I discovered this while scrolling through research papers such as this one:

https://files.eric.ed.gov/fulltext/EJ1064259.pdf

https://towardsdatascience.com/gentle-start-to-natural-language-processing-using-python-6e46c07addf3

Youtube NLTK Python Tutorial: https://www.youtube.com/watch?v=XFoehWRzG-I

Text Extraction and Preprocessing:

Tokenization at 9:07
Is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units are called tokens. It works by separating words using spaces and punctuation.

Additional sources:
https://www.analyticsvidhya.com/blog/2019/07/how-get-started-nlp-6-unique-ways-perform-tokenization/

N-grams at 9:30
Is a simple language model that assigns probabilities to sequences of words and sentences. Helps in search engines and word autocompletes like in Gmail as it identifies the probability of these sequences of words showing up.

This is a sentence is an example of a unigram

This is a sentence is an example of a bigram

Additional sources:
https://towardsdatascience.com/understanding-word-n-grams-and-n-gram-probability-in-natural-language-processing-9d9eef0fa058

Stop word removal at 10:20
Stop words are words that have little to no meaning and significance in the text (“a”, “an”, “and”, “or”, “the”). Helps improve computation speed.

Stemming at 11:53
Removes suffixes and prefixes to find the root word. There are many types of stemming algorithms.

Lemmatization at 12:37
Is similar to Stemming but uses more context. It requires analyzing the POS tag of the word and its surrounding

POS Tagging at 13:04

Classifies and attributes words to their word type (nouns, adjectives, verbs, article)

https://www.youtube.com/watch?v=r9QjkdSJZ2g

https://www.youtube.com/watch?v=xvqsFTUsOmc
https://www.youtube.com/watch?v=X2vAabgKiuM
https://www.youtube.com/watch?v=nla4C-VYNEU&list=PLQVvvaa0QuDf2JswnfiGkliBInZnIC4HL&index=15

Embedding is an important concept in identifying meaning:
At 6:04
https://www.youtube.com/watch?v=Y_hzMnRXjhI&list=PLQY2H8rRoyvzDbLUZkbudP-MFQZwNmU4S&index=3

Might need a recurrent neural network in order to analyse the semantics of the sentence because my project requires context.

Check out LSTM (long short-term memory)

https://www.youtube.com/watch?v=A9QVYOBjZdY&list=PLQY2H8rRoyvzDbLUZkbudP-MFQZwNmU4S&index=5

I need to pad my data, just like in CNNs, we need all data to have the same dimensions and size

Might explore Latent Semantic Data Analysis

Look into the research paper: https://files.eric.ed.gov/fulltext/EJ1064259.pdf
They put the text into a vector and an attribute-value table
They also used cosine similarity to detect if there was cheating involved
They also measured the term frequency: TF-IDF

https://stackoverflow.com/questions/8897593/how-to-compute-the-similarity-between-two-text-documents

Use TF-IDF and cosine similarity to determine if documents are similar:
Turn the text into vector space

Cosine similarity is a technique used to determine how similar two vectors are

https://www.sciencedirect.com/topics/computer-science/cosine-similarity#:~:text=Cosine%20similarity%20measures%20the%20similarity,document%20similarity%20in%20text%20analysis.

We also need to explore data mining:
I need to measure the accuracy and how to properly search for keywords

Logic should be: The student's text is analyzed, with a list of keywords. The cosine similarity is applied to multiple texts. Data mining is enabled using the keywords.

I need to determine some quantitative value to show how the confidence score (Level of cheating).

Need to print out a list of 5 possible sources that the student could have referred to.

Overlap coefficient is another metric used to determine if a document is similar

https://towardsdatascience.com/overview-of-text-similarity-metrics-3397c4601f50 A very detailed list of text similarity metrics in NLP

https://rxnlp.com/what-is-text-similarity-nlp/ Got paraphrase identification and semantic similarity terms from this article.

https://www.researchgate.net/publication/339487284_Paraphrase_Identification
Dice's coefficient, length ratio are other text metrics

"Techniques which utilized
syntactic, semantic or hybrid features in addition to lexical features resulted mostly in performance higher than simple lexical matching based techniques" - 2.3 Summary

What are lexical terms?

https://medium.com/@bedigunjit/simple-guide-to-text-classification-nlp-using-svm-and-naive-bayes-with-python-421db3a72d34
Another interesting article.

