**In-Depth Analysis of Abstracts within Computer Science, Ecology, Chemistry, and Biology**

Jeffrey Cruz

Pace University: ENG 201

Dr. Kristen di Gennaro

Genre Analysis Project

**Introduction:**

The abstract is an essential and arguably the most valuable part of any article or research paper. It is often the only part of a paper that will get read, as most people tend to just skim them to get the main idea of what the paper is about. Even if it is not the only part that is read, it is most definitely one of the first parts that will get read, and will determine if the reader dedicates the time to go forward in the paper. For this reason, I see it prudent to do an in-depth analysis of abstracts across various disciplines. For this analysis, I have collected abstracts of research papers in the fields of Computer Science, Ecology, Chemistry, and Biology. In this analysis, I will calculate various metrics of these abstracts see how these metrics vary across the disciplines. This analysis will also include various machine learning models that will be trained to classify which discipline an abstract is from. The accuracy of these models will tell us which metrics or features are most unique or predictive of each discipline. Since the goal of this project is to analyze the structure and contents of abstracts across disciplines and not machine learning, the details of the machine learning model and algorithm themselves will not be explained. It might help for any reader to just imagine it as a black box that takes in the features of an abstract, and gives out a prediction for its belonging discipline.

The overall goal of this analysis is to determine whether or not abstracts are distinctive among these different disciplines. If the abstracts are not distinctive enough, then the machine learning model will have low accuracy prediction, meaning it cannot tell apart abstracts of different disciplines. However, if we find that certain metrics are unique to certain disciplines, then the machine learning model will be able to make accurate decisions.

**Methodology:**

My research is done using a collection of abstracts from various sources. I use 3323 Computer Science Abstracts from papers submitted to the Conference on Neural Information Processing Systems, 995 Biology Abstracters from papers submitted to pubmed.gov, 3000 Chemistry Abstracts from the Journal of the American Chemical Society, and 601 Ecology Abstracts from crowdflower.com (Which itself collected the abstracts from various online databases such as PubMed). This totals to 7919 abstracts. All of my analysis and calculations are done using the Python programming language, which makes processing the vast amount of abstracts extremely easy to do. For each abstract all the metrics I can think of that have to do with the construction of an abstract will be calculated. These metrics include number of words, number of unique words, number of stop words (filler words that add no extra meaning to a sentence such as “a” and “the”), and number of punctuations.

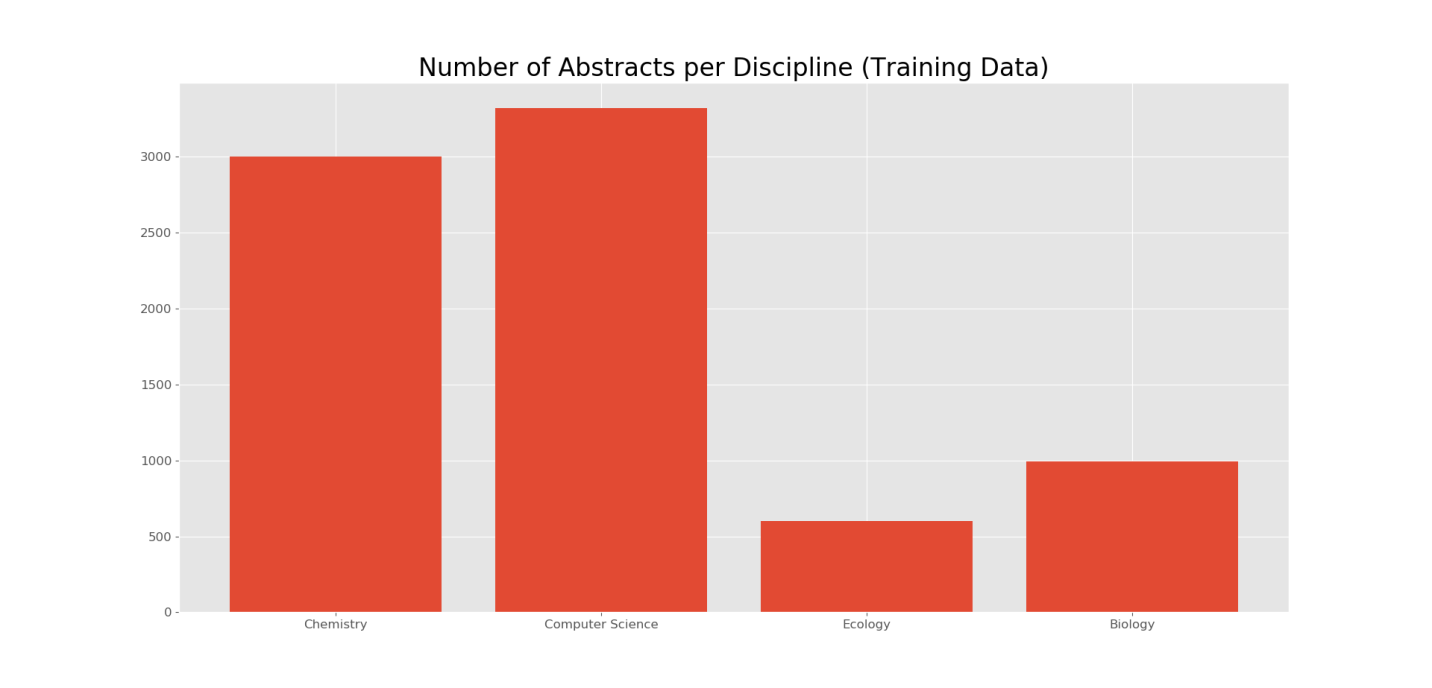
Using these metrics as features, I construct a machine learning model using a python module called XGBoostClassifier. The machine learning model is fed the metrics from all of my abstracts, herein forth called “Training Data”, and is then passed a separate set of abstracts, herein forth called “Testing Data”. The testing data contains 601 Computer Science Abstracts, 399 Biology Abstracts, 401 Chemistry Abstracts, and 201 Ecology Abstracts. This totals to 1602 abstracts in the testing dataset. All the sources for the abstracts in the testing data are the same as the ones in the training data for each discipline. The goal of the machine learning model is to correctly predict which discipline each abstract from the testing data belongs to. It is not known to the model beforehand the discipline of the models in the test set, it must make the decisions based solely on the metric calculations. The metrics mentioned above will also be done on the testing set, and this is what the machine learning model will base its predictions on. This will tell us if the metrics are predictive of the disciplines, for example if a high number of punctuation is indicative of a biology abstract, or if a low number of unique words is indicative of a computer science abstract. This is the kind of results we want to see.

