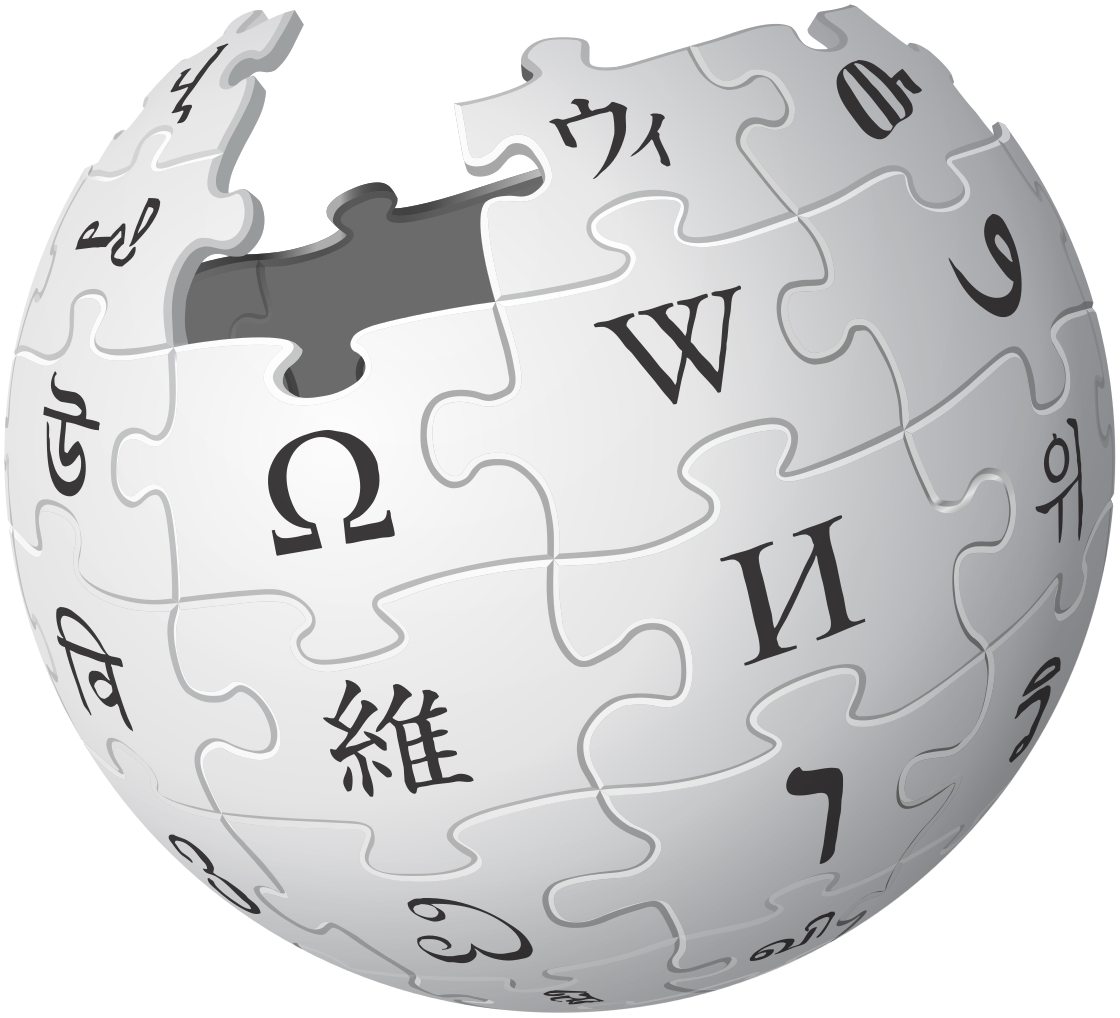
Title: WikiClassify™
Group Number: 11
Group Members: Nathan Kjer, Luke Wielgus, Adam Massoud, Brian Faure, Wanshu (Wayne) Sun, Brian Chu



First Report (Part 1)

Project Website: https://github.com/nathankjer/wikiclassify

Responsibility Matrix

	Adam	Nathan	Wayne	Brian F.	Luke	Brian C.	Viswanathan
Project Management (10 pts)							
Sec. 1: Customer Statement of Requirements (9 pts)							
Sec. 2: System Requirements (6 pts)							
Sec. 3: Functional Requirements (30 pts)							
Sec. 4: User Interface Specs (15 pts)							
Sec. 5: Domain Analysis (25 pts)							
Sec. 6: Plan of Work (5 pts)							
Total Points:							

