Ted Talk Clustering

Noah Carver
Email: ncarver1@umbc.edu

Jason Seaman
Email: jass2@umbc.edu

David Martirossian
Email: dm16@umbc.edu

Kyle Nehman
Email: nehman1@umbc.edu

I. Introduction

This projects aim is to create a system capable of understanding intrinsic relationships between Ted talks. We developed an AI to group a corpus of Ted talks hierarchically based only on their text transcription.

II. Background

A. Unsupervised Learning

Unsupervised Learning is a type of machine learning in which the algorithm draws inferences from only the data provided to it. The main difference between this and supervised learning is the absence of labeled data. Without this baseline to compare data to, an unsupervised algorithm must build groupings of similar data points from scratch, as opposed to simply placing data points into prescribed categories. Unsupervised learning creates relations that are intrinsic, based on the basic properties of the data. Supervised learning ascribes relations that are extrinsic, or defined by an external set of rules. For example, Hawaii isn't intrinsically very similar to Maine. It isn't until the extrinsic quality of being part of America is applied that they can be grouped.

B. Tokenization

The text-classification library that we are using uses tokenization in order to sift through monologue delivered by the speaker. Tokenization is taken from lexical analysis and, in this context, is the process of splitting up a sentence or paragraph into a descriptive token with which that sentence or paragraph can be related to other sentences or paragraphs. This project explores using exclusively nouns as tokens during comparisons. Other studies have suggested using nouns in combination with prepositions and comparative adjectives in order to decrease ambiguity. For example, if a speaker was giving a presentation on Dwayne The Rock Johnson appearing in a movie, this other approach would consider The Rock as a noun as opposed to various rocks appearing in movies. Before we tokenize, we represent a document as an unordered set of words, called a bag-of-words. In addition to disregarding word order,, we also disregard grammar by removing stop words, such as the and is, since they do not convey any useful meaning. After doing so, we used each unique word as a token in a vector so that it could be easily compared to other tokens. Without this numerical representation of our data set, we would be unable to map the centroid values of each speech in a universe of TED Talks.

C. Hierarchical Clustering

Hierarchical Clustering is used in order to take the set of tokens created by the tokenization process and group them into collections of similar documents. These tokens form a multidimensional vector in a vector space with as many dimensions as unique words. These multidimensional vectors are compared and condensed using cosine similarity. Using this relation, hierarchical agglomerative clustering repeatedly merges similar groups, or clusters, of data points by proximity until all points have been merged into one cluster, as seen in figure 1. The sub-clusters form a hierarchy that is best represented with a Dendrogram, explained more in section II-D. The two forms of hierarchical clustering are agglomerative and divisive. Agglomerative Clustering is a form of hierarchical clustering with that constructs its similarity model from the bottom-up. Each cluster begins as its own vector and is subsequently condensed as the cluster gets more general. Divisive clustering is a type of hierarchical clustering that uses a top-down approach. The vectors begin as one large cluster and then proceed to split up as the tree is traversed until each vector is in a cluster of its own.
In this application of our project, hierarchical clustering was used to represent the resulting values of the cosine similarity through agglomerative clustering. We found that using a subset of data points was the most effective way of visualizing our clusters since our model analyzed around 500 individual data points.

D. Dendrograms and Diagramatic representation

Dendrograms are particularly useful tools for visualizing hierarchical groupings (figure II-D). Dendrograms are commonly used to represent evolutionary relations in what are called Phylogenic trees (figure II-D). The farther from the root of the tree a connection is made, the more similar that connection is. For example in Figure II-D, it can be observed that Streptococcus Pneumoniae is more similar to Streptococcus Cremoris than Bacillales is to Lactobacillales.

E. K-Means

K-means clustering groups similar points or documents into k clusters, rather than just one like with hierarchical clustering. However, the number of clusters must be specified, which can be difficult when working with a sparse dataset, such as the Ted Talk corpus. K-means also assumes that the variance of the distribution of each feature is spherical and equal across all features. The algorithm starts off by creating k centroids and assigning each datapoint to its closest centroid.

