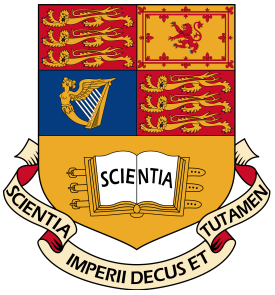


Imperial College London

C401

MEng Computing Final Report

NewSumm

News Aggregator and Summariser

Author:
Kunal M L Wagle

Supervisor:
Anandha Gopalan

May 29, 2017

Abstract

Abstract to be written

Acknowledgements

Acknowledgements to be written

Contents

List of Figures

List of Tables

1 Introduction

1.1 The Problem

1.1.1 The Media

The Fourth Estate is currently enduring one of its toughest passages of time to date. Some of the challenges it faces are familiar, whilst some are new ones that have been created by the onset of the digital age.

The term 'fake news' is now at the forefront of our minds in an atmosphere that is highly charged. The Digital News Report 2016 found that in the United Kingdom only 50% of readers trust what they say in the news, whilst only 29% say that they trust journalists (see Figure 1.1). This statistics appear to be even more dire for the media in the Untied States of America, when the percentages are 33% for news channels, and 27% for print journalists.

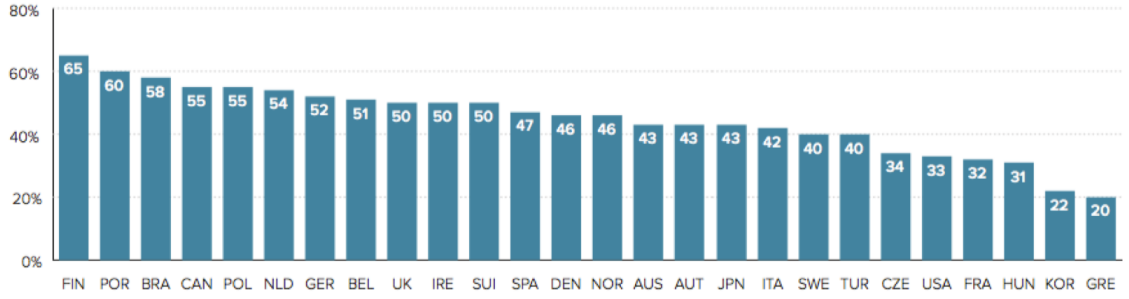


Figure 1.1: *Digital News Report findings showed that when it came to trust in the news, in only a few countries was trust as high as 50% in the accuracy of a piece of news. In countries such as the United States it is far lower.*

In addition to this, the media have been submerged in a deluge of websites that also claim to disseminate news to readers. In an increasingly digital world, anyone with access to the internet could write a blog about a news story. The popularity of social networks such as Facebook and Twitter also mean that these amateur 'journalists' have a platform on which they can compete with professionals from the industry.

In 2016, the concept of fake news dominated the general election in the United States. Shortly after the country went to the polls, Hillary Clinton's campaign became dogged in rumour after more than a million tweets were sent out about the 'pizzagate' scandal. The fake news went both ways. In July 2016, it was reported

widely on blogs that Donald Trump had said in an interview with *People* in 1998 that:

> *'If I were to run, I'd run as a Republican. They're the dumbest group of voters in the country. They believe anything on Fox News. I could lie and they'd still eat it up. I bet my numbers would be terrific.'*

The claim was later found to have been fabricated.

Fake news and the people's trust is not the only major issue facing the media at the moment. Another key is sheer volume. Given the abundance of media outlets in the world at the moment, and perhaps more importantly (due to the internet) the accessibility around the globe, there is a lot of competition between outlets.

1.1.2 The Consumer

As a result of a lot of the aforementioned accessibility, consumers also have a challenge in that they can't efficiently read all the news that is out there. If I were to search Google News on the 24th May 2017 for news about Donald Trump, I'd find a story about the President's meeting with the Pope at the Vatican (see Figure 1.2). On further inspection, I can see that there are several pages of results about the meeting itself, from dozens of different news sources. A consumer has little to no hope of reading all of these.

So which news sources should a consumer prioritise? Often, that comes down to political opinion. In fact, the Digital News Report 2016, conducted by the Reuters Institute for the Study of Journalism, says that only 34% of consumers in the United Kingdom believe that the media is 'free from undue political influence'. In the United States, which doesn't have a state broadcaster like the United Kingdom does in the BBC, the corresponding statistic is significantly lower at 21%.

The immediate consequence of statistics like this is that if someone hasn't read all the sources in a piece of news, they are reasonably likely to develop some form of political bias on the issue (perhaps without even realising it). For example, someone who reads about a story in a left-leaning publication such as *The Guardian* may feel differently when compared to someone who reads the corresponding story in a right-leaning publication such as *The Daily Mail*.

The simple solution to this would be reading the story in as many sources as possible before forming a final opinion. However, as mentioned before, this is just simply not feasible. In July 2016 a study was conducted by the Statistic Brain Research Institute that said that the average attention span of a person is now as low as 8.25 seconds. In addition to this, the Institute's findings also said that the average

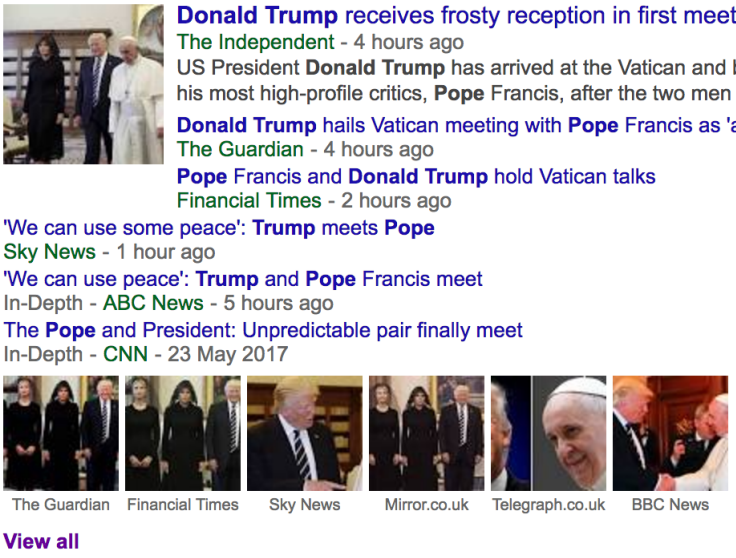


Figure 1.2: *A Google News search for stories about President Trump's visit to the Vatican on 24th May 2017*

person only reads 28% of a webpage of average length (593 words). Another key finding was that users spend only 4.4 seconds to cover each additional 100 words on a webpage. It's simply not possible for every news source to capture a user's attention, especially if they are all broadly covering the same story.

1.2 News Aggregator and Summariser: A potential solution

The fundamental aim of my project would be to create a project that directly addresses the consumer's concerns mentioned above, whilst indirectly aiding the concerns of the media.

The project would be a News Aggregator, that would allow users to see the news from multiple mainstream (and therefore 'trusted') sources in one place. In addition to this though, the aggregator would also act as a summariser, so that a user doesn't need to read as much as before to gain the same opinion.

However, a key aim has to be to address the issue of there just simply being too much news for someone (who is reading from multiple sources) to take in. To try and achieve this, the News Aggregator will need to identify articles across sources that are about the same topic (for example, Donald Trump's visit to the Vatican), and summarise them. This will mean that a user only has to read a single article to

get the full information from multiple sources about a news piece.

In an indirect attempt to allay the concerns of the media and the public about the phenomena that is fake news the News Aggregator will need to find a way to counteract it. One obvious way to do this would be through the limitation of the number of outlets to only those that are 'trusted'. In addition to this, a clear indication when reading the summary of what facts different outlets all agree on, and what facts they either differ on or omit, will allow the user to easily corroborate parts of the story.

A further secondary aim to this could be to allow a user to customise their summary, taking in information from the outlets of their choice (from the original 'trusted' list). The advantage to this would potentially be that if a user decides that too many issues from one source are uncorroborated, they can remove them, and see only information from the other sources that have written about a topic. The disadvantage to this could be that if a user dismisses an entire section of the political spectrum as fake continuously, they'll still pick up a political bias on the facts of the topic. Although, it could be said that the Digital News Report 2016 findings suggest that this is a feature that won't be used much. One of the more interesting findings in the report was that more than 60% (in the United Kingdom) were concerned about the potential impact of personalising their news feeds. They felt that it could lead to both 'missing information' and 'missing challenging viewpoints'.

2 Market Research

2.1 News Reading Habits

2.1.1 Digital News Report 2016

Conducted by the Reuters Institute for the Study of Journalism[24] in Oxford University[31], the Digital News Report[23] is an annual study that serves to investigate how people access and find out about news in various countries across the globe, including the United Kingdom.

The study attempts to cover, amongst other things how people get the news, how they use social media, what types of news they trust, and both the reasons to use and concerns about news aggregators. Findings of the report relevant to my concept are listed below:

Reasons to use News Aggregators

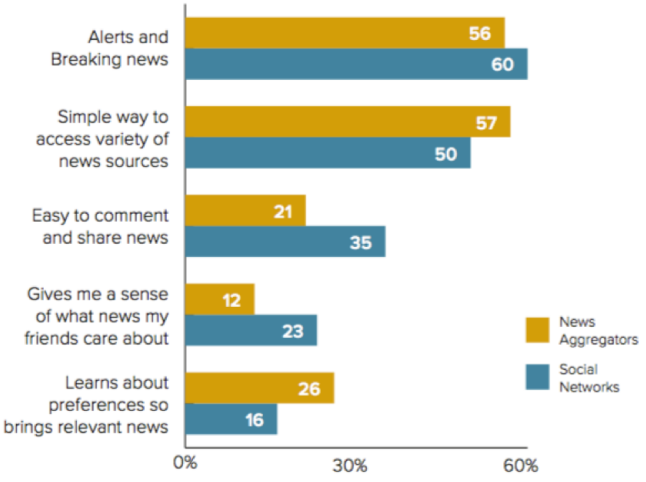


Figure 2.1: *A graph showing why people use News Aggregators and Social Networks as news sources*

As shown in the graph in Figure 2.1 the key to the popularity of a News Aggregator is that it's simplicity and speed. It's important to bring news as soon as it happens, and it needs to provide access to a variety of different news sources. Of lesser importance is the social aspect - the need to comment and share the news, although the fact that 26% of people want their preferences to influence the news that they're interested in should not be discounted.

Concerns about Personalised News

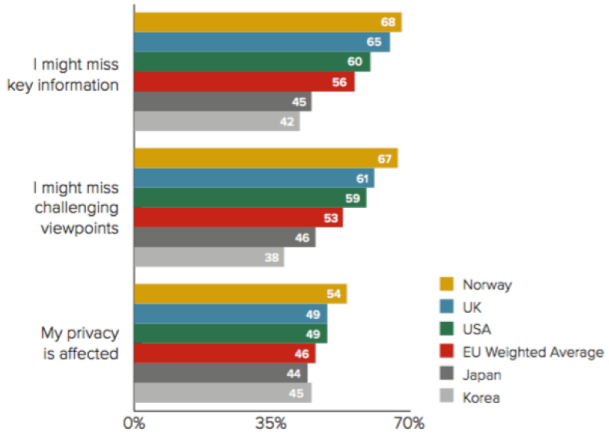


Figure 2.2: *A graph showing the main three concerns that people have with Personalised News Feeds determining what they read.*

Key concerns, especially in the United Kingdom, with Personalised News is that information might be incomplete. It is not difficult to see why this may be a concern - if you limit your news to a subset of the sources that might be available then it's possible to miss parts of the story - the parts that challenge aspects of the articles you're reading. It's therefore important that the project reflects this concern, which is why it is listed as a secondary aim (See section **??**).

What should determine what someone reads next

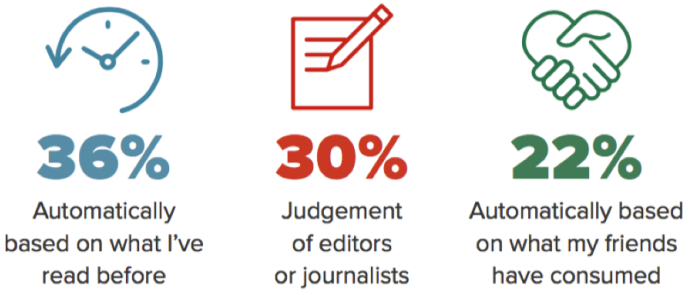


Figure 2.3: *Key statistics in response to the question: Is this a good way to get the news?*

Least popular are algorithms that present news based on what people's friends have been reading. In general, according to the Reuters Institute[24], 'People think *they*

are the best judge of what's important to them.' More interesting, is that people are now less inclined to allow journalists or editors to push articles to the top of reading lists than having an algorithm suggest similar stories to what they've just consumed. This could perhaps indicate that they aren't as trusting in journalists and editors as before - something that the Reuters Institute touches upon in an essay near the end of the report.

