589)

[Part III 5](#_Toc512341590)

[0. Creating from Scratch 5](#_Toc512341591)

[1. Simple SELECTS (on the parents table) 5](#_Toc512341592)

[2. JOINS 7](#_Toc512341593)

[3. Aggregate functions, numerical logic and grouping 7](#_Toc512341594)

# 

# Introduction

**Due Date:** 24rd April 2018

**Student’s Name:** Diego Sapunar

**Student’s Cal ID:** 013109070

**Student’s Data-X GitHub:** [**Github**](http://www.github.com/dasapunar/dasapunar_data_x_s18)

# Part I

## Identify the first feature

The dataset shown below represents bank customers with 3 features and the label corresponding to each customer identifies whether they’ve defaulted or not.

Description:

**HasJob:** Binary value, equal to 0 when a person has no job and 1

otherwise.

**HasFamily:** Binary value, equal to 0 when a person has no family and 1 otherwise.

**IsAbove30years**: Binary value, equal to 0 when a person's age is 30 or below and 1 otherwise.

**Defaulter** is also a binary valued label which is equal to 1 if a person is a defaulter and 0 otherwise.

Use this dataset to **identify the best feature to do the first split** in a binary decision tree, so as to maximize the information gain in the next split. Show your calculations.

1. Formulas - Reference: PUC Chile Artificial Intelligence Course (IIC2613):
   1. (1)
   2. (2)
   3. (3)
   4. Convention:
      1. Defaulter = D
      2. HasJob = J
      3. HasFamily = F
      4. IsAbove30years = A
      5. Entropy(S) = E(S)
      6. Entropy(T,X) = E(T,X)
2. Calculate Entropy of target, Defaulter in this case. Using (1):
3. Calculate the entropy for each attribute (branch). Using (2). And then the Gain with (3):
   1. **HasJob (J):**
   2. **HasFamily (F):**
   3. **IsAbove30years (A):**

**(By definition)**

**(By definition)**

1. Choose attribute with the largest information gain as the decision node

We see that the largest information gain is HasFamily (F) with **,** so is the best feature to do the first split.

## Entropy of S

Given a signal of three symbols **S = (A, B, C)** and P(A)=0.7, P(B)=0.2, P(C)=0.1, What is the entropy of S? What does it mean according to the *Source coding Theorem*?

1. Formula
2. Calculate:

**1.1567795**

1. This means that by theory, the minimum number of bits necessary to represent each symbol is 1.1567795

# Part II

## What is the difference between a Bag Of words Model in NLP and a Word2vec Model, discuss advantages of one over the other?

Bag-of-words, for a given document extract the unigram words to just create a random list of words. It doesn’t take care of the positions, the syntax, semantics, anything, it just put them in a list or array.

In the other hand, according to [Towards Data Science](https://towardsdatascience.com/word-to-vectors-natural-language-processing-b253dd0b0817), we have Word to vector model, which is a more sophisticated way to aim NLP. Basically, given a bag of words that you got from the document, you create a real vector, as a feature vector, where each feature is a word and the feature’s value is a term weight.

Also, using Bag of Words you assign word frequency to document-term matrix element and in Vector Space Model document-term matrix elements are quite general as long as operations (dot product) in vector space make sense, getting the word position.

## What is a word vector? What is a word Embedding? On what factors does the word embedding of a word depend (explain it from a NLP perspective)?

A mathematical way of representing words is as vectors. Where these vectors are in a continuous space.

Word Embedding, is a method where all the words and phrases of your vocabulary are related to a vector of real numbers. So, by theory, it implies the space with one dimension by word to a continuous vectoral space with less dimensions.