A 2nd machine learning model will be made using the same python module, however the data it is trained on will have one extra layer of depth added. I will create a one hot bag of words encoding of the 100 most common words among all of the abstracts in the training data (not including stop words). What this means is that within in the training dataset, there will be 100 extra columns added representing the 100 most common words, that have a value of 0 or 1 in each row, which indicate whether that word exists or not in that abstract. This calculation will also be done on the testing data, still using the 100 most common words from the training data. This will tell us if certain words appearing are indicative of a certain discipline. I expect this to boost the accuracy of the machine learning model considerably, since each discipline should have distinct vocabulary. If this machine learning model performs considerably better than the first, then this would tell us that the overall structure of abstracts among disciplines is similar for the most part, with the only difference being the content of the abstract itself, which is what the first machine learning model didn’t have knowledge of.

**Analysis:**

This is the graph showing the number of abstracts per discipline in the training data. We see that it consists abundantly of Computer Science and Chemistry abstracts, while having significantly less of Ecology and Biology. While relative to each other, the first two disciplines may dwarf the latter 2, I believe that there enough abstracts in each to be representative of their discipline, the higher amount of papers in Chemistry and Computer Science can only benefit, but should not take away from the results for Ecology and Biology.

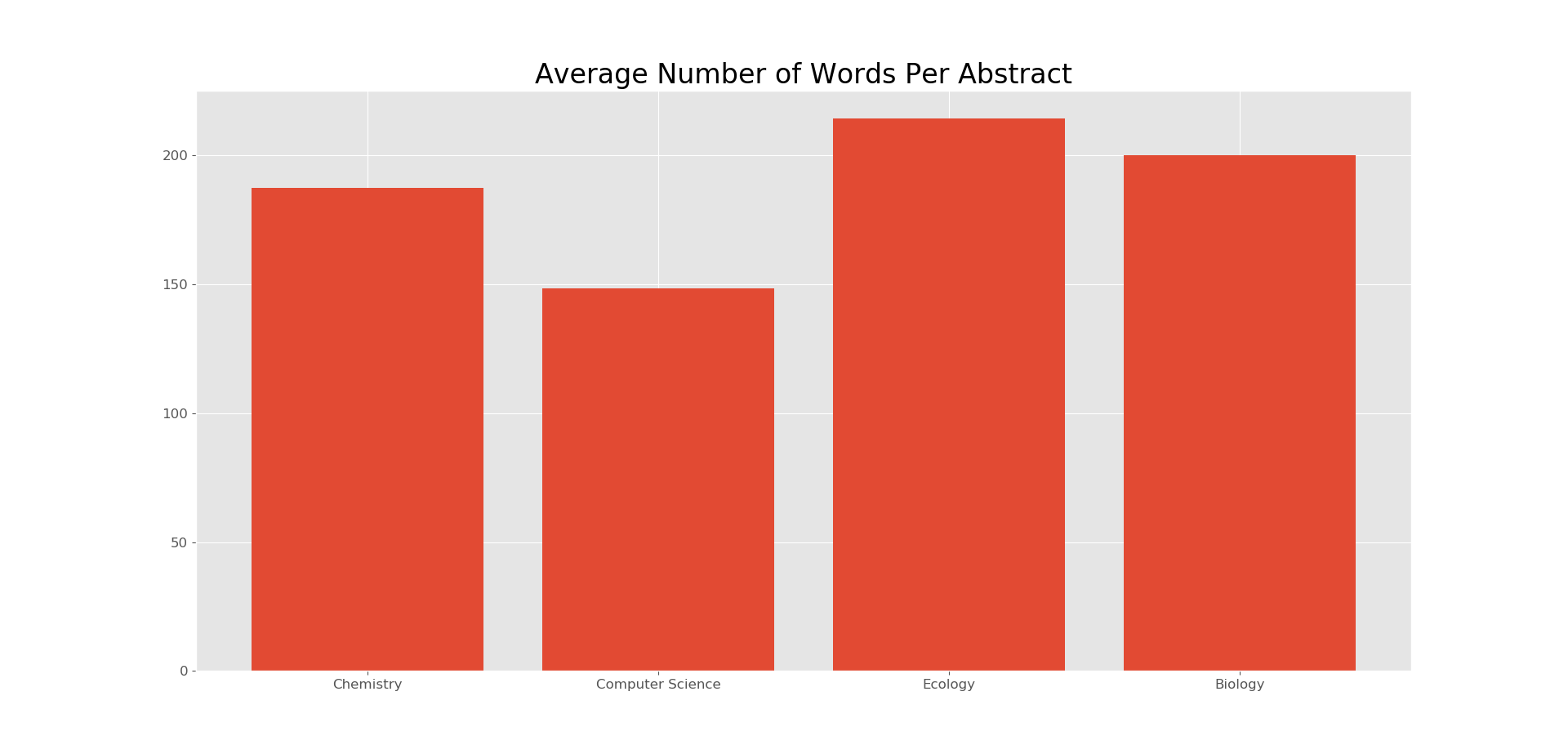
Figure 1: Number of Abstracts per Discipline:

****

These are the graphs showing the average of these metrics for all of the abstracts of each discipline in the training set. Please note the scale of the y-axis for each graph, they were deliberately changed in order to keep the relative size of each graph the same.

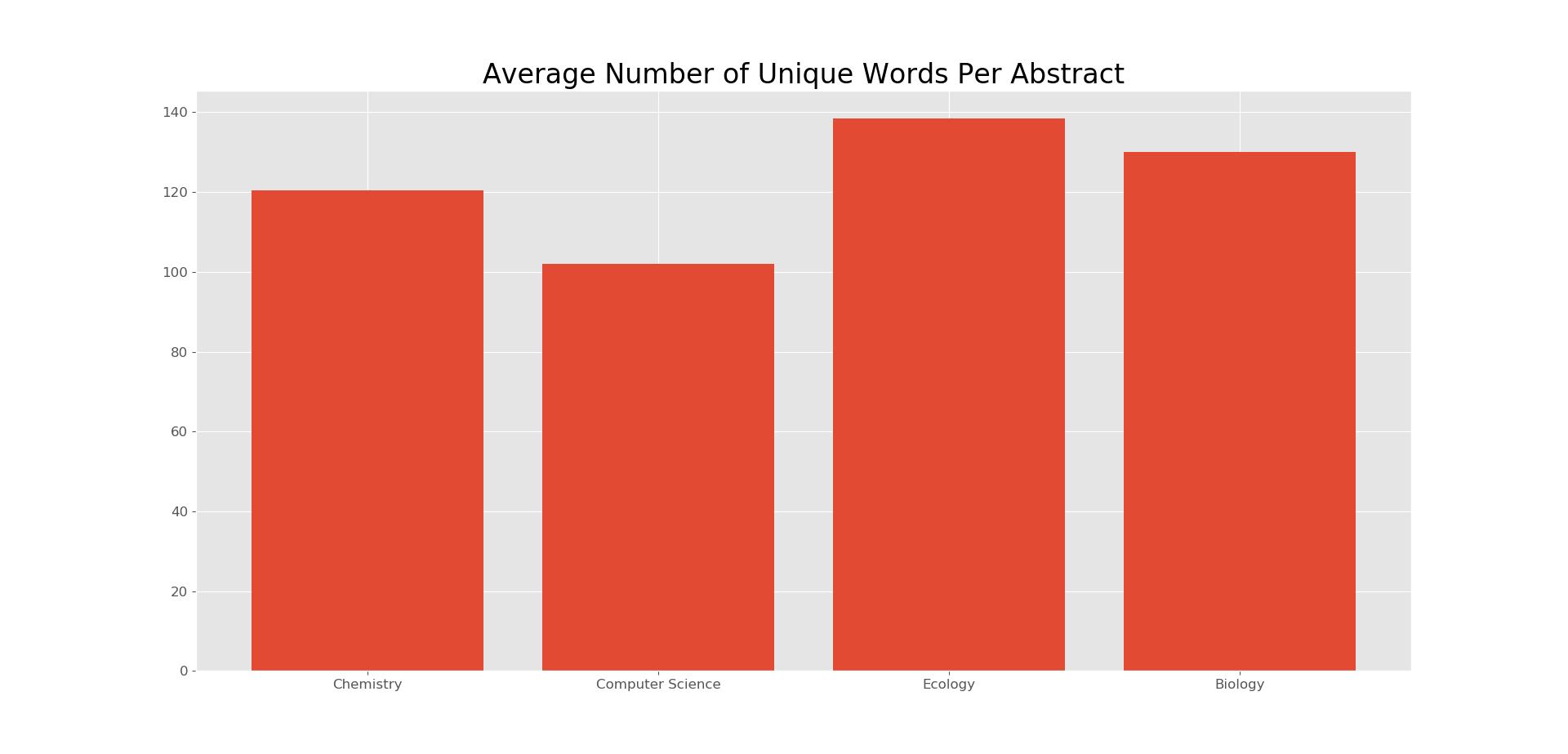
The average number of words per abstract are 187.42 for Chemistry, 148.26 for Computer Science, 214.34 for Ecology, and 200.1 for Biology. This may indicate an abstract with a low word count is indicative of Computer Science, however for the other disciplines the average number of words is relatively similar.