Conclusions

The Reuters Institute for the Study of Journalism is a trusted and respected organisation, and this study has been supported by institutions varying from media outlets such as the BBC[3], news search engines such as Google, and other Universities, such as University of Canberra[30]. Importantly, it's also supported by regulators, including Ofcom[22], which is the regulator in the United Kingdom.

The survey was conducted on their behalf by YouGov[39], which is a highly respected polling company in the United Kingdom, and it surveyed more than 50,000 people, including over 2000 in the United Kingdom. In addition, the report is coupled with numerous essays on various key points that arose from the study. The writers of these essays varied from the Director of Research at the Reuters Institute to the Chief Executive Officer of The New York Times[29].

As a result of this, I deemed the Digital News Report 2016[23] to be a very reliable source of information with regards to how people read the news, and their preferences in the field.

2.1.2 Potential User Survey

Following this initial research I decided to conduct a survey of potential users to delve into more specific aspects of News Reading Habits. I was more focused on the number of sources people use to ascertain the facts about a piece of news. I also asked questions regarding the summarisation of news and news digests.

I was interested in the questions about summarisation and digests after reading some data from a study conducted by the Statistic Brain Research Institute[25]. Statistic Brain conducted a study in July 2016 that suggested that the average attention span of a human is now only 8.25 seconds[16]. They coupled the presentation of this data with statistics about people browsing the internet taken from the paper *Not Quite the Average: An Empirical Study of Web Use*[33]. The statistics suggest that the average user only reads 28% of the words on the average webpage, and that for each additional 100 words beyond that, the user would only spend an additional 4.4

seconds on the page.

Admittedly, the paper was written in 2008, and so might not reflect average web browsing activity, but it raised a question worth answering in my opinion. Do users actually read full articles on the news, and if not, would summarising articles help make sure they didn't miss out on key points, thus counteracting an issue discovered in the Digital News Report 2016[23] (Section 2.1.1)?

The survey was presented on Google Forms[10] and was answered by 96 respondents. Further findings are shown below:

1. **How often do you read the news?**

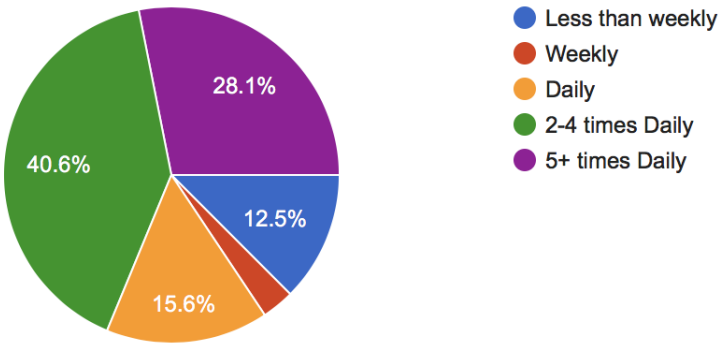


Figure 2.4: *A pie chart showing how often people read the news.*

The first question (Figure 2.4) on the survey was designed to ascertain how often people read the news. Overall, well over 80% of respondents said that they read the news at least daily, and nearly 60% more than once per day.

2. **How many sources do you read?**

Next I asked about how many sources a user read. It was quite rare for someone to use only one source for a piece of news. In fact, over 90% said that they used two or more sources, with three and five sources being the most prominent selections.

3. **Are these the same sources every time?**

The respondents were a lot more split on the question of consistency of sources, although more than half said that they used the same sources every time, with

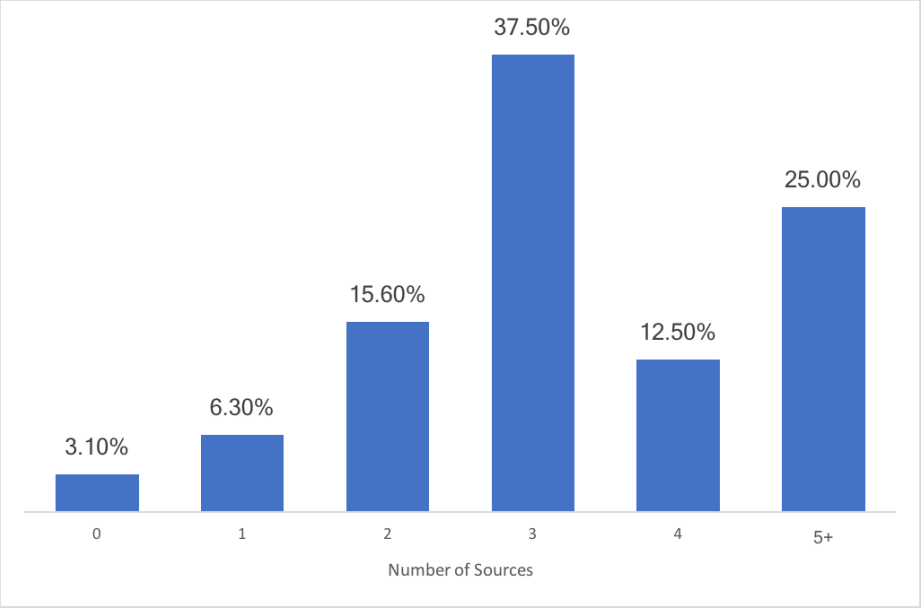


Figure 2.5: *A bar graph showing how many sources people normally use for a source of news.*

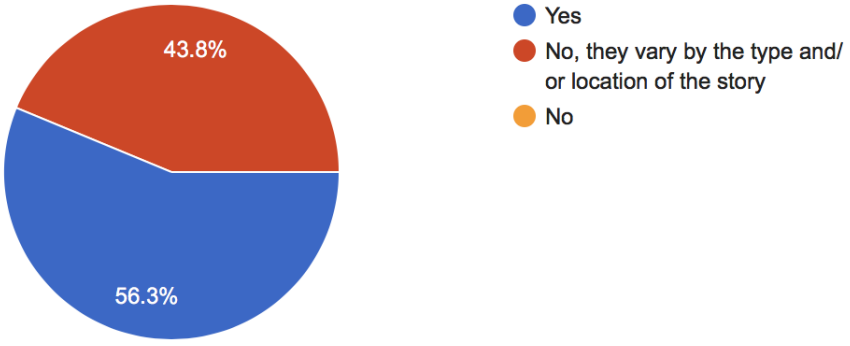


Figure 2.6: *A pie chart showing whether people use the same sources every time.*

43.8% saying that they vary their sources based on the type and/or location of the news story.

4. **How much of an article do you normally read?**

   I further expanded upon the initial reading I had done about attention spans by asking potential users how much of articles they normally read. Only 18.8%

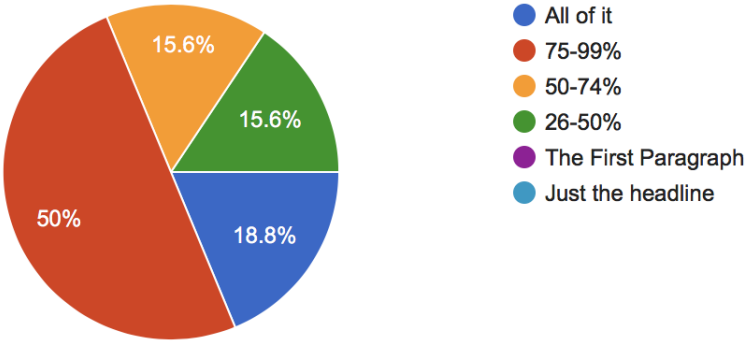


Figure 2.7: *A pie chart showing how much of articles people normally read.*

of respondents said that they read the entirety of articles (Figure 2.8), and as much as 15.6% confess that they read less than 50% of an article on average. It's plausible that the users could be missing key facts through not reading the entire article. On reflection, I could have also asked if this was a concern to those users.

5. **If an article was presented as a shorter summary, would you read more/all of it?**

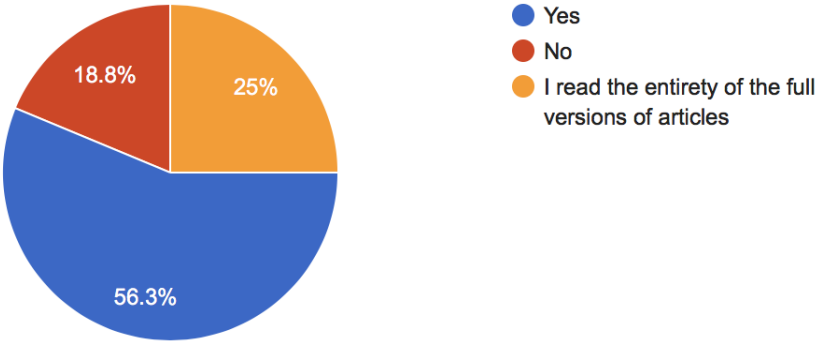


Figure 2.8: *A pie chart showing what percentage of people would read a shorter summary of an article.*

Another question asked was whether reading a shorter summary would result in users reading more or all of the article. Encouragingly, more than half

(56.3%) of respondents said that they would read a shorter summary.

6. **Would you read a summary of different articles on the same topic combined?**

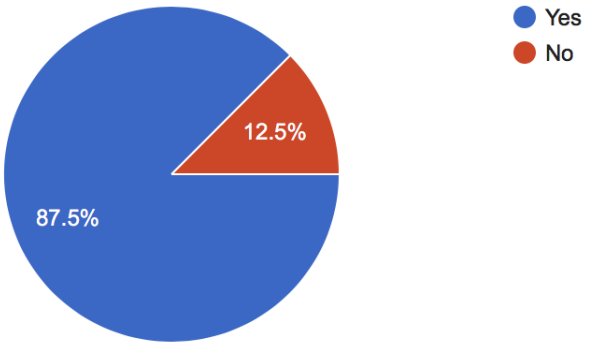


Figure 2.9: *A pie chart showing whether people would read a summary of different articles on the same topic combined.*

With this question (Figure 2.9) I aimed to obtain an idea from potential users as to whether combining articles about the same topic from different sources and then summarising it would be considered useful. The question got a resounding response, with as much as 87.5% saying that they would read a summary like this.

7. **Would you use a News Aggregator that summarised articles?**

Consolidating the findings of the previous question was the fact that 81.3% said they would use a News Aggregator that summarised articles (Figure 2.10). Amongst those who said they wouldn't, there were comments along the lines of 'I'm not great with technology' given as explanation. An interesting comment however, said that 'nuance could be lost in the summarisation'. This is a valid point, and so a key point of the evaluation process will have to be focused on the summarisation of articles itself, to ensure it's not losing important information at any stage.

8. **How often do you search for news on a specific topic?**

For the questions in Figure 2.11 I tried to get an idea of how much people search for news on a specific subject of interest to them. In general, this came

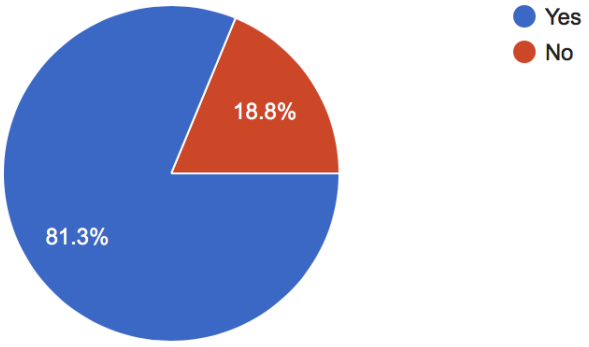


Figure 2.10: *A pie chart showing answers to the question: 'Would you use a News Aggregator that summarised articles'*

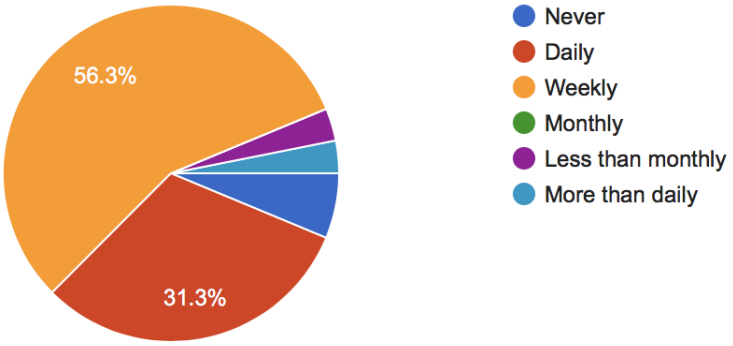


Figure 2.11: *A pie chart showing how frequently people search for news on a specific topic.*

out to be less frequent then reading the news itself, with just over half the respondents saying that they search for a specific topic weekly. However, a healthy proportion (31.3%) search daily for specific topics, and some search even more often than this.

9. **Should the news aggregator have the ability to search for a specific topic?**

   As much as 81.3% of respondents agreed that the News Aggregator should have a function to search for specific topics.