1. Customer Statement of Requirements (CSR):

1.1 Problem Statement

Wikipedia is a crowdsourced online encyclopedia founded in 2001. With the goal of centralizing all of human knowledge, it attempts to be a neutral, global, and uncensored source of free information. As of 2016, Wikipedia is ranked number seven on the world's most popular websites, holds a top priority on search engines, and has been estimated[2] to be worth hundreds of billions of dollars. Furthermore, articles on Wikipedia have been cited[3] in hundreds of court cases. However, the most notable problem with Wikipedia is the quality of information in articles. Reliable or not, Wikipedia is one of the most heavily used online encyclopedias as a result of its high-quality content across a broad range of subjects.

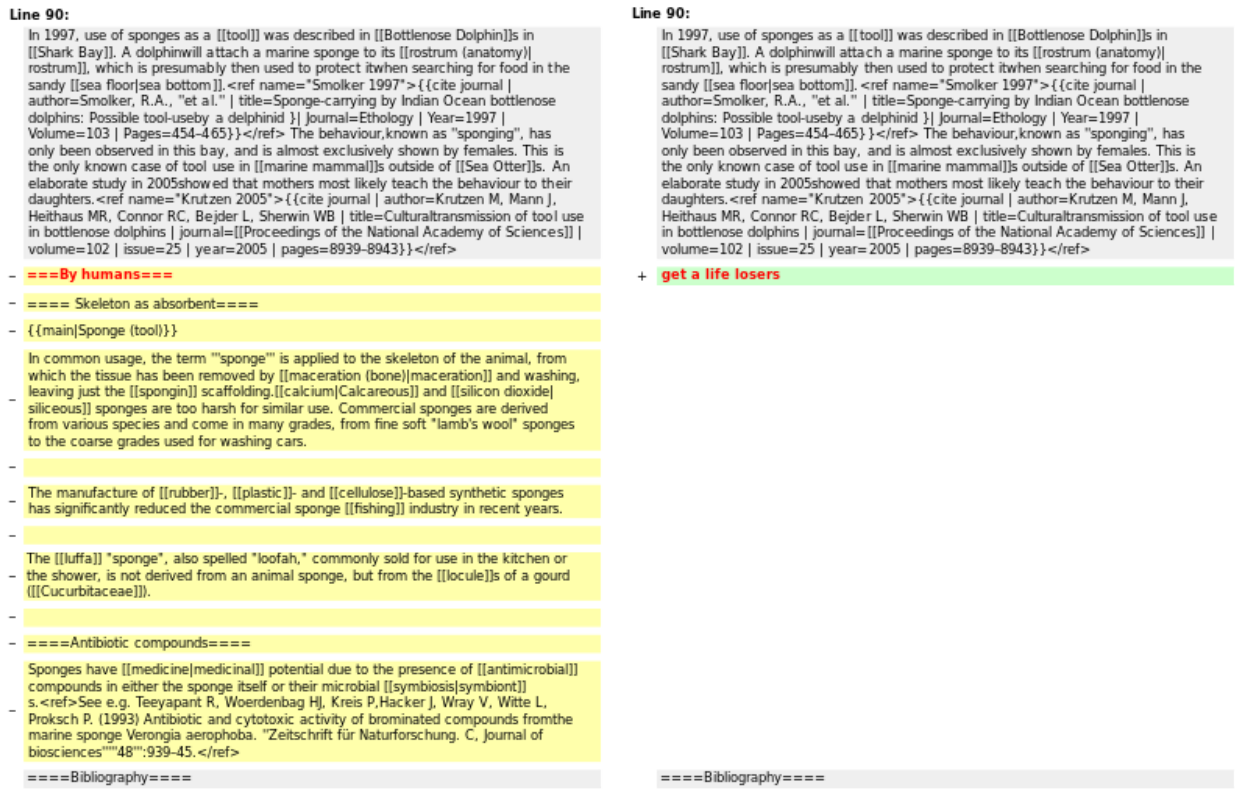
Anyone can edit Wikipedia, which is the primary reason why it is one of the largest online encyclopedias in the world. This results in almost an "infinite" source of information, and is widely used for schoolwork, research, and by anyone wishing to learn more about a certain subject or object. However, this has numerous tradeoffs. As anyone can edit articles, one can put misinformation or delete legitimate information on new or existing articles. In some cases, certain editors may not be as knowledgeable about the subject they are making edits about on the corresponding Wikipedia article. This will also result in misinformation, regardless of whether or not it was intentional. Destructive editing is a common problem that can go unnoticed for extended periods of time. There is even a list[4] that Wikipedia has created of bad article ideas, which is a result of destructive editing. Obviously, this decreases and sometimes destroys the condition and legitimacy of articles on Wikipedia. In addition, articles are frequently cited with unreliable sources, or not cited at all. This results in articles that may or may not be accurate and will negatively affect the credibility of articles and Wikipedia as a whole. Furthermore, Wikipedia has editor bias; 90% of Wikipedians are male, 40% of the world's population has access to the internet (Wikipedia can only be edited from a selection of 40% of the world's population), editors are often from white collar backgrounds, have stronger technical abilities, are more frequently from the Northern Hemisphere, and common languages and Western culture are more heavily represented. With such an editor bias, there is a larger probability of the content being misrepresented or not being of acceptable quality, as certain editors will input information that they feel is correct, but may only be a product of how opinionated they are about the subject they are editing about in the article.