Figure 2: Average Number of Words per Abstract

****

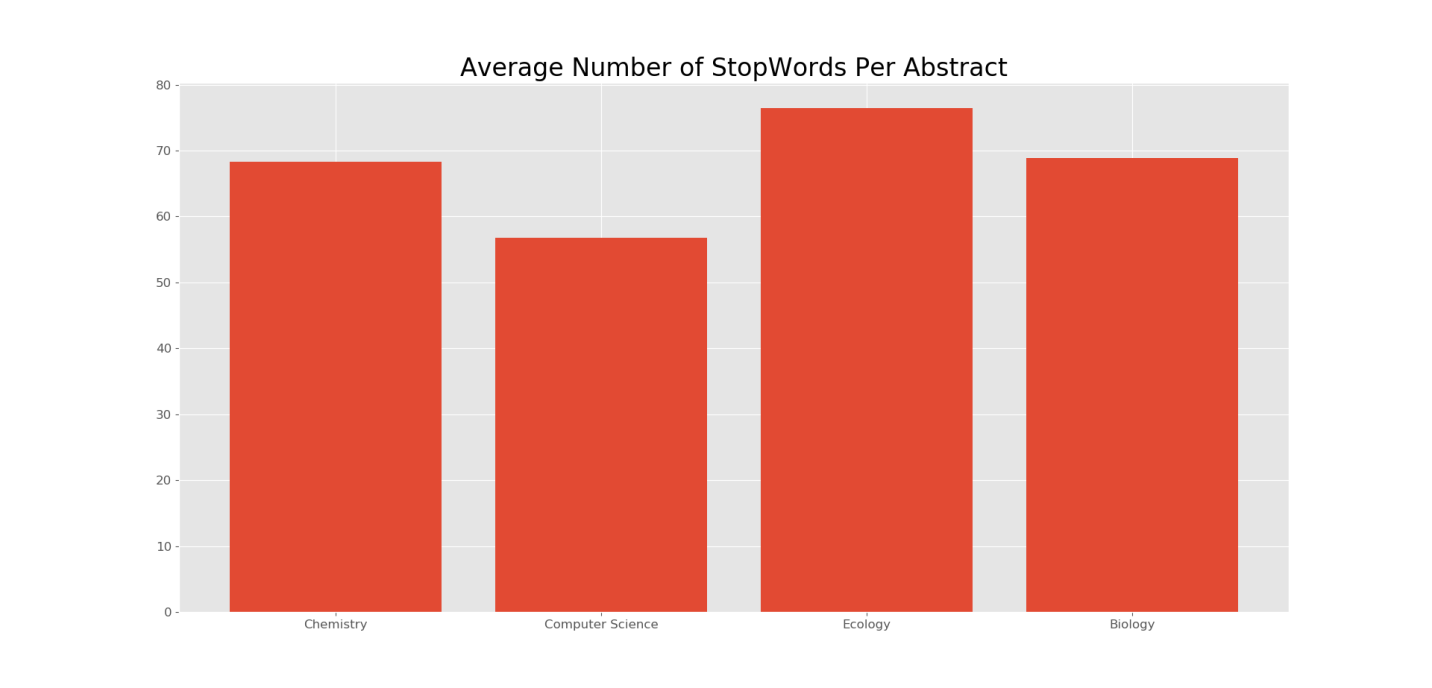
The number of unique words per abstract for each discipline is 120.381 for Chemistry, 102.03 for Computer Science, 138.51 for Ecology, and 130.08 for Biology. The distribution of this metric similar to the first graph, meaning this may be a redundant feature since its implications are the same as the number of words. A low amount of unique words may be indicative of computer science, while for the other abstracts the values are relatively similar.