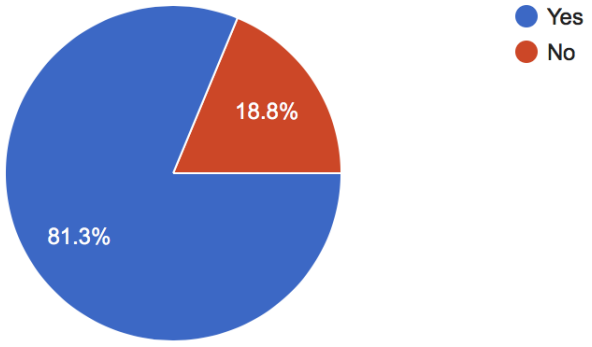


Figure 2.12: *The answers to a question asking if the News Aggregator should have the ability to search for a specific topic.*

10. **Do you like services such as news digests that prepare lists of articles each day that apply to a topic you may be interested in?**

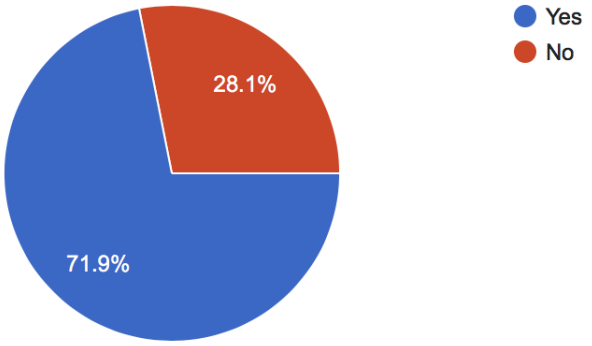


Figure 2.13: *A pie chart showing people's attitudes towards News Digests.*

The next question, shown in Figure 2.13 centred around News Digests. Nearly 72% of respondents said that they liked services that provide news digests.

11. **How often should these news digests be updated?**

People who were in favour of news digests in general felt that these should be updated at least daily (47.8%), with a further 34.8% on top of that saying it

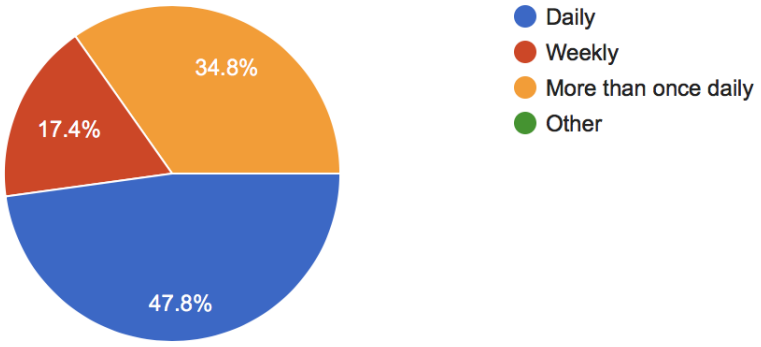


Figure 2.14: *Responses regarding how frequently news digests should be updated.*

should be updated more than once.

12. **What platform should the aggregator be developed for?**

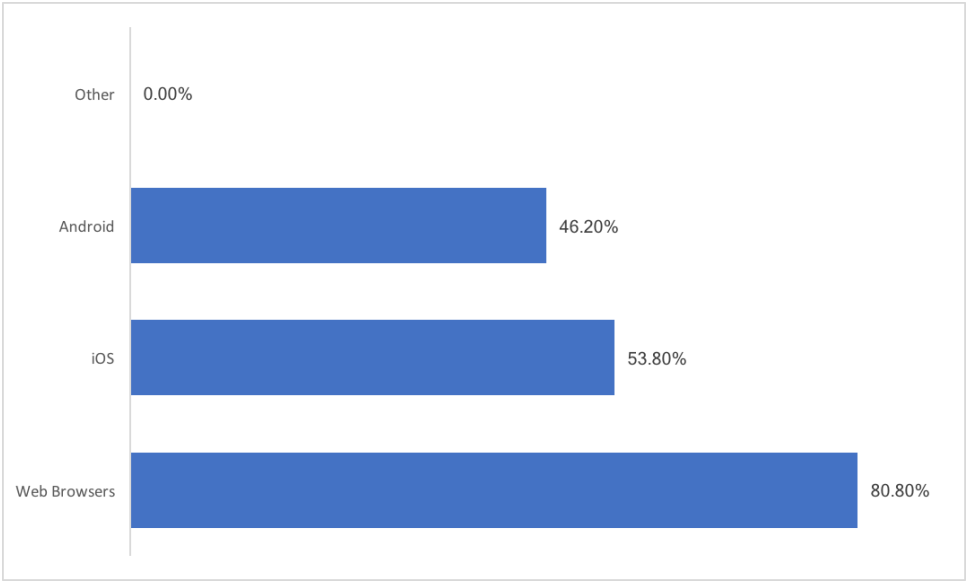


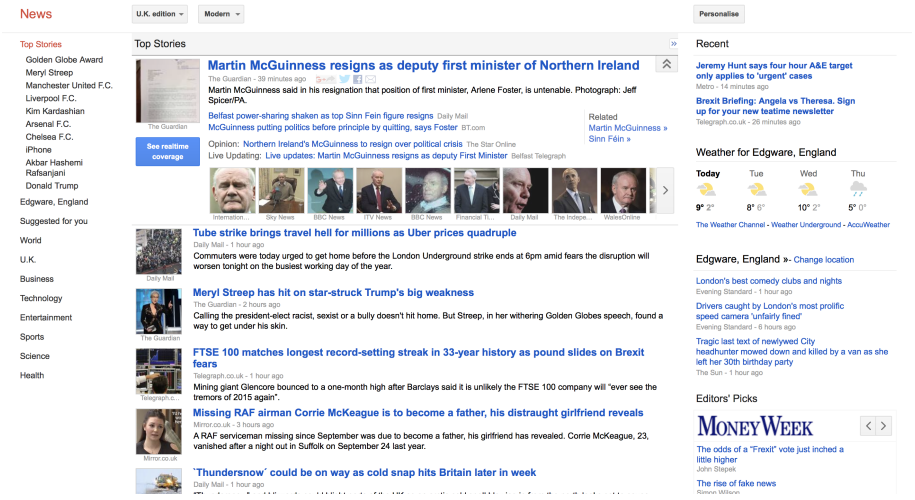
Figure 2.15: *A bar chart showing responses when asked which platform they'd rather see the news aggregator developed for.*

Finally, respondents were asked which platform they would prefer to see the news aggregator developed for. 80.8% said they'd want a version for web browsers, and 53.8% for iOS and 46/2% for Android. This would suggest that

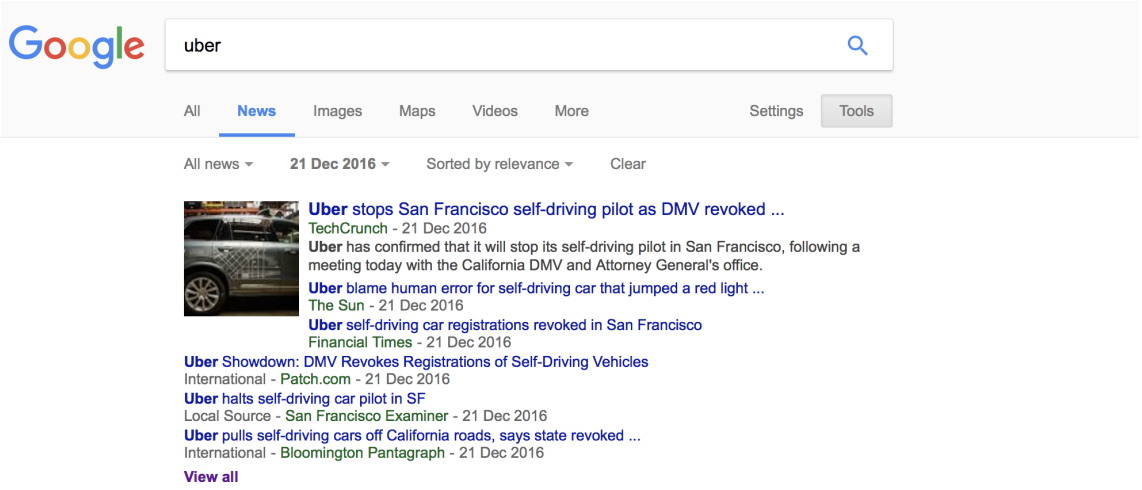
I should develop a web solution and then move to a mobile platform afterwards if there is time.

2.2 Related Products

2.2.1 Google News



(a) A section of the homepage for Google News.[11]



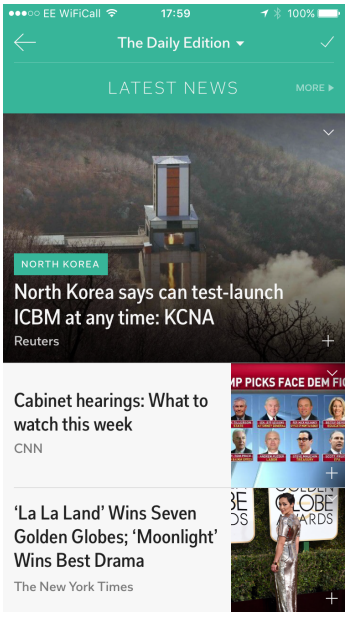
***(b)** Google News uses clustering techniques to group articles about the same topic together.*

***Figure 2.16:** Screenshots from Google News*

Initially developed early in the century and released in 2006, Google News[11] is a free-to-use news aggregator. Google News operates in a similar manner to the traditional Google[9] search engine, thus making it a go to aggregator when searching for a specific topic. Google News also groups 'similar' articles together using Clustering techniques[15]. Google News operates purely as a go-between - when clicking on an article the user is taken straight to the media source itself, rather than being able to read the article on Google News itself.

2.2.2 Flipboard

(a) Users can subscribe to 'magazines' that they may be interested in.

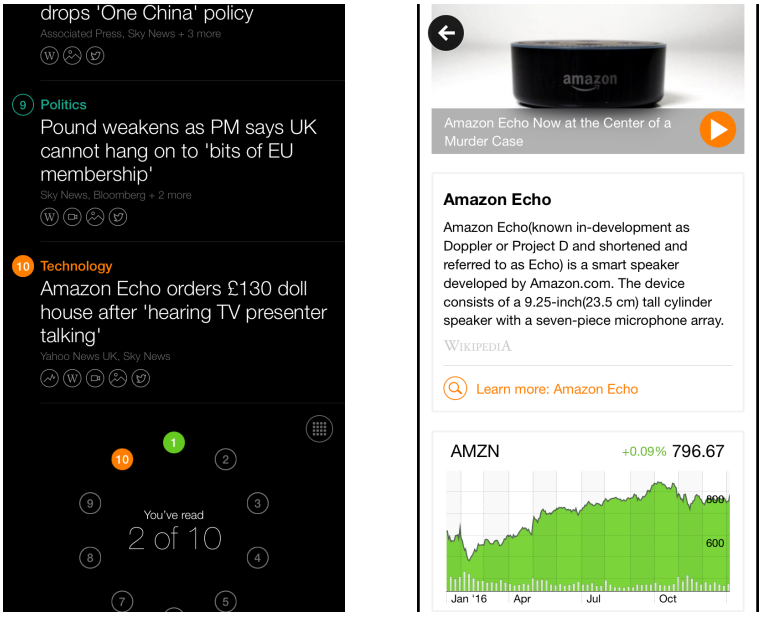


(b) All users are subscribed to The Daily Edition*, which puts together the latest most popular news.*

Figure 2.17: *Screenshots from the Flipboard app*

Flipboard[6] is a much more recent attempt at a news aggregator (developed in 2010), and relies on the concept of users subscribing to 'magazines' on different topics. There's a central 'cover page' on the home page that shows the most recent stories from across all a user's subscriptions. Like Google News, links from the desktop website send the user to the original media source, while links from within the mobile applications open a browser within the app itself.

