HITCH TECH INC

USE OF AI AND ML TO REDUCE EDUCATIONAL INEQUITIES AND IMPROVE DIGITAL LEARNING

Authored by

Mark Buchner
Professor

William Parmenter
Researcher

Colleen McCann
Research Project Manager

Parth Darji
Researcher

TRADEMARKS AND DISCLAIMERS

This project was funded and supported through NSERC and OCE, executed by Conestoga College's SMART Centre in partnership with Waterloo's HITCH Tech Inc.

ACKNOWLEDGEMENTS

Special thanks are given to the following reviewers of this paper, for their valuable time and insights.

Uche Onuora, HITCH Tech Inc.

Steve Veerman, HITCH Tech Inc.

Wallace Trenholm, Sightline Innovation

Conestoga College, 299 Doon Valley Dr, Kitchener, Ontario, N2G 4M4

TABLE OF CONTENTS

1. USE OF AI AND ML TO REDUCE EDUCATIONAL INEQUITIES AND IMPROVE DIGITAL LEARNING

1.1 ABSTRACT

Hitch Tech Inc., in conjunction with Conestoga College, is committed to tackling educational inequities by designing and building an information system to revolutionize the connection of printed text and digital content supporting the delivery of high-quality education. The benefits of this system will serve under-privileged countries, regions, and peoples in its ability to leverage the existing work of educators across the globe, and to bridge language gaps, in order to link specialized regional and ethnic learnings to formal educational curricula.

The system deconstructs existing educational content - defined as curricula, textbooks, and video files accessible across all platforms (YouTube, Vimeo, Facebook, etc.) - into a molecular unit called an IDEA. These units are cleansed, summarized, vectorized, tagged, and then stored. Artificial Intelligence (AI) and Machine Learning (ML) is then applied to classify and match up the IDEAs. The overall goal is to easily, and contextually, connect a curated video with a course lesson plan, or a textbook chapter. Our vision for this system is to support educators and publishers in modernizing and digitizing their classroom content, as well as adapting to the needs of individual learners. By making the IDEAs openly available and easy to consume through cloud computing, synchronized with affordable offline delivery mechanisms, this system can radically improve equity in education, enhance applied learning relevance, and boost overall comprehension and quality.

We envision this system supporting learners globally – from emerging markets to developed ones - offering educators the ability to populate lesson plans in an automatic time saving way, while also providing the ability to link and thus preserve cultural context and underserved local languages in the learning process.

2. TACKLING EDUCATIONAL INEQUITIES

Education is the most powerful weapon which you can use to change the world
- Nelson Mandela

According to the United Nations, 85% of global wealth is owned by just 10% of individuals. The geographic and social boundaries our societies have constructed over the millennia are clearly counterproductive to making our world a better place. They cannot stop pandemics, climate change, or ignorance. Unchecked, our current socio-economic systems only reinforce the zero-sum notion that the rich get richer, and the poor get poorer. Collectively we recognize that humanity must do better, must work together to fix the existential problems we have created.

Few will argue that the greatest enabler for a better and more sustainable human future is ***education***. It has the power to lift populations out of poverty, remove barriers for individuals, and create a path to a more financially stable future. But the educational waters have been muddied - advertising, political propaganda, extremism, click fodder, and banal entertainment are not education, although they often get passed off as such. Despite the most noble of intentions education faces both visible and hidden inequalities. These factors can be local - teachers in one school board not having access to the tools of other boards; regional - income disparities and other inequities faced by specific communities; and national – different standards of educational quality in remote communities versus urban cities. They are observed with gender, race, immigration status, language, dialog, and culture.

Throughout our history access to education has been weaponized. Denying opportunities to specific social groups, all while maintaining barriers designed to accelerate the concentration of wealth and power. While some may make the habitual argument for *trickle-down economics*, suggesting that wealth in the hands of the few eventually winds up spread out across the many, no one has ever proposed *trickle-down education*. Knowledge in the hands of the few has never benefited society as a whole.

3. HITCH'S PARTNERSHIP AND MISSION

Our partner in this project tackles global educational inequality as its mission. Part of their solution includes harnessing the power of IT, cloud computing, and Artificial Intelligence to create a compelling platform for educators, students, and parents. The digitization of educational content – defined as textbooks, lectures, videos, labs, quizzes and assignments - coupled with the networking power of the Internet and IOT enables a fundamental transformation of the education sector. This is not unlike large scale transformations which have occurred in other sectors like music, movies, retail, taxi, travel and tourism. Here, the power of the cloud can be harnessed for good to break down barriers and match quality education content to those who need it.

The idea of the internet as a transformative force within education is certainly not novel, but a critical analysis of today's various platforms and tools show that there are many obstacles. HITCH has a particular passion in deploying technology to reduce educational barriers in Nigeria and other African countries, with an eye to doing the same for remote Canadian and Indigenous communities.

3.1 HITCH TODAY

HITCH offers an extensive library of top-quality curated videos focusing on both the topics of importance for the formal primary (1st to 6th grade) and secondary (7th to 12th grade) school curriculum in West Africa, and also providing supporting content for secondary school WAEC (West Africa Examinations Council) exit exams, such as WASSCE. These videos cover all the curriculum-defined performance objectives and learning outcomes, encompassing exam preparation tracks which cover all the right material for any given exam. The HITCH library also delivers educational videos covering general skills development, applied learning and academic topics of interest that can be explored for self-directed learning. The videos are available on-demand for almost any device, and for remote schools the full library of videos can even be supported offline when the internet connection is not operational. At each level different stakeholders can realize many powerful benefits.

