Dataset	Accuracy	sentences	Authors
Becker-Posner	0.82	26922	2
GC-TF-PK	0.67	11984	3
MD-TF-PK	0.70	13422	3
MD-GC-PK	0.66	13448	3
MD-GC-TF-PK	0.61	15584	4

Limitations of the Baseline System

We can see that no deep NLP features are used for the task. A bag of words model is a weak model to be able to discriminate between the authors. Also, the accuracy of the final stage of classification depends on the chunk (V) of sentences picked from the individual authors to form the merged document. Changing that parameter (V) from 200 to 50 reduces the final accuracy from 82% to 49%. Training on segments and testing on sentences is not such a good idea as the whole bottleneck for achieving high accuracy is the clustering algorithm.

Proposed Methodology

The main idea is to quantify the differences of the grammatical writing styles which the earlier baseline model was lacking and use this information to build paragraph clusters. So, by doing this, what kind of sentences can we decompose? An example shown below illustrates this. Consider the two sentences below:
S1: My chair started squeaking a few days ago and its driving me nuts.
S2: Since a few days my chair is squeaking-its simply annoying.

The above sentences are semantically similar, although they differ way too much syntactically(as shown in the figure below) and a bag of words model, which just relies on the occurrences of the words/word counts cant distinguish between these two sentences as they have more or less similar kinds of words. The main idea is to quantify those differences by calculating grammar profiles and to use this information to decompose a collaboratively written document.

What is PQ Grams ?

Similar to n-grams that represent the subparts of given length n of a string, p-q grams extract substructures of an ordered labelled tree. The size of p-q gram is determined by stem (p) and base (q). P defines how many nodes are included vertically, and q defines the number of nodes to be considered horizontally. For example, a valid p-q gram with p=2 and q=3 starting from PP at the left side of the tree (S2) shown in the above figure would be [PP-NP-DT-JJ-NNS]. The p-q gram index then consists of all possible p-q grams of a tree. In order to obtain all p-q grams, the base is shifted left and right additionally. If less than p nodes exists horizontally, the corresponding place in the pq-gram is filled with *, indicating a missing node.

As seen from the flow chart above, paragraphs are extracted from text and each sentence is extracted from the paragraphs. For each sentences, a parse tree is formed using the standard StanfordParser. We call this the Grammar tree. From this Grammar tree, we extract the PQ Gram indices of these sentences. p-q gram index of a sentence is all possible p-q grams of a sentence , whereby multiple occurrences of the same p-q grams are also present multiple times in the index. By combining all p-q gram indices of all sentences, a p-q gram profile is created which contains a list of all p-q grams and their corresponding frequency of appearance in the text. For our experiment, we have used p=2 and q=3. Finally, each paragraph-profile is provided as input for clustering algorithm, which are asked to build clusters based on the p-q grams contained. Also the labels are POS tags of Penn Treebank.

Bottleneck

When applied to the data set Becker-Posner dataset (26922 sentences), we encountered many long sentences which the parser was not able to parse.(out of memory). Below is a screenshot of the memory error we got. The sentence is 215 characters long.
One way out of this was to ignore all such sentences and proceed on the remaining smaller sentences, but that would have defeated the purpose of the baseline method, on whose results we were trying to improve upon. Our observation from the above experiment is that, intuitively its a good method to capture the different syntactic aspects of writers. Although, it pushes the dimensionality of the feature space to quite high compared to the baseline. For 1000 sentences, we were getting a pq-gram profile size of 9200(and a training time of 120minutes), which will be quite high when run for more sentences. The baseline methods feature size were much smaller and simpler and faster.

References

A generic unsupervised method for decomposing multi-author documents. Navot Akiva and Moshe Koppel.2013. *Journal of the American Society for information Science and Technology,* 64: 2256–2264. Navot Akiva and Moshe Koppel.2013. *Science* 208: 1019–1026.
*Unsupervised Decomposition of a Multi-Author Document Based on Naive-Bayesian Model.*Khaled Aldebei, Xiangjian He and Jie Yang. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 501505, Beijing, China