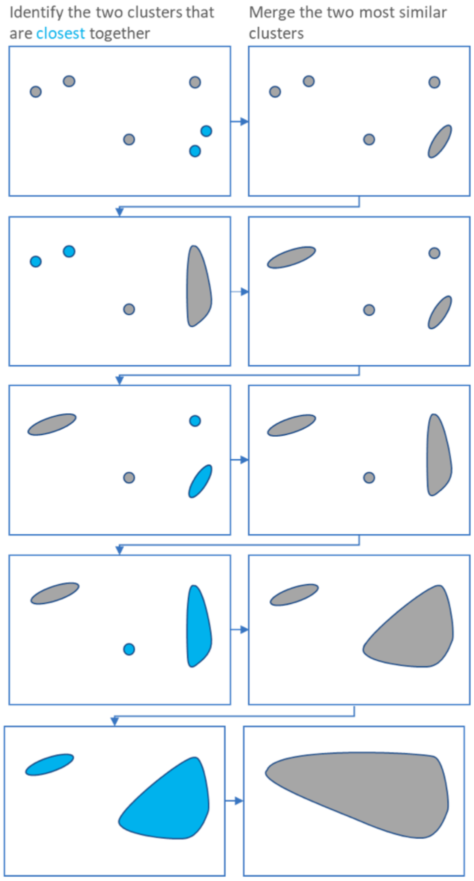

Identify the two clusters that are closest together
Merge the two most similar clusters