I COULD USE SUPPORT VECTOR MACHINES!!! THEY USE FEATURES

What is Wordnet?
A lexical dictionary that seems to identify information and provide synonyms, a bit of context, roots and "synsets".

It is capable of identifying the roots and relations of words

https://www.aclweb.org/anthology/I05-5001.pdf
What are the features that they used for the SVM?

- **Morphological variants** (involves stemming and lemmatization)

- **String similarity**
- **https://stackabuse.com/levenshtein-distance-and-text-similarity-in-python/** (another interesting article)
- involves **edit distance** (also known as **levesheistien distance**) which is the minimum number of operations in order to make the strings similar.
- What I could do is stem all the words. Put them in a string. Apply edit distance. But since some sentences might be short, I divide the edit distance by how many words there are in the sentence. Depending on the value, we could use it as another feature to identify academic dishonesty
- This also involves how many shared words are in the sentence

- **Wordnet lexical matchings** (do the words have the same roots and have common words that relate to them?)

Need to learn what a document term matrix is and how it is involved in cosine similarity

What the new logic should be:

1. Tokenize, Stem, Lematize and remove stop words from the doc
2. Put the text in a Bag of Words or in a document term-matrix
3. Apply cosine similarity and other text similarity metrics to other documents
4. Apply the features noted from the SVM paper (edit distance, wordnet, etc)
5. Need to create an algorithm to compare the student's previous work with this one
    a. I will look into the sentence variety of the student (S, CX, CP C-C sentences) Could use google cloud's syntactic api
    b. The type of vocabulary used. (Compare teacher's previous vocab mark with this essay)

    https://edintegrity.biomedcentral.com/articles/10.1007/s40979-017-0021-6 (additional file 1 and I think 2)
6. Repeat steps 3 to 4 on data mining.
7. Log student's essays into the database.

Feb 14: New logic:

- Preprocessing: Stem, remove stop words, lemmatize. Could use wordnet to search up root words and synonyms. Remove special characters
- Perform Cosine Similarity and edit distance
- Only perform edit distance on sentences that only seem suspicious
- Perform SVM (singular value decomposition) on the vectors

- Possibly use google cloud's syntactic api to analyze sentence variety. Also use the teacher's vocabulary rating as a comparison of the student's previous work.
- Perform data mining

https://www.researchgate.net/publication/329610376_A_Latent_Semantic_Analysis_LSA_approach_for_plagiarism_detection_in_Arabic_documents

Forget everything that was mentioned before: Going to use this one:
https://copyleaks.com/blog/natural-language-processing-for-plagiarism-checker/

Could use spaCy as a good library: https://spacy.io/usage/spacy-101

https://machinelearningmastery.com/develop-word-embedding-model-predicting-movie-review-sentiment/ implementing neural networks for Semantic analysis

Could use Knn:
https://www.ijstr.org/final-print/apr2016/Plagiarism-Detection-Using-Artificial-Intelligence-Technique-In-Multiple-Files.pdf

They used POS tagging
Then Wordnet and Semantic
Then used Machine learning with "Weka"?

https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/

Word embeddings are important? TF-IDF is a simple word embedding technique.

https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/

Thing that mom sent me (Gmail):
It also computes the how similar the n-grams are front the documents
If this word is always followed up by this one, then they are similar.

Word2vec can make accurate guesses on the word's meaning analyzing its context

Have to consider that the documents are of varying lengths:
https://www.aclweb.org/anthology/P18-1218.pdf
What is doc2vec?

Could also use word mover's distance
https://towardsai.net/p/nlp/word-movers-distance-wmd-explained-an-effective-method-of-document-classification-89cb258401f4

Another example good example:
https://medium.com/swlh/sentiment-classification-using-word-embeddings-word2vec-aedf28fbb8ca?source=email-7354e6547acb-1613814492060-digest.reader------0-59----------57ffcdb5_3283_42b8_ae2a_86d64dacc5c7-1-6c816f7e_5935_430d_92ae_eeca4ec9b585----

I would have to make a classification model for each assignment for possibly, each student.
https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/

Doc2vec:https://radimrehurek.com/gensim/auto_examples/tutorials/run_doc2vec_lee.html
It is like word2vec but creates a doc vector that is a compressed or summarised form of the word vectors

I created the doc2vec model. I now need to research how to properly adjust the hyperparameters

I also tested three sentences/docs from this site:

Legitimate paraphrase to plagiarised version = a negative result
Acceptable to legitimate = a 40%
Acceptable to plagiarised = 34ish%

https://medium.com/@amarbudhiraja/understanding-document-embeddings-of-doc2vec-bfe7237a26da

https://israelg99.github.io/2017-03-23-Word2Vec-Explained/

Now I need to add syntax analysis for analyzing sentence structure and to compare a student's previous work. I could also perform a unique word count to show if the student used vivid vocab. By analyzing sentence variety and vocab variety of the student, we could give an approximation of their writing level (Could be a metric or number). We then apply the same system to the current assignment. If the metrics are relatively the same (threshold distance is 10%) then it passes this "non-cheating" criteria.

Now I need to edit distance for sentence structure. In this case, I could rehash my doc2vec knowledge and just apply it to sentences. Suspicious sentences would be put into a dictionary then have edit distance applied to them.

- When comparing sentences I need a “for” loop to iterate over the sentences (I could review my previous colab for that algorithm/system)
- I need to apply cosine similarity to them

I need to explore data mining soon.

I also need to lookup training methods in order to properly tune my hyperparameters Additionally, I need to validate my methods and make some sort of “loss” metric. Depending on the loss, I could tell the computer to redo the training process until it is good enough.\

I might need to visualize the document vectors. I could use T-SNE or PCA:
http://csmoon-ml.com/index.php/2019/02/15/tutorial-doc2vec-and-t-sne/
What I think this does is that it squishes all of the document vectors into a single vector. Like it does, it performs a dimensional reduction. Using this technique, I could just “plug in” an entire class’ essays into the doc2vec neural net to create the vector representations and produce a graph.

But according to this article, we could use tensorflow’s projector tool:
https://towardsdatascience.com/detecting-document-similarity-with-doc2vec-f8289a9a7db7
Documentation about it

Revisiting Latent Semantic Analysis as only using doc2vec would be pretty bad.
https://www.datascienceassn.org/sites/default/files/users/user1/lsa_presentation_final.pdf
https://towardsdatascience.com/latent-semantic-analysis-deduce-the-hidden-topic-from-the-document-f360e8c0614b

It is known as a topic modelling classifier:

Uses BoW then applies a dimension reducing SVD on it. Sees if the sent or doc contributes to either topic 1 or topic 2.
Separate docs that are related to each topic
We then apply the same thing on the separated groups of docs/sents

Another one uses LDA and Jensen-Shannon Distance (JS distance). LDA is another topic model :
Text Similarity Computing Based on LDA Topic Model and ...www.atlantis-press.com › article

https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2

LDA learns words and assigns them topics or labels. On unseen docs it will learn if the words inside of it could be assigned to any document topic.

https://stackabuse.com/implementing-lda-in-python-with-scikit-learn/
According to this doc we could also use PCA if the data is distributed irregularly.

If two docs belong to the same topic, then that is really bad and suspicious!

I learned what lda2vec is from this
https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05
https://towardsdatascience.com/lda2vec-word-embeddings-in-topic-models-4ee3fc4b2843
https://www.slideshare.net/ChristopherMoody3/word2vec-lda-and-introducing-a-new-hybrid-algorithm-lda2vec-57135994
It learns word, document and topic vectors. Toooooo baaaadddd there is not really any documentation!
Very interesting! Could put in conclusion

LDA is a way to figure out readable topics.

Since I have a code that POS tags, I could use the pos tagging information to change all words to its hyponym. Using wordnet, I could specify what POS tag it is and obtain a root word.

Very interesting article:
https://pub.towardsai.net/natural-language-processing-nlp-with-python-tutorial-for-beginners-1f54e610a1a0#e9b8
My project involves syntactic (sentence structure), semantic (LDA, Doc2vec and TF-IDF) and a bit of lexical analysis (pre-processing and wordnet especially).

Should we lemmatize and remove stopwords from corpus for doc2vec? What is the best way to preprocess text?
https://stats.stackexchange.com/questions/374209/pre-processing-lemmatizing-and-stemming-make-a-better-doc2vec

Fed the doc vectors into a logistic regression model

They experimented with certain hyper parameters here:
https://ep.liu.se/ecp/131/039/ecp17131039.pdf

I can't use logistic regression because it is supervised

https://radimrehurek.com/gensim/models/doc2vec.html#gensim.models.doc2vec.Doc2Vec.get_latest_training_loss You can track the training loss maybe? And adjust the vectors?