2.2.3 Yahoo News Digest



(a) *A section of the home screen, which displays the top ten stories for the day.*

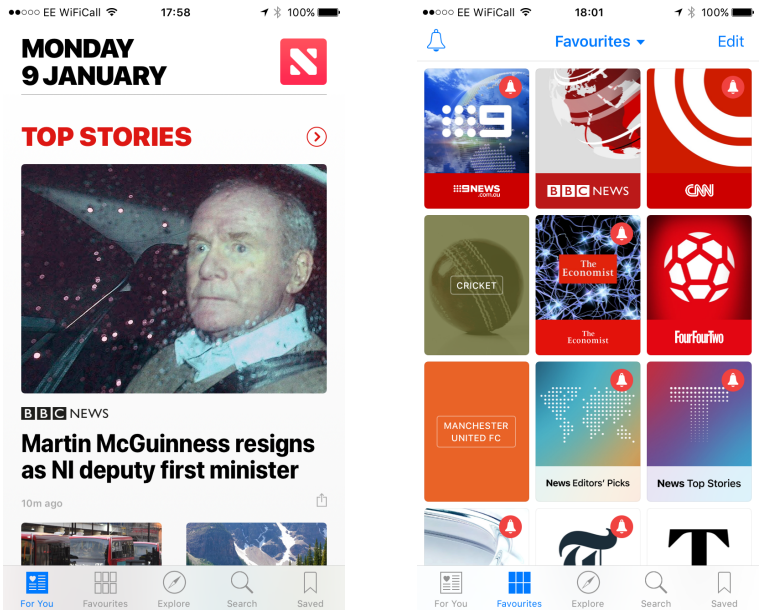
(b) *The Yahoo News Digest app provides links and infographics that are relevant to the article.*

Figure 2.18: *Screenshots from the Yahoo News Digest app[38]*

Yahoo News Digest[38] is a direct evolution from Summly[26], which was an app that summarised news. Yahoo News Digest is a phone application that creates two digests a day: one in the morning and one in the evening. Each digest contains the ten leading articles from the previous 12 hours. Each article provides a summary of the story and links to relevant other pages - such as articles from Wikipedia[36].

2.2.4 Apple News

Apple News[2] is an application that is installed by default on all recent iOS devices. Apple's default attempt at a news aggregator allows users to select their preferred news sources, and from a selection of topics. Apple[1] then presents on a home screen news from those sources and topics. Clicking on each article keeps it in the native application, rather than sending the user to the media source itself.



(a) The home screen for the Apple News app

(b) Apple News allows users to select sources and topics to be their favourites.

Figure 2.19: *Screenshots from the Apple News app[2].*

2.2.5 Comparing the existing products

The existing products are compared in Table 2.1 on page 23.

Product	What it does well	What it doesn't do well
Google News	• Groups similar articles together • Similar to a traditional search engine and so is easy to use • Obtains articles instantly, thus providing most up-to-date information when searching	• Doesnt́ host articles itself - therefore making it harder to navigate than perhaps could be possible. Although, this could be to avoid any copyright issues.
Flipboard	• Provides a home page that allows a user to see the most popular stories at the time related to the user's subscriptions. • Available as a mobile application	• Topics are much broader than Google News, thus meaning that users can't necessarily search for topics specific-enough for them. • Flipboard requires registration before being able to read articles
Yahoo News Digest	• Yahoo News Digest won the 2014 Apple Design award[37]. • The articles also provide key infographics, quotes, and other information potentially relevant to the article.	• There are only ten articles available per digest, and no capability for searching by topic. • Digests are only produced in the morning and the evening, so the news articles presented could be out of date.
Apple News	• Allows selection based off both topics and the news sources themselves. • The home screen for the app allows users to see a lot of headlines at a glance.	• Navigation on the application is not simple. If a user has accessed many articles from push notifications, then the user could have to press the back button several times to get back to the home screen. • The topics that a user can subscribe to can be quite limited, and aren't reactive to current affairs - for example, if a natural disaster occurs, you couldn't then subscribe to that natural disaster as a topic.

Table 2.1: *Comparing the relative benefits and drawbacks of Google News, Flipboard, Yahoo News Digest and Apple News*

3 Background Research

3.1 Machine Learning Techniques

3.1.1 Topic Modelling Techniques

Topic modelling is a subsection of Machine Learning that aims to determine what topic a given document is about. The topics wouldn't be named at this stage, they would simply be given generic names such as Topic A and Topic B. Assigning names to topics will be done at a later stage (see Section 3.1.2).

Latent Semantic Indexing

Also known as Latent Semantic Analysis[19], Latent Semantic Indexing (LSI) was one of the initial forerunners in the field of topic modelling. It uses singular value decomposition to locate patterns in the text of a document and thus form a basis on which to categorise the document. A major benefit of LSI is that it is fast to train, but in general it has lower accuracy when compared to models that are probabilistic, such as Latent Dirichlet Allocation[35].

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probablistic model that assumes that the topic distribution has a Dirichlet prior. The theory behind LDA is that a document of words contains a mixture of different topics, and this is reflected in the final answers given for the algorithm.

A rough algorithm for performing LDA[34]:

1. **Set n to be the number of topics there are in the document.** We can do this by trial and error.

2. **Assign every word w in the document d to a random topic.** These topics are temporary. At this stage we can remove function words, such as 'the'. However, we keep duplicates - in fact, at this stage they could be in different topics.

3. **Check and update topic assignments.** To do this, we loop through each word in the document, taking note of how prevalent the word is across documents, and prevalent those topics are in the document. These two probabilities are then passed to a sampling algorithm to generate a new topic for the word. This step is usually completed using the statistical model *Gibbs Sampling*.

4. **Repeat step 3 until there are no more topic-reassignments**.

3.1.2 Topic Labelling Techniques

Topic Labelling techniques in the project for labelling the topics that are generated from the LDA analysis of the document in Section 3.1.1. When conducting research into this topic specifically, I found a paper (*Automatic Labelling of Topic Models* by Lau, Grieser, Newman and Baldwin in 2011[20]) that documents the creation of an algorithm that labels topics using Wikipedia[36] titles. This could work very well in my project, as Wikipedia titles as title headings would allow users to search for topics that are both broad and specific.

A rough version of Lau, Grieser, Newman and Baldwin's algorithm:

1. **Calculate the top 10 topic terms.** This is done by finding the marginal probabilities of each word from the original topic models, and taking the top ten. The marginal probability of a term is the probability of that term being randomly selected given a topic t.

2. **Search Wikipedia using these terms.** We also search Google[9] using a site restricted search (to Wikipedia) and take the top eight results from each. These are called the *primary labels*.

3. **Isolate all 'noun chunks' from the terms.** In this case noun chunks are combinations of words from the terms that appear next to each other. For example, with the term 'Summer Olympic Games' the noun chunks would be 'Summer', 'Olympic', 'Games', 'Summer Olympic', 'Olympic Games' and 'Summer Olympic Games'. Note that 'Summer Games' is not a noun chunk as the words don't appear juxtaposed. These noun chunks are added to the primary labels from step 2 and are deemed *secondary labels*.

4. **For each noun chunk:**
   - Check to see if the noun chunk is the title for a Wikipedia article
   - Remove the noun chunk if it doesn't correspond to a Wikipedia article

5. **Calculate the *Related Article Conceptual Overlap* scores.** Related Article Conceptual Overlap (RACO), developed by Grieser et al in 2011[12], is a calculation designed to identify the strength of relationship between terms by

inspecting the category overlap between the terms' corresponding articles. We do this for each secondary label still remaining - details on how to calculate the RACO scores are explained in further detail below.

6. **Discard all secondary labels with RACO score of less than 0.1.**

7. **Add five highest topic terms to the list.** Now we return to the original list of topic terms from step 1 and add the five highest to the remaining candidates.

8. **Perform candidate ranking.** There are multiple ways to do this, but the original paper recommends using a variety of statistical methods, based around a T-test, the Chi-squared test and a log-likelihood test. The aim is to estimate how closely related the candidate is to all the terms in the topic. We then take the top candidate as our final answer.

Calculating the *Related Article Conceptual Overlap*

Related Article Conceptual Overlap (RACO) was first introduced as a concept in the paper *Using Ontological and Document Similarity to Estimate Museum Exhibit Relatedness*[12] by Grieser, Baldwin, Bohnert and Sonenberg in 2011. It compares two terms and estimates their similarity by comparing the two terms' similarity on Wikipedia. The core calculation for RACO goes as follows:

$$Category - Overlap(a, b) = \left| \left(\bigcup_{p \in O(a)} C(p) \right) \bigcap \left(\bigcup_{p \in O(b)} C(p) \right) \right|$$

In this equation, *O(a)* represents the outlinks (the links to other articles from the Wikipedia article) of an article a and *C(p)* represents the set of categories that article p is a part of.

An issue with the Category-Overlap calculation as it is, is that there's a bias for articles that are larger than others as they will have more outlinks but won't necessarily be in more categories. As a result it's normalised using Dice's coefficient to produce the final RACO equation:

$$sim_{RACO}(a, b) = \frac{2 \times \left| \left(\bigcup_{p \in O(a)} C(p) \right) \bigcap \left(\bigcup_{p \in O(b)} C(p) \right) \right|}{\left(\bigcup_{p \in O(a)} C(p) \right) + \left(\bigcup_{p \in O(b)} C(p) \right)}$$

3.1.3 Clustering Techniques

Cluster analysis is a machine learning technique that is used to put items that are similar to each other into groups. In practice it's used by Google News[11] to put articles about the same topic together for a user[15]. There are multiple commonly used types of cluster analysis:

Centroid Clustering

In Centroid Clustering[4], also known as k-means clustering, there are k clusters. A vector is calculated for each article in the list. The article is then assigned to the cluster that is closest to it's vector score. A major downside to this method however is that it requires k to be defined in advance. In the context of this project, that's not applicable as we don't know how many different articles we are clustering based on.

Density Clustering

Density Clustering[4] also involves calculating a vector score for each article. Once these have been calculated, they can be graphed, and the areas of the graph that have highest density are chosen as the clusters. An advantage of this over the centroid clustering methods is that we don't need to know the number of clusters beforehand. However, density clustering can become less accurate as it requires areas of sparse density on the graph to precisely separate the different groups, which isn't always possible.

Hierarchical Clustering

Hierarchical Clustering[14], which is also known as Connectivity Clustering, is based on the idea of using distance measures between articles to identify which ones are most similar. There are two approaches to Hierarchical Clustering:

- *Agglomerative*, which is a bottom-up approach, assigns each item to its own cluster and then merges pairs that are closer together.
- *Divisive*, a top-down approach, that begins with all items in a single cluster, and then proceeds to split it into multiple clusters.

Creating a vector score for each item

The first step in each of the three clustering techniques is to create a vector score

for each item. As this will play a big part in the final results, it's important to get this stage right. This can be split into two steps:

1. **Find definitive terms within the article.** This can be done using techniques such as the popular *Term Frequency-Inverse Document Frequency* (TF-IDF) , which is explained in further detail below. Proper nouns would be useful in this step, as they are more likely to be relevant to what the article is specifically about.

2. **Create a vector.** This vector, based from the definitive terms from the first step, would be a set of keywords and corresponding weights.

Term Frequency-Inverse Document Frequency

TF-IDF[27] is designed to identify terms that appear frequently in one article that don't occur a lot over the entire set of articles (also known as the *corpus*). Variations of it are commonly used by search engines to identify search results that are most relevant to a query. The calculation for TF-IDF is as follows:

Term Frequency

Term Frequency can be most simply calculated as the frequency of a term in a document. However, this could result in a bias towards terms that appear in longer articles. A more accepted way to calculate the Term Frequency therefore is to use a normalising function, called augmented frequency:

$$tf(t,d) = 0.5 + 0.5 \times \frac{f_{t,d}}{max\{f_{t',d} : t' \in d\}}$$

Inverse Document Frequency

Inverse Document Frequency is used to check in how many documents of a corpus D a given term t occurs. It is an inverse function, so as to minimise the value for the term when it is common amongst various different documents. It is also logarithmically scaled. It is calculated using the following formula:

$$idf(t,D) = \log \frac{|D|}{|d \in D : t \in d|}$$

Term Frequency-Inverse Document Frequency

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

3.2 Summarisation Techniques

Summarisation techniques will be (predictably) used for summarising the merged articles that were identified by the processes of Topic Modelling , Topic Labelling and Clustering. There are two types of summarisation:

- **Extractive Summarisation**, which consists of taking sentences that are important from the original text, and discarding the rest.[13]
- **Abstractive Summarisation**, which aims to generate a piece of text using natural language techniques. A key to this is that some words in the final summary may not have been the original piece of text.

3.2.1 Extractive Summarisation Techniques

LexRank and TextRank

There are two well known examples of extractive summarisation: LexRank and TextRank.

LexRank and TextRank have very similar methods for extracting a summary[5]. Initially a graph is constructed, that consists of one node for each sentence in the corpus. Then a clustering algorithm is applied, using a TF-IDF calculation to determine similarities.

There are a couple of key differences between LexRank and TextRank. The first arises in the calculation. Both use a TF-IDF calculation, but they are varied slightly. LexRank uses a cosine similarity function in order to weight the final calculation, whereas TextRank uses a more simple logarithmic weighting to perform the calculation.

Both use Google's famous PageRank algorithm to then rank the importance of each sentence based on the calculations in the previous step. With d being a damping factor, the PageRank of a node u is given as:

$$p(u) = \frac{d}{N} + (1 - d) \sum_{v \in adj[u]} \frac{p(v)}{deg(v)}$$

After this has been calculated, the top ranked sentences are taken by the TextRank algorithm to form the summary. However, in the LexRank algorithm sentence position and length are also taken into consideration. Also, when adding a sentence to the summary, the LexRank checks the sentence against the summary to ensure that it won't be redundant. As a result of this extra step, LexRank is considered more suitable than TextRank when summarising multiple documents, whereas TextRank is only normally used for summarising a single document.