Fig. 1. Agglomerative Clustering

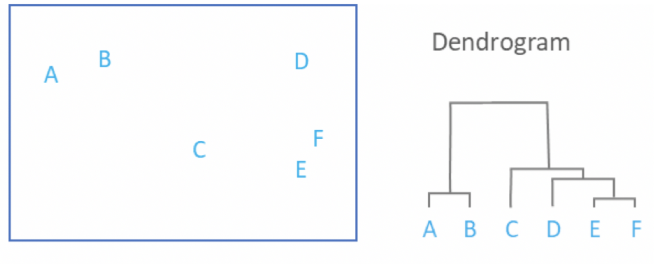


Fig. 2. Dendrogram Example of a Vector Space Grouping

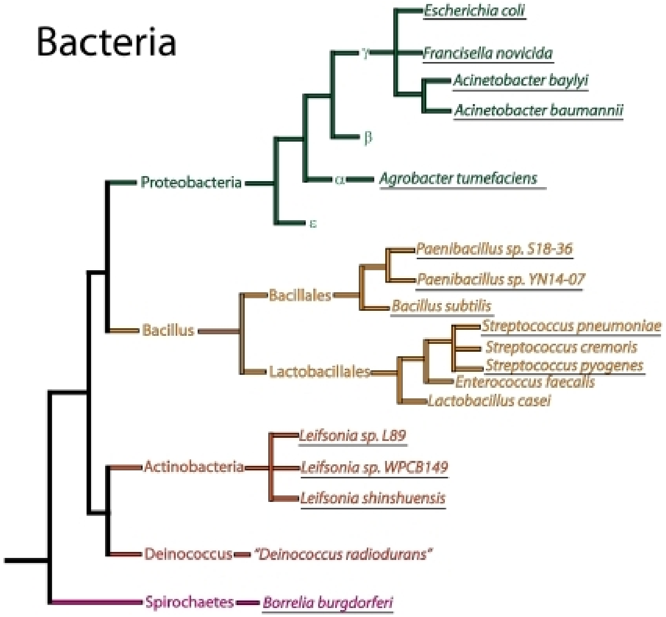


Fig. 3. PHylogenic Dendrogram Example

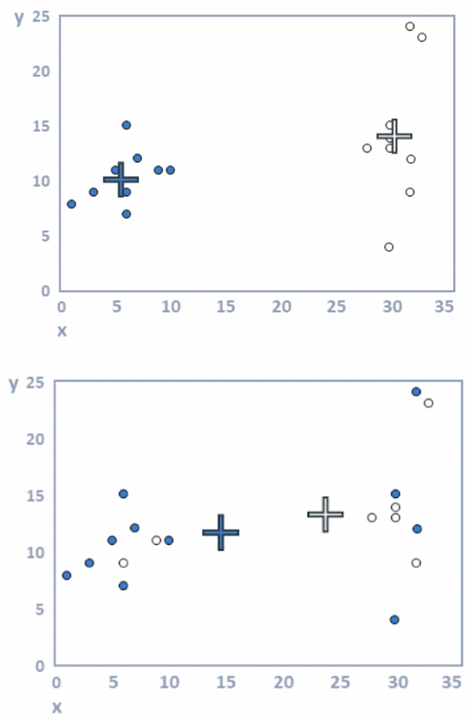


Fig. 4. Iteration of k-means clustering

After assignment, the centroids are updated to the position representing the mean of the surrounding points. Iteration continues until a certain exit condition is met. For example, after points are no longer being assigned to different centroids. The algorithm will always reach this criterion no matter the initial centroids; the complexity of this algorithm is $O(n^{nd})$ in d-dimensional space. We used k-means clustering to create a 2D vector space representation of related documents.

F. Term Frequency-Inverse Document Frequency

The tf-idf value is a statistical weight that reflects the importance of a word in a document in a collection or corpus. Term Frequency is simply a measure of how frequently a word appears in a document and is normalized by the total number of terms in said document. The Inverse Document Frequency provides a more accurate measure of the importance of a word relative to all other words. When calculating TF, all terms are equally important. However, because certain words may appear more frequently, these words are weighed down in the IDF and normalized by the number of times they appear. To summarize, TF measures how often a word appears in a document, while IDF measures how rare the word is throughout the corpus. So, the product of these two values provides us with the importance of a word. We are using the tf-idf value to determine the main message of a document, so we can later compare documents with each other. TF and IDF are defined by:

$$TF(i,j) = \frac{\text{Term } i \text{ frequency in document j}}{\text{Total words in document } j} \tag{1}$$

$$IDF(i) = \log_2(\frac{\text{Total documents}}{\text{documents with term } i}) \tag{2}$$

G. Cosine similarity

The cosine similarity metric is a similarity used to relate two non-zero vectors. The process of relating these two vectors is contingent on the cosine of the inner product space of the two vectors. To find the cosine similarity of these two vectors, the Euclidean DOT product and vector magnitude is used to find the cosine of the two given vectors. The resulting similarity will always be in between zero and one. In our project, the cosine similarity metric was necessary to relate distances between the centroids of our speech vectors. We are using cosine similarity to determine the distance between the tf-idf vectors of documents. With this, we can calculate how similar one document is to another.

$$\begin{aligned} similarity &= \cos\theta \\ &= \frac{\vec{a}\cdot\vec{b}}{||\vec{a}||\cdot||\vec{b}||} \end{aligned} \tag{3}$$

H. Manifold Learning

Manifold learning is a process of nonlinear dimensionality reduction. In essence, manifold learning is a technique to visually represent data points that exist in multidimensional planes. These data points must be mapped into a lower dimension so that additional pattern recognition algorithms can be applied. The process of manifolding can be explained through the hidden layers of neural networks. A vector of inputs may have to be condensed into smaller dimensionality to be recognized by other layers of the network.

In our project, manifolding was necessary to represent the unsupervised results of our experiment into a two dimensional graph or tree. Without the process, it would be much more difficult to communicate the findings of our model on the dataset of TED talks.

I. Performance Evaluation

It is difficult to accurately measure the performance of unsupervised learning. Unlike measuring the performance of supervised learning, there is no binary correct answer. To work around this, K-fold cross-validation can be used to test for consistency in the results. To do this, all of the test data is split up into k subsets. $k-1$ subsets are used as training data and the remaining set is used as test data. This process is repeated so that each subset of data can be trialed as the test data. Once the results are gathered, each result from the test data can be compared to the rest for consistency.

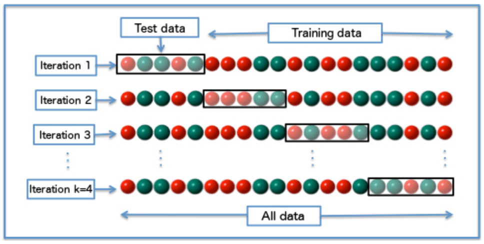


Fig. 5. K-fold Cross evaluation

III. Databases

A corpus of 514 TED Talk transcriptions, on a variety of different topics (the TED-LIUM Corpus [1]) has been chosen as our primary data-source for training. A second smaller corpus of TED talks and transcriptions, with some possible overlap will also be used in conjunction with the first for input.