Parents benefit from top quality curated educational content, trustworthy, and distraction-free. Homework and assignments are more engaging with video content that can be pre-loaded and watched anywhere, and tools are provided to help with standard exams.

Students can go beyond the pages of a textbook, immersing themselves with engaging and useful instruction through video and other multimedia content which is customizable to their learning style and convenience.

Teachers have access to a categorized, searchable video library that lets them find relevant examples quickly during lesson preparation. It brings the lessons to life and increases student engagement. In addition, it helps the educator continue to develop their resource bank and teaching methods, with support for shareable insights and professional development material.

3.2 CURRICULUM ANALYSIS

HITCH provided our research team with 350 worldwide unique curricula for analysis during the course of the project. One of the first objectives was to take the various curriculum and parse them using a Python tool to discover their structure and themes. Naturally, the project gravitated to analyzing the Nigerian K-12 (kindergarten to grade 12) curriculum, and contrasting that to Ontario's curriculum. This approach provided the greatest familiarity for the research team, and the initial aim was to seek out some overall basic terminology that we could use as equalizers amongst the 350 worldwide curricula. The subsequent analysis revealed a stark contrast in structure, semantics, and style summarized as follows.

> **Nigeria:** lacks a clear set of unified education policies and is characterized by regional differences in quality standards, documentation level and funding. Curriculum - where defined and where we analyzed it - is highly prescriptive. Language is dry, procedural, and formal. Visuals are not intuitive and are generally basic line drawings. Content is highly prescribed, "hard coded" and structured, suggesting the rote memorization approach.
>
> **Ontario, Canada:** K-12 education in Canada falls under provincial jurisdiction, and it is the district school boards who then administer the programs. While decentralized from the federal government, the quality levels are consistently high across all provinces. Furthermore, funding is allocated for the districts by each province and is not directly related to local property taxes or wealth. This serves to somewhat reduce funding disparities between school boards. The various curricula for K-12 are in stark contrast to the Nigeria curriculum, in that they generally focus on the needs of individual learners, prioritizing innovation and the learning process. They use vastly different syntax and semantics. The language is engaging and conversational, visuals are easily understood, and content is targeted towards learning outcomes, leaving open the methods of engagement for the student.

Our conclusion was therefore curriculum, syllabus, or class vary too broadly across domains and thus do not, and cannot, provide a common denominator for educational content equalization. We realized instead that our tools and processes must focus on breaking down the curriculum far deeper than just the subject and grade level. A molecular educational unit was needed as the common denominator applied not only across global curriculums, but also textbooks and other content.

We were inspired by an automotive sector project where they similarly had to come up with a definition of the smallest possible sub-component of a “part”. In the automotive sector they choose to call this an “each”, abbreviated as EA. Accordingly, we decided to call our molecular educational unit an **Interactive Digital Education Asset**, pronounced as “idea” and abbreviated as IDEA. Our tooling was then oriented to take the educational content - such as a curriculum, lesson plan, textbook, video or, assignment - and then discovering what IDEAs were embedded therein.

3.3 CURRENT ISSUES: SOURCES OF EDUCATIONAL CONTENT AND CONTROL

Driven by the rapid adoption of “education technology”, and a plethora of players in the educational content production sphere, considerable but still siloed libraries of quality digital material and services continue to grow exponentially. This style of educational content delivery inadvertently perpetuates the very same systemic discovery and access problems. The content is often proprietary and hosted on a paywalled platform. Notwithstanding the explosion of e-learning and remote education demand brought by the global COVID-19 pandemic millions of parents, students and teachers still face significant challenges discovering and accessing affordable, interactive, and engaging digital learning experiences tailored to their specific contexts or requirements.

On the other hand, “traditional" educational content producers like textbook publishers struggle with the digital learning transition. Changing user behavior, lost sales from increasing digital piracy and attitude shifts regarding digital asset ownership amount to sales losses of over $1 billion annually. These circumstances clearly demonstrate an urgent need for an integrated strategic approach and comprehensive model that expands educational content sourcing and control. This new paradigm must evolve in a way that fosters learning on-demand, by intelligently pulling relevant educational content from multiple curated sources, according to customizable constraints and in response to personalized use-cases.

Indeed, by examining the prominent IT vendors and brands operating within the education sector - and their role in providing content - we begin to see the inherent limitations created because their goals and values are not fully aligned with ours. Each platform was created to serve a particular purpose, and our tool does not fit well into the current models.

CATEGORY	PURPOSE	WEAKNESS
Google, Facebook, Youtube	Generate clicks and ad revenue	Not focused on the educational quality or value of content
Amazon, Chapters	Generate revenue through online retailing of SKUs	Primarily focused on product revenue and SKU sales over educational content delivery
MOOC (Coursera, Udemy, LinkedIn Learning)	Focused on the structure and control of content	Unit of education is an overall course, not a piece of content
LMS (D2L, Blackboard, Moodle, Google Classroom)	Focused on class and student administration such as enrollment, tests, assignments and grading	The revenue model is based on control of licensed users
Publishers (Nelson Learning, Pearson)	Generate revenue based on proprietary content	High degree of protection of IP, unit of education is a course not a chapter or component
Public Library	Make information available to the general public	Ability to access, funding/infrastructure, regionality, indexing
Wikipedia/wiki	Public compendium containing information on all branches of knowledge	Crowdsourcing used to produce a single web page and set of links

Table 1 - comparison of current educational content delivery platforms

3.4 HITCH TOMORROW

This project focused on building and providing the components of a system designed to support the identification, tagging, transacting and matching of educational content units (IDEAs) regardless of whether they belong to videos, assignments, textbooks, course outlines, or class plans, notwithstanding their originating source. The content should be highly context-relevant and language-based to a known set of IDEA controls. This paves the way for material to be sourced from multiple producers and platforms, linked and shared across cultures and learning dynamics to remove language barriers and improve learning. In this way educational content can be both translated and preserved in the case of unique regional or Indigenous dialects. Such a foundation shapes the domain in which HITCH fulfills its unique role by identifying, analyzing, matching, and transacting in the basic building blocks of educational value.