3.2.2 Abstractive Summarisation Techniques

There are six common methods for abstractive summarisation, which can be split evenly into two distinct categories[17]. These two categories centre around the creation of a representation of the given document:

- **Structure based methods** consist of techniques that involve determining the important information in a document by considering its structure[18]. Ways of doing this include fitting the given document to a template, or converting the text in a tree-like structure.

- **Semantic based methods** involve building a semantic representation of the given document and then feeding that into an algorithm that generates natural language that forms the final summary.

Structure based summarisation

Tree based summarisation

With tree based structuring, the first step is to create a dependency tree to represent the document. A dependency tree is a tree with a node for each word in a sentence, the links between the trees show which words depend on each other. For example, given the sentence *This is an example*, we can construct a tree as in figure 3.1:

In the tree (figure 3.1), the word *This* is dependent on *is* and so is linked. *an* is linked to *example* in a similar way, and *an example* is dependent on the verb *is* and so also has a link to it.

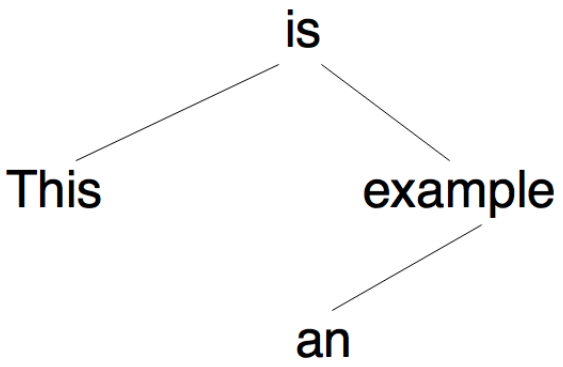


Figure 3.1: *A possible dependency tree for the sentence* This is an example

Once dependency trees have been created for each of the sentences a similarity algorithm is used to find sentences that are similar. The common phrases between these similar sentences are then taken and form the basis of the final summary. A language generator then combines the common phrases and arranges it to create a final set of summary sentences.

An obvious disadvantage to this method is that if only common phrases are taken from the sentences, context could be lost from some of the sentences. However, on the other hand, the use of a language generator means that a more coherent, more grammatically correct summary is formed.

Rule based summarisation

In rule based summarisation[8], the process is centred around a list of pre-determined categories. With each of these categories, there is a pre-determined set of questions to be answered.

For example, with a theoretical category *Product Launch* we could have the following questions:

1. What is the name of the product?
2. What company is launching it?
3. What type of product is it?
4. What's new about it?
5. What price is it retailing at?

This represents only a subset of the possible questions we could have for this category.

To perform rule based summarisation, the first step is to analyse the document and determine which category it fits best. Once that's been determined, the next stage is to analyse the text to find answers to the questions corresponding to that category. These answers are then fed in to a natural language generator that forms sentences, and thus the summary.

Results for this method of abstractive summarisation have been promising, but the method has a major disadvantage in that the list of categories and questions needs to be pre-determined. As a result, it might not be an optimal algorithm to use for the ever-changing world of news reporting.

Ontology based summarisation

Methods of ontology based summarisation have been developed, primarily using domain based ontology. In this method a 'domain expert' defines a domain ontology for a news event. Each new document is then classified into a topic using these domain ontologies. Important phrases are determined by how close they are to items in the ontology. These phrases are then passed into a natural language generator to form the final summary sentences.

A key disadvantage to this method is that a lot of the domain ontologies has to be manually determined by the 'domain experts' and so can be very time consuming. As a result this may also not be particularly optimal for summarisation of news events.

Semantic based summarisation

Multimodal Semantic summarisation

Multimodal semantic summarisation can work on documents that contains both text and images. First, a semantic model is built to represent the document. This model is made up of different concepts that are surmised from the text. For example, the sentence *Multimodal Semantic summarisation is an example of an algorithm that performs abstractive summarisation* could form the concept shown in figure 3.2:

Concepts are gradually filled with more information as the entire text is analysed. Links are also added between concepts that share some relationship. For example in figure 3.2 if there was another sentence that surmised a concept called *Abstractive Summarisation* then there would be a link from *Algorithm1* to that new concept.

The next step is to rank the concepts. This is done by taking into account the completeness of the concept, and the number of links that the concept as to others. This

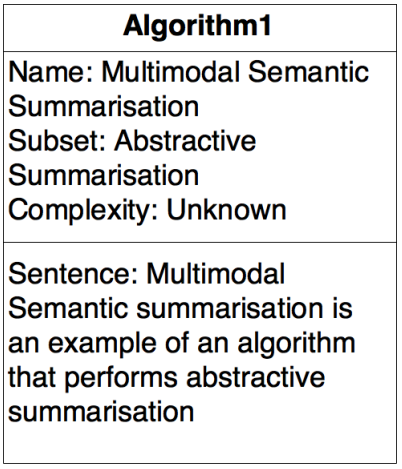


Figure 3.2: *A possible concept created from the analysis of the sentence* Multimodal Semantic summarisation is an example of an algorithm that performs abstractive summarisation

way, the concepts are ranked by which is most important to the original document. Once the key concepts have been identified summary sentences can be generated featuring these concepts.

Information Item based method

Information Item based summarisation[7] relies on the content of the summary being determined from an abstract representation of the original document, rather than the sentences from the document themselves. To do this, the document is first scanned so that Information Items (InIt) can be generated. An information item is defined as being 'the smallest element of coherent information in a text or sentence'.

Once the information items have been created, they are then ranked using frequency analysis to find the most important predicates and entities from the original document. This step is near identical to the term frequency stage in extractive summarisation (Section 3.2.1). These information items that are ranked highly are then combined and fed into a natural language generator to form the final summary sentences.

Semantic Graph summarisation

Semantic Graph summarisation centres around a rich semantic graph (RSG). The document is first converted into a RSG. Each node in the RSG represents a noun or verb in the document, and the links between the nodes represent the semantic and

topological relations between these nouns and verbs. In the second stage, heuristic rules are used to reduce the semantic graph to a more minimalistic version. This will form the basis of the final summary. In the final step, the minimised RSG is passed into a generator that creates the final summary sentences.

The steps are outlined in the flowchart provided in figure 3.3.

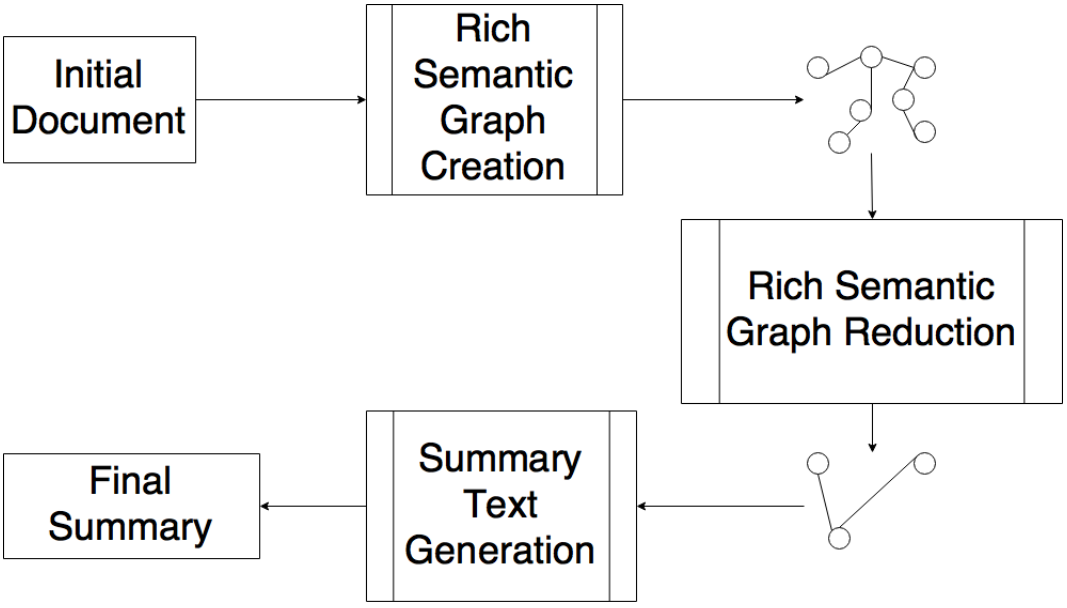


Figure 3.3: *This shows the process of semantic graph summarisation in a flowchart based on those provided in [17, 18]*

.

Semantic Graph summarisation has had success with producing an abstractive summary that has fewer redundant sentences, and is also good at producing grammatically correct sentences. However, it's only designed for use with a single document as input. As a result it might not be suitable as a method for summarising the multiple documents needed for the aggregator, unless it's combined with another method.

3.3 Natural Language Processing Libraries

A lot of the summarisation techniques specified earlier, especially for abstractive summarisation, would require a full analysis of the semantics of a body of text. As a result, it's clear that a Natural Language Processing library will be needed for this task.

3.3.1 Aspects of Natural Language Processing

There are several different tasks that a Natural Language Processor. A few key ones that are most likely to be required for the project are explained briefly below:

- **Tokenisation** is the splitting of words in a given document.
- **Sentence Segmentation** is the splitting of sentences in a given document.
- **Part-of-Speech (POS) tagging** takes a list of tokens and analyses them, returning tags for each. A tag represents the grammatical function of the token in the sentence (for example if it is a noun, verb or adjective).
- **Named entity extraction** identifies the proper nouns within a document, such as people, dates or locations, as well as other categories of noun.
- **Chunking** takes a list of tags from a POS tagger and groups sets of tokens by their function in a sentence. Examples can include noun phrases and verb phrases.
- **Parsing** takes a sentence and develops a tree showing the functions of each section of the sentence.
- **Coreference Resolution** identifies noun phrases within a document that are the same. This can commonly be used for replacing pronouns with the original noun phrase.

4 Design

4.1 Front End Architecture Diagram

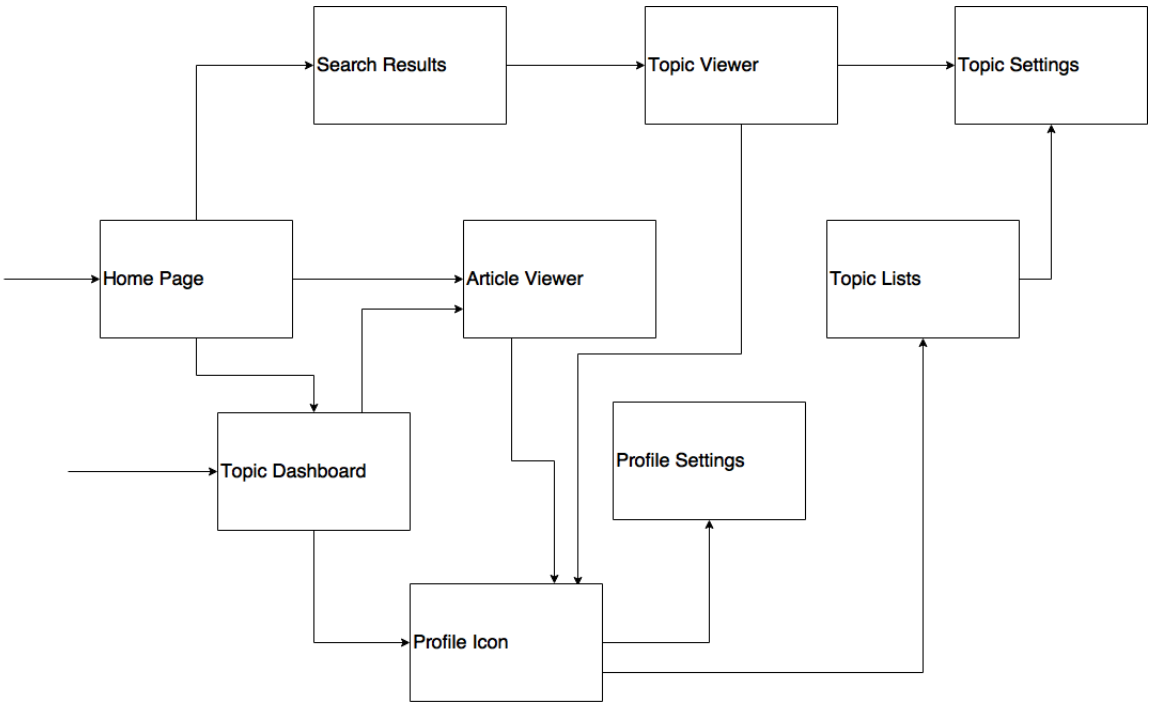


Figure 4.1: *This diagram doesn't show all anticipated connections between each facet of the application's front end. For example, the profile icon is in theory accessible anywhere when the user is logged in.*

4.2 User Story

4.2.1 Home Page

If a user has not logged in, the first page they will see on opening to the website will be the Home Page. The key to the design of the home page is that it's easy to get started for a user. They are presented with a large search box that they can use without having to sign in. There'll also be a row of trending articles displayed below the search bar. This row will consist of icons consisting of images and titles below, as shown in the wireframe. There's a navigation bar at the top, that is present on all screens, with a button to register and a button to sign in on the top right, and a link to information about the application itself on the top left.