IV. Data interpretation & results

The final hierarchical model for our model listed out all 514 points of data related to each other through a dendrogram and the titles of each talk in this model were too hard to distinguish between the smallest leaves. The full dendrogram is available in appendix B. To overcome this problem, we chose a subset of 40 Data points from the full 500 to represent them in a more legible diagram. To check our accuracy, we compared the most closely related leaves and cross-referenced those TED talks with their topic that they spoke on. For example, in the subset diagram, the two data points NicholasNegroponte_2006 and ThomMayne_2005 are a pair of talks that the model designated as closely related that are boxed in red in figure 6. Nicholas Negroponte is a Greek American architect & founder of MITs media lab. He gave his 2006 TED talk about a program called One Child One Laptop with an aim to better connect the world through the internet. Thom Mayne is also an architect by profession and his 2005 TED talk was titled How Architecture Can Connect Us where Mayne details how his vision for modern architecture is heavily influenced by the roots of the community in an attempt to bring together the community that interacts with the building every day.

This association is interesting for a few reasons. Both of the talks were given by architects by profession which can be a justification for a relation. Another interesting association between the two talks is that although the specifics of what the talks about were vastly different, both had an overarching theme of community and connectivity. These two associations are interesting because, under a supervised model, these two talks may not have been associated together at all. In fact, the keywords listed under each talk were Children, Design, Education, Entrepreneur, Global Issues, Philanthropy, Social Change, Technology and Architecture, Cities, Culture, Design, Invention. This makes the association of the TED talks all the more interesting. This unsupervised model gave us a new perspective on these talks when tasked with relating them. The association of these talks on the basis of community and connectivity is justified because of this small sample dendrogram in proportion to the full list of results.

To make sure this association isnt just confirmation bias or a fluke, we can look at another random leave in the hierarchical tree. Arbitrarily selected from the remaining leaves, is a talk labeled KRamdas_2009. Given by Kavita Ramdas in 2009, her talk is entitled Radical Women Embracing Tradition, and is highlighted blue in figure 6. Ramdas explains how women in many different cultures across the world are becoming leaders in their communities and mixing Western-style empowerment of women with the strong traditional values of their respective cultures. What is interesting about this example, is the tags associated with the video: Culture, Feminism, Social Change, Women, and India. When compared to the tags listed above, Ramdas talk shares one tag in common with each aforementioned talk: Culture in common with Maynes talk and Social Change in common with Negropontes talk.

In a supervised model, and using the TED video tags as labels, Ramdas talk would appear more closely related to Mayne and Negroponte than Mayne and Negroponte would be to each other. This is a notable difference between the unsupervised and supervised model.

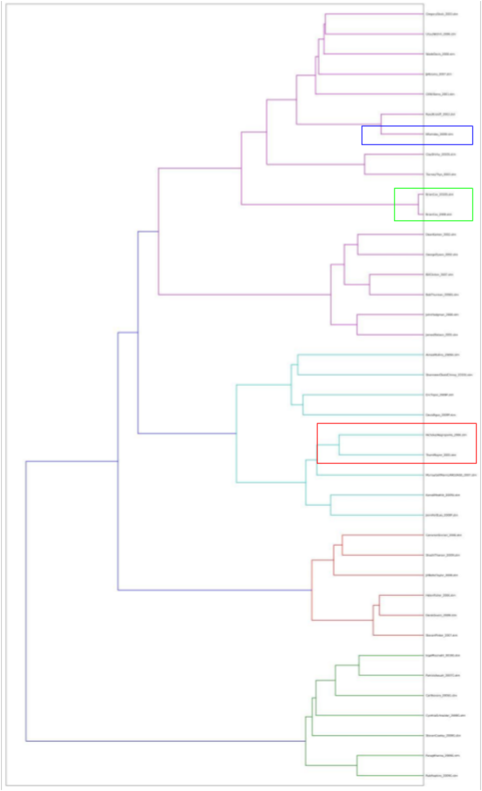


Fig. 6. Dendrogram of 40-Ted talk Corpus

We used k-means clustering to create a 2D vector space representation of related documents. As shown in FIGURE, each cluster is represented with its own color and the legend shows which words the documents are related by. For example,