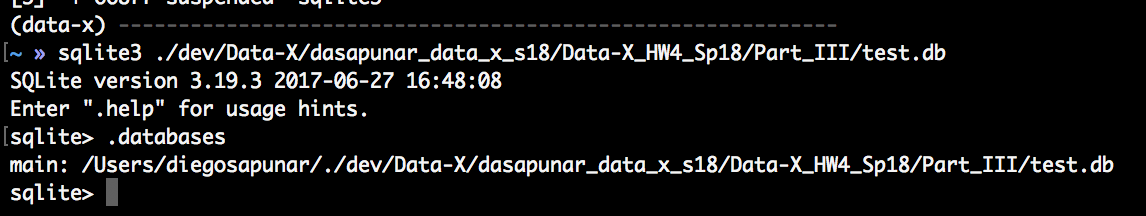
## What is a corpus in NLP? How is the vocabulary of a model different from the corpus?

Corresponding to [**StackOverFlow**](https://stackoverflow.com/questions/6700652/is-this-the-correct-definition-of-a-corpus) forum, a corpus, in linguistics, is any coherent body of real-life text or speech being studied. So yes, a book is a corpus. The fact that it's in one string doesn't matter, as long as you don't randomly shuffle the characters. So, **Corpus** basically means a body, and in the context of Natural Language Processing (NLP), it means a body of text. On the other hand, a vocabulary are the specific words of your string, such as a long array.

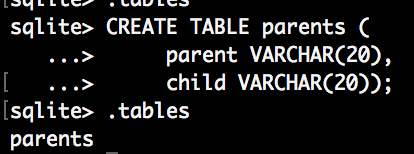
# Part III

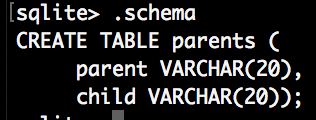
## Creating from Scratch

1. I created a DB called *test.db* in my GitHub corresponding folder.

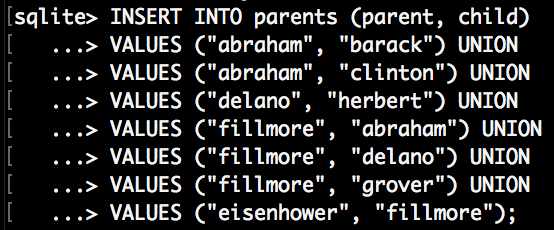


1. I created the table Parents.





1. I inserted data to the table Parents.

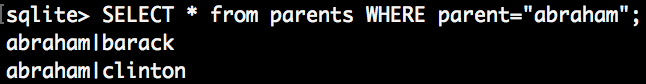


## Simple SELECTS (on the parents table)

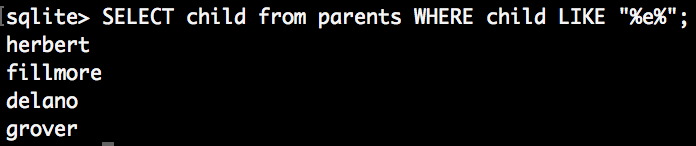
1. SELECT all records in the table

****

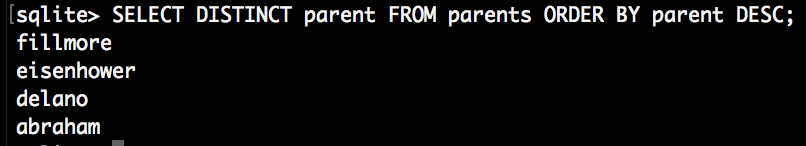
1. SELECT child and parent where Abraham is the parent



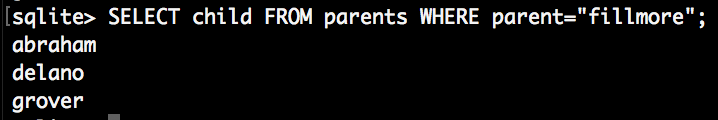
1. SELECT all children that have an ‘e’ in their name (hint: use LIKE ‘%e%’).

