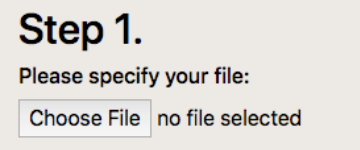
“How To” Guide

The Linguistic Analyzer app is designed to compare two documents and give insightful scores relating to the relevance or similarity of the documents. We use the Yule’s K and Yule’s I algorithm, in addition to a keyword scores algorithm. The results of the calculations are displayed graphically to the user.
The Google NLP API is utilized to scrub and provide “keywords” relevant to a particular document. In addition, a salience score of each keyword is provided which measures the importance of each keyword in regards to a document.

Home Page

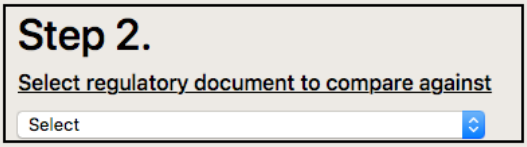
Step 1: Please specify your file

Currently, the app only supports .pdf file extensions.



1. Select a pdf version of a document to analyze

Step 2: Select a regulatory document to compare against:

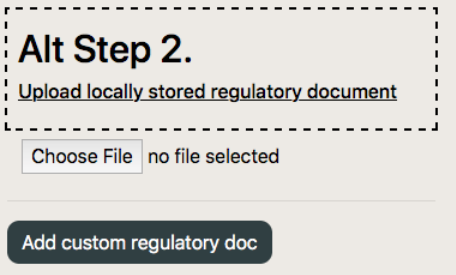


1. Select from a list of documents that have already been loaded/analyzed into the app.

2. If a regulatory document does not appear in the list, it can be added using “Alt Step 2.”

Alt Step 2: Upload locally stored regulatory document

Currently, the app only supports the uploading of .pdf files.

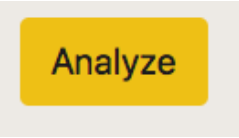


1. Select a pdf version of a regulatory document to upload.

2. Once the file is selected, click "Add custom regulatory doc". The home page will refresh and the selected file will be included in the list of regulatory documents that can be found in Step 2.

3. Repeat Step 1 and Step 2.

Step 3: Analyze

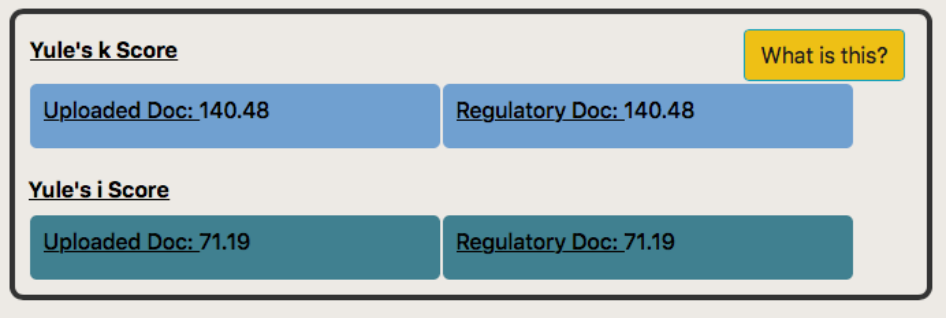


1. Click this button to begin analysis of documents

***Note: Depending on the size of the document, it could take a little while for the analysis to complete. If there is an error in the process, a page will display stating that "an unknown error" has occurred. Please be patient and allow the analysis to complete. Upon completion, a score page will be displayed.
If a regulatory document is added to the app, the initial analysis of the documents will take longer than normal. This is because the newly added regulatory document is being analyzed for the first time. Upon completion of analysis, any re-use of that particular document will not experience delayed processing time ***

Score Page

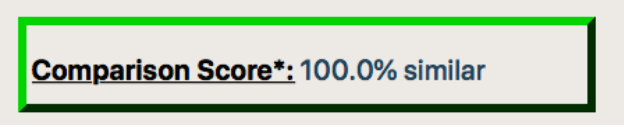
Yule's K and I scores



These are calculated values that represent the semantic richness of a given text. We utilize this algorithm because semantic richness is one benchmark by which technical writes can measure the effectiveness of what they have written. The score is largely useful as a way to compare an uploaded document's significance against the significance of a regulatory text.