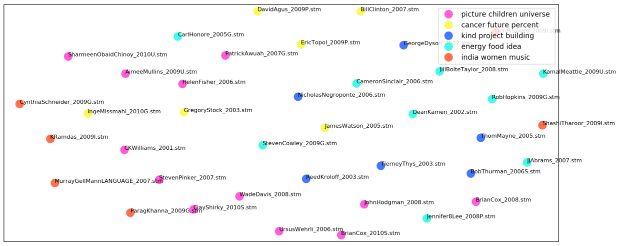


Fig. 7. K-Means Clustering of 40-Ted talk Corpus

the orange cluster is generalized with the words, india, women, and music. Looking at the transcripts of these documents, it is evident that they all are, in some shape or form, related to these topics.

V. Related Works

A. NLTK

The Natural Language Toolkit is a suite of Python libraries and programs used for NLP. These libraries can be used for tokenization, parsing, stemming, and semantic reasoning, making it the perfect tool for classifying text. [2]

B. Group Average similarity metric

The Group Average Similarity Metric compares all documents in a cluster, avoiding the pitfalls of both single-link and complete-link clustering, in which only the closest and farthest documents are compared. It is represented by the following equation:

$$\text{GA-SIM}(\omega_i, \omega_j) = \frac{1}{(N_i + N_j)(N_i + N_j - 1)} \sum_{d_m \in \omega_i \cup \omega_j} \sum_{d_n \in \omega_i \cup \omega_j, d_n \neq d_m} \vec{d_m} \cdot \vec{d_n} \tag{4}$$

where $\vec{d}$ is the document vector, N_i and N_j are the numbers of documents in groups ω_i and ω_j respectively and $\vec{d_m} \cdot \vec{d_m}$ is the dot product of documents $\vec{d_m}$ and $\vec{d_n}$.

C. Supervised Learning

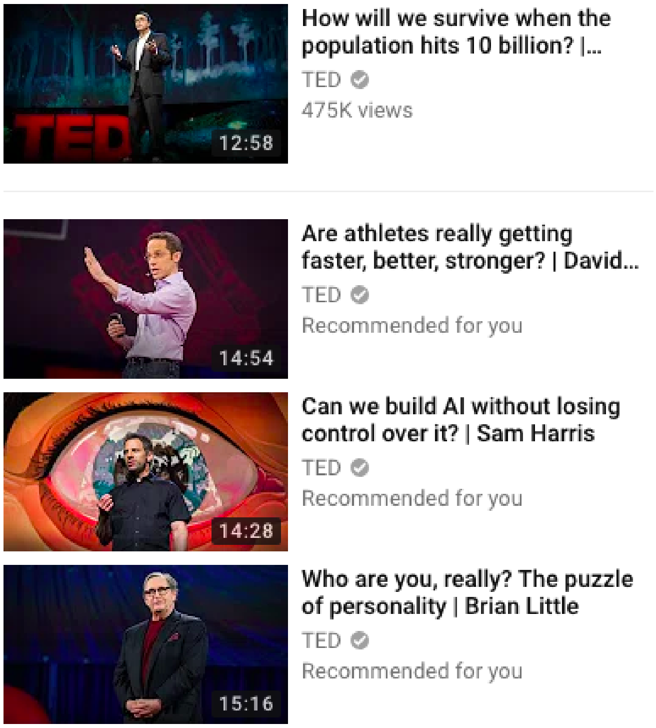
Supervised learning is a variant of machine learning where the outputs of a function are defined by labels applied to outputs in the test data. The goal of supervised learning is to produce an AI that can accurately infer when met with entirely new sets of data [3]. The benefit to this system is that the results will always be precise with exact decision boundaries. This project does not explore classifying TED talks by classification through supervised learning. The justification for this is that the amount of topics that are covered by TED talks far surpass that of a useful classification function. Unlike subjects in elementary school, TED talks cover a much larger variety of topics from activism to 3D printing [4]. A disadvantage to this approach is that if a new topic, say on esports, were to be covered, an unsupervised approach may plot this talk in between sports and video games. This would let the reader infer without listening to that talk what it may be about a topic in between. A supervised approach would only ever classify this new topic as an old one. This would lead the model to be outdated over time. A supervised learning model would be a useful reference for our unsupervised model should each current topic listed be labeled as a valid talk type.