Thus, although Wikipedia is one of the largely used online encyclopedias, despite having a substantial source of accurate articles, many articles are not of satisfactory quality due to the reasons stated above. Therefore, users of Wikipedia come across the problem of whether or not the article they are viewing has correct information. Being that Wikipedia is designed to provide users with information, any hint of inaccuracies or slight bias would be completely contradictory to the initial website intentions.

In order for users to combat this, software must be developed that can analyze and rate the quality of Wikipedia articles. This way, whenever a user views a Wikipedia

article, he or she will be able to tell if the information being viewed is of adequate quality or not. For the sake of convenience, having an extension or application created to implement this would be an ideal solution. From now on, we will refer to this software as WikiClassify. The idea is that whenever a user is viewing a wikipedia article, WikiClassify will activate. WikiClassify will analyze the content of the article and rate the quality of it, giving the user a good idea of whether or not he or she is looking at an article with quality information.

Wikipedia is a popular and useful online encyclopedia, but faces the issue of articles having information of poor quality. This is primarily due to destructive editing, cited sources that are not reputable, and the bias of the editors. In order to combat this, software will be created that will rate the quality of wikipedia articles, giving users a better idea of what they are looking at. The ultimate goal is for this software to be used not only for the benefit of Wikipedia users, but for the entire online encyclopedia to be "filtered" to the point that eventually all articles will be of satisfactory quality.



Line 90:

In 1997, use of sponges as a [[tool]] was described in [[Bottlenose Dolphin]]s in [[Shark Bay]]. A dolphinwill attach a marine sponge to its [[rostrum (anatomy)|rostrum]], which is presumably then used to protect itwhen searching for food in the sandy [[sea floor|sea bottom]].<ref name="Smolker 1997">{{cite journal | author=Smolker, R.A., "et al." | title=Sponge-carrying by Indian Ocean bottlenose dolphins: Possible tool-useby a delphinid }| Journal=Ethology | Year=1997 | Volume=103 | Pages=454–465}}</ref> The behaviour,known as "sponging", has only been observed in this bay, and is almost exclusively shown by females. This is the only known case of tool use in [[marine mammal]]s outside of [[Sea Otter]]s. An elaborate study in 2005showed that mothers most likely teach the behaviour to their daughters.<ref name="Krutzen 2005">{{cite journal | author=Krutzen M, Mann J, Heithaus MR, Connor RC, Bejder L, Sherwin WB | title=Culturaltransmission of tool use in bottlenose dolphins | journal=[[Proceedings of the National Academy of Sciences]] | volume=102 | issue=25 | year=2005 | pages=8939–8943}}</ref>

\- ===By humans===

\- ==== Skeleton as absorbent====

\- {{main|Sponge (tool)}}

\- In common usage, the term ""sponge"" is applied to the skeleton of the animal, from which the tissue has been removed by [[maceration (bone)|maceration]] and washing, leaving just the [[spongin]] scaffolding.[[calcium|Calcareous]] and [[silicon dioxide|siliceous]] sponges are too harsh for similar use. Commercial sponges are derived from various species and come in many grades, from fine soft "lamb's wool" sponges to the coarse grades used for washing cars.

