## Objective

Every writer has a unique style of writing which is independent of the content/genre. This project explores the style aspect of technical papers.

We want to come up with a style score/style metric of each technical paper irrespective of what the paper is about. This has many applications like author disambiguation, determining which style of writing would lead to more number of citations, readability of the paper and many more.

-- What we are trying to achieve: Plot of Content Score VS Style Similarity Score of papers with respect to a reference paper.

## Data

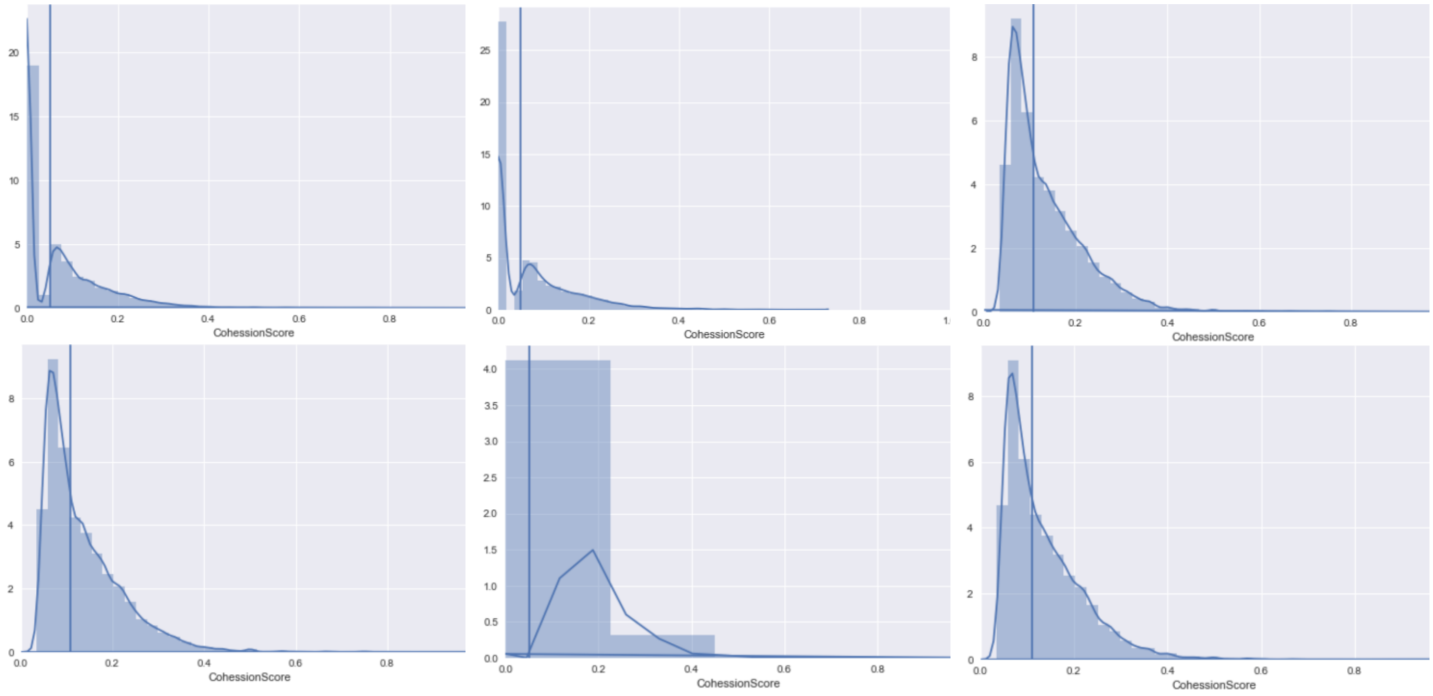
We had explored abstract of technical papers from Pubmed.