VI. Future Applications

In the data interpretations and results section, we discuss the topics and tags listed beneath the each individual TED talk as a substitute for results from a supervised model. A future application of this data would be to create that specified supervised classification model and contrast it with the tags listed on the TED website to see if our claims in the results section hold. This label process means that the these associations between talks are extrinsic. Our unsupervised model gives intrinsic models of value for each talk from word content. If the tags were not generated by another ai on the TED website, it is possible that the supervised model is observing different trends and relations, but is restricted to the labels that were fed to it by a developer.

Since the relation clusters were based on this intrinsic mathematical value, this implementation could serve useful in combination with a supervised model to produce a recommendation algorithm on the TED talks website for which TED talk to watch next. The suggestions found by sorting by proximity in an unsupervised model would provide a more diverse set of results for a user to engage in. Since the motives of a TED talk viewer to be watching a TED talk are not uniform across all viewers, a more diverse set of recommendation videos is necessary to maximize user engagement on the website.

Despite the corollary between the two related talks, the topic of their relationship can only be hypothesized by the observer of the data. However, the relationship and proximity that is given to us between the two talks is still important and could yield additional information outside of just the topic of the conversation. As mentioned in the sources of error, this model could have picked up on the diction or vocabulary of a given subset of speakers. Even though this additional data could yield unrelated associations in speaker topic, the model could be used as reference to how be an effective speaker, or how to recognize speaking trends in different guests. This information could also be helpful with looking at linguistic characteristics of speakers that do not speak English as a native language.


How will we survive when the population hits 10 billion? |...
TED
475K views
12:58

Are athletes really getting faster, better, stronger? | David...
TED
Recommended for you
14:54

Can we build AI without losing control over it? | Sam Harris
TED
Recommended for you
14:28

Who are you, really? The puzzle of personality | Brian Little
TED
Recommended for you
15:16

Fig. 8. A Recommendation System

VII. Sources of Error

One potential source of error in our trials is the vocabulary of the speaker giving the TED talk. For example, when speakers are hung up on certain adjectives, verbs, or phrases. Many speakers use words that appear throughout most talks, such as world, new, and life. This can affect our results through word frequencies. In our tf-idf, nouns and adjectives are not weighted more or less than each other. Therefore, by coincidence, a two speakers could be closely related to each other in part to the way they describe their TED topic as opposed to the topic itself. This source of error may only be applicable to interpretations of our data that look for relation based on the topic of the TED talk itself. Nonetheless, even though this may be an unintended result, there is still a justification for the talks being related.

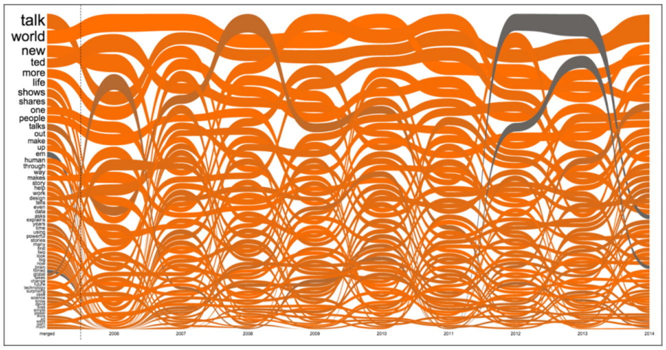


Fig. 9. Visualization of Most Common Words in Ted Talks

Another source of error in our model is the weighting of proportionally lengthed TED talks and transcripts. The tf-idf model takes the frequency of each word and proportionalized it in relation to the entire transcript of the document. Some TED talks were as short as two minutes in length and others exceeded 25 minutes. If a speaker on a smaller video gives a brief introduction about themselves, the words used in that introduction may be proportionally inaccurate and misrepresented to the content of the brief talk.