\-

\- The manufacture of [[rubber]]-, [[plastic]]- and [[cellulose]]-based synthetic sponges has significantly reduced the commercial sponge [[fishing]] industry in recent years.

\-

\- The [[luffa]] "sponge", also spelled "loofah," commonly sold for use in the kitchen or the shower, is not derived from an animal sponge, but from the [[locule]]s of a gourd ([[Cucurbitaceae]]).

\-

\- ====Antibiotic compounds====

\- Sponges have [[medicine|medicinal]] potential due to the presence of [[antimicrobial]] compounds in either the sponge itself or their microbial [[symbiosis|symbiont]] s.<ref>See e.g. Teeyapant R, Woerdenbag HJ, Kreis P,Hacker J, Wray V, Witte L, Proksch P. (1993) Antibiotic and cytotoxic activity of brominated compounds fromthe marine sponge Verongia aerophoba. "Zeitschrift für Naturforschung. C, Journal of biosciences""48"":939-45.</ref>

====Bibliography====

Line 90:

In 1997, use of sponges as a [[tool]] was described in [[Bottlenose Dolphin]]s in [[Shark Bay]]. A dolphinwill attach a marine sponge to its [[rostrum (anatomy)|rostrum]], which is presumably then used to protect itwhen searching for food in the sandy [[sea floor|sea bottom]].<ref name="Smolker 1997">{{cite journal | author=Smolker, R.A., "et al." | title=Sponge-carrying by Indian Ocean bottlenose dolphins: Possible tool-useby a delphinid }| Journal=Ethology | Year=1997 | Volume=103 | Pages=454–465}}</ref> The behaviour,known as "sponging", has only been observed in this bay, and is almost exclusively shown by females. This is the only known case of tool use in [[marine mammal]]s outside of [[Sea Otter]]s. An elaborate study in 2005showed that mothers most likely teach the behaviour to their daughters.<ref name="Krutzen 2005">{{cite journal | author=Krutzen M, Mann J, Heithaus MR, Connor RC, Bejder L, Sherwin WB | title=Culturaltransmission of tool use in bottlenose dolphins | journal=[[Proceedings of the National Academy of Sciences]] | volume=102 | issue=25 | year=2005 | pages=8939–8943}}</ref>

\+ get a life losers

====Bibliography====

→ Comparison of a satisfactory Wikipedia article to an example of destructive editing
→ https://en.wikipedia.org/wiki/Reliability_of_Wikipedia[5]

1.2 Solution

At Wikipedia we utilize a rating system for our articles. This system allows Wikipedia users to automatically identify whether or not the article they are reading is verified by the Wikipedia community. The problem with this system is that it is extremely time-intensive to maintain and the process to apply for a better rating is long and tedious. Due to its inherent issues, we are seeking an automated approach to label and maintain labels for our entire library of articles.

Specifically we are looking for an algorithm which can use machine learning to implement the the assignment of featured, good, or null to every article following the criteria below:

https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria[6]
https://en.wikipedia.org/wiki/Wikipedia:Good_article_criteria[7]

By analyzing the articles with the good and featured label, in tandem with the criterion above, we hope that you will be able to reverse engineer a 'formula' to label more articles accurately in the future.

Along with the rating system, Wikipedia also classifies articles into different subgroups. Classifications includes the following categories:

Maintenance category (Class)	Occurrence in text	Number of Articles	Notes
Stub Articles	stub}}	1954458	Article is very short and is not developed.
Cleanup	{{Cleanup	20678	Article requires general cleanup.
Advert	{{Advert	16462	Article is written like an advertisement.
Update	{{Update	13661	Article is outdated.
Tone	{{Tone	8307	Article does not match the tone of Wikipedia.
Featured Articles	{{Featured article}}	4659	Article is among the best on Wikipedia.
Plot	{{Plot	4090	Article is long/excessively detailed.
Essay	{{Essay-like	3716	Article is written like an opinion essay.
Peacock	{{Peacock	3555	Article overly promotes its subject.
Technical	{{Technical	3042	Article is overly technical.
Confusing	{{Confusing	2336	Article is confusing.
Overly Detailed	{{Overly Detailed	1724	Article is overly detailed.