Other hyperparameter questions asked on stack overflow and github:
https://gensim.narkive.com/k4BeeCNG/gensim-6160-how-to-pre-process-data-tune-hyperparameters-of-doc2vec

https://github.com/RaRe-Technologies/gensim/issues/2983
https://groups.google.com/g/gensim/c/xKvUv-yZI2U

https://stackoverflow.com/questions/62801052/my-doc2vec-code-after-many-loops-of-training-isnt-giving-good-results-what-m That my code used from the other example is bad because essentially the learning rate is being adjusted towards the wrong direction

https://gensim.narkive.com/Ew2q1Q86/gensim-6126-how-to-get-the-document-vector-from-doc2vec-in-gensim-0-11-1

I am having trouble specifying the number of topics in a document. If I can specify a correct amount for LDA, the best case scenario would be that there are 25 separate topics (meaning everyone has original work). The Worst case is 2 or 3 meaning everyone copied from 2 people. I could use HDP to calculate how many topics are there in a set of documents.

HDP ---> number of topics suggested (denoted as x)

LDA(num of topics = x) -----> I investigate thoroughly the student work that is grouped on the same topic.

https://towardsdatascience.com/unsupervised-nlp-topic-models-as-a-supervised-learning-input-cf8ee9e5cf28

A research paper explaining HDP with more detail:
https://people.eecs.berkeley.edu/~jordan/papers/hierarchical-dp.pdf

HDP documentation: https://radimrehurek.com/gensim/models/hdpmodel.html

I had 150 topics for only 3 sentences! Yeash! Well, according to this stack overflow question, I might need to evaluate the quality of the topics presented (aka, topic coherence). This article can maybe compute it:
https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/#13viewthetopicsinldamodel

https://radimrehurek.com/gensim/models/coherencemodel.html

Reading this article, I might use a Mallet LDA model!

Example of LDA mallet implementation:
https://www.tutorialspoint.com/gensim/gensim_creating_lda_mallet_model.htm

While judging coherence score, it is important to look at qualitative and quantitative factors:https://medium.com/square-corner-blog/topic-modeling-optimizing-for-human-interpretability-48a81f6ce0ed

Good lda article:
https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0

https://towardsdatascience.com/nlp-extracting-the-main-topics-from-your-dataset-using-lda-in-minutes-21486f5aa925

https://stackoverflow.com/questions/43357247/get-document-topics-and-get-term-topics-in-gensim referenced this for get document topics method()

https://stackoverflow.com/questions/52876014/gridsearch-for-doc2vec-model-built-using-gensim gird search for hyperparameter tuning

Could summarize docs with TextRank algorithm. I learned this from here:
https://www.machinelearningplus.com/nlp/gensim-tutorial/#17howtocreatedocumentvectorsusingdoc2vec and here: https://rare-technologies.com/text-summarization-with-gensim/

Officials text rank paper:
https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf

Doc2vec is no longer good when the corpus is a few docs long

https://par.nsf.gov/servlets/purl/10091276

http://m-mitchell.com/NAACL-2016/NAACL-HLT2016/pdf/N16-1108.pdf
https://medium.com/@adriensieg/text-similarities-da019229c894

I could also use a naive bayes classifier.

From pluralsight, we learned stylometry. It analyzes the writing of an individual: It is actually a combination of lexical and syntactic analysis!

https://www.researchgate.net/publication/324201958_Stylometry-based_Approach_for_Detecting_Writing_Style_Changes_in_Literary_Texts

What I can do is use the burrow's delta method.
Analyze sentence structure of the student
Analyze vocabulary and any distinct words
Also analyze function words

Could also create a multi label classification model which, as input, are a vector containing a bunch of the features named above.