## Data Exploration/Feature Selection

### Experiment 1: Look at the distribution of cohesive score of each word in the abstract.

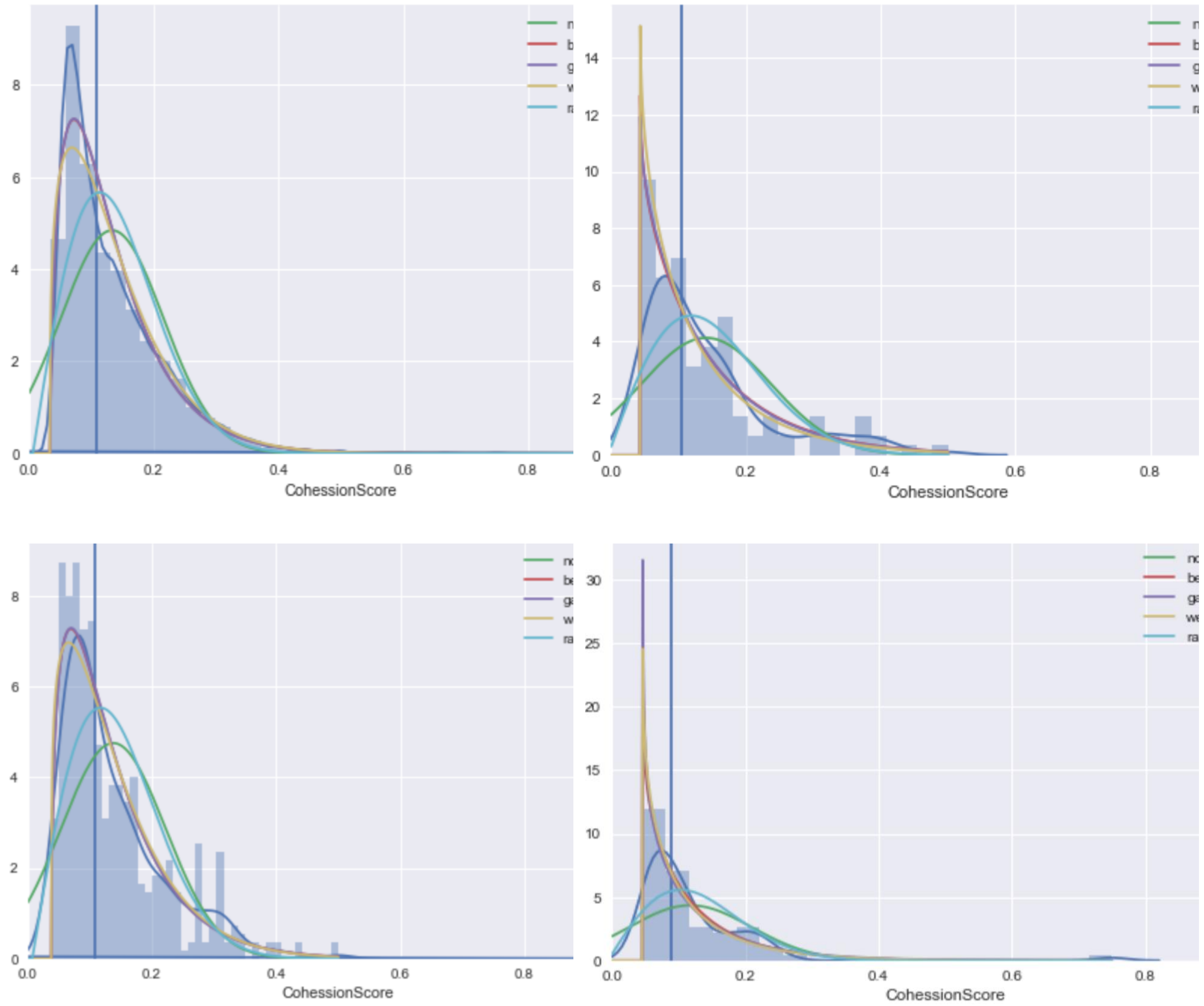
***Cohesive score***: Tells us if the word is highly specific to a topic or more general. As adapted from this paper. A word with high cohesive score tells us that it is specific to the topic of the article or a topic word. On the other hand, a word with low cohesive score is more general and are taken into consideration for determining style of writing we will call such words as *Function words*.