→ Source: https://en.wikipedia.org/wiki/Wikipedia:Maintenance[8]

Having a system that is able to automatically classify every article and place them under a subgroup is very desirable. This type of system will help us target which articles need to be improved and the type of improvement that is needed. We believe the optimal solution would be either a chrome extension or a website, either of which would place the power of the machine learning process in the hands of the user. The specifics of the platform (precompiled database vs. real time algorithm etc.) are completely up to you, the only stipulation being that the interface should be quick and simple.

1.3 Summary

This project aims to rate and classify every article on Wikipedia using the known examples that have been assigned by our team. The features of the system should include the following:

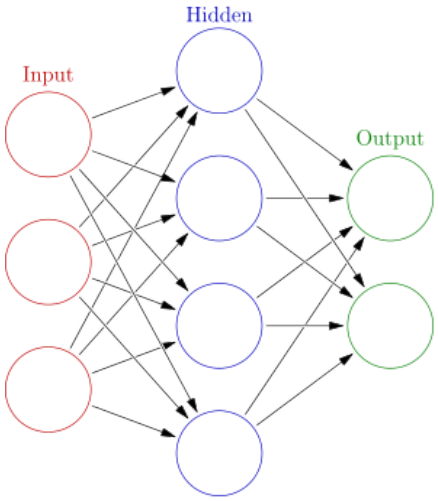
1. Input text and receive output parameters that classify the text based on our machine learning algorithm.
2. Uncover biases hidden in Wikipedia rating schema.
3. Compare accuracy of multiple input text articles.
4. Create procedurally generated articles on the fly.
5. Check text for historical accuracy/bias.
6. Compare different portions of Wikipedia based on article quality.

1.4. Glossary of Terms

Activation Function: The function that a neural network uses within its nodes.

Article: A single Wikipedia entry. Similar in structure and length to an article found in an encyclopedia.

Artificial Neural Network (ANN): Artificial neural networks, often simplified as “neural networks”, are a class of models used in machine learning. They are inspired by the mechanisms used by neurons in the brain.



An illustration of weight correlations within an ANN
Credit: https://commons.wikimedia.org/wiki/File:Colored_neural_network.svg

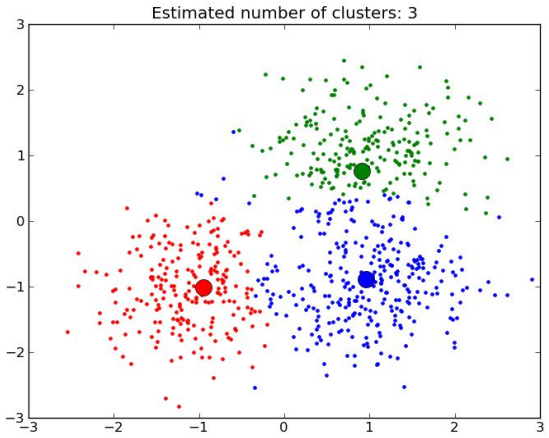
Backpropagation: A learning method that iteratively finds the gradient of the error to estimate a better solution of weights.

Bias: An underlying feature of text that skews it's accuracy according to the beliefs of the writer. An article with strong bias may present only partial facts and opinions in an attempt to sway the readers ideology towards that of the writer. An article without much bias will present information without undue weight to any side. This allows the reader to form their own opinions independent of the writer. Preventing article bias is extremely important to the Wikipedia community, as well as news sources and other credible sources of information.

Classification: The act of placing a selected sequence a category.

Class: A category. This is distinctly different than an OOP class.

Clustering: The relative grouping of article criteria pulled from the same classification, arises when the ML algorithm has found some relationship between the data that links articles of the same classifier. In a broad sense, clustering is the task of grouping objects in such a way that objects of the same cluster share more attributes than objects of differing clusters.

