# Types of Files

All matches in this section of the downloads page are Maximal matches, as defined [here](http://shakespearestext.com/can/counting.htm).

For each play, collocation and N-gram search results are provided in three formats, explained separately below.

## HTML

These files can be opened in any browser. They are for casual browsing and to allow qualitative analysis of the results, to get an idea of which other plays appear to have the strongest links with the play whose search results the HTML page presents.

The results are ordered according to a ranking formula, with the best matches at the top. The ranking formula is the same for all plays and for both collocations and N-grams.

The HTML file does not contain all results, owing to the typically very large number of them. For collocations, the best 2.5% of results are shown, while for N-grams the best 25% are shown. You need to the consult the corresponding CSV file to view the full set of results.

A filter box is provided on the HTML page. This is useful if you want to filter the results to show only matches with a specified play, or for specified words and phrases. For example, typing *paris* in the filter box and pressing the Filter button will cause only the matches with *The Massacre at Paris* and *The Arraignment of Paris* to be shown, as well as any text which happens to contain the word *Paris*. Filtering is not case-sensitive. To clear a filter, delete everything in the Filter box and press the Filter button again.

## CSV

This file can be opened in Excel and other spreadsheet programs. It contains the full set of results, to be used for your own analysis.

If you intend to work often with a CSV file then you will find it very much more convenient if you save it as an Excel spreadsheet. This will allow you to resize and format the columns as you wish.

The results are ordered according to a ranking formula, with the best matches at the top. The ranking formula is the same for all plays and for both collocations and N-grams. The same formula is used in both the HTML and CSV files, so all the results on the HTML page appear at the top of the CSV file in the same order.

The table below lists the columns in the CSV file and explains what they mean. The CSV files for collocations and N-grams contain the same columns.

| Excel Column | Name | Meaning |
| --- | --- | --- |
| A | Rank | The rank of this match, according to the ranking formula. The best match, having the rank 1, is displayed at the top, the second-best one below it, and so on. |
| B | Play | The play or division whose search results are in this file. This is of course the same in every row. It is provided to make it easier to merge rows from different files without losing the knowledge of which file they came from. |
| C | Year | The year, according to the SHC and Folger websites, in which the play was written (which may be much earlier than the date of publication). These dates are not always reliable. The dates in the latest authority (Wiggins) may differ. |
| D | No. of Matching Lemmata  | If the phrase *kind hearts* in one text is matched with the compound word *kind-hearted* in the other, then there is only one matching word (*kind*), but two matching lemmata. This column shows the number of matching lemmata, not the number of matching words. |
| E | Matching Lemmata | The matching lemmata are separated by the pipe (|) character. |
| F | Text  | A snippet of text from the play or division whose search results this file presents, containing the matching phrase. |
| G | Text in Matching Play | A snippet of text from the matching play or division, containing the matching phrase. |
| H | Matching Play | The play or division the matching text has been found in.  |
| I | Matching Play Year  | The year, according to the SHC and Folger websites, in which the matching play was written (which may be much earlier than the date of publication). These dates are not always reliable. The dates in the latest authority (Wiggins) may differ. |
| J | No. of Plays Found In  | **This is an important column to understand.**The number of plays the matching lemmata have been found in. This includes the two plays named on this line; therefore, this number is always at least 2. If it is exactly 2, it means the match is unique to these two plays. Suppose you are viewing results for some play P and it contains the phrase *clear blue sky*. Play A also contains the phrase *clear blue sky*, while play B only contains the phrase *blue sky*. Suppose that no other plays contain either of these phrases.In the CSV file for play P you will see one row showing a match between P and A for *clear blue sky*. The “No. of Plays Found In” value will be 2, as this is a unique match.You will also see one row showing a match between P and B for the phrase *blue sky*. You will **not** see a row for the match between P and A for *blue sky*, because that match was subsumed in the larger match for *clear blue sky*. Nevertheless, the row for the match between P and B for *blue sky* will tell you that the “No. of Plays Found In” is 3. This is correct, because *blue sky* is indeed found in 3 plays.Even if a matching phrase has been found in more than one division of a play, it is counted just once for that play. This is to ensure that the ranking index for this match is *invariant under division*. If we did not do this, the index would vary according to our choice of divisions, which would prevent a like-for-like comparison between competing authorship attribution theories that required different divisions of a play. |
| K | Matching Lemmata Weights  | This can be ignored. It is a value used only by the ranking formula. |
| L | No. of Proper Nouns  | The number of proper nouns detected in the matching lemmata. The detection is not wholly accurate and some proper nouns may fail to be recognized as such. This failure has no effect on the search results themselves, only on the rank they are assigned. Therefore, this column can be ignored. It is used only by the ranking formula. |
| M | Index for Ranking | The index calculated by the ranking formula and used to derive the Rank in the first column of the CSV file. It can be ignored. |