4. OUR VISION FOR THE SOLUTION

The key problem this project faces, and aims to tackle, originates from the granularity of content required to deliver our vision of supporting on-demand education, and the preponderance of educational content silos that are so prevalent online. Today's products are course based - textbooks, credentials, certificate centered SKUs that do not focus on individual learning outcomes.

Our vision is for any learning resources, for example an eBook, to be fed into a Machine Learning application which then accesses libraries of digital content, syllabi, and assessments to find the most relevant companion material. Whether the desire is to learn a skill, language, or topic in or out of school, the IDEA bridges content fragmentation across educational platforms in a simple way which rewards content owners, while allowing access to an affordable menu of the best content, or service, offerings tailored to a consumer's learning context.

In delivering the best IDEAs for a specific use-case there will be a need to source content from multiple, often paywalled, educational content providers. For the educational providers the major concerns are piracy and compensation for their content. While, for the content aggregator the major concerns will be seamlessly receiving the content within their app, or learning environment, and the cost of à la carte content. For both these parties the complexities of negotiating monetization agreements and per unit costs of developing independent integration infrastructure will be the major concern. For consumers however, the major concern is accessing the best content, based on their needs, to learn from. Thus, we further envision an intelligent layer that interconnects the global digital education content ecosystem, while aligning the interests of all stakeholders in perpetuity.

Essentially, by ripping apart educational inputs to understand and classify each granular piece, content can be contextually and relevantly matched to specific learning outcomes rather than a specific curriculum. The solution will have the capability of picking apart and understanding multiple forms of educational inputs (see Figure 1). Decomposing each one in order to identify the basic IDEAs contained within will serve two purposes; first the ability to tag and organize each IDEA, and then be able to mine those tags quickly and easily to re-compose a new lesson plan automatically with a wide array of unique and quality controlled supporting content.

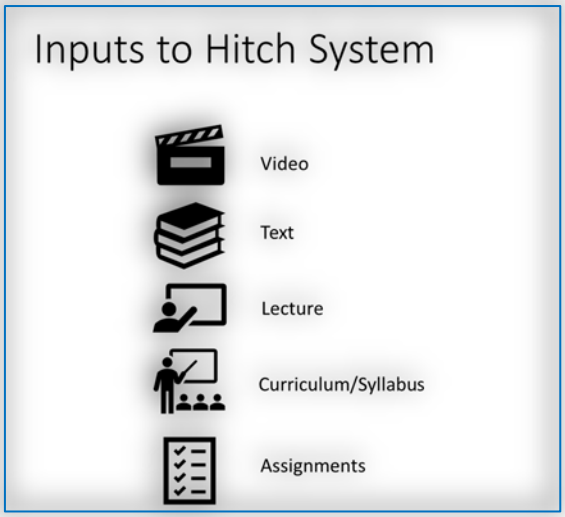

Figure 1 - Proposed system inputs

A good way to picture the solution is through the use of some creative analogies.

4.1 THE MUSIC ANALOGY

Not so long ago it used to be that if you wanted to listen to your favorite song, you first had to go to the music store and buy the whole LP/album. You had to know what the song title was, the name of the band, and what album it was released on. Then you had to pay for it, but now you also had a handful of other songs you may or may not have wanted. Maybe you got lucky, and your song was available as a single. The other option of course - you could hear your song on the radio if it made it to the top 40, but you had to be willing to wait and listen to the commercials.

Today, if you want to listen to your favorite song you don't even have to know its name! Just Shazam it and the app will tell you the artist and song, then you can download it or just listen through a streaming service. You can even piece together random songs in a playlist that suits your mood or taste, or you can have your streaming service make recommendations for you. You do pay for the service, but you don't have to listen to the commercials. The IP is taken care of by the streaming service, so you aren't doing anything illegal.

4.2 THE EDUCATION ANALOGY

Not so long ago it used to be that if you were a teacher and wanted to deliver a course, you would have to know the curriculum's prescribed textbook for the class and then go get it, learn the content, and figure out how best it can be used to support the learning objectives you were mandated to deliver. You may, or may not, have known or heard about alternate or superior content in other textbooks. For evaluations such as assignments, quizzes and tests, you were mostly on your own, but you may have been able to collaborate with fellow teachers or reuse content from prior terms. Instructor guides may or may not have been available. You needed to be careful protecting your evaluations, because tests and their answers had a way of making the rounds amongst students rendering your assessments worthless.

Today, if you want to deliver a course you just look at the curriculum and expected learning outcomes. For each learning outcome the HITCH system will match and recommend online content that meets the needs, according to diverse contextual and local parameters, and then you are free to incorporate that into your class plan.

4.3 USER STORIES AND PROBLEM ILLUSTRATIONS

For the purposes of illustration, let's examine a series of five user stories representative of the various stakeholders within the HITCH domain. Each story will highlight the weaknesses in our present-day environment and propose solutions through the use of the HITCH system.