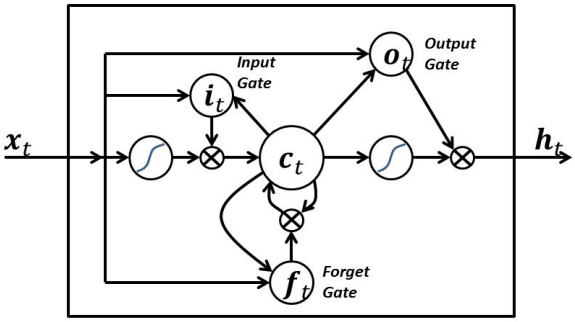


Clustering in two dimensions, with 3 classes
Credit: http://scikit-learn.sourceforge.net/0.5/modules/clustering.html

Error: The magnitude of difference between a model and its desired output.

Featured Article: A label given to certain Wikipedia articles. See the following link for more information: https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria[6]

Long short-term memory (LSTM): An artificial neural network architecture that is used to interpret and predict temporal data. It consists of a "memory cell" that contains a series of gates to try to solve the problem of vanishing weights.



Credit: https://commons.wikimedia.org/wiki/File:Long_Short_Term_Memory.png

Machine Learning: A programming technique wherein you provide the inputs and outputs and the algorithm does the grunt-work of finding a reliable relationship (or path) between the two. Effective if there is a large quantity of data pertaining to inputs and outputs. In our case, the inputs are the Wikipedia articles and the outputs are the article classifications. After an algorithm has been trained on a substantial quantity of input/output data, it can be used to predict outputs given only the inputs.

Noise: Data that only serves to hinder the progress of the algorithm. Noise pertains to, among other things, the html artifacts parsed out in the initial stages of the program.

Parse: The act of filtering and removing certain parts of something. In our case, we are sifting through the content of our data dumps and removing the html artifacts (formatting cues etc.)

Patterns: Relationships between certain pieces of data which serve to either substantiate, or invalidate prior assumptions. In our case, the ML algorithm finds underlying patterns in the articles which it can use as evidence that the article it is currently observing fits in some category.

Preprocessed Data: Classified text sequences (strings) are truncated into fixed lengths and converted into a numeric matrix format character-wise and preprocessed.

Raw Data: Pertains to the data directly after being downloaded and before passing through the parsing stage. At this point the data may contain junk formatting artifacts that must be removed.

Sampling: The process of extracting useful data from a body of text.

Supervised Learning: Machine learning method that builds a model based off of data with known inputs and outputs. This model can then predict outputs of new input data.

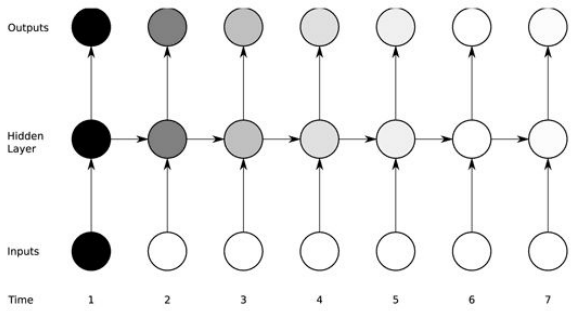
Tag: An element in the text which can be used to help in the identification process once it has been parsed.

Target Data: Consists of the Machine learning desired output data; algorithm is trained to make the connection from input data (text) to the output target data.

Text: Pertains to the 'body' of the Wikipedia article once it has been parsed.

Transformed Data: The text once all of the formatting artifacts have been removed; plain, lower-case text.

Vanishing Gradient Problem: An observed problem with machine learning algorithms, when models cannot retain information throughout their architecture or over time.



Credit: Graves, Alex. A Novel Connectionist System for Unconstrained Handwriting Recognition. 2009.

Visualization: How the results of the Machine learning algorithm can be viewed and interpreted by humans, this step is essential in making sense of the entire process.

2. System Requirements:

2.1 Enumerated Functional Requirements

Identifier	Requirement	PW
REQ1	The system shall use a model to classify sequences of text.	5
REQ2	The system shall use a model that can be trained, given examples of sequences with desired labels.	5
REQ3	The system shall have pre-labeled example data available, with methods for importing it.	4
REQ4	The system shall visualize sequence classifications for user interpretation.	4
REQ5	The system shall read in given text and classify based on given parameters.	4
REQ6	The system shall allow the user to notify a Wikipedia administrator if an article is found to be of very poor quality.	2
REQ7	Given an input article (url or file), the system shall return various values associated with that article (either cached on server or compiled in real time).	3