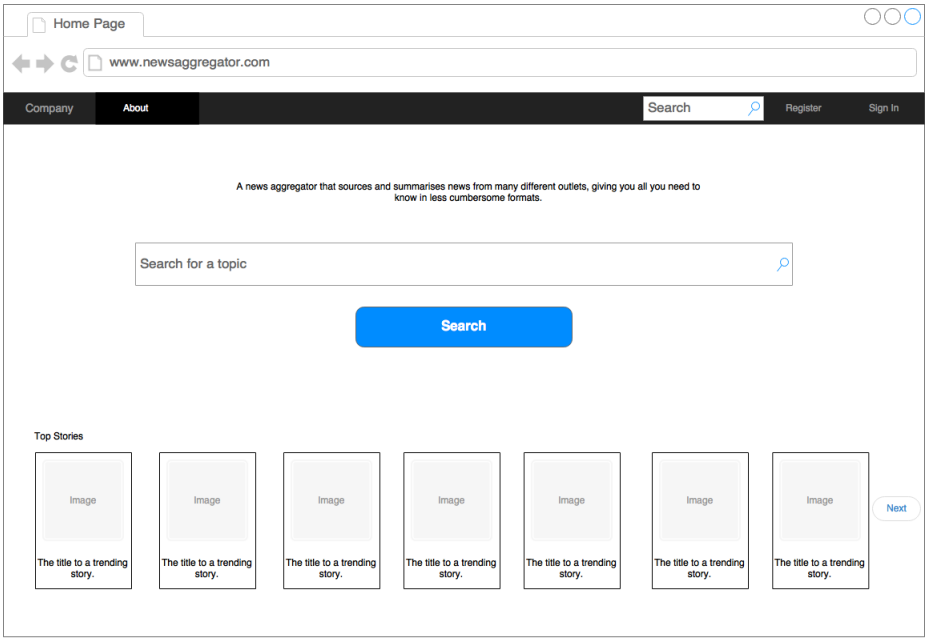


Figure 4.2: *A key focus of the home page is that the user can easily search for articles without having to sign in or register.*

4.2.2 Topic Dashboard

For a user that is logged in, or alternatively a user that has logged in (or signed up) from the home page will first see the topic dashboard screen that shows sets of articles corresponding to each topic they are subscribed. For users who don't have any topics that they are subscribed to, they will instead go back to the home page. The topic dashboard initially shows leading articles from across all the topics they are subscribed to. They can drag to one direction (akin to a sideways swipe on a phone or tablet) to see articles pertaining to the next topic.

Articles themselves are presented as panels. On a panel is an image for the article, with the title overlaid. In the top right hand corner are icons corresponding to the sources that the article has been summarised from. Leading articles (those out of the most recent that have the most sources attached) are placed in the larger panels on the top left and bottom right of the screen, as shown in the diagram.

4.2.3 Search Results Page

The search results page is designed to allow users to see as much as possible and do as much as possible without having to leave the page. The results are presented in

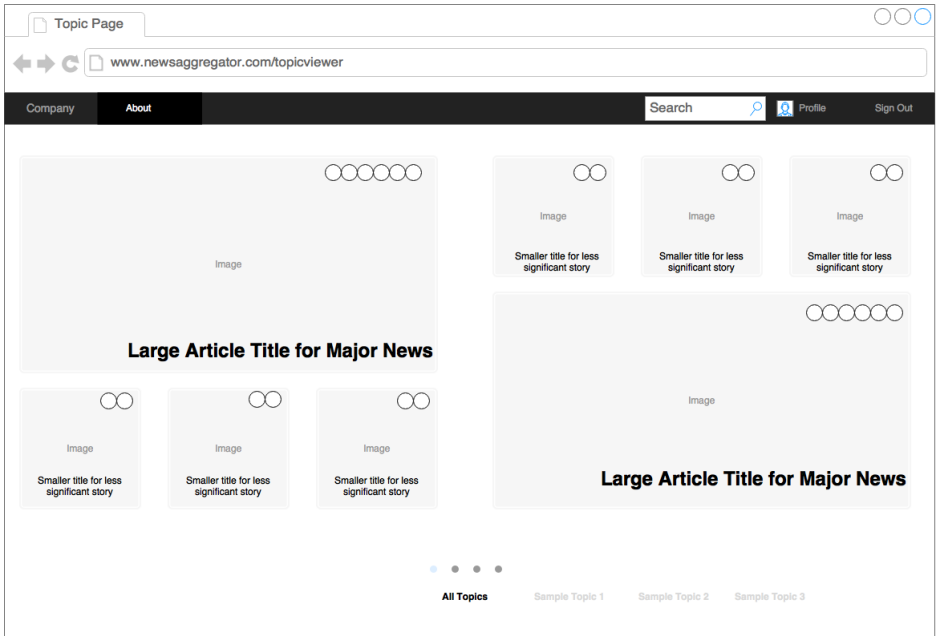


Figure 4.3: *The topic dashboard has the ability to show articles across all topics, or pertaining to a specific topic.*

the application's panel design, with an image corresponding to each panel. Below the panel, is a description of the topic, that is likely to be created from the first paragraph of the Wikipedia[36] article about that particular topic. Below this are two buttons, one to view the topic's articles, which will go to a topic viewer page (shown later), and a subscribe button. If logged in, the user is taken to the topic settings page, where they can choose any particular preferences they have for this specific topic. If the user isn't logged in, they are taken to a page allowing them to log in or register, before being taken to the preference page.

It's necessary at this stage to have a user sign in if they want to subscribe to a topic. This way, they can access their topics from anywhere. In addition to this, the search bar is still prominent on the search results page, so that users can edit their queries easily in case of minor errors.

4.2.4 Topic Viewer

The topic viewer is accessed if a user has clicked on the view button from the search results page. This screen is similar to that of the topic dashboard. Minor changes include the page control at the bottom of the screen being replaced by page statistics, including figures about how often articles are created, the number of subscribers,

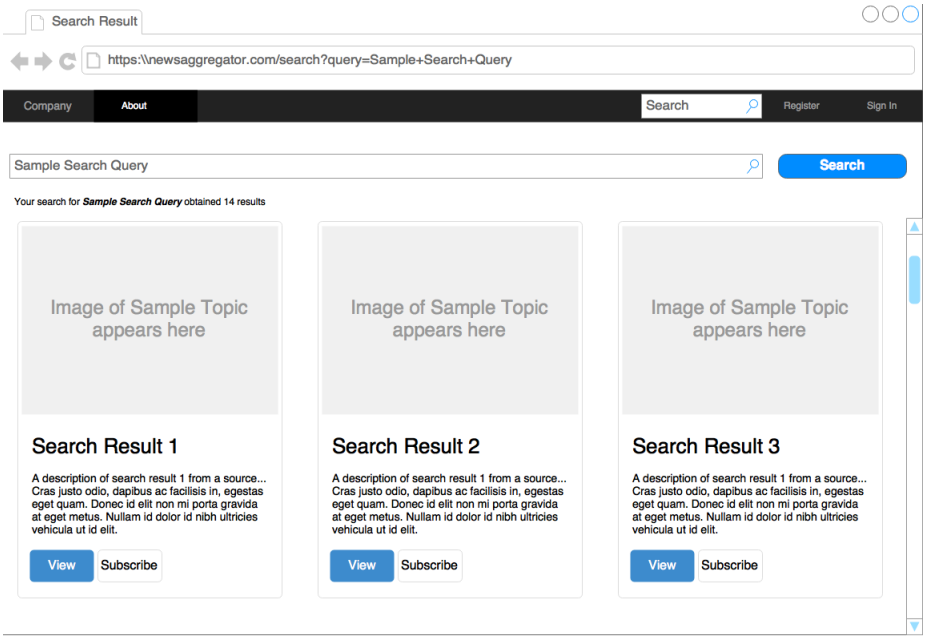


Figure 4.4: *The Search Results page has a focus on showing users as much information as possible about each topic.*

and the total number of hits that the page has had. The other major difference is the presence of a subscribe button in the top right hand side of the screen. As with the search results page, this leads to the topic preferences screen if logged in, and to a page prompting a user to sign in or register when not already logged in.

The remaining aspects of the screen are the same - namely the panelled articles, and the icons in the top right of each panel that indicate which sources the article originated from.

4.2.5 Topic Settings

The aforementioned Topic Settings page sets out two preferences from the user for a specific topic. These are namely:

- **Should the topic's articles be included in the user's email digest?**

  This is answered by a simple on/off switch.

- **Which sources should be used to construct summaries**

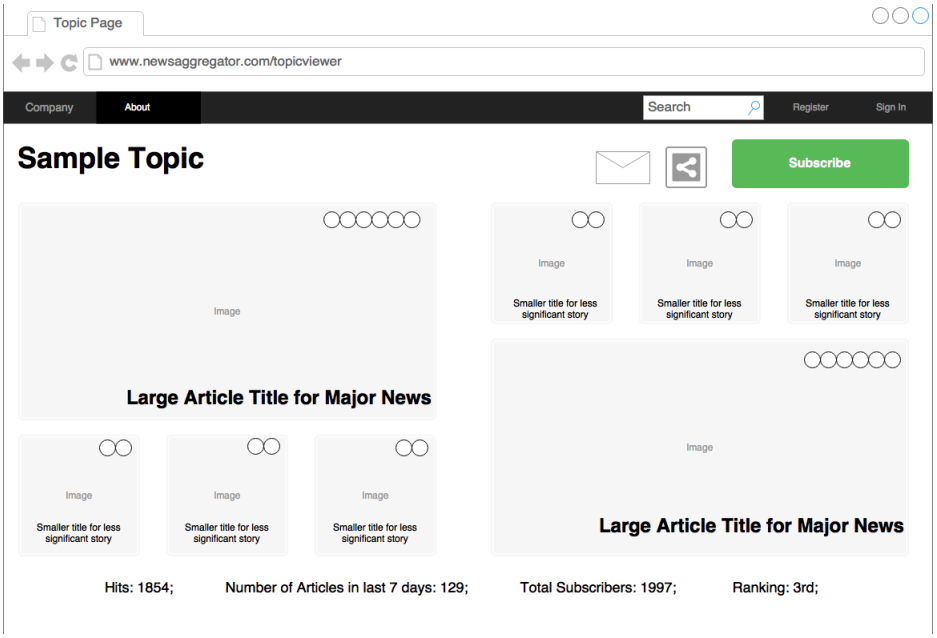


Figure 4.5: *The topic viewer is similar in structure to the topic dashboard, but with the addition of statistics and a subscribe button.*

This aspect is done using the panelling effect used in other screens. Each news outlet's logo forms the basis of the panel, with the title of the outlet below. These panels form buttons that the user can click on to either select or deselect an outlet. The panel will either have a coloured outer rim, or an embedded effect to indicate selection.

4.2.6 Article Viewer

When the user clicks on an article panel in the topic viewer or topic dashboard they are presented with the summarised article itself. This screen is modelled on standard news interfaces, and produces the headline, image and article body on the left hand side of the page. On the right hand side are links to the original articles that the summary was generated from, and links to other articles that are in the same topic.

Like most other screens in the application, there is also a share button on the top right, and an email button.

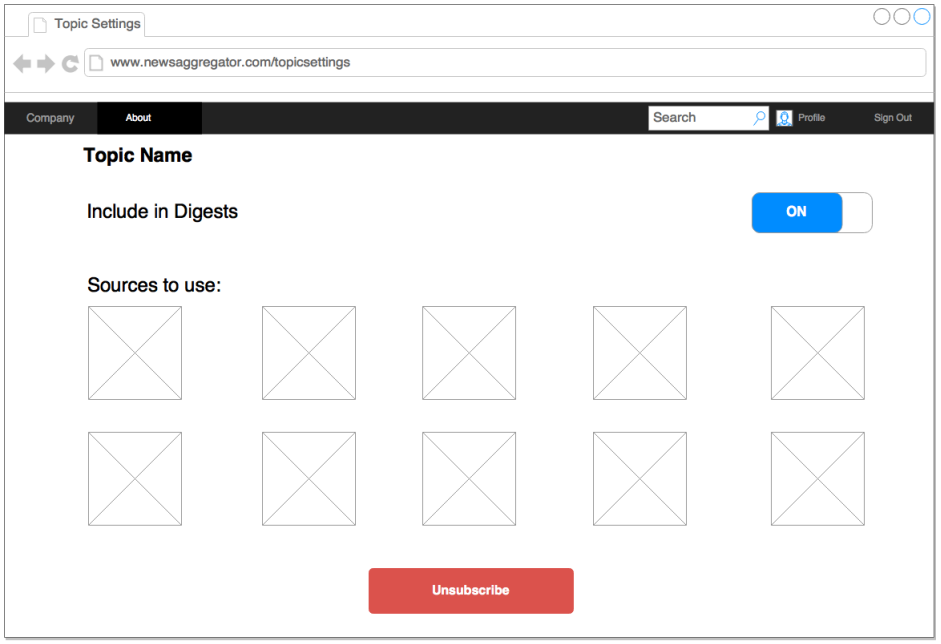


Figure 4.6: *The Topic Settings page allows users to specify key preferences about the topic.*

Figure 4.7: *The Article Viewer is modelled on standard news outlets' interfaces.*

4.2.7 Topic Lists

The Topic List page is accessed by clicking on the profile button when logged in to the application. A panel appears, listing two choices - My Topics, and My Settings. This page is presented on the selection of the My Topics button.