## Summary Files

Summary files are different for collocations and N-grams. They are intended to provide totals rather than listings of results.

**It is important to understand how to use summaries. If used incorrectly they may lead you to cite incorrect numbers in your work. Most importantly, the notes below show you how you can correctly combine data from different rows of a Summary file.**

### N-grams Summary File

The table below tells you what the columns mean in the N-grams Summary files.

| Excel Column | Name | Meaning |
| --- | --- | --- |
| A | Play | The play or division whose search results are summarised in this file. This is of course the same in every row. It is provided to make it easier to merge rows from different files without losing the knowledge of which file they came from. |
| B | Year | The year, according to the SHC and Folger websites, in which the play was written (which may be much earlier than the date of publication). These dates are not always reliable. The dates in the latest authority (Wiggins) may differ. |
| C | No. of Words | The number of words in the play or division, excluding speech prefixes but including stage directions. |
| D | Matching Play | The play or division whose matches are summarised in this row.  |
| E | Matching Play Year | The year, according to the SHC and Folger websites, in which the matching play was written (which may be much earlier than the date of publication). These dates are not always reliable. The dates in the latest authority (Wiggins) may differ. |
| F | No. of Words in Matching Play | The number of words in the matching play or division, excluding speech prefixes but including stage directions. |
| G | No. of Unique Matches | The number of matching N-grams which are unique to the two plays or divisions named in this row. |
| H | Weighted No. of Unique Matches | Each ‘weighted’ is the value in the column immediately to its left, divided by the sum of the numbers of words in the two plays or divisions. Weighted values allow fair comparison between plays (especially divisions) of different lengths. |
| I | No. of All Matches | The number of all matching N-grams between the two plays or divisions named in this row. |
| J | Weighted No. of All Matches | Each ‘weighted’ is the value in the column immediately to its left, divided by the sum of the numbers of words in the two plays. Weighted values allow fair comparison between plays (especially divisions) of different lengths. |
| K | No. of Unique Trigram+ Matches | The number of matching N-grams, with n>=3, which are unique to the two plays or divisions named in this row. |
| L | Weighted No. of Unique Trigram+ Matches | Each ‘weighted’ is the value in the column immediately to its left, divided by the sum of the numbers of words in the two plays. Weighted values allow fair comparison between plays (especially divisions) of different lengths. |
| M | No. of Trigram+ Matches | The number of all matching N-grams, with n>=3, between the two plays or divisions named in this row. |
| N | Weighted No. of Trigram+ Matches | Each ‘weighted’ is the value in the column immediately to its left, divided by the sum of the numbers of words in the two plays. Weighted values allow fair comparison between plays (especially divisions) of different lengths. |
| O | No. of Unique Tetragram+ Matches | The number of matching N-grams, with n>=4, which are unique to the two plays or divisions named in this row. |
| P | Weighted No. of Unique Tetragram+ Matches | Each ‘weighted’ is the value in the column immediately to its left, divided by the sum of the numbers of words in the two plays. Weighted values allow fair comparison between plays (especially divisions) of different lengths. |
| Q | No. of Tetragram+ Matches | The number of all matching N-grams, with n>=4, between the two plays or divisions named in this row. |
| R | Weighted No. of Tetragram+ Matches | Each ‘weighted’ is the value in the column immediately to its left, divided by the sum of the numbers of words in the two plays. Weighted values allow fair comparison between plays (especially divisions) of different lengths. |
| S | Sum of Ranking Indices | The sum of all the ranking index values given in the CSV file for this play or division.  |
| T | Weighted Sum of Ranking Indices | Each ‘weighted’ is the value in the column immediately to its left, divided by the sum of the numbers of words in the two plays. Weighted values allow fair comparison between plays (especially divisions) of different lengths. |

The rows in the Summary file are sorted by column H, the weighted number of unique matches. This means that plays which have the greatest number of unique matches with the play that this Summary file is for are shown at the top, after weighting to take account of the differing lengths of the plays.

