CS 495 - Introduction to Web Science

Fall 2014

Assignment 10

by

Eric Littley
UIN: 00821698

December 11, 2014

Instructor

Dr. Michael Nelson

Department of Computer Science
Old Dominion University

Honor Pledge

I pledge to support the Honor System of Old Dominion University. I will refrain from any form of academic dishonesty or deception, such as cheating or plagiarism. I am aware that as a member of the academic community it is my responsibility to turn in all suspected violations of the Honor Code. I will report to a hearing if summoned.

Signed: Eric Littley

Contents

1 Introduction **1**

2 Design **1**

3 Extra Credit **3**

1 Introduction

The purpose of this assignment was to introduce techniques used to filter documents. As such, an Introductory to classifiers was in order. This assignment used code from the book that included an implementation of the Fisher classifier[1]. This classifier was used to try and guess the category that entries from a blog might fall into.

2 Design

100 blogs were downloaded from http://uoacomputerscience.blogspot.com/feeds/posts/default a blog written by Ian Watson about Computer Science. After perusing through the blogs, it was determined that seven categories could be used to classify the blogs. Those seven categories are shown below.

Categories
AI
History
Education
Software
Internet
Blog
Hardware

A list of 100 blogs was downloaded and each entry was manually categorized. The first 50 blog entries were used to train the classifier. The next 50 blogs were attempted to be classified by the classifier. The classifier was modified so that "none" was returned if all the categories returned the same Fischer probability. Also, "none" was returned when the probability for the most likely candidate was less than 0.5. The table below shows the cprob and Fischer probability of the correct category not of the classifier in the guess column. Cprob was 0 for many of the classified entries because a 2-gram or 3-gram was used frequently and the small dataset made it unliked that those features had appeared before.

Table 1: 50 trained 50 classified

Title	Gram	Cprob	Fisherprob	Guess	Actual
Time travel with Google Street View	google street view	0.0	0.855	Software	Software
Alan #Turing infographic	turing	0.397	0.412	none	History
Keeping Secrets: Privacy and Security in the Information Age	security	1.0	0.875	Internet	Internet
Farewell to Microsoft XP	microsoft	0.636	0.602	Software	Software
Croudfunding for a computer history display - IBM 5080	computer history display	0.0	0.747	History	History
Facebook Introduces 'Hack,' the programming language of the future	facebook	1.0	0.75	Internet	Internet
Compare: How London Looks on Google vs. Paintings From the 1700s	looks on google	0.0	0.399	none	Software
Apps to Get Your Kids Coding on the iPad	kids coding	0.0	0.743	Education	Education
	robots are now	0.0	0.205	none	AI
This day in history... the first tweet	the first tweet	0.0	0.343	none	History
Tim Berners-Lee didn't expect kittens to take over the web	web	0.56	0.54	Internet	Internet
Home automation	automation	0.0	0.5	none	Hardware
The Big Data Brain Drain: Why Science is in Trouble	big data	0.0	0.174	Internet	Education
The world's largest photo service just made its pictures free to use	photo	0.0	0.5	none	Software
Inbox zero - progress report	inbox zero	0.0	0.743	Internet	Internet
20 Resources for Teaching Kids How to Program & Code	teaching Kids	0.0	0.886	Education	Education
The Return of the ZX Spectrum	the return of	0.0	0.358	none	History
Computer-generated fake papers are flooding academia	academia	0.0	0.5	none	Education
The IT History Society	it history	0.0	0.875	History	History
Common Lisp: The Untold Story	story	0.268	0.301	AI	History
Why Watson and Siri Are Not Real AI	ai	0.0	1.0	none	AI
This day in computing history	computing history	0.0	0.474	none	History
Turing's Halting Problem	halting problem	0.0	0.513	Hardware	Education
Facebook at 10: Zuckerberg hails 'incredible journey'	zuckerburg	0.0	0.5	none	Internet
Computer Science in Sculpture	computer science	0.0	0.217	none	Education
After Setbacks, Online Courses Are Rethought	courses	0.0	0.5	none	Education
Bob Marley's birthday is a national holiday in New Zealand	national holiday	0.0	0.14	none	History
This day in history...	this day in	0.0	0.358	none	History
Google Acquires AI Startup DeepMind For More Than $500M	ai startup	0.0	0.25	Internet	AI
Steve Jobs Unveils Mac at Boston Computer Society, Unseen Since 1984	unveils mac	0.0	0.597	none	History
Play games and help scientists	play games	0.0	0.427	none	Education
Blogging and web automation - #IFTTT	blogging	0.0	0.5	none	Internet
Happy birthday #Macintosh!	birthday	0.0	0.5	none	History
Your smartphone replaces the roomful of equipment	replaces	0.0	0.5	none	History
Sweet solution? Google tests smart contact lens for diabetics	tests	0.0	0.5	none	Hardware
Douglas Adams' last post on his online forum was about excitement over Mac OS X	online forum	0.0	0.51	AI	Internet
The Precision Dynamics Discovery Shed	discovery	0.0	0.75	History	History
Take command of your email	email	1.0	0.833	Internet	Internet
Even my dog has wearable tech!	tech	0.0	0.25	Internet	Hardware
30th anniversary of the Macintosh	anniversary	1.0	0.75	History	History
The Powerhouse Museum's totalisator	museum	0.417	0.444	AI	History
The robots are coming	the robots are	0.0	0.225	none	AI
It's official - Alan #Turing is pardoned!!!	turing is pardoned	0.0	0.531	Blog	History
ELSIE hiding in Dunedin	elsie	0.0	0.5	none	History
New Zealand Computer Museum	computer museum	0.0	0.346	none	History
Strandbeest - kinetic sculptures	sculptures	0.0	0.5	none	Education
Earn college credits whilst working for Facebook	college	0.0	0.167	Internet	Education
#Apple's wise maps decision	maps	0.0	0.5	none	Software
Computer history on display at Auckland University	history on display	0.0	0.959	History	History
Rock paper scissors robot wins every time!	robot wins	0.0	0.799	AI	AI

Precision, Recall and F1 are shown below. Precision was the total number of correct guesses over the total number of guesses (where "none" does not count as a guess). Recall is the total number of correct guesses over the number of guesses and the "none" classifiers.

```
Precision = 0.58
Recall = 0.28
F1 = 0.38
```

3 Extra Credit

One part of the extra credit was to reclassify the blogs this time traing 90 and testing 10. The results are tabulated below like the first test.

Table 2: My caption

Title	**Gram**	**Cprob**	**Fisherprob**	**Guess**	**Actual**
The Powerhouse Museum's totalisator	museum	0.699	0.671	History	History
The robots are coming	the robots are	0.0	0.225	none	AI
It's official - Alan #Turing is pardoned!!!	turing is pardoned	0.0	0.403	Blog	History
ELSIE hiding in Dunedin	elsie	0.0	0.5	none	History
New Zealand Computer Museum	computer museum	0.0	0.438	none	History
Strandbeest - kinetic sculptures	sculptures	1.0	0.75	Education	Education
Earn college credits whilst working for Facebook	college	0.447	0.46	none	Education
#Apple's wise maps decision	maps	0.0	0.5	none	Software
Computer history on display at Auckland University	history on display	0.0	0.991	History	History
Rock paper scissors robot wins every time!	robot wins	0.0	0.809	AI	AI

```
Precision = 0.80
Recall = 0.40
F1 = 0.53
```

The Precision, Recall and F1 ratings improved in the second trial. This was expected since there was more training data, which inevitably means better results.

References

[1] T. Segaran, *Programming collective intelligence*, first ed., O'Reilly, 2007.