Audience participation is another potential divergence in our

model. Some TED talks are not exclusively a monologue, some talks, such as Arthur Benjamins talks about Mathemagic, rely on the naivety and responsiveness of the audience to adequately convey the speakers message. This dialogue is recorded in the transcripts of the talks and gravitate the TED talks with audience participation closer to each other in the results than they otherwise would have been.
Another source of error is not implementing stemming and lemmatization in our model. Stemming is the process of reducing a word to its stem. For example, remove ed, ing, and ly from words to get their base term. This would have improved our model because a words frequency throughout documents would be more accurate. This is also the case with lemmatization, which consists of reducing a word to its core message, such as changing studies to study.

VIII. Conclusion

We used unsupervised learning to create clusters of TED Talks based on the content of words in their respective transcripts and found unique similarities between those TED Talks independent of their topic labels on the TED Talks website. The similarities found in our dataset can prove useful across multiple disciplines and potential future applications.

References

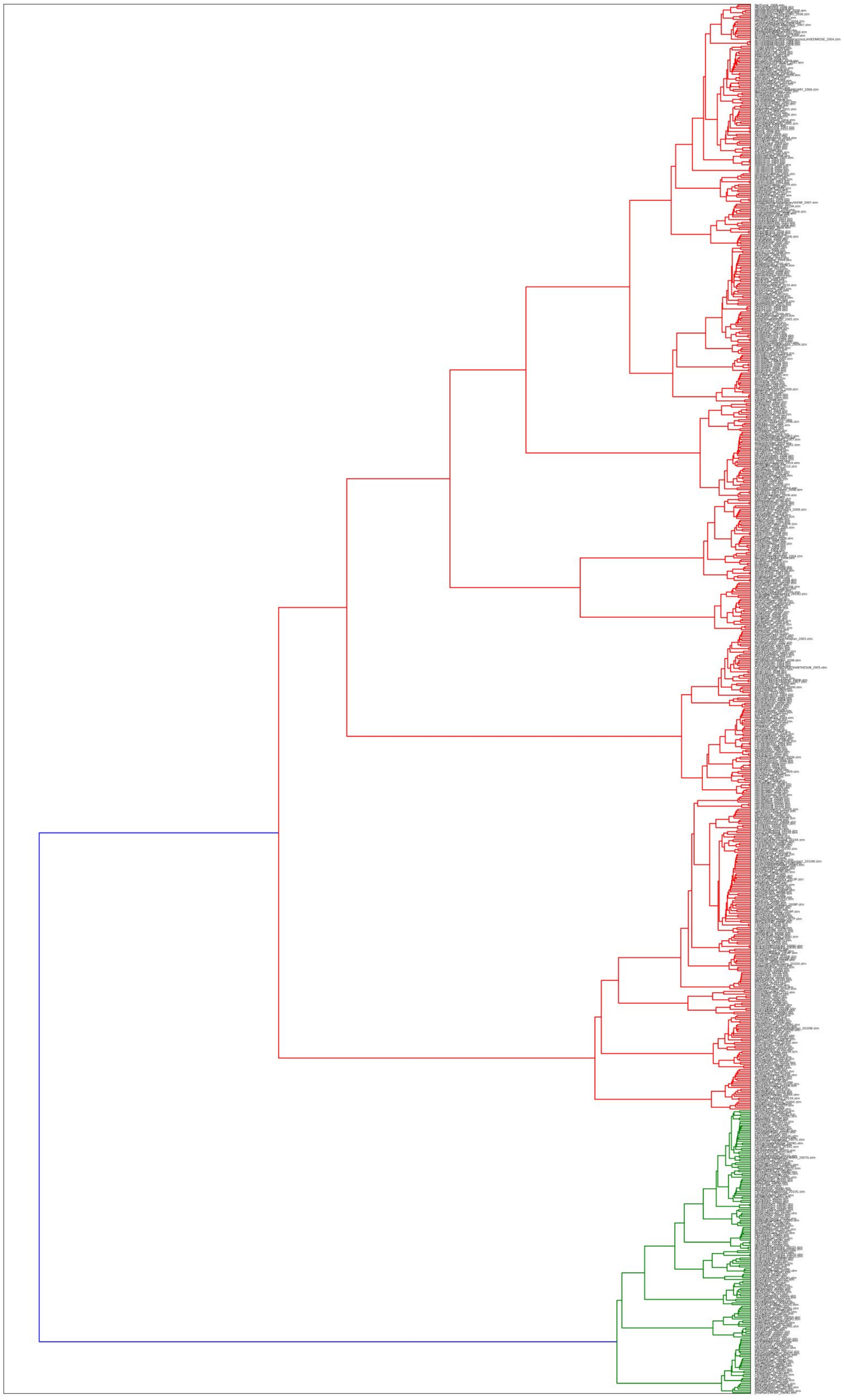
[1] A. Rousseau, P. Delglise, and Y. Estve, “Ted-lium: an automatic speech recognition dedicated corpus,” in *Proceedings of LREC 2012*, may 2012.