- Get a sample of paper abstracts.
- Tokenize each of the abstracts and get the cohesive scores of each of the tokens.
- Plot the distribution of the cohesive score of these words.



### Experiment 2: Fit the distribution with a known probability distribution.

In statistical modelling we try to fit the distribution of our data with some known probability distribution. We need to choose a distribution whose interpretation makes sense.

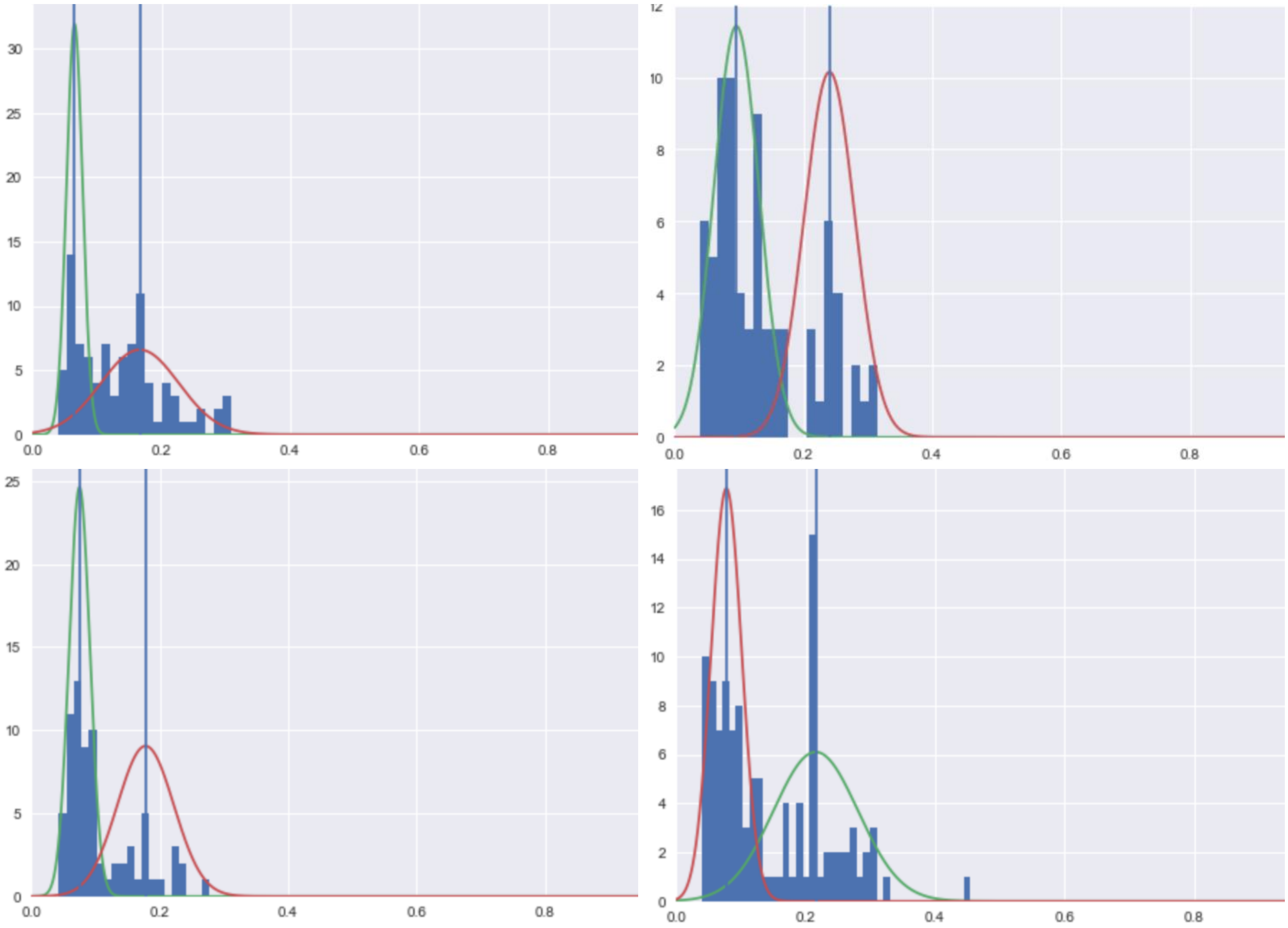


In some plots beta seems to fit the data well and in some plots Weibull seems to fit well. But the fit does not look convincing enough also the
interpretation of these distributions do not make sense in this context.