USER STORY	DESCRIPTION
User Story 1	*I'm a Kindergarten teacher in Ontario's Catholic School Board. Because of COVID-19 I'm forced to deliver online learning for my classes. I want to find suitable interactive assignments using Google Docs which I can use within my D2L classroom environment. It's for an Earth Day learning unit, and I really want to maintain full engagement with my students.*
User Story 2	*I'm a high school teacher in Lagos, Nigeria. I want to incorporate digital learning content and video assignments that can enhance my prescribed, old-fashioned Grade 8 chemistry textbook. I need to be able to customize the content to the needs of my individual learners, and there are many students from different classes sitting together in one classroom.*
User Story 3	*I'm an Education Director for a First Nations region in Northern Ontario. I would like to create a Grade 3 language arts curriculum that enables our students to access high quality content, but it has to use local language dialects and customs for each community. This curriculum and its content need to be available offline since internet infrastructure is unreliable.*
User Story 4	*I'm a grade 6 teacher in Ontario, and I'm very passionate about teaching the sciences. I have received praise from parents, peers and supervisors for some video tutorials I created for my Grade 6 Physics class. I found the content especially useful for engaging girls in the subject matter, and I would like to freely share this work with other educators around the world.*
User Story 5	*I'm a female Grade 7 student from a rural village in Nigeria. My mother must work to feed our family, and I can't attend school anymore because I need to take care of my younger siblings. I loved going to school and learning. I want to go to university and become a teacher in my community to help other kids have a better life. I want to continue learning and advance my education without having to physically attend the class.*

Table 2 - representative examples of user stories

User story 1 illustrates how real-life dynamics can influence the teaching arena, leading to a sudden or unplanned need to pivot teaching delivery online, or enhance the classroom experience. Large school boards use a controlled LMS system with many teachers and classes. While we can assume that some teachers or consultants within the board have appropriate content, or know where to find it, there is just no suitable way to archive it, log it and find it. The material needed would generally be part of a larger course, book, curriculum or program and not easily separated. Furthermore, the ability to find suitable content in a *just-in-time* fashion is not possible because of too many structural barriers.

User story 2 illustrates the need to improve educational content so that it's more engaging for students. Additionally, it underscores the necessity of being easily adaptable to the needs of individual students, without a lot of extra work on the teacher's part. Often, a teacher is not in a position to change the course plan or prescribed textbooks for many reasons. Yet they

are still motivated to improve and modernize the learning experience from the old fashioned well-used textbook. Here, class material can be enhanced by linking its component IDEAs to engaging content that matches it, regardless of a teacher's geography or economics. As well, the teacher and learner can fully use the material in a way that suits their level or pedagogical style.

User story 3 illustrates the urgency to improve education and learning content with respect to supporting and including aboriginal languages or regional dialects. This has never been more crucial to removing the barriers of inequity faced by minority communities, and the sub-standard infrastructure and outdated resources they find themselves faced with. While it may be premature to support fully automated cultural transformation of content, it is feasible to provide tools and supports to significantly improve the productivity of professionals in Indigenous studies. In this case, popular teaching materials can be deconstructed in order to understand its context. Reduction of the learning unit from text or course-based to IDEA saves time and increases efficiency.

User story 4 perfectly illustrates the intellectual property challenges associated with educational content today. Often the property of publishers or individual schooling institutions, the IP is protected and copyrighted limiting its ability to be shared. Software creators for example can contribute to the knowledge base by uploading their work in a controlled *open-source* standard, enjoying the freedom that comes with GNU licenses, though there are limits to such freedoms for content providers. A growing movement of Open Educational Resources (OERs) across a variety of content formats and use-cases have been promoted and implemented by a variety of stakeholders. However, there is now an urgent need to comprehensively standardize and accelerate this across different content silos and for different contexts. Having digital education content "go viral" as in the case for YouTubers, Tik-Tokers, or Instagramers is not the objective of this user story. Here, the system prescribes a framework for top quality content to be curated, registered, matched, and tracked. Social media means are then used to provide feedback loops to AI systems to better "match" that content.

User story 5 illustrates the realities faced by millions of young learners globally. The need for alternate and flexible learning pathways is very real, and these alternate streams can be supported by the use of targeted technology and the internet. In-class synchronous learning content can be fully complemented with remote asynchronous learning content to provide a flexible and hybrid learning experience. Digital device access then requires a basic internet access to all populations.

5. SOLUTION CHALLENGES

The system we build will consist of three major functional components:

1. Break down a provided file into its lowest common denominator (IDEA). The file can be:
    a. A textbook or a curriculum (.pdf)
    b. A transcript file that corresponds to a video (.vtt)
2. Summarize an IDEA and store it in a database
    a. Summarize the content and provide a meaningful output
    b. Reduce the summary to a vector (numerical representation) as input to AI systems
    c. When appropriate or available, allow a human curator to tag the IDEA
3. Find matches between IDEAs – (curriculum, textbook and videos)
    a. Linear regressions analysis
    b. Clustering of like IDEAs

5.1 FILE BREAKDOWN

Our educational content needs to be in the form of a text file or stream, so that we can process it using Python and regular expressions. Typically, the raw input is not in the form of a text stream format as we want it, and thus it will need some pre-processing to cleanse the content.

The structure of the input document needs to be fully understood so that the lowest common denominators are identified.