Could also use a k-means or k-NN and try to represent all of the features into one single data point on graph space.
This does work according to this article:
https://www.ai.uga.edu/sites/default/files/inline-files/hollingsworth_charles_d_201208_ms.pdf

List of features:

- Delta method: https://programminghistorian.org/en/lessons/introduction-to-stylometry-with-python
- POS trigrams and bigrams: https://github.com/jabraunlin/reddit-user-id, https://www.ai.uga.edu/sites/default/files/inline-files/hollingsworth_charles_d_201208_ms.pdf https://mail.google.com/mail/u/1/#starred?projector=1
- Frequent chunks: https://mail.google.com/mail/u/1/#starred?projector=1
- Frequent function words: https://arxiv.org/pdf/1406.4469.pdf proved to be useful. It also almost achieved a perfect score according to this one: https://www.researchgate.net/publication/282309148_Using_Function_Words_for_Authorship_Attribution_Bag-Of-Words_vs_Sequential_Rules and this one: https://era.ed.ac.uk/bitstream/handle/1842/6638/Horton1987.pdf?sequence=1&isAllowed=y
- This explains function words in detail too: https://www.researchgate.net/publication/301404098_Function_Words_in_Authorship_Attribution_From_Black_Magic_to_Theory
- A compound form N-grams: https://mail.google.com/mail/u/1/#starred?projector=1
- Largest subsequence: https://mail.google.com/mail/u/1/#starred?projector=1
- Context free grammar and : https://www.researchgate.net/publication/262241380_Stylometric_analysis_of_scientific_articles
- Edit distance: SVM research paper
- Punctuation, mean word length, mean sentence length: https://www.researchgate.net/publication/324201958_Stylometry-based_Approach_for_Detecting_Writing_Style_Changes_in_Literary_Texts

Vocab richness:

- https://github.com/Hassaan-Elahi/Writing-Styles-Classification-Using-Stylometric-Analysis
- I could use Yule's K characteristic which measures lexical diversity within a document: https://swizec.com/blog/measuring-vocabulary-richness-with-python
- Or Measure of Textual Lexical Diversity: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3813439/

https://www.aclweb.org/anthology/W05-0207.pdf

Could use stanford's parser NLP

First test:
Took the top 5, excluding sample F in this case.

Went pretty 50/50

- It did catch the first one (target C) but went wrong in a bunch of other areas. It could have been due to the fact that syntactic and lexical based approaches were weighed just about the same as semantic features.

- Jaccard and TF-IDF thresholds were maybe too high (both 50%). Could reduce to 30% or 40%
- Topic, TF-IDF, Jaccard and will be weighted
- Next, I will adjust the weights so that some will say 0.5 and others, just a 1
- Lexical diversity needs a metric change or needs to be replaced. The differences vary a lot.
-

C: Gardening in mixed beds is a great way to get the most productivity from a small space. Some investment is required, to purchase materials for the beds themselves, as well as soil and compost. The investment will likely pay-off in terms of increased productivity.

D: Performance management involves more than just conducting annual performance evaluations. In fact, many companies have done away with formal performance appraisals altogether. Instead, they opt for one-on-one dialogues between managers and employees on a quarterly or monthly basis.

```
[False, [3.9801044390438367], [1.0], 68.09347455488026, 5, -1, 1.0, 0,
3, 0.31872860911209655, 0.3225806451612903, False]
topic: 1.0
True 5
```

E: If you don't have a lot of space for a garden, raised beds can be a great option. Gardening in mixed beds is a great way to get the most productivity from a small area. Some investment is required. You'll need to purchase materials for the raised beds themselves, as well as soil and compost. The investment will pay off, though, in the form of increased productivity.

```
[False, [0.8809726386541934], [-0.6666666666666661],
132.88462668437455, -9, -5, 0.0, -2, -34, 0.09404396943942184,
0.08571428571428572, False]
topic: 1.0

 True 5
```

F: When your focus is to improve employee performance, ongoing dialogue between managers and their direct reports is essential. While performance management often involves conducting annual performance evaluations, it does involve more than just that.

```
[False, [5.847546194724712], [-3.666666666666666], 117.5681124286343,
8, 0, -3.666666666666666, 1, 8, 0.09404396943942184,
0.08571428571428572, False]
topic: 0.0

 False 3
```

B- When your focus is to improve employee performance, it's essential to encourage ongoing dialogue between managers and their direct reports. Some companies encourage supervisors to hold one-on-one meetings with employees as a way to facilitate two-way communication.

[False, [3.8475461947247123], [-5.666666666666666], 98.76716004768195, 3, 1, -5.666666666666666, 1, 4, 0.46459393544478556, 0.3559322033898305, False]
topic: 0.0

False 4

B - target

Many companies don't do formal performance appraisals anymore. Instead, they encourage one-on-one dialogues in which supervisors hold meetings with employees as a way to facilitate two-way communication and one-on one dialogue. These are done on a quarterly or monthly basis.