The aim of the topic list is self-explanatory. It lists all topics that a user is subscribed to. Upon clicking on a topic in the list, a user would be taken to the Topic Settings page for that topic.

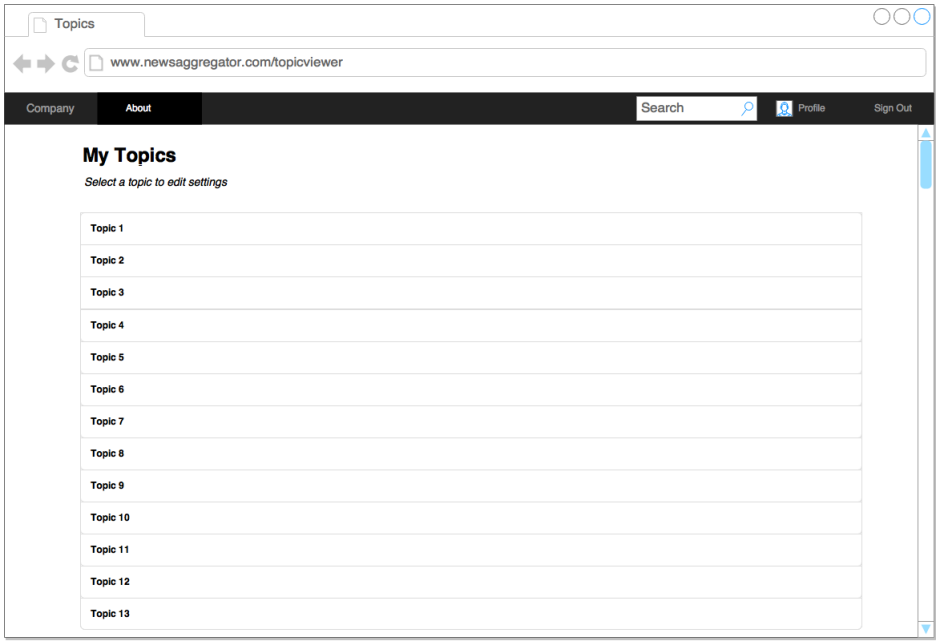


Figure 4.8: *The Topic List page allows users to select and adjust the settings of a topic.*

4.2.8 Profile Settings

The other option upon clicking on the profile button in the navigation bar of the application is to access the Profile Settings page. Here, the user is able to change their profile picture using the button in the top half of the page. Also present on this page are the user's digest settings. Here the user can set whether they want to receive a news digest in the morning, the afternoon, at both times, or not at all. They can also specify the email address that they wish the news digest to be sent to.

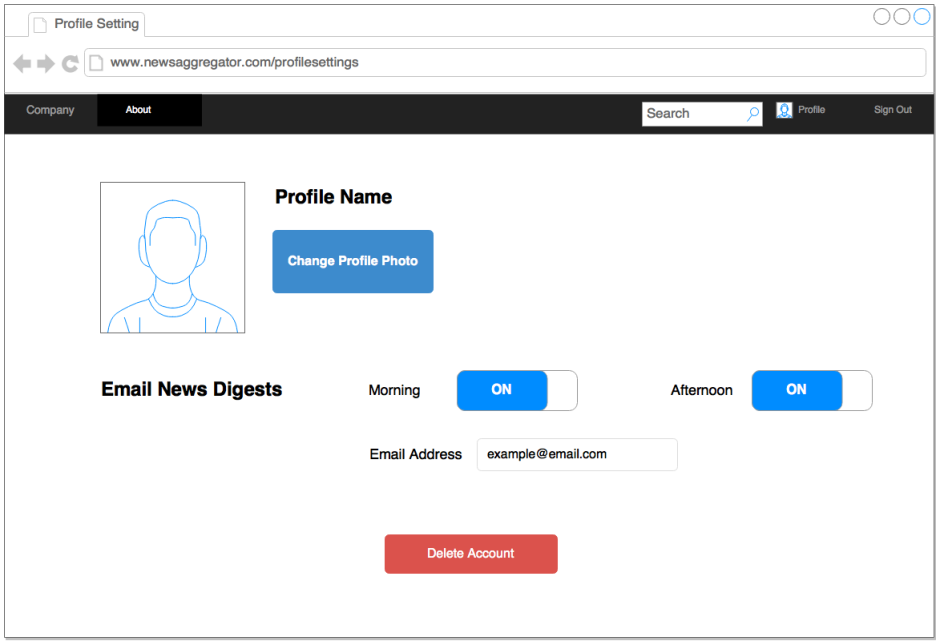


Figure 4.9: *The profile settings page allows the user to adjust their profile picture and digest settings.*

4.3 Back End Flow Diagram

In Figure 4.10 the expected flow of the machine learning aspects of the back end is presented. The process of an article being summarised is as follows:

1. The **Article Fetcher** queries the various News APIs used for new articles.

2. The **Article Curator** takes the response from the News APIs and performs any scraping that may be necessary, before passing the article to the Machine Learning phases.

3. The article is then passed to the **Topic ModellingTopic Modelling** phase, which identifies which topics it consists of.

4. The **Topic Labelling** section takes the results of the Topic Modelling phase, and proceeds to label its topic

5. **Clustering** then occurs. The article, along with its topic label is passed to

this phase, and the program pulls articles from the database that have already been assigned the same topic. These articles are then clustered, and the cluster containing this article is passed to the next phase.

6. Now that the cluster has been passed, **Summarisation** occurs on the article, which is then put in the Summarised Articles database, ready to be called for a user to read.

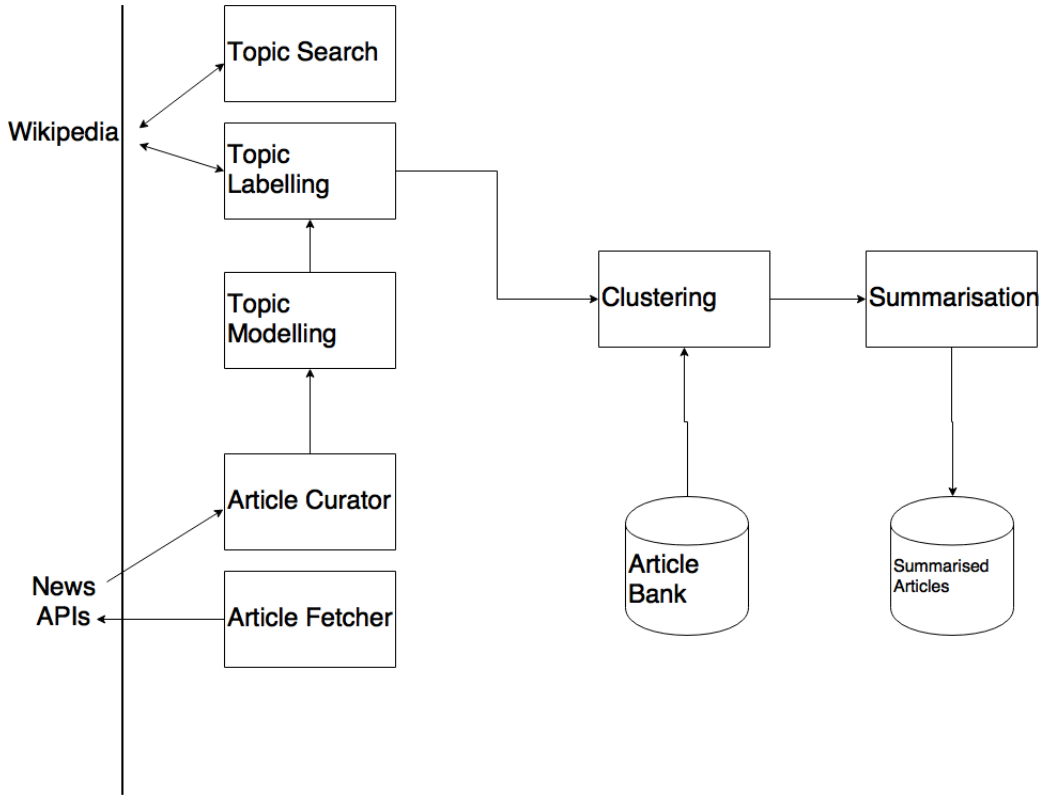


Figure 4.10: *The expected flow of the Back End.*

4.4 Language and Platform Choices

4.4.1 Front End

The survey that was conducted in Section 2.1.2 clearly presented the result that a website would be the preferred option for the application, followed by a mobile

application in either iOS or Android. Therefore I've decided to follow this and create a WebApp as a primary platform. If time permits, I'll then develop the application for a second platform, which will be a mobile platform.

For the WebApp I will primarily use Bootstrap for the Interface. One design framework I used to great effect during my Industrial Placement in Summer 2016 was React and Redux. React is a JavaScript library developed by Facebook that is designed for building scalable and reusable user interfaces. Redux is a design pattern for JavaScript that is designed around a central predictable state container that is the single source of truth for a User Interface.

The benefit of this framework is that it allows for very clean code in the user interface. This is because the concept of a single source of truth allows for data flow that is purely in only one direction, making it easier to understand.

4.4.2 Back End

The back end will be constructed using Java. The reasoning behind this is that I have experience from my Industrial Placement using Java as the basis of a server and a back end.

5 Libraries and APIs

5.1 Potentially Useful APIs

5.1.1 News Outlets

Readership statistics for selected outlets in the United Kingdom

Statistics below are taken from the *Digital News Report 2016*[23] by the Reuters Institute for the Study of Journalism[24]. Viewership is given as a percentage of people who say they use the outlet at least weekly.

Outlet	Viewership	Political Stance
BBC	51%	Neutral
Mail Online	17%	Right
Huffington Post	14%	Left
The Guardian	14%	Left
Sky News Online	11%	Neutral
The Telegraph Online	9%	Right
The Independent	6%	Centrist

Table 5.1: *Readership statistics for selected outlets in the United Kingdom*

The Guardian[28]

The Guardian is a UK newspaper that commands about 14% of the viewership as outlined in Table 5.1. It is also known for having an API that allows a user full and free access to all articles, including the article text itself. The API can access any article dated on or after 1999, and as long as not used for a commercial purpose, is completely free to use.

To obtain an article from *The Guardian* there are two calls that need to be made. First, we make a call to the search endpoint, which returns a JSON object similar to this:

```
{
    "response": {
        "status": "ok",
        "userTier": "free",
        "total": 1,
        "startIndex": 1,
```

```
        "pageSize": 10,
        "currentPage": 1,
        "pages": 1,
        "orderBy": "newest",
        "results": [
            {
                "id":
                  ↪ "politics/blog/2014/feb/17/alex-salmond-speech-fir
                "sectionId": "politics",
                "sectionName": "Politics",
                "webPublicationDate":
                  ↪ "2014-02-17T12:05:47Z",
                "webTitle": "Alex Salmond speech - first
                  ↪ minister hits back over Scottish
                  ↪ independence - live",
                "webUrl":
                  ↪ "https://www.theguardian.com/politics/blog/2014/fe
                "apiUrl":
                  ↪ "https://content.guardianapis.com/politics/blog/20
            }
        ]
    }
}
```


From this, we extract the apiUrl, and call this, with the API key in the request headers. This will return the raw text for the article.

The Financial Times

The Financial Times is an internationally renowned, UK based broadsheet that provides news primarily on politics and the financial sector. Their API is similar to that of *The Guardian*, but with a key difference that could prove very useful. While *The Guardian* allows a user to query the API by date, *The Financial Times* does better. It has a notification endpoint, which allows a user to query with a specific date and time. They are then returned a list of 'notifications', showing articles that have been created, updated or deleted since the time provided, as shown in the example below.


```
{
  "requestUrl":
    ↪ "http://api.ft.com/content/notifications?since=2014-05-06T10:00:
```

```
  "notifications": [
    {
      "type":
        ↪ "http://www.ft.com/thing/ThingChangeType/UPDATE",
      "id":
        ↪ "http://www.ft.com/thing/7348edd8-0403-11de-845b-000077b0765
      "apiUrl":
        ↪ "http://api.ft.com/content/7348edd8-0403-11de-845b-000077b07
    }
  ],
  "links": [
    {
      "href":
        ↪ "http://api.ft.com/content/notifications?since=2014-05-06T18
      "rel": "next"
    }
  ]
}
```

Using this, the notifications of type UPDATE are kept, and we query for the raw data using the apiUrl in a similar fashion to before in the example for *The Guardian*.