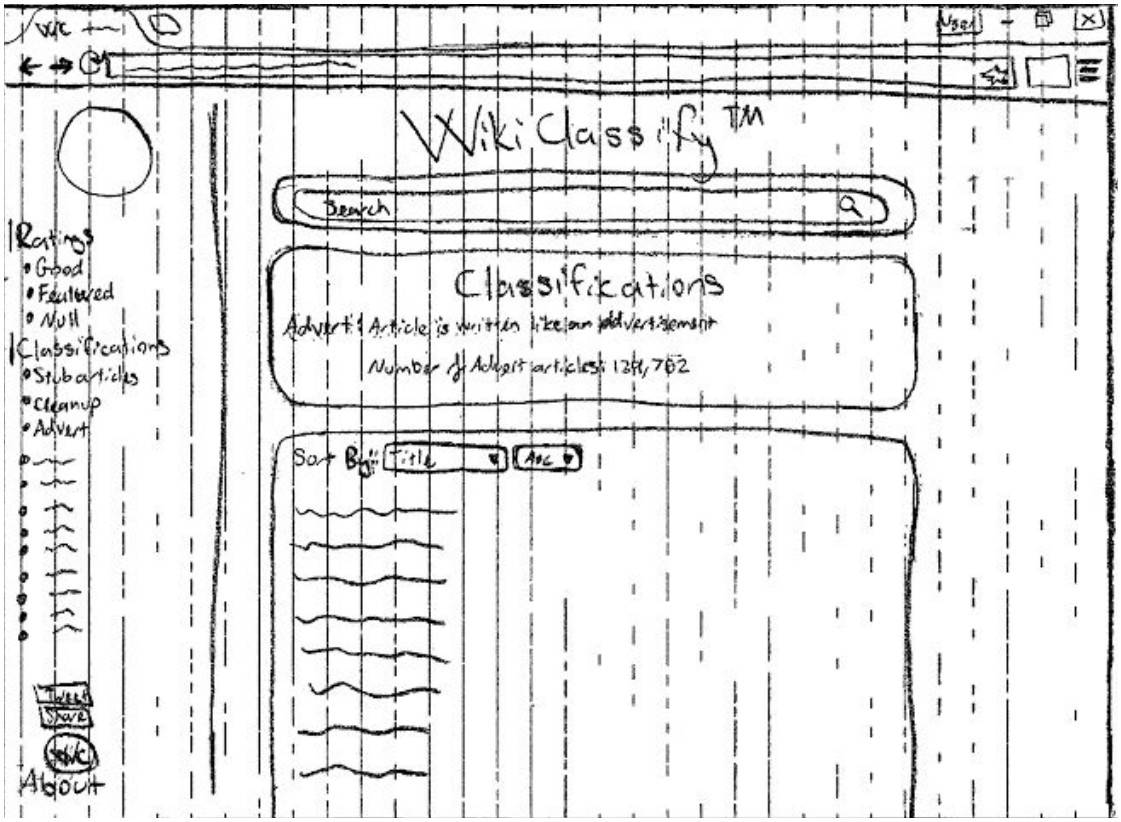
2.2 Enumerated Nonfunctional Requirements

Non-functional requirements are a more descriptive than practical listing the qualities of our system. These requirements are based on FURPS+, which are functionality, usability, reliability, performance, supportability, and other various attributes. These requirements are mainly concerned with quality attributes such as capability, compatibility, security, responsiveness, availability, efficiency, and maintainability.