Figure 3: Average Number of Unique Words per Abstract

****

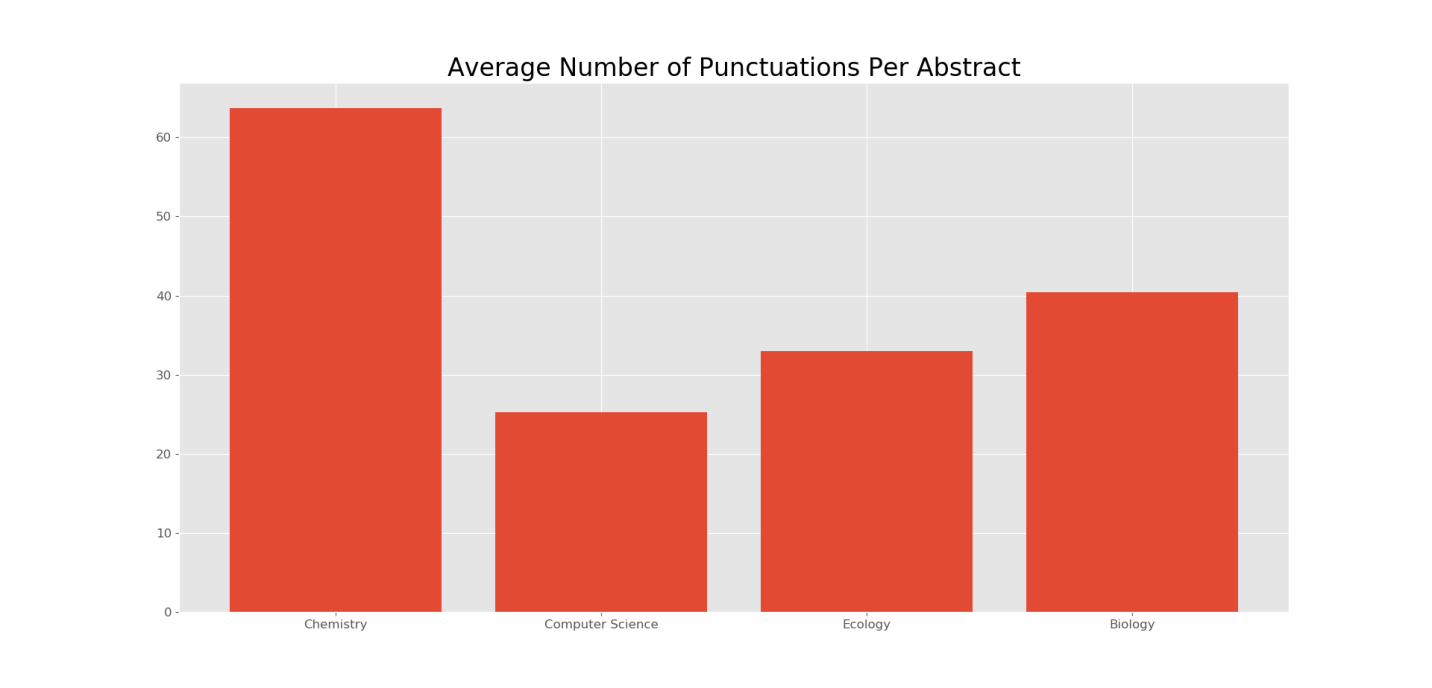
Stop words are natural language words which have very little meaning, such as "and", "the", "a", "an", and similar words. The average number of stop words per abstract for each discipline is 68.28 for Chemistry, 56.78 for Computer Science, 76.48 for Ecology, and 68.84 for Biology. Once again, we see that this graph is similarly distributed to the first two. A low amount of stop words may be indicative of computer science, while for the other disciplines they are similar.

Figure 4: Average Number of Stop Words per Abstract

****

The average number of punctuations per abstract is 63.72 for Chemistry, 25.26 for Computer Science, 32.96 for Ecology, and 40.44 for Biology. This graph tells us that a high amount of punctuation may be indicative of Chemistry abstracts.