*Observation:* All these plots are bimodal.

### Experiment 3: Fit two normal distributions.

Let us see if it is a combination of 2 normal distributions, so we try to fit the two normal distributions using EM (Expectation-Maximization)
algorithm.



**Interpretation**: After looking at the words that lie in the left normal distribution and those lying in the right normal distribution we could
interpret the two normal distributions as the distributions of function words (left) and distribution of topic words (right).

### Experiment 4: Start with a straightforward approach to plot style similarity vs content similarity.

Let the reference paper be **Pr**
Let the set of n papers we are trying to plot be **{Pi}**

For each function word we have defined proportional frequency
**PF(word w) = Number of occurrences of word w / Total number of words**

**Proportional Frequency in all papers**

$PF_{All}(w)$ = Number of occurrences of word w / Total number of words in all papers

**Proportional Frequency in reference paper Pr**

$PF_{Pr}(w)$ = Number of occurrences of word w in Pr / Total number of words in reference paper

**Proportional Frequency in paper to be compared/plotted Pi**

$PF_{Pi}(w)$ = Number of occurrences of word w in Pi / Total number of words in Pi paper

$PF_{All}(w) - PF_{Pi}(w)$ : **Comparison of proportional frequency of word in Pi to that of All papers.**
**Greater than 0 if** $PF_{Pi}(w) < PF_{All}(w)$ **else Less than 0**

$PF_{All}(w) - PF_{Pr}(w)$ : **Comparison of proportional frequency of word in Pr to that of All papers.**
**Greater than 0 if** $PF_{Pr}(w) < PF_{All}(w)$ **else Less than 0**

