Project Proposal

Ranking News Articles based on Prestige

Paul Horan, x12421032, x12421032@student.ncirl.ie

BSc (Hons) in Computing

Specialisation : Data Analytics

Date 26/10/2017

# Objectives

The objective of this project is to go through the internet using the likes of a Web-Crawler and taking out news articles. I would like to be able to rank the News Articles on Prestige, in theory this could help to take out fake news sites. With the use of API’s I am trying to search through the web and take out these fake news sites to allow end users access to real news. The likes of “Click-Bait” websites that use fake news to draw people to view their sites, these are what I would like to take out of the searches with this project.

The way I plan on making this a reality I will need to screen scrape News websites and take all of the tags from the sites to categorize the news. This will be done using a web-crawler and it will link all of the news via the tags on the articles webpage.

These Articles will then be ranked using the tags. I will be looking to rank each article by its integrity. Using something like Google’s Search Algorithm is how I would like this to run through. Having all of the ranked Articles towards the top of the search and all of the fake news articles around the bottom, hopefully remaining out of the end users view.

There will need to be a database storing all of the site tags, recent searches and categories.

A search engine should be created to allow ease of use for the end user.

# Background

Facebook and Twitter are currently looking into similar ideas to this project. I recently read about how they are looking into finding fake accounts. Facebook is taking aim at Fake News, and a part of this is discovering fake accounts. Currently it is believed that Facebook are currently rolling out updates to its systems to make it more difficult to create these fake accounts that promote fake news. The aim is to recognize patterns of activity, such as repeated posting of the same content. Facebook have said that changes have allowed it to identify and eliminate more than 30,000 fake accounts in France alone. If we were to look into this on a global scale the amount of accounts across the two social media platforms would be in the millions.

Fake media outlets are also an issue, many times i have been searching online for help or to read stories/articles and unrelated websites and pages appear with similar titles.This can be very irritating when it happens time and time again. But at the moment you just have to keep an eye out for pages like that.

I found this topic very interesting and wanted to see if I could create an application that could essentially decipher between Real and Fake news. This may be able to aid these companies in they’re search to disable fake accounts and news.

# Technical Approach

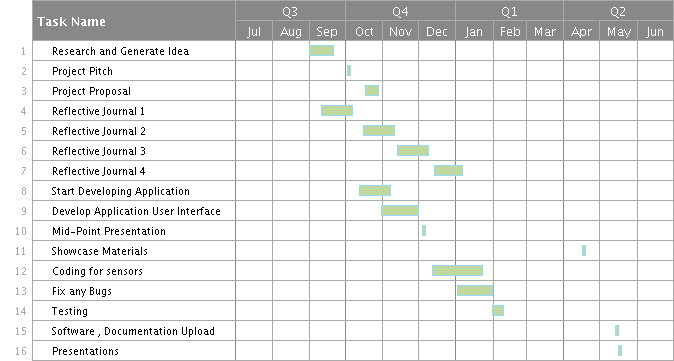
Languages such as Java, Javascript, HTML, CSS, Bootstrap and MySQL. These will be used for the overall GUI of the application. Java will be used for registering the users. MySQL then for the Database.

Research Technologies:

1. Python
2. BeautifulSoup
3. API Services
4. Apache Solr
5. ApacheLucene

Each to be researched for the Search Engine aspect of the project. Python and BeautifulSoup are being researched for the indexing of the data gathered from each webpage.

# Project Plan



# Evaluation

To evaluate the system, it must first be completed. Its main purpose is to eradicate the fake news providers. So every time the application is used the fake news websites should not appear within the search. Testing the system both offline and online. Offline by testing the search engine against a common database of the documents/Indexes. Then again when the system is online and we can see how it fairs against the offline test. We are looking for the online test to be faster than the offline test. More ideas will be further looked into during the making of the project. There will be a long haul of hard work throughout so testing will be needed and will be researched further as the project develops.