1. **Curriculum (.pdf content)** Our system imagines the IDEA to be the leaf of the tree in any curriculum. Thus, to understand and catalog the IDEAs we must analyze the structure of the curriculum to its smallest components. Parent nodes along the way will give clues as to the hierarchy of the information and subject matter.

2. **Textbook (.pdf content)** If the file is a book we look to a chapter as its equivalent IDEA. The text component of the PDF needs to be scraped from the document. An effort is made to extract text and meaning from all tables. The system will ignore images and pictures at this time.

3. **Video (.vtt or .mp4)** The automated system will not have the ability to process MP4 files at this time, while the human-assisted tagging application may process these files. It is the VTT file, representing the content of the video in transcript form, that our system takes interest in. There are multiple problems and distractions with the VTT file format. First, the VTT transcript may not be properly punctuated, so the resulting sentence or paragraph structure may not be fully understood. A second problem is that the VTT may contain incoherent text strings that not relevant to what we want to extract. The ability to divide up a single longer video into multiple IDEAs is a function planned for future releases. Video transcripts may be prone to syntax and semantic errors, especially in comparison to a textbook, so the first step will be attempting to format the transcript into readable paragraphs. Correcting language issues resulting from errors in Natural Language Processing or voice translation is out of scope.

4. **Tag (.csv)** An optional data capture feature is provided for manual curation and tagging of IDEAs. This data is useful both for filtering of candidate IDEAs and for future use as a training dataset for Machine Learning models. This module presents a video to a human curator who is asked to categorize and subjectively evaluate the video. The system presents a word cloud of the video to the curator in order to expedite their work.

5.2 SUMMARIZING THE IDEA

The precondition for the next set of use cases is that we now have a useful text stream representing an IDEA. Now we'd like to understand the IDEA, and reduce it to a set of values that describe its content. Our intermediate formats, which are human-readable forms include:

1. A **word cloud.** Word clouds are surprisingly effective ways to communicate important information at a glance. The word cloud helps to pinpoint specific words that appear in texts and accounts for their frequency. Then, based on an algorithm, displays the words in various colors, sizes, boldness, or angles. Our use of the word cloud will serve two functions, one as a user interface for the human curator/tagger to help improve their productivity and, second as a labels-tree intermediate form used for our vectorization of the IDEA. Project researchers reviewed several python-based word cloud tools in order to select the most appropriate one.
2. A **summarized text**. Various text summarization services are available for free use on the internet. The concept with this is to reduce the text or phrases to a

summary using AI/ML. The research team analyzed several available tools for use with Python before selecting the most appropriate one.

The intermediate forms of the word cloud (labels trees) and summarized text are then used as inputs to the final step which is to provide a numerical (vector) representation of our IDEA. This vector representation is only understandable by the computer:

3. **Vectors**. A major feature of our AI system will be the vectorization of the document. For AI/ML processing, numerical representation of the document is far more valuable than text, thus the vectorization process is important in creating a numeric representation of the IDEA. As an input to the process, we can use a "bag of words" such as our word cloud or summarized text. The vectorized content is excellent input for sentiment analysis and analogical reasoning, all related to calculating distances between matching pairs of words.

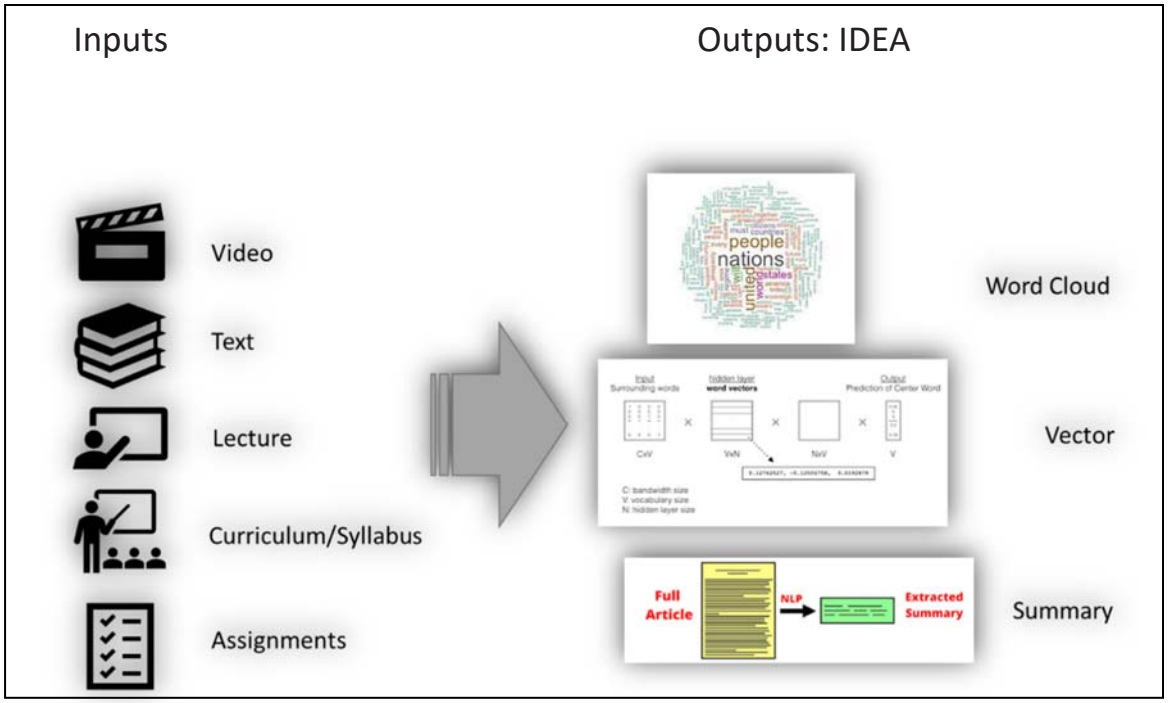

Figure 2 - Translating inputs to outputs

5.3 FINDING MATCHES BETWEEN IDEA

After processing, our system now has four sets of features associated with an IDEA. From these we are able to apply ML to further refine the handling of IDEAs.

1. A word cloud labels tree
2. Summarized text
3. A numerical vector
4. Optional tags, created by human curators

5.4 MACHINE LEARNING

The system shall now use machine learning processes to assess the relationships between IDEAs. The initial IDEAs we would like to match in the first release include:

- Curriculum IDEA ←→ Video IDEA: Used to see how well a video matches a curriculum and vice versa
- Textbook IDEA ←→Video IDEA: Used to see how well a video matches a text and vice versa
- Video IDEA ←→ Video IDEA: Used to identify videos that are related to each other

Rather than trying to codify "intelligent" strategies, statistical machine learning seeks to model the probabilistic occurrence frequencies within large data sets.

The key concept: *if you observe a pattern, you don't need to understand it, you only need to be able to detect, replicate, and predict it.*

Statistical machine learning relies heavily on probabilistic relationships amongst abstract variables in multidimensional spaces. The basis of everything within statistical machine learning are the probability distributions and parameter spaces, both of which are represented using vectors. Each element of a vector is a different parameter or data dimension in the problem being modeled. Thus, machine learning relies heavily on Linear Algebra (vectors and matrices), Calculus (functions, derivatives, gradients, vectors and matrix calculus), and Probability and Statistics (rules and axioms, discrete variables, distributions and Bayes Theorem) to help it create results that approach, or even surpass, that of a human.

When it comes to machine learning, there two general forms of learning (see Figure 3):

Supervised Learning, where the system identifies a function to predict the next value. It can create the function once it is trained with enough data, and this requires a lot of data. There is paired input and output data with the goal to minimize error between network output and desired output. Typically, tasks suited for supervised learning are regression analysis (function

approximation) and classification (pattern recognition).

Unsupervised Learning, where the system is driven purely by the data and identifies commonalities in that data, such as clustering or dimensionality reduction. Unsupervised learning occurs when an algorithm learns from plain examples without any associated response, leaving the algorithm to determine the data patterns on its own.

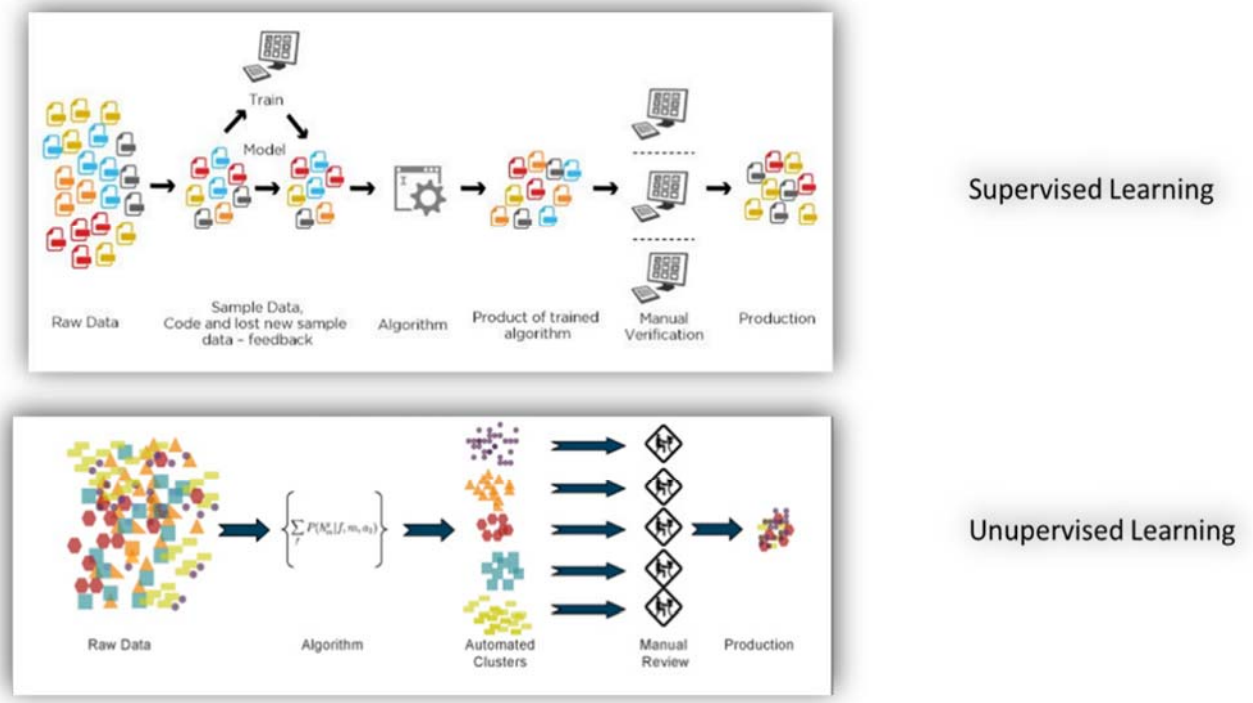

Figure 3 - Graphical representation of machine learning styles

An unsupervised learning algorithm tends to restructure the data into something else, such as new features that may represent a class or a new series of uncorrelated values. The resulting data are quite useful in providing humans with insights into the meaning of the original data, and new useful inputs to supervised machine learning algorithms.