News API

A major issue is that most newspapers (for example, *The New York Times*[29]) only give the first paragraph to an article in their API, and don't give out their full articles as standard. There are solutions, such as Webhose[32], that offer feeds into this outlets, but at either a cost, or with a very limited number of requests per month.

News API[21] is a company that provides an API to get headlines for over seventy different news outlets. A call returns data in the JSON format as follows:

```
{
"status": "ok",
"source": "the-next-web",
"sortBy": "latest",
"articles": [
{
        "author": "Abhimanyu Ghoshal",
        "title": "Are these ridiculous headphones the way
```

```
            ↪ forward for music tech?",
9           "description": "All the amazing tech we now have
              ↪ at our disposal for enjoying music is
              ↪ closing us off from other people instead of
              ↪ bringing us together. Is there hope yet?",
10          "url":
              ↪ "https://thenextweb.com/gear/2017/01/26/can-music-tech-mak
11          "urlToImage":
              ↪ "https://cdn3.tnwcdn.com/wp-content/blogs.dir/1/files/2017
12          "publishedAt": "2017-01-26T13:12:10Z"
13  }
14  ]
15  }
```

We can use the url provided for each of these results to get to the webpage of the corresponding article. However at this stage I'll need to create a scraper to extract the raw content of the article. There could be copyright implications for this step, even though the final production summary won't look the same as the raw content. These implications will be fully examined in the final report.

News API counts amongst its 70 sources the following:

- BBC News
- Mail Online
- *The Telegraph*
- *The New York Times*
- Associated Press
- *The Independent*
- *The Daily Mirror*

News API also has *The Financial Times* and *The Guardian*, although these are left off the list above as they have their own fully accessible APIs.

5.1.2 Wikipedia

Wikimedia (the parent company for Wikipedia[36]) provides documentation for the Wikipedia API on its website MediaWiki website. Here, I can search using the search term that a user has used (perhaps in searching for a topic), and get a result that is as follows:

```
{
    "query": {
        "searchinfo": {
            "totalhits": 4152
        },
        "search": [
            {
                "ns": 0,
                "title": "Albert Einstein",
                "snippet": "&quot;<span
                    ↪ class=\"searchmatch\">Einstein</span>&quot;
                    ↪ redirects here. For other uses, see
                    ↪ <span
                    ↪ class=\"searchmatch\">Albert</span>
                    ↪ <span
                    ↪ class=\"searchmatch\">Einstein</span>
                    ↪ (disambiguation) and <span
                    ↪ class=\"searchmatch\">Einstein</span>
                    ↪ (disambiguation). <span
                    ↪ class=\"searchmatch\">Albert</span>
                    ↪ <span
                    ↪ class=\"searchmatch\">Einstein</span>
                    ↪ (/?alb?rt ?a?n?ta?n/; German:",
                "size": 124479,
                "wordcount": 13398,
                "timestamp": "2015-05-10T12:37:14Z"
            },
            ...
}
```

The search query can be altered to provide different information, such as the URL for the lead image, which can be used to show previews of various items to users. The Wikipedia API can also be used in the topic labelling aspect of the Machine Learning techniques (see Section 3.1.2).

5.2 Libraries

5.2.1 Mallet

Mallet (Machine Learning for Language Toolkit) is a java library that provides an interface for various natural language-based machine learning tasks. These tasks include:

- Document Classification
- Sequence Tagging
- Clustering
- Topic Modeling
- Information Extraction

Mallet in Topic Modeling

Mallet is particularly useful for Topic Modeling. Mallet uses a corpus in the form of a list of strings in order to train a set of topics. The library trains the topics using Gibbs Sampling and Latent Dirichlet Allocation.

Mallet also provides an inferencer that allows a user to submit a new document and receive a list of probabilities. Each of these probabilities shows the likelihood that the corresponding topic is an accurate classification of the document we are testing. Thus the user can then infer the actual topics that the document is about by simply taking the topics with the highest probabilities.

5.2.2 Natural Language Processing

There are several Natural Language Processing libraries available. The following are designed for use in Java:

Apache OpenNLP

OpenNLP is an open source library developed by Apache. The aim of the project is to support the basic aspects of natural language processing.

OpenNLP supports all of the basic tasks that were specified in section 3.3.1 with varying degrees of success, plus ‘document categorisation’.

Document Categorisation, as the name suggests, takes a document and categorises it. This could potentially be used in the topic modelling phase, but has the disadvantage that it doesn't come with a model for training, and the user of the library has to create their own. This is a contrast to the Mallet library, which can create a model when the user passes in the raw data.

When it comes to the other key tasks of Natural Language Processing, OpenNLP provides flexibility on models. They provide a default model in a variety of languages, including English. However, there is at least one different model file for each task, so space constraints need to be considered.

Name Finder

For using the name finder, there are different model files, depending on what types of proper nouns a user is looking for. These are:

- Date
- Location
- Money
- Organisation
- Percentage
- Person
- Time

Space Constraint Table

Table 5.2 shows each model and their respective sizes. As Apache's recommended method to bring the files in to the program is through the Java class FileInputStream their sizes can be important, as they make use of the Java heap.

Model Name	Task	Size (MB)
en-chunker	Chunking	2.6
en-ner-date	Name Finder: Dates	5.0
en-ner-location	Name Finder: Locations	5.1
en-ner-money	Name Finder: Money	4.8
en-ner-organization	Name Finder: Organisations	5.3
en-ner-percentage	Name Finder: Percentages	4.7
en-ner-person	Name Finder: People	5.2
en-ner-time	Name Finder: Time	4.7
en-pos-maxent	POS Tagging	5.7
en-sent	Sentence Detection	0.01
en-token	Tokenisation	0.44
Total if all used		**43.55**

Table 5.2: *A table showing the space required for the various models provided for use with OpenNLP*

Stanford CoreNLP

The Stanford CoreNLP is a toolkit provided for Java by the Stanford Natural Language Processing Group. Like Apache OpenNLP, CoreNLP also offers the same functionalities that were discussed in section 3.3.1. However, Stanford go above and beyond this list in what they can provide. Their coreference and name recognition solutions are 'competition winning', whilst they also provide functionality for semantic analysis of text.

As a result, CoreNLP could have a major benefit for abstractive summarisation, through its coreference resolution.

Extracting elements from a document

The following is a code snippet, taken from the Stanford CoreNLP documentation that demonstrates how to perform certain Natural Language tasks from a given document.

```
1 {
2 // these are all the sentences in this document
3 // a CoreMap is essentially a Map that uses class objects
    ↪ as keys and has values with custom types
4 List<CoreMap> sentences =
    ↪ document.get(SentencesAnnotation.class);
5
```

```
for(CoreMap sentence: sentences) {
  // traversing the words in the current sentence
  // a CoreLabel is a CoreMap with additional token-specific methods
  for (CoreLabel token: sentence.get(TokensAnnotation.class)) {
    // this is the text of the token
    String word = token.get(TextAnnotation.class);
    // this is the POS tag of the token
    String pos = token.get(PartOfSpeechAnnotation.class);
    // this is the NER label of the token
    String ne = token.get(NamedEntityTagAnnotation.class);
  }

  // this is the parse tree of the current sentence
  Tree tree = sentence.get(TreeAnnotation.class);
}

// This is the coreference link graph
// Each chain stores a set of mentions that link to each other,
// along with a method for getting the most representative mention
// Both sentence and token offsets start at 1!
Map<Integer, CorefChain> graph =
  document.get(CorefChainAnnotation.class);
```

Wordnet

6 Implementation

6.1 Database

6.2 Restlet

6.3 Server Tasks

6.4 Front End

6.5 Key Classes

7 Optimisation

7.1 Speed Optimisations

7.1.1 Topic Modelling

7.1.2 Topic Labelling

7.1.3 Clustering

7.2 Memory Optimisations

8 Evaluation

8.1 Machine Learning and Summarisation

8.2 Summary Analysis

8.3 User Interface Evaluation

9 Conclusion

10 Future Work

10.1 Foreign Languages

10.2 Further Optimisations

10.3 Other Apps

References

[1] *Apple*. www.apple.com.

[2] *Apple News*. www.apple.com/uk/news/.

[3] *BBC*. www.bbc.co.uk.

[4] *Cluster Analysis*. URL: https://en.wikipedia.org/wiki/Cluster_analysis (Retrieved Dec. 29, 2016).

[5] Gunes Erkan and Dragomir R Radev. 'LexRank: Graph-based Lexical Centrality as Salience in Text Summarization'. In: *Journal of Artificial Intelligence Research* (December 2004).

[6] *Flipboard*. www.flipboard.com.

[7] Pierre-Etienne Genest and Guy Lapalme. 'Framework for Abstractive Summarization using Text-to-Text Generation'. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (June 2011).

[8] Pierre-Etienne Genest and Guy Lapalme. 'Fully Abstractive Approach to Guided Summarization'. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (July 2012).

[9] *Google*. www.google.com.

[10] *Google Forms*. www.forms.google.com.

[11] *Google News*. www.news.google.com.

[12] Karl Grieser et al. 'Using Ontological and Document Similarity to Estimate Museum Exhibit Relatedness'. In: *ACM Journal ACM Journal of Computing and Cultural Heritage* (2011).

[13] Vishal Gupta and Gurpreet Lehla. 'A Survey of Text Summarization Extractive Techniques'. In: *Journal of Emerging Technologies in Web Intelligence* (2010).

[14] *Hierarchical Clustering*. URL: https://en.wikipedia.org/wiki/Hierarchical_clustering (Retrieved Dec. 29, 2016).

[15] *How does Google News cluster stories?* URL: https://www.quora.com/How-does-Google-News-cluster-stories/answer/Bharath-Kumar-M?srid=Qord (Retrieved Jan. 8, 2017).

[16] Statistic Brain Research Institute. *Attention Span Statistics*. URL: http://www.statisticbrain.com/attention-span-statistics/.

[17] N. R. Kasture et al. 'A Survey on Methods of Abstractive Text Summarization'. In: *International Journal for Research in Emerging Science and Technology* (November 2014).

[18] Atif Khan and Naomie Salim. 'A review on abstractive summarization methods'. In: *Journal of Theoretical and Applied Information Technology* (January 2014).

[19] *Latent Semantic Analysis*. URL: https://en.wikipedia.org/wiki/Latent_semantic_analysis#Latent_semantic_indexing (Retrieved Dec. 22, 2016).

[20] Jey Han Lau et al. 'Automatic Labelling of Topic Models'. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (June 2011).

[21] *News API*. URL: www.newsapi.org (Retrieved Jan. 27, 2017).

[22] *Ofcom*. www.ofcom.org.uk.

[23] *Reuters Institute Digital News Report 2016*. Reuters Institute for the Study of Journalism, 2016.

[24] *Reuters Institute for the Study of Journalism*. www.reutersinstitute.politics.ox.ac.uk.

[25] *Statistic Brain Research Institute*. www.statisticbrain.com.

[26] *Summly*. www.summly.com.

[27] *Term Frequency - Inverse Document Frequency*. URL: https://en.wikipedia.org/wiki/Tf%E2%80%93idf#Example_of_tf.E2.80.93idf (Retrieved Dec. 29, 2016).

[28] *The Guardian*. www.theguardian.com.

[29] *The New York Times*. www.nytimes.com.

[30] *University of Canberra*. www.canberra.edu.au.

[31] *University of Oxford*. www.oxford.ac.uk.

[32] *Webhose*. URL: https://webhose.io (Retrieved Jan. 27, 2017).

[33] Harald Weinreich et al. 'Not Quite the Average: An Empirical Study of Web Use'. In: *ACM Transactions on the Web* (February 2008).

[34] *What is a good explanation of Latent Dirichlet Allocation?* URL: https://www.quora.com/What-is-a-good-explanation-of-Latent-Dirichlet-Allocation/answer/Edwin-Chen-1?srid=Qord (Retrieved Jan. 8, 2017).

[35] *What's the difference between Latent Semantic Indexing and Latent Dirichlet Allocation*. URL: https://www.quora.com/Whats-the-difference-between-Latent-Semantic-Indexing-LSI-and-Latent-Dirichlet-Allocation-LDA/answer/Joseph-Turian?srid=Qord (Retrieved Dec. 22, 2016).

[36] *Wikipedia*. www.wikipedia.org.

[37] *WWDC Apple Design Award Winners for 2014*. URL: http://www.macworld.com/article/2358481/wwdc-apple-design-awards-winners-for-2014.html (Retrieved Dec. 29, 2016).

[38] *Yahoo News Digest*. www.uk.mobile.yahoo.com/newsdigest/.

[39] *YouGov*. www.yougov.com.

Appendices

A Source Code

B API

C User Guide

Index