“Analyse” output specifications

General considerations

* CSV format
* Include a header row. The contents of this row should give some hint about the contents of each column. It’s not crucial for the header row to look exactly like the model file.
* The order of the columns is not crucial.
* An easily modifiable script is way more useful than a script that collects all and only the information described below – so feel free to focus on any subset of the remainder of this document.

Notes on columns

1. Name of file
2. Wordcount for the file, excluding punctuation
3. Type-Token Ratio of 500 words (either the first 500 words, as in sample script, or 500 random words). Expected output is a decimal number between 0 and 1.
4. Percent of finite verbs that are present tense (see sample script; expected output is a decimal between 0 and 100)
5. Count of passive constructions (countPassives in sample script)
6. Percent of clauses in passive voice (see sample script; expected output is a decimal between 0 and 100)

G-N. Count of words ending in each of the following (see sample script, re nominalizations):

tion

tions

ment

ments

ness

nesses

ity

ities

Next batch of columns: Count of each of the following words:

|  |  |  |  |
| --- | --- | --- | --- |
| additionally | actually | accordingly | afterwards |
| also | admittedly | consequently | eventually |
| alternatively | anyhow | hence | finally |
| further | anyway | naturally | first |
| furthermore | conversely | otherwise | firstly |
| likewise | however | so | last |
| moreover | instead | therefore | lastly |
| namely | nevertheless | thus | next |
| similarly | still |  | second |
|  | though |  | secondly |
|  | yet |  | third |
|  |  |  | thirdly |

Next batch of columns: Count of each of the following bigrams:

|  |  |
| --- | --- |
| above all | after all |
| as well | by comparison |
| for example | by contrast |
| for instance | in comparison |
| in addition | in contrast |
| that is | in fact |
|  | in reality |
|  | other hand |
|  | then again |

Last batch of columns: Count of each of the following trigrams:

|  |
| --- |
| all things considered |
| as a consequence |
| as a result |
| in that case |