Figure 5: Average Number of Punctuations per Abstract

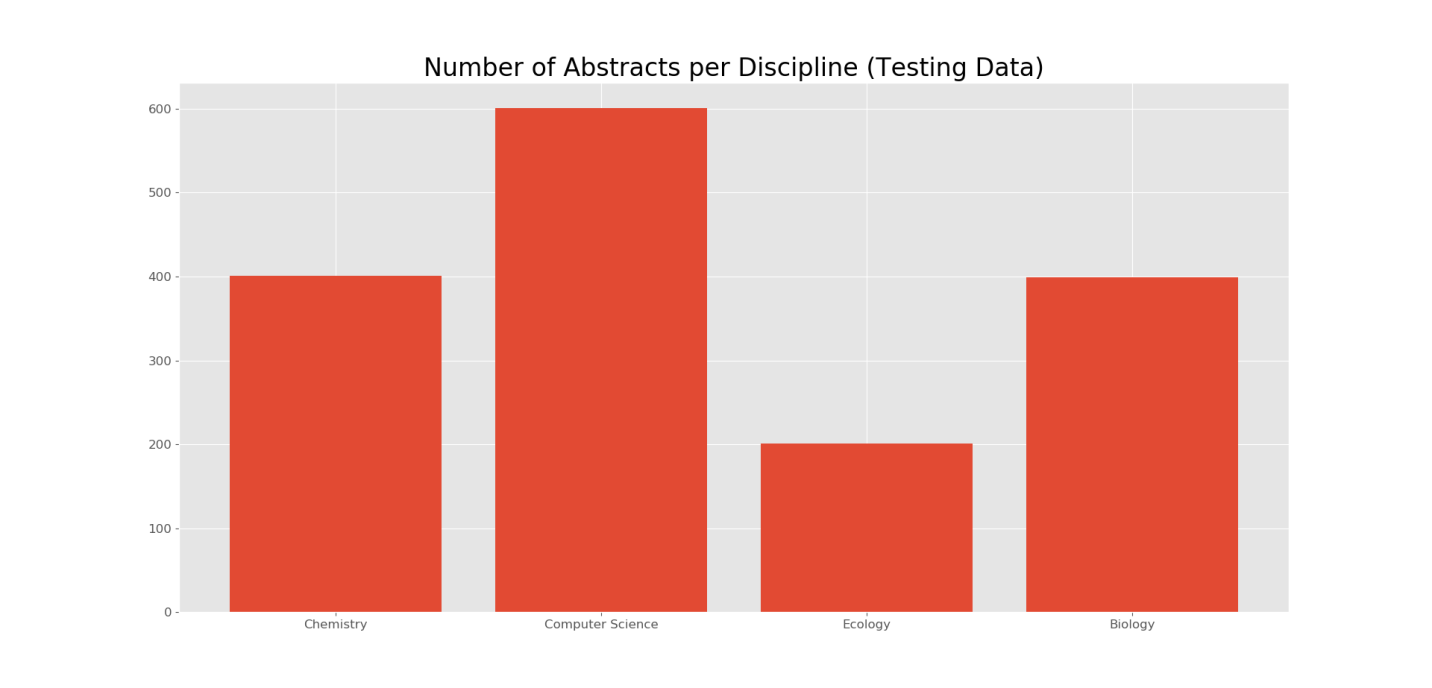
****

For the second machine learning model described earlier, I calculated the word counts for every word among all of the 7919 abstracts in the training data and used the top 100 words as features. The 100 most common words among all of the 7919 abstracts in the training data were:

['DATA', 'MODEL', 'LEARNING', 'TWO', 'SHOW', 'USING', 'RESULTS', 'BASED', 'STATE', 'ALSO', 'ALGORITHM', 'HIGH', 'STRUCTURE', 'MODELS', 'METHOD', 'TIME', 'PROBLEM', 'NEW', 'ONE', 'SPECIES', 'ANALYSIS', 'METHODS', 'USED', 'REACTION', 'APPROACH', 'WELL', 'DIFFERENT', 'STUDY', 'FIRST', 'LARGE', 'HOWEVER', 'FUNCTION', 'OBSERVED', 'COMPLEX', 'ALGORITHMS', 'PAPER', 'NUMBER', 'RATE', 'PRESENT', 'PROPOSE', 'CLIMATE', 'CO', 'MAY', 'ENERGY', 'LOW', 'PROVIDE', 'USE', 'NOVEL', 'TEMPERATURE', 'PERFORMANCE', 'CHANGE', 'NON', 'PROPOSED', 'MOLECULAR', 'LINEAR', 'ELECTRON', 'BOND', 'DEMONSTRATE', 'FOUND', 'STUDIES', 'EXPERIMENTS', 'PROCESS', 'EFFECTS', 'CELLS', '10', 'FORMATION', 'INFORMATION', 'ORDER', 'SINGLE', 'DISTRIBUTION', 'GROUP', 'MANY', 'PROBLEMS', 'ACID', 'NETWORK', 'COMPLEXES', 'PROTEIN', 'FRAMEWORK', 'WATER', 'FORM', 'SOLUTION', 'THREE', 'CELL', 'NETWORKS', 'VIA', 'EFFICIENT', 'TRANSFER', 'PROPERTIES', 'NEURAL', 'SEVERAL', 'CONDITIONS', 'ADDITION', 'EXPERIMENTAL', 'CHANGES', 'STRUCTURES', 'POTENTIAL', 'INFERENCE', 'OPTIMIZATION', 'FUNCTIONS', 'FEATURES']