A = test

Can't use Jaccard distance on sentences because it takes too long! Nor genism's text rank summarize! Takes 3 minutes to process one result for a webpage. It is due to the length of the webpages.

[True, [-1.8896391165871567], [0.982905982905983], 31, 32, 17, 0.982905982905983, 5, 120, 0.13996406259648272, 0.14046822742474915, {'Many children get made fun because of their clothes.': 'Many children get made fun because of their clothes.', 'One of the most important things among students is style.': 'One of the most important things among students is style.', 'There are some people who cannot afford to buy the latest designer clothes which results in negative comments, feelings low self worth and ultimately it may lead to bullying.': 'There are some people who cannot afford to buy the latest designer clothes which results in negative comments, feelings low self worth and ultimately it may lead to bullying.', 'People often claim it should not matter what you wear but unfortunately it does.': 'People often claim it should not matter what you wear but unfortunately it does.', 'This discrimination based on clothing choices may be just the starting point to not fitting in or it may be the catalyst that breaks a junior high career; either way it can be avoided entirely with the help of a common wardrobe.': 'This discrimination based on clothing choices may be just the starting point to not fitting in or it may be the catalyst that breaks a junior high career; either way it can be avoided entirely with the help of a common wardrobe.', 'With formal school attire, there is no

discrimination and intimidation.': 'With formal school attire, there is no discrimination and intimidation.', 'People may stand out for other reasons but the clothing is not the reason, so hopefully some bullying will be diminished.': 'People may stand out for other reasons but the clothing is not the reason, so hopefully some bullying will be diminished.', 'Clothing also often reflects who you are and who you socialize with.': 'Clothing also often reflects who you are and who you socialize with.', 'You may even sometimes avoid a particular person because of what they are wearing.': 'You may even sometimes avoid a particular person because of what they are wearing.', 'Some clothing ultimately defines you and not in the best way.': 'Some clothing ultimately defines you and not in the best way.', 'A common wardrobe will encourage individuals to find another way to reflect their personality and hopefully people will find out more about you than your clothing choices.': 'A common wardrobe will encourage individuals to find another way to reflect their personality and hopefully people will find out more about you than your clothing choices.', 'Uniforms require students to get to actually know each other by finding out about whom they really are and, lesson the actual amount of judging that is going on in schools nowadays.': 'Uniforms require students to get to actually know each other by finding out about whom they really are and, lesson the actual amount of judging that is going on in schools nowadays.', 'Not only do the uniforms create equality, they can also make you look more sophisticated.': 'Not only do the uniforms create equality, they can also make you look more sophisticated.'}, 0.003932744926211875, False]

Another set of tests: this time, more structured and documented. Docs:

Sets that are +1 each 9 (set 1):
Goal: All non plagiarised docs should be excluded from the final list.
Adi = g2pC (Copy) task A
Carl = g2pA (Non)
Bill = g4pC (Copy)
Mike task = g2pB (Non)

John = g2pE (copy heavy revision)

Ben - g4pD (light)

Tyler= gp0pA (Non)

Findings:

- Turns out Jaccard needs to be set from 0.4 to 0.3, as it could not detect sus sentences because of the threshold.
- Model was correct in detecting the source of Ben and Fred.
- Model was incorrect on keeping certain documents alone like Detecting Harry and such .
- Too many different keywords I got to set a cap
- In the original to Ben example, even though it was copied, the difference was really too high like -20. I will set it to 20 or 10 next time
- WMD distance is a good indicator as it can separate it from the rest. Same with the TF-IDF vals. WMD might need more testing because of its inconsistencies (will save modifying for another time).
- I should take the previous suggestion to weigh certain points. TF-IDF, jaccard, WMD should be weighed the highest. Another reason to weigh them because just adding +1 to each section makes the results inaccurate and could accidentally catch “innocent” docs
- I should put all keywords to 5 as the keyword extraction varies a lot

Test with points set 2:

Findings:

- I should change the list of function words to 3 as, in reality only a student might use 3.
- The values still vary a lot
- Might add bias/weights to more features

Mike to Hitch

Many children get made fun of…

Having a crash…

Scores were too high because the syntactic and lexical features weighed just as much as the other ones. The above example was denoted True despite being very opposite because of the weight features.

I ran the test again and with weighted features

Adjusted the weights so that the semantic are weighted a lot more.