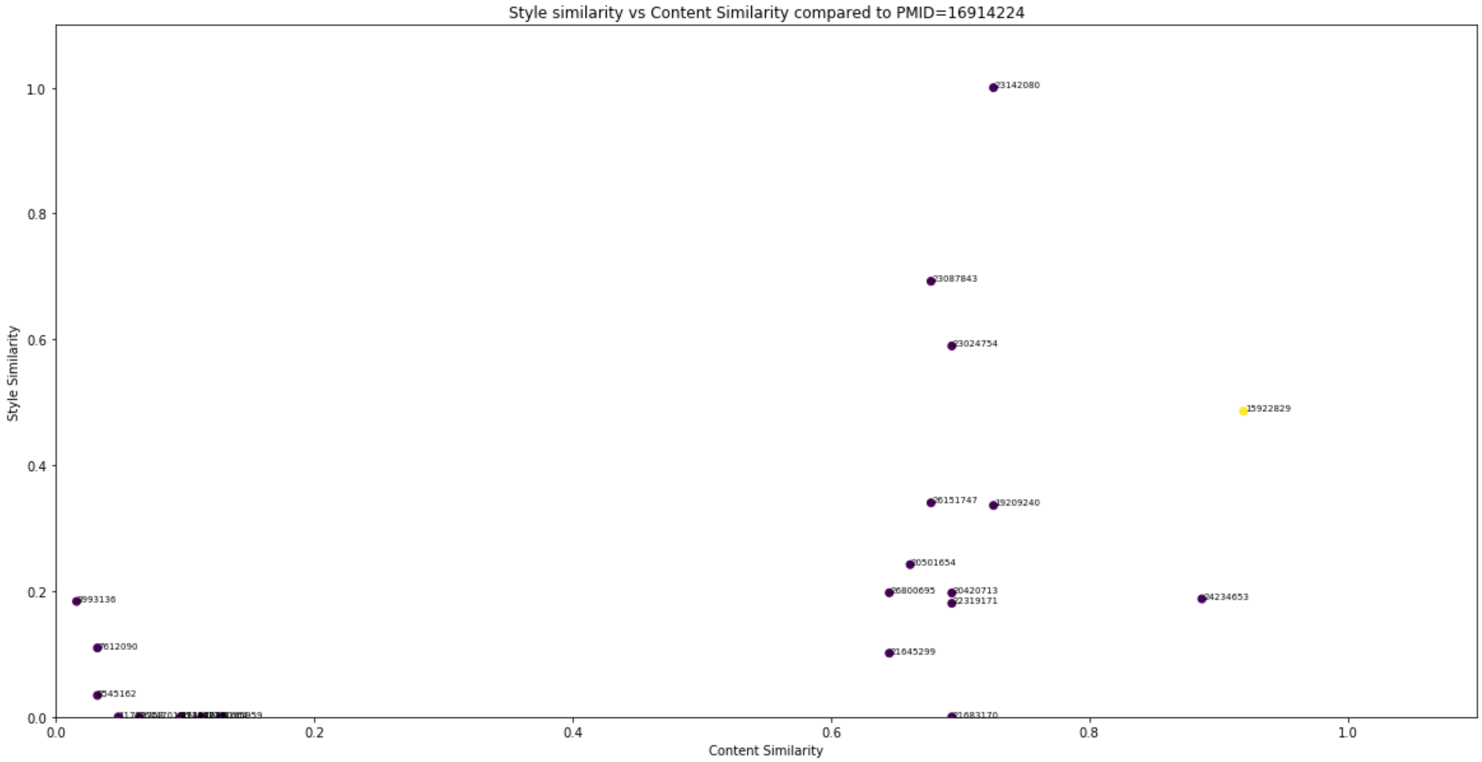
$$(\boldsymbol{PF_{All}(w) - PF_{Pr}(w)}) * (\boldsymbol{PF_{All}(w) - PF_{Pi}(w)}) = \begin{cases} > \mathbf{0}, \ \ \textit{if } PF(w) \textit{ for both } Pi \textit{ and } \mathrm{Pr} \textit{ is less} \\ \quad \textit{or both greater than } PF(w) \textit{ overall} \\ \qquad\qquad < \mathbf{0}, \qquad \textit{otherwise} \end{cases}$$

To calculate similarity score of paper i, Pi we use the following formula
**Style Similarity Score** = $\sum_{\boldsymbol{w \ in \ both \ \mathrm{Pr} \ and \ Pi}} (\boldsymbol{PF_{All}(w) - PF_{\mathrm{Pr}}(w)}) * (\boldsymbol{PF_{All}(w) - PF_{Pi}(w)}) * \boldsymbol{Cohession(w)}$