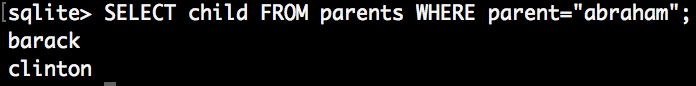


1. SELECT all unique parents (use SELECT DISTINCT) and order them by name, descending order (i.e fillmore first).

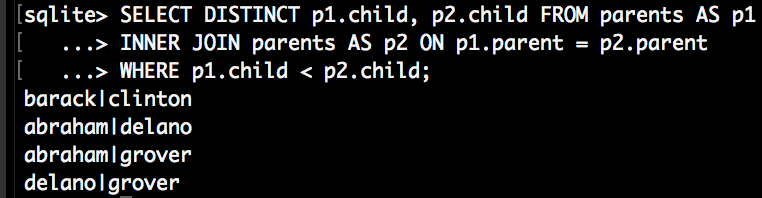


1. SELECT all dogs that are siblings (one-to-one relations). Only show a sibling pair once. To do this you need to select two times from the parents table.



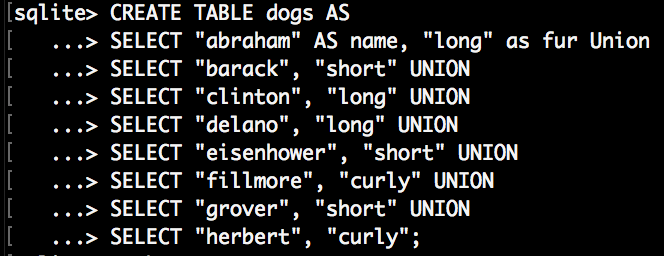


**OTHER METHOD WITH JOINS:**

****

## JOINS

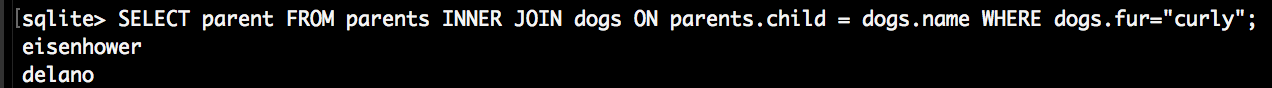
1. Created the new table dogs.



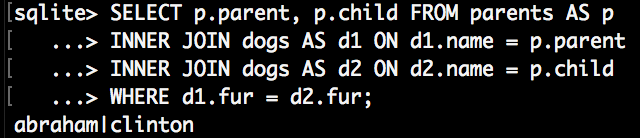
1. COUNT the number of short haired dogs.



1. JOIN tables parents and dogs and SELECT the parents of curly dogs

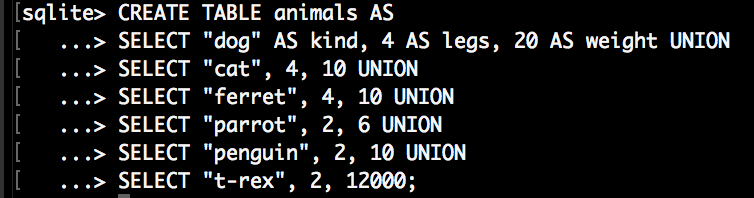


1. JOIN tables parents and dogs and SELECT the parents and children that have the same fur type. Only show them once.



## Aggregate functions, numerical logic and grouping

1. Created the new table animals.



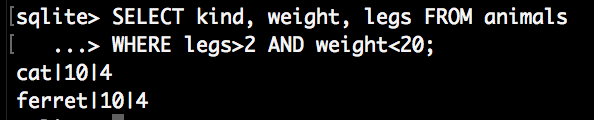
1. SELECT the animal with minimum weight. Display kind and min\_weight.



1. Use the aggregate function AVG to display a table with the average number of legs and the average weight.



1. SELECT the animal kind(s) that have more than two legs, but weight less than 20. Display kind, weight and legs.



1. SELECT the average weight for all the animals with 2 legs and the animals with 4 legs (by using GROUP BY).