You are free to sort the rows in a different order. For example, if you wish to see the plays with which we have the greatest number of tetragram+ matches, then sort by column Q or column R, depending on whether you want to apply the weighting. Remember always to sort in descending order, so you see the best matching plays at the top:



### Combining Summaries for N-grams

Below is a screenshot from the N-grams Summary CSV file for the play called *Albovine King of the Lombards*. I have filtered the Matching Play column to show matches only with the four divisions we have for *The Two Noble Kinsmen*. For clarity I have coloured some of the columns.



Suppose you are not interested in the divisions and you just want to see the matches with *The Two Noble Kinsmen* as a whole. What you must do is to add up the numbers in the green columns, as shown below.



It is safe and correct to add up values in the green columns. What you must never do is to add the values in the ‘Weighted’ columns, shown in red. You must recalculate these by dividing the value in the green column by the sum of the numbers of words in the two plays (the values in columns C and F). The blue text in the screenshot above is just a visual reminder that this row is not part of the CSV file but has been added by me.

You can also combine rows for different plays. For example, suppose you wanted to summarise all matches between *Dr. Faustus* and the works of Samuel Daniel. To do this, open the Summary file for *Dr. Faustus* in Excel and filter on the plays you believe were written by Daniel, as shown below, adding the values in the green columns:



By doing this we can see, for example, that *Dr. Faustus* has 9 unique N-gram matches with the works of Samuel Daniel. Of these, 6 are tetragrams or better and 8 are trigrams or better, so we can deduce that 2 must be trigrams and 1 must be a bigram. If we want to view these matching N-grams, we can do so in the corresponding CSV file.

As before, you can add the green values but you must recalculate the weighted values yourself by dividing the value to its left by the sum of columns C and F.

### Collocations Summary

The table below tells you what the columns mean in the collocations Summary files.

| Excel Column | Name | Meaning |
| --- | --- | --- |
| A | Play | The play or division whose search results are summarised in this file. This is of course the same in every row. It is provided to make it easier to merge rows from different files without losing the knowledge of which file they came from. |
| B | Year | The year, according to the SHC and Folger websites, in which the play was written (which may be much earlier than the date of publication). These dates are not always reliable. The dates in the latest authority (Wiggins) may differ. |
| C | No. of Words | The number of words in the play or division, excluding speech prefixes but including stage directions. |
| D | Matching Play | The play or division whose matches are summarised in this row.  |
| E | Matching Play Year | The year, according to the SHC and Folger websites, in which the matching play was written (which may be much earlier than the date of publication). These dates are not always reliable. The dates in the latest authority (Wiggins) may differ. |
| F | No. of Words in Matching Play | The number of words in the matching play or division, excluding speech prefixes but including stage directions. |
| G | No. of All Matches | The number of all matching collocations between the two plays or divisions named in this row. |
| H | No. of Unique Matches | The number of unique matching collocations between the two plays or divisions named in this row. |
| I | Weighted No. of All Matches | The number of all matches divided by the sum of the numbers of words in the two plays. Weighted values allow fair comparison between plays (especially divisions) of different lengths. |
| J | Weighted No. of Unique Matches | The number of unique matches divided by the sum of the numbers of words in the two plays. Weighted values allow fair comparison between plays (especially divisions) of different lengths. |
| K | Sum of Ranking Indices | The sum of all the ranking index values given in the CSV file for this play or division. |
| L | Weighted Sum of Ranking Indices | Each ‘weighted’ is the value in the column immediately to its left, divided by the sum of the numbers of words in the two plays. Weighted values allow fair comparison between plays (especially divisions) of different lengths. |

### Combining Summaries for Collocations

The principles are the same as for N-grams, as explained above.