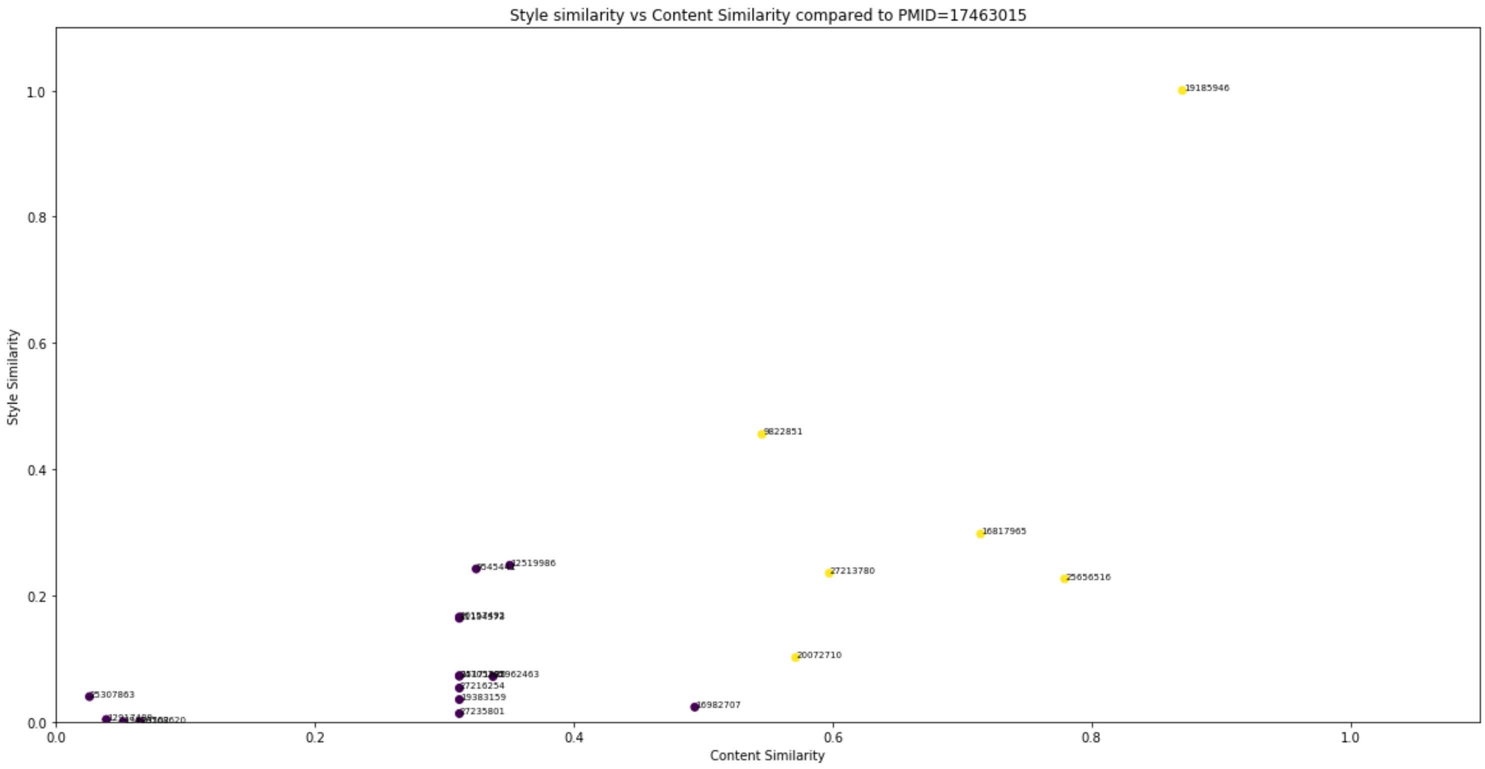
**Content Similarity Score =** We get from Absim

We now plot the papers Pi, Style Similarity Score on y axis and Content Similarity score on x axis.

Results of two such experiments are below.



Plot showing Style similarity vs Content similarity of papers as compared to pmid:16914224.
Papers with same author(s) as 16914224 are shown in yellow



Plot showing Style similarity vs Content similarity of papers as compared to pmid:17463015.
Papers with same author(s) as 17463015 are shown in yellow

## Experiment 5: Style similarity score.

**BM25:** Is used to rank documents on how similar they are to a query. If we integrate length of the document as well then the
formula looks like this:

$$BM25\ score = IDF * \frac{\left((k+1) * tf\right)}{k * (1 - b + b * L) + tf}$$

IDF: Inverse Document Frequency
tf: Term Frequency
L: length of the document
k, b: constants. These are parameters that need to be tuned for optimal results.

