Measuring Sexually Explicit Content
Ethics Approval

Christopher Forbes

February 24, 2013

Introduction

The over all aim of the project is to investigate means of measuring the explicitness of language in textual documents. This document has been written to request ethics approval to undertake tests in which participants will be asked to read a collection of short texts and rank these by the degree/quantity of sexually explicit language. The purpose of these tests is to provide a set of rankings from human readers in order to compare the rankings generated from the software system being developed as part of this project.

Hypothesis

I aim to prove that the software package currently in development will rank documents in a similar order to human readers.

Experiment Participants

Ideally i would like between six and eight participants aged 20+ (due to the material requiring analysis see test data).

Participation agreement

Each participant in these tests will be given a short description detailing the aims of the tests and the overall project objectives, then they will be asked to read and sign the below agreement, giving their explicit agreement to participate in the tests.

I agree to take part in this experiment of my own free will, I understand that I will be provided with ten paragraphs of potentially explicit text and be asked to read the text then pass judgement on the explicitness of the language. I also understand that I have the right to terminate my participation in the experiment

at any point. By signing this document i agree to participate in this experiment and have been made aware of content provided as part of the experiment. No personal information will be gathered during this experiment, the only data gathered will be the order in which documents will be placed.

_____   _____
Participant   Date

_____   _____
Experiment Co-ordinator   Date

Participant information

Upon arriving each participant will also be asked to read a small description of the experiment, the information can be found below.

The aim of the experiment is to look at which order people place the documents provided in. The results of the experiment will then be compared to the results of the software currently in developement. The experiment should last approximately ten minutes during which time you will be left alone. No personal data will be gathered during the experiment exluding your signature on the participation form. The only data gathered will relate to your rankings of the documents. If at any point you feel uncomfortable with the text or wish to stop please say to the coordinator and the test will cease no questions asked. If you have any questions please feel free to ask before the test begins.

Variables

When asking for participants from the computer science faculty i must be aware that many students from the CIS departments spend much of their time online, as such may have built up a mental tolerance to particular words or phrasings. Subsequtnely skewing the results to the lower end of the spectrum. In an attempt to avoid this from happening i will ideally be able to get participants from some of the other years (not only fourth year).

Parameters

Each participant will be given ten paragraphs of text (each approxamatley eight to ten lines of text) asked to read each paragraph then place each of the paragraphs into order ranging from the least explicit text to the most explicit.

Intended use for data

I intend to use the data gathered from the experiments to order each of the ten documents, based on the average position it is placed in by the participants. I then intend to run each of these documents through my software and work out the order which it places them based on the keyword density assigned to each document.

Finally I will compare the positions calculated by the keyword density algorithm and compare them to the positions they are placed in by the participants.

Content

The content provided to participants is of a variently explicit language.

Educational the test data will have at least one example of text from sexual education websites, this is to allow people a gentle introduction to the text.

Control Case The texts will also include a control case. A benign text using the contextual ambiguity of English wording to it's advantage. This case will allow the differences between software and human readers to become apparent in terms of the context of the language.

Chat Logs The final type of content provided will be chat logs, as they are the basis of my test data the participants will be provided with excerts from the test data used entirely out of context with all names & ages removed.

Possible Issues

Due to the nature of the data being presented to the participants, their is a small possibility that people may feel uncomfortable with the material and as such I intend to make it clear to them that if at any point they feel uncomfortable with the text, they may stop the experiment and leave no questions asked.

To protect the participants I have also removed any information from the chat log text's which could cause distress, such as the ages / names of those involved in the chat.