WMD was always less that 1 it was around 0.003 for the suspicious ones

The function word frequencies did not require changing
Number words were changed from 2 to 3 threshold

Due to the random frequencies I get from documents, I will consider using Google's pre-trained word2vec model.

I need a way to visualize (graph my data). Only features that could be put into vector format are vocab/sent count, std dev, mean, maybe punctuation frequency and function words, tf-idf vectors.

Some noise could have been created from the vocab richness. I thus adjusted the minimum count to 0, stopwords to english, and strip accents. Might as well do the same for TF-IDF and for word2vec.

Additionally, I will also take measures to use a pre trained word2vec model instead of a customized one. This could be the cause of why the WMD values are ranging too much. Another reason is, after reflecting, like in every Machine Learning Scenario, I don't have a lot of data (only 2 documents)! Need to test on maybe three documents to see if Google or a custom one is better.

Could use T-SNE for plotting for word2vec vectors.

Perform preprocessing on keyword extraction, mover's distance, LDA, TF-IDF and Countvectorizer.

Keywords are non dependable if the document is too short.

Adding preprocessing makes WMD more accurate and went from 0.05 to 0.03 on the law documents. After testing, I received readings of 0.15 and 0.04 on similar docs. The rest were around 0.25. I will set the threshold to 0.2. I did more testing and research and just noticed that I should not preprocess because it removes the context words required for the widow parameter used in the mode. I did more testing and found that adjusting the vector size to 300 or 100 did not make any difference.

- More testing revealed that dissimilar documents present readings of 0.1 or higher in the word mover's distance

Jaccard similarity: Docs that were really similar presented a score for 0.5 or higher. Harder cases were 0.35. I'll set the threshold to 0.3

TFIDF: Docs that were really similar were either 0.4 or above. The hardest docs were, like Jaccard, 0.35 or 0.33 similarity courses.

Decoded a bug: Sentences in the documents were similar.

I noticed that some words like "previous martial" were one word and could not use an "if" statement to detect if martial was actually in the word.

Setting the function word threshold to 3 did have a substantial impact on the result. Made it inaccurate, thus I will change it back to 4

Turns out TF-IDF scores and Jaccard are really good indicators of plagiarism. Will set them to +2. Did not do it with WMD (just a +1 for now) because of its inconsistencies.

Changing Jaccard to +1.5 now. It's because of documents that have the same start and stop.

Test with points set 3:

Findings:

- They are values are little bit more accurate normalized

- Still misdetects documents because the score might be too high?
- Also, topic modelling should be shot down to 0.5, as if this was a real test case and all students are answering a text, the actual topic/subject they are responding to might be the same
- I will adjust the everything to 0.25, Jaccard and TF-IDF will be reduced to 1.5
- I will set the increase the final point threshold to 5 instead of 4 because all the sscords were too high
- Word count might be set to 10 to -10
- Jaccard should be put back to 0.4 for sentences. It detected sentences that were not suspicious.

Test with points set 4:

- It is finally accurate, and leaves the non copied docs alone while detecting Ben and Fred to the originals.
- I will test it with other docs

Other test with set 4 with the following names:
6/7

Adi = g2pC (Copy) task A
Carl = g2pA (Non)
Bill = g4pC (Copy)
Mike task A = g2pB (Non)

John = g2pE (copy heavy revision)

Ben - g4pD (light)

Tyler= gp0pA (Non)

Was 100% Accurate. Did not include the non copied texts almost all of the copied texts. But what is concerning is that some that are partially (light and heavy revision) were attributed to the near copy documents.

Other test with Task B:

Original (TASK B)

Adi = g0pA (Copy, heavy)
Ben = g0pB (Non)
John = g0pC (Non)
Mike = g1pD (Copy)
Fred = g4pD (Copy)
Dan = g0pD (Copy of a few sentences)

Caroline = g4pE (Copy)

Again, 100% accuracy. Same results, and same concerning factor. Partial ones are being compared to partial ones and entire ones.

Test 3 Oxford papers:

100%, did leave the non plagiarised examples alone. A bit concerning because some plagiarised texts were not paired up with the “source” doc, but paired with other copied/cheated texts.

Test 4. The only plagiarism example provided to me by a teacher:

- The TF-IDF values were low
- The WMD was far
- But it did catch a lot of copied sentences.
- I actually might give Jaccard a +3 or a + 2 to compensate for this. Like, if two sentences are detected give a +2 or +3. And it goes higher for each time there are more similar sentences in the documents.