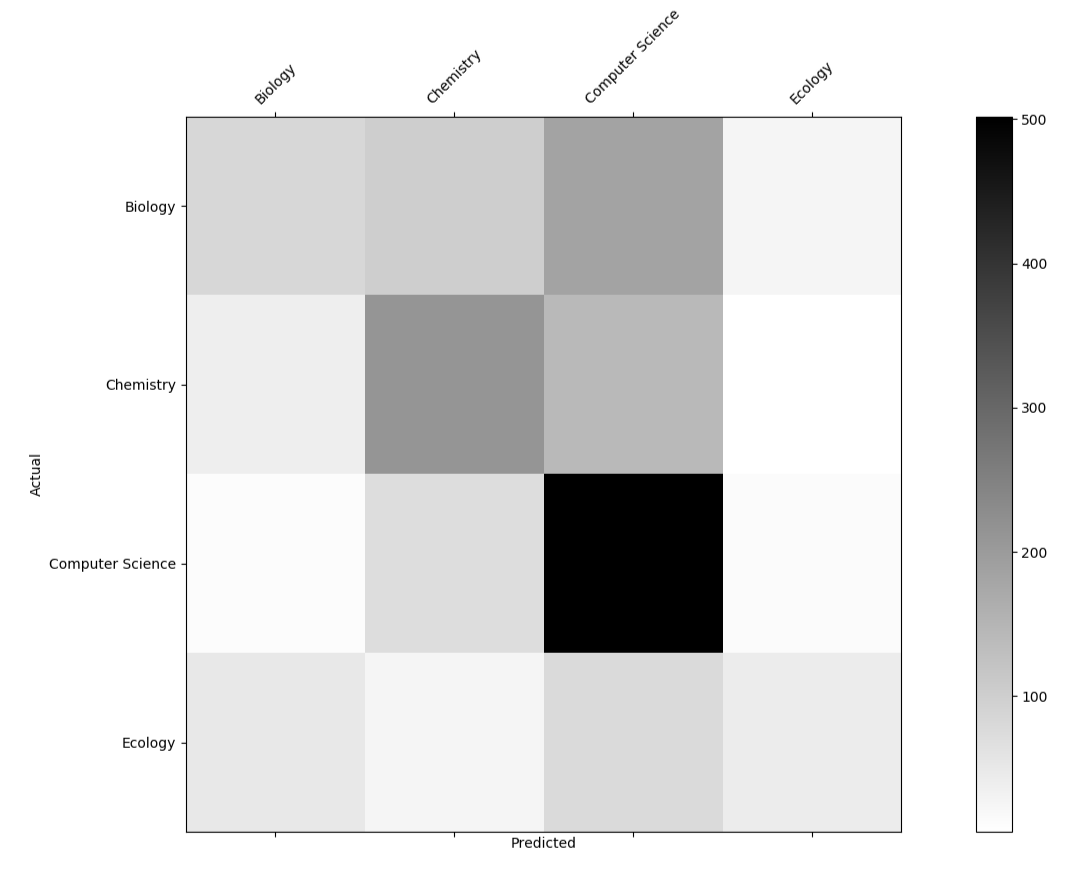
It makes sense that most of the words we see here are words that would be associated with computer science or chemistry, since those are the most abundant disciplines in the training set, so they would contribute more to the most common words. However we still see the most abundant words from the other disciplines such as “Climate”, “Water”, “Protein”, etc. The one hot encoding of these words will be added to the training and testing data for the 2nd machine learning model to see how it affects its accuracy. If my initial hypothesis was right; that the structure of all the abstracts across the disciplines are similar, then the 2nd machine learning model should perform significantly better than the 1st model that does not have knowledge of the contents of the abstracts.

The test set contains a total of 1602 abstracts; I tried to keep the distribution similar to the training set’s distribution. The training set that the models are trained on are labeled, so the model knows what discipline they belong to. However, the testing data is not labeled, since it is the data that we are using for our experiment which the models must make predictions on.