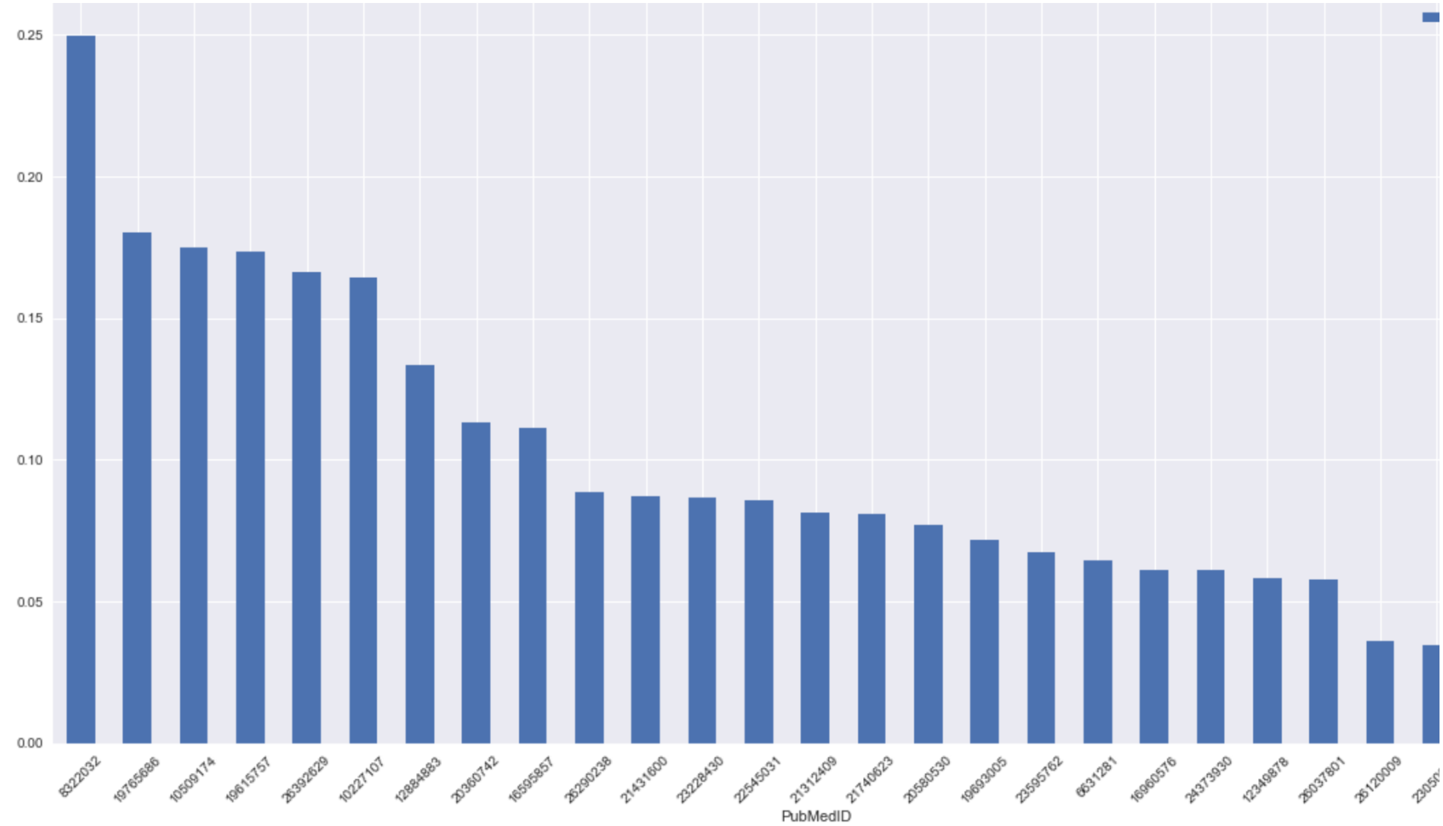
The BM25 score is inversely proportional to L (assuming b > 0).

Inspired by BM25 we devised a similar formula for ranking documents based on style similarity.