Did a test with case 1 (task a in the corpus) with set 1, 2, 3, and the finalized set.

1 and 2 and 3 all included Carl and Tyler as copiers despite being tagged as no plagiarism.

Other test with Bowdoin examples:

- Set 1: Very inaccurate Why? Because of syntax features. LIke the deviation feature. Even though the difference is zero, they are not the same and it adds more points to it.
- Set 2, 3 and final: Accurate

Why can I go grocery shopping, but I'm not allowed to gather with family and friends?

As provinces across the country continue to reopen, we're hearing from many Canadians who are wondering why some activities are allowed while others remain off-limits.

Lara B. wrote to ask why it is "acceptable to gather with hundreds of strangers inside a grocery store" but that meeting with family and friends is still prohibited.

It's important to note that in some provinces, such as New Brunswick, Nova Scotia and Newfoundland and Labrador, you can choose one family to be in physical contact with, but only if you agree to be exclusive.

In Lara's home province of Ontario, however, households are still being asked to stick to themselves. So why is family contact a no-go but grocery stores are OK? The answer comes down to varying risks of contact.

"The kind of contact you have with people who are your family and friends tends to be much closer and [more] prolonged than walking by someone in a [grocery] store," said Dr. Lisa Barrett, a professor at Dalhousie's medical school and an infectious disease researcher.

"That casual contact is much less risky than the kind of contact we have with family and friends."

Infectious disease specialist Dr. Isaac Bogoch agrees. "A lot of the data that's emerged show the greatest risk of getting this infection is in indoor environments where people are close together," he said.

Most grocery stores have implemented measures to minimize the spread of the coronavirus, such as limiting the number of people they let into the store at a time, installing Plexiglas in front of cashiers and placing arrows on the floor to direct traffic and enforce distancing.

If you're seeing a smaller grocery store with "100 or more people" you should let somebody know, Barrett says, because that is "too many people."

Can flatulence carry COVID-19?

By now you've probably heard about the importance of keeping our coughs and sneezes to ourselves, but what about our other gases?

Jill G.'s grandson wants to know if flatulence can carry the virus. "After laughing, we agreed that it was a valid question," she wrote.

The question also made some of our experts smile, but they all agreed that it would be unlikely for farts to spread COVID-19.

Infectious diseases physicians Dr. Sumon Chakrabarti and Dr. Zain Chagla both said "very small" amounts of the virus can be found in the stool of a few people.

Dr. Lynora Saxinger, an infectious diseases expert at the University of Alberta, said new research suggests some of that virus could be potentially viable or cultivatable.

But the real question, she said, is whether there is enough virus in stool — and thereby in flatulence — for you to inhale.

"The volume of emissions [when someone passes wind] is much lower than … the air from your chest," said Saxinger. Your clothes may also provide an extra layer of protection.

"There is a bit of literature on this with bacteria, suggesting that underpants will filter bacteria out of flatulence."

Ultimately, she said, the risk would be "really negligible and not be a big concern."

Can I get coronavirus through an open cut?

Janice R. wrote in asking if you can get the virus from a cut.

"No, it's not possible to contract it this way," said Chakrabarti, who is an infectious diseases physician at Trillium Health Partners.

"The virus has the ability to enter the body only through respiratory mucous membranes, which are not present with a cut."

Simply put, the types of cells the virus can bind to are "quite specific," and skin or blood cells likely won't set up an infection, Saxinger said.

"The initial steps of infection generally involve the virus contacting a respiratory, eye, mouth or nose membrane surface, binding itself to specific cell receptors and entering the cells to set up infection," she said.

While the virus — or traces or fragments of the virus — can "maybe be found in the blood," Saxinger said that it doesn't appear to be a blood-borne infection.

"People need to focus on protection from respiratory droplet spread, like breathing in close quarters."

She said it's important that people are mindful that the virus can travel from their hands to their face, and therefore come in contact with membranes in their eyes, nose and mouth.

OVERALL CONCLUSIONS.

- Can detect direct copying and docs that have had partial copies
- For paraphrased examples, the model finds difficulty to properly identify them.
- It is also stil faulty, as show when “Mike” was apparent even with the final threshold list,
- I require more testing in regard to my threshold values and possibly re-shaping the model architecture itself.
- The accuracy of my model is roughly 70%