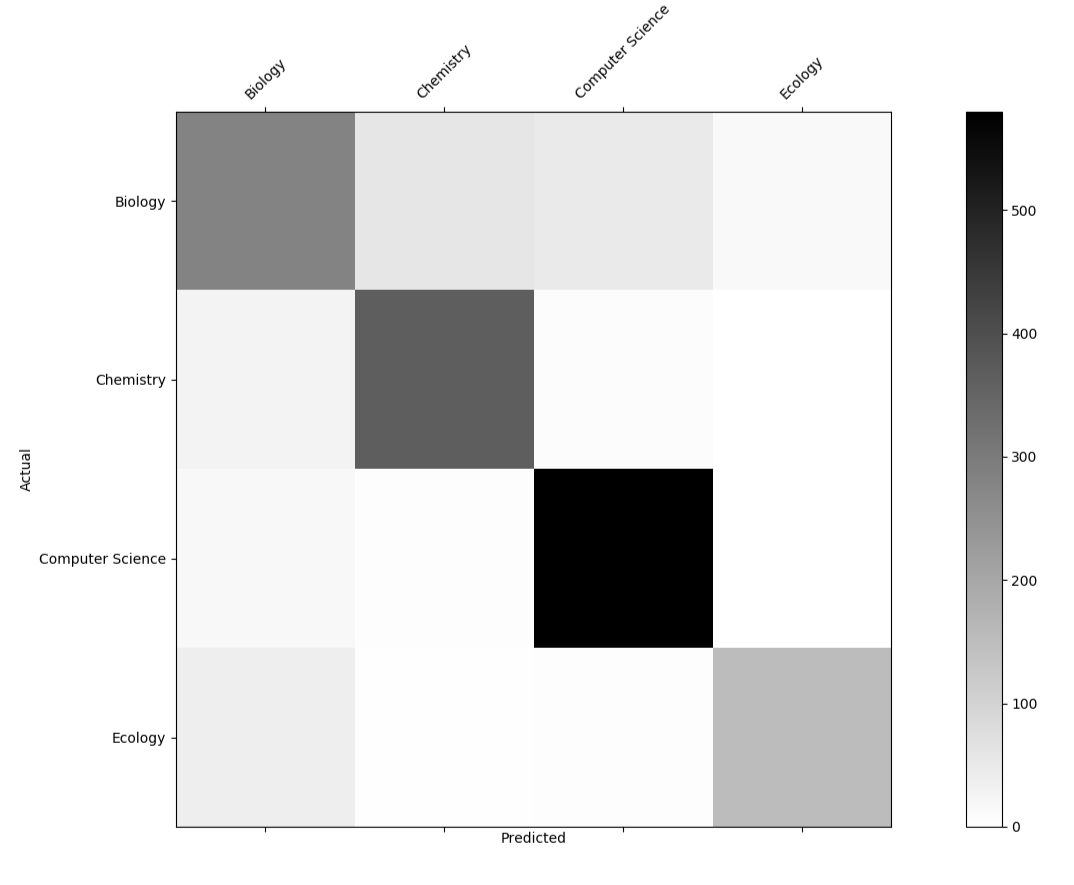
****

**Results and Analysis:**

Confusion Matrix for 1st Machine Learning Model:



Confusion Matrix for 2nd Machine Learning Model:



Above are the confusion matrices for the machine learning models. What a confusion matrix represents is the genre to genre accuracy of the models. The genres on the left are the true genres of the abstracts, while the ones on the top are the predictions of the genre. The darker the color in the square, the more times the genre on the left got predicted as the genre on the top. Ideally we would want to see darker shades on the diagonals, since that would mean that the model is predicting the true genre most of the time. The 1st machine learning model had 52% accuracy on the test set, meaning it correctly guessed the discipline of 843 abstracts in the testing data, while it incorrectly guessed on the rest. The confusion matrix for the 1st model shows that it had almost perfect accuracy when labeling the computer science abstracts, while for chemistry and biology it did somewhat ok, and for biology it could differentiate it from the rest. This makes sense when we consider the graphs of the metrics we calculated that the 1st machine learning model used. Aside from computer science, the metrics for the other three disciplines were all similarly distributed. Meaning from the metrics alone, it can only really precisely detect computer science abstracts. The discipline it performed the 2nd best in was chemistry, which also makes sense since the last metric (punctuation) showed that Chemistry uniquely had a high amount of punctuations. For ecology, and especially biology, we see light shades in all of the boxes for those rows, meaning that the model was very confused for these genres and could not tell them apart.

The 2nd machine learning model did significantly better than the 1st, having an accuracy of 82%. Meaning it only failed on 238 of the 1602 abstracts. The dark on the diagonals shows that very little confusion happened in this model, it was able to accurately tell apart the abstracts of different disciplines.

**Limitations and Recommendations for Further Research:**

My research contained an abundance of data which is good since it means that should be very representative of the genres that I used in the analysis. However, as seen in figure 1, there were vastly many more chemistry and computer abstracts than there were biology and ecology. While I this should not have impacted the results of skewed the results in any way, it would be better if I could have found more ecology and biology abstracts in order to match the amount that chemistry and computer science did in order to erase any doubt that the distribution difference may cause. Another limitation is that for a fair amount of the abstracts, especially those in Chemistry, the way there originally written used some special kind of text (For example when writing chemical formulas, they were written in some specially encoded way that did not translate perfectly when extracted to raw text). The way these encoded texts were translated may have caused a slight skew when calculating metrics such as punctuation. In future repetitions of this experiment, I would filter out abstracts that included such encoded text. Aside from these two slight limitations, I am glad to say that I did not find encounter any major limitations during this analysis. If I were to recommend a way to further the research done in this project, I would include a wider variety of genres in the analysis, and especially genres outside of hard sciences since this analysis mostly included hard sciences. It would be interesting to see the results of an abstract analysis that included both hard sciences such as chemistry, biology, physics, and soft sciences such as psychology, philosophy, etc.

**Datasets/Source Code:**

[**https://www.kaggle.com/benhamner/nips-papers**](https://www.kaggle.com/benhamner/nips-papers)

[**https://www.kaggle.com/mathewsavage/jacs**](https://www.kaggle.com/mathewsavage/jacs)

[**https://www.figure-eight.com/data-for-everyone/**](https://www.figure-eight.com/data-for-everyone/)

[**https://www.ncbi.nlm.nih.gov/pubmed**](https://www.ncbi.nlm.nih.gov/pubmed)

[**https://github.com/dmlc/xgboost**](https://github.com/dmlc/xgboost)

[**https://github.com/CruzJeff/Abstract\_Classifier**](https://github.com/CruzJeff/Abstract_Classifier)