[2] E. L. Bird, Steven and E. Klein, *Natural Language Processing with Python*. OReilly Media Inc., 2009.

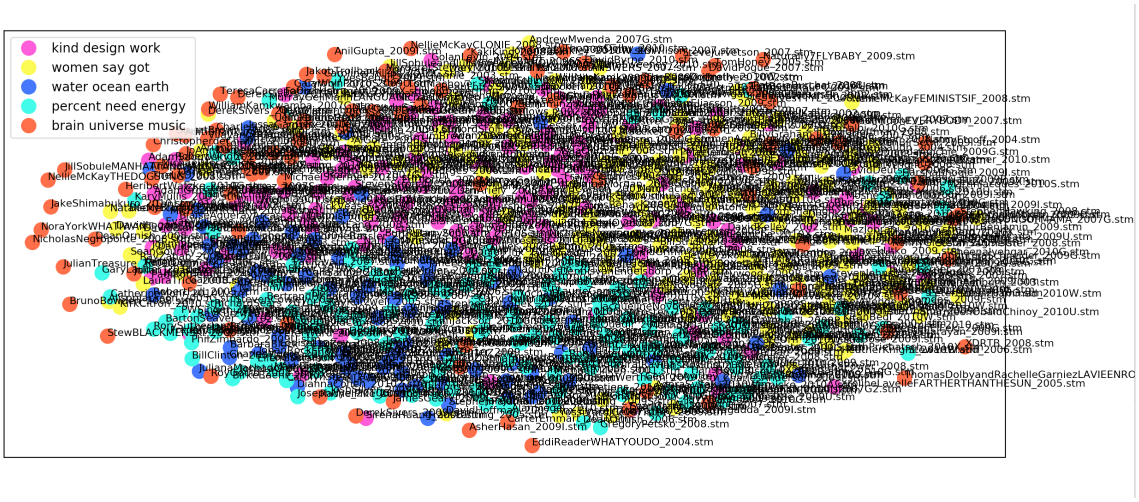
[3] “Supervised and unsupervised learning.” [Online]. Available: dataaspirant.com/2014/09/19/supervised-and-unsupervised-learning/

[4] “Topics.” [Online]. Available: www.ted.com/topics

APPENDIX A
FULL DENDROGRAM



Appendix B
Full K-means Clusters


● kind design work
● women say got
● water ocean earth
● percent need energy
● brain universe music
AndrewMwenda_2007G.stm
AnilGupta_2009I.stm
NellieMcKayCLONIE_2008
FLYBABY_2009.stm
FEMINISTSIF_2008.stm
DerekSivers
JillSobuleMANHATTAN
NellieMcKayTHEDOGSONG
JakeShimabukuro
NoraYorkWHAT
NicholasNegroponte
JulianTreasure
BrunoBowden
StewBLACKMEN
BillClinton
ThomasDolbyandRachelleGarniezLAVIEENRO
ThomasDolbyandRachelleGarniezFARTHERTHANTHESUN_2005.stm
KDRTB_2008.stm
AsherHasan_2009I.stm
GregoryPetsko_2008.stm
EddiReaderWHATYOUDO_2004.stm