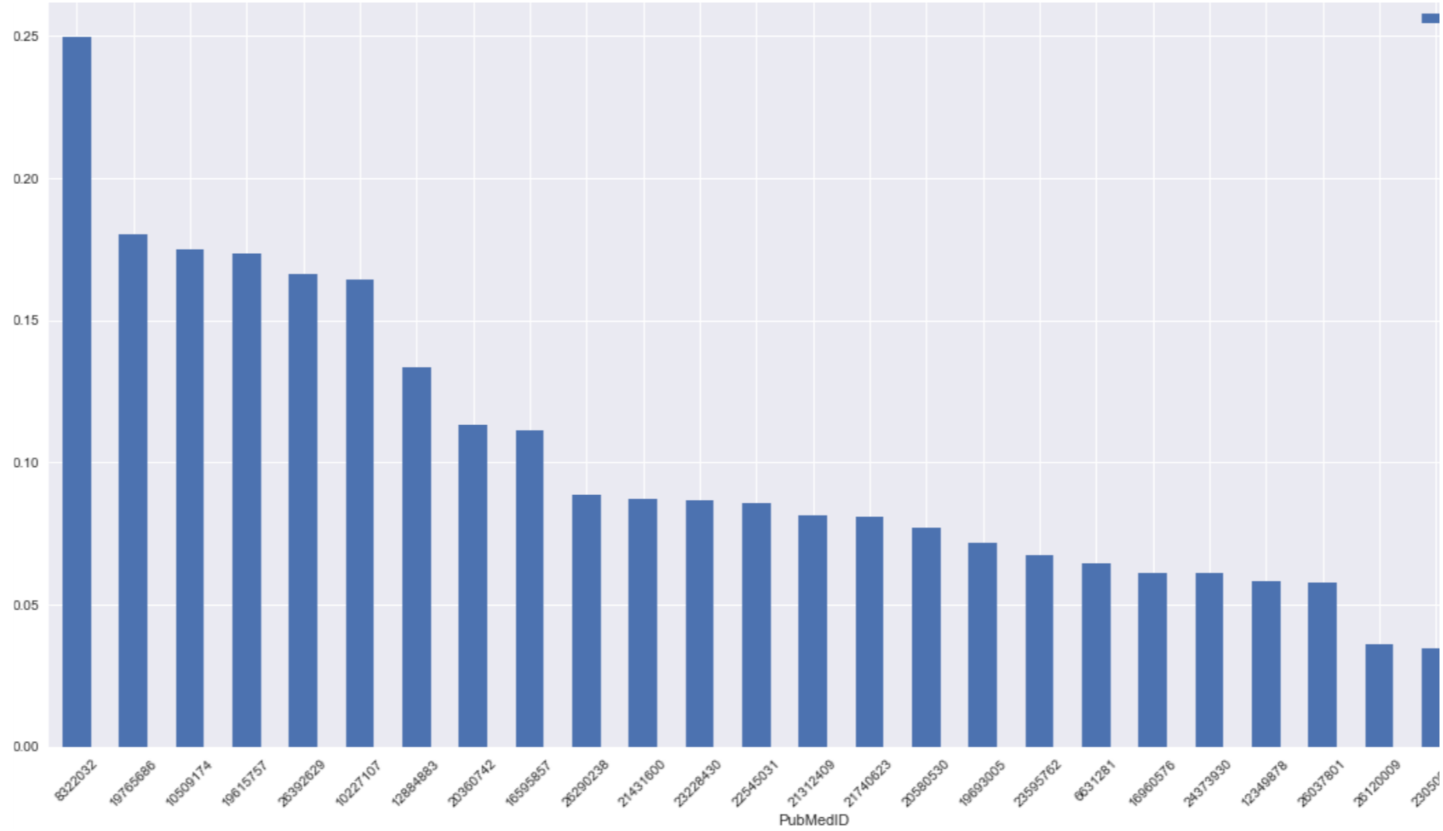
$$style\ score = IDF * \frac{\left((k+1) * tf^{p}\right)}{k * (1 - b * (1 - \alpha)^{b}) + tf^{p}}$$

IDF: Inverse Document Frequency
tf: Term Frequency
α: proportion of the document under the first normal curve (function words distribution).
k, p, b: constants. These are parameters that need to be tuned for optimal results.

Style score is inversely proportional to the term 1 – α which is the proportion of the document not in the first normal curve (function
word distribution).



Above plot shows the similarity with respect to a query paper (PMID: 10719138)



Above plot shows the similarity with respect to a query paper (PMID: 11738834)

The highest ranked papers (high scores) are those authored by the same person, this is an indication that the similarity score would be
a good starting point for measuring style.