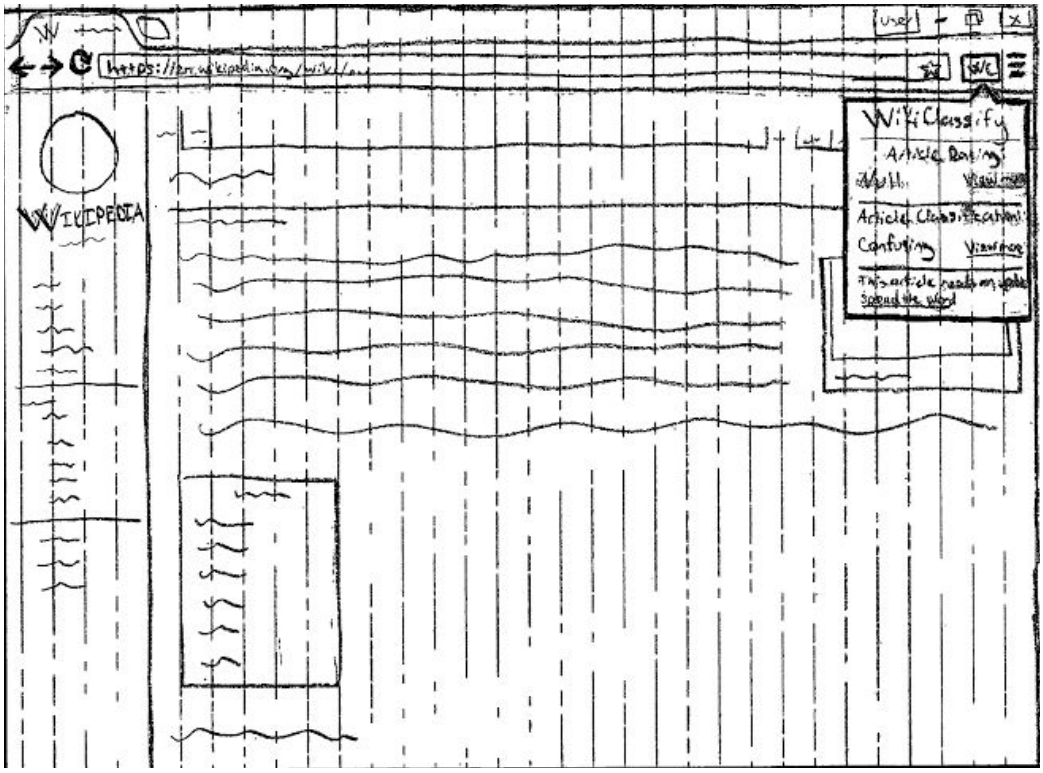
FURPS+ (User Stories)

Identifier	Requirement	PW
US 1	As a user, the system should be able to classify different articles found on Wikipedia.	5
US 2	As a user, I want the system to learn and be trained so that it can continue to work for articles written in the future.	5
US 3	As a user there should be documentation on how to use the program to the best of its abilities.	4
US 4	As a user I want a nice enough user interface so that I will be able to easily and quickly see data on the classification	4
US 5	As a user I would not want to be limited to just Wikipedia and instead would like to be able to apply the program to a wide range of texts.	4
US 6	As a user, if I find a problem with the article or the classification I should be able to notify an administrator to look into it.	2
US 7	As a user I would like the program to work quickly so that I do not have to wait for the program to run.	3

2.3 On-Screen Appearance Requirements



(Rough sketch of the user interface layout on the webpage)



(Rough sketch of the Chrome extension)

Identifier	Requirement
OSA1	Webpage: The webpage shall include an easy to maneuver layout. The left side of the page will include be the same for every page and will be static on scrolls. It will include the Wikipedia logo, the lists of the different ratings and classifications, social media sharing compatibility, our logo, and a link to our About page. The center of the page will always include “WikiClassify™” at the top along with the search. Depending on what current page you are on, the page will show the meaning of the rating or classification value you are searching through along with the number of articles for that rating or classification. Under that will the list of all the articles, where the user will also be able to sort the results.
OSA2	Chrome Extension (if time applicable): The Chrome extension should automatically pull the rating and classification of the current Wikipedia page from the server and display it as a popup. If the article needs to be updated, this extension will notify the user at the bottom and include a link for which the user can report the page.

References:

1. https://computation.llnl.gov/casc/sapphire/overview/overview.html
2. http://infojustice.org/wp-content/uploads/2013/10/band-gerafi10032013.pdf
3. http://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1055&context=yjolt
4. https://en.wikipedia.org/wiki/Wikipedia:List_of_bad_article_ideas
5. https://en.wikipedia.org/wiki/Reliability_of_Wikipedia
6. https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria
7. https://en.wikipedia.org/wiki/Wikipedia:Good_article_criteria
8. https://en.wikipedia.org/wiki/Wikipedia:Maintenance

Pictures:
https://commons.wikimedia.org/wiki/File:Colored_neural_network.svg
http://scikit-learn.sourceforge.net/0.5/modules/clustering.html
https://commons.wikimedia.org/wiki/File:Long_Short_Term_Memory.png