The research team selected the most suitable AI/ML and deep learning algorithms to help tackle the various challenges of the HITCH System.

6. SOLUTION DESIGN

Our proposed solution design requires the following functions, which were provided and made available as microservices in our python-based system. Figure 4 below represents the use case diagram.

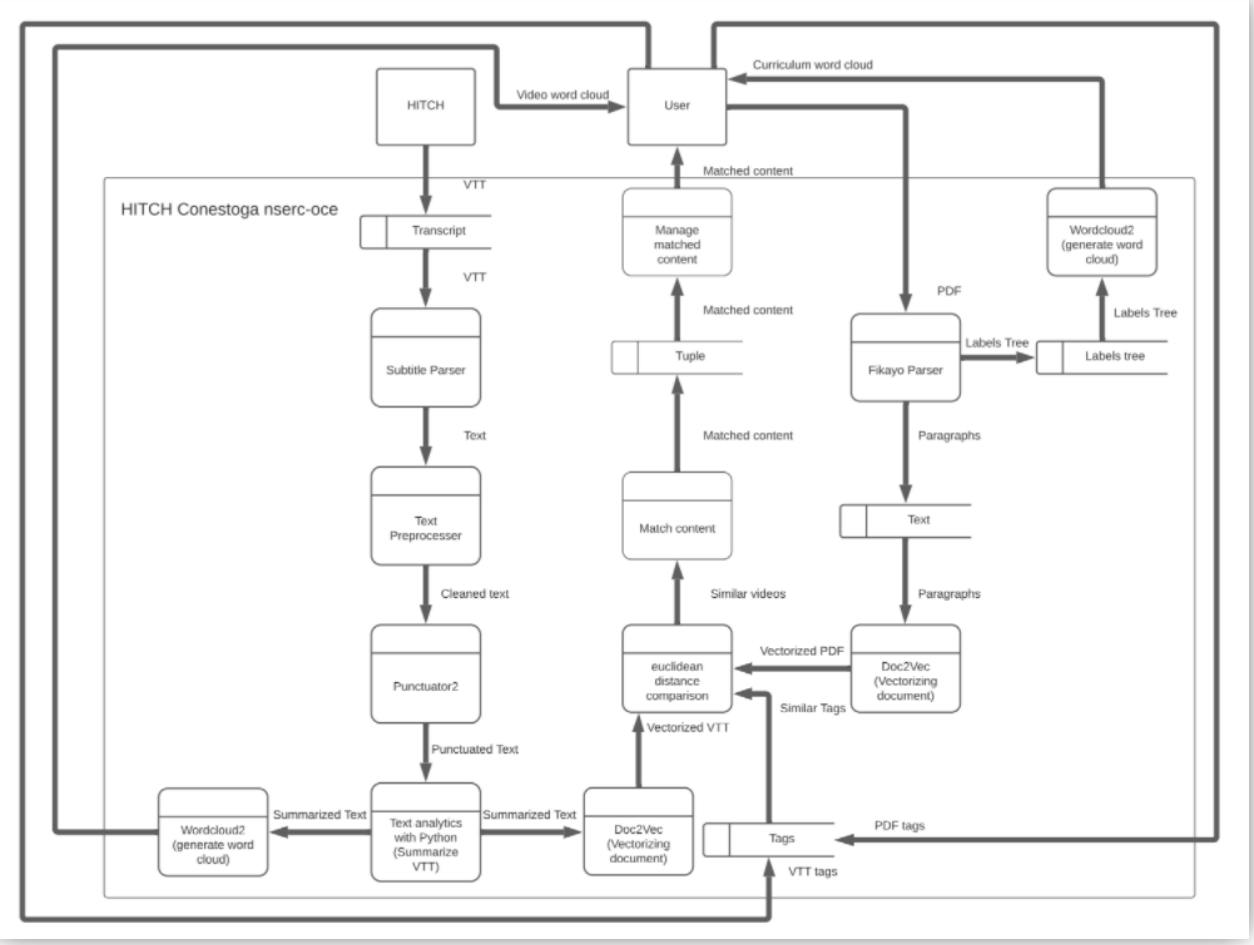

Figure 4 - Use case diagram of our solution

6.1 PARSING A VTT TRANSCRIPT

A VTT file is a text file saved in the Web Video Text Tracks (WebVTT) format. It contains supplementary information about a web video, including subtitles, captions, descriptions, chapters, and metadata. VTT files do not contain any video data. Our system will need to store the association between a VTT and the corresponding video in a key-value pair. Various steps will be involved in parsing this file and extracting actual content, and are outlined below:

1. Text Preprocessor: Filter out tags, timing info, blanks, irregularities to access raw content
2. Subtitle Parser: Analyze the tags to identify segment or chapter breaks
3. Punctuator: Analyze time stamps to assess words and sentences that belong together

Data Storage: Analyze the data and store it using MongoDB database

Retrieve Stored Data: Saved data can be retrieved using MongoDB's Compass utility as well as it can perform query selection using programming languages.

6.2 SUMMARIZING AN IDEA

A summarized IDEA is represented by a Word cloud. The Word cloud takes a summarized IDEA as input and returns key-value pairs of words and frequency. It shows words on the word cloud based on their frequency highest to lowest. Figure 5 below shows an example of word-cloud that generated for specific education content.

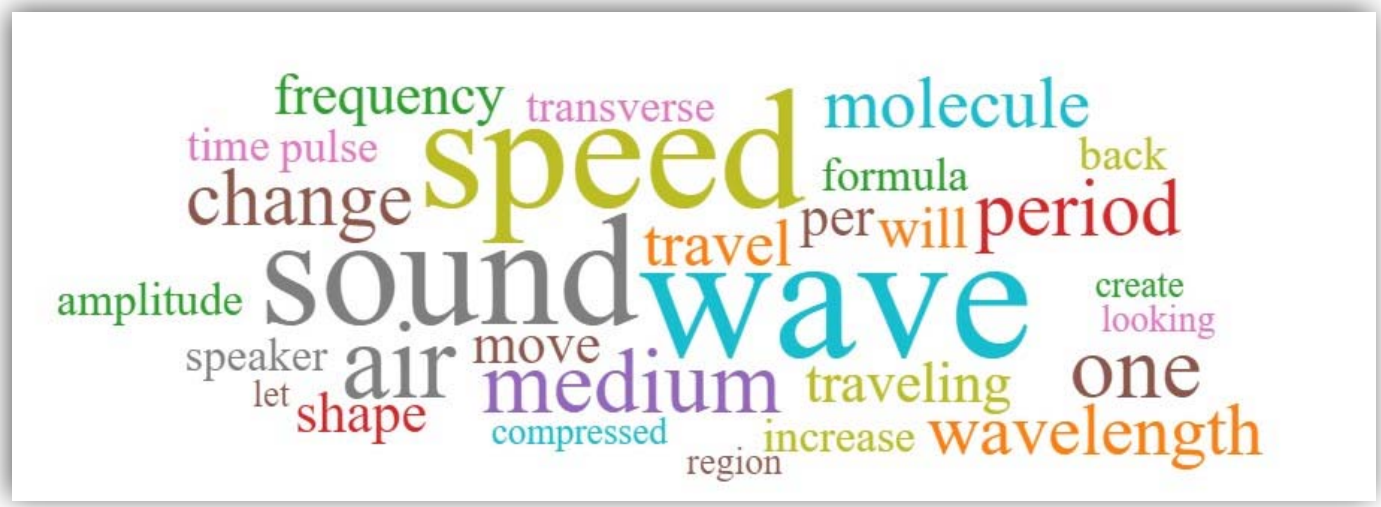

Figure 5 - Word cloud example

6.3 GENERATE VECTOR REPRESENTING IDEA

Once an IDEA has been summarized the system can vectorize the most important aspects of the IDEA. Machine learning algorithms are unable to understand sequences of words in the same manner that humans can, so to allow a machine learning system to perform an analysis on language, the data must be represented in a numerical manner. The process of generating a vector representation of an IDEA allows the system to present IDEAs to the machine learning algorithms in a way that they can understand.

Through vectorization, the system can perform meaningful analytics of the content and context of an IDEA. Once many IDEAs are vectorized, we are left with a two-dimensional array where the rows are instances of IDEAs, and the columns are the different features of the IDEA. We must then shift the thought of language from a sequence of words to a point in space that occupies a high-dimensional semantic space. These points can be close together, clustered, further apart, or evenly spaced. The closer the points indicate IDEAs that are more similar than IDEAs that are further apart. These points can now be used in calculations.

6.4 MATCHING OF IDEA VECTORS

The method that this project employs to vectorize text is called term frequency-inverse document frequency (TF-IDF). TF-IDF calculates the importance of a word to a document in a collection of documents. TF-IDF contains two parts, term frequency (TF) and inverse document frequency (IDF). TF is how frequently a word (term) appears in a document or collection of documents. Some words are frequently used but hold little to no context, such as; the, be, to, of. The frequency of these words is higher than words that aid in understanding the context of a document. IDF is used to mitigate the biased weightage of the mentioned common words. TF-IDF is a multiplication of TF and IDF, resulting in common words having a lower weightage and more unique words having a higher weightage.

```
# TF-IDF method convert text into numeric form
# tfIdfVectorizer=TfidfVectorizer(use_idf=True)
tfIdfVectorizer=TfidfVectorizer(max_features=4000, analyzer='word', min_df=20, max_df=0.80, stop_words=stopwords.words('english'))
tfidf = tfIdfVectorizer.fit_transform(df_vtt_cleaned_text)
```

Figure 6 -TF-IDF implementation

Once vectors represent the IDEAs, the system can perform calculations to determine the similarity of the IDEAs. For the system to calculate similarity, it uses the cosine similarity function to calculate the similarity between vectors. Typically, the cosine similarity calculation results range from -1 to 1, with -1 meaning completely different and 1 meaning the same. However, because TF cannot be negative, the system will output results ranging from 0 to 1.

7. SUMMARY OF OUR FINDINGS

In partnership with HITCH Tech Inc, the Conestoga College applied research team has produced a *first-of-its-kind* system to support the identification, tagging, transacting and matching of educational IDEA units regardless of their source of origin. At its core the system consumes VTT file transcripts created from educational videos, PDF documents of textbooks and/or curriculum as inputs. These are subsequently examined and parsed into IDEAs (Interactive Digital Educational Assets) representing the most basic molecular educational concepts. Next, they are punctuated, summarized and ultimately represented as a word cloud. Word clouds from individual IDEAs are vectorized and the distances between them computed using Cosine or Euclidean distance. The functional output is therefore an AI/ML backed recommendation system that truly knows no educational boundaries.

The system has successfully begun its first beta testing during the course of the project, and the team looks forward to analyzing the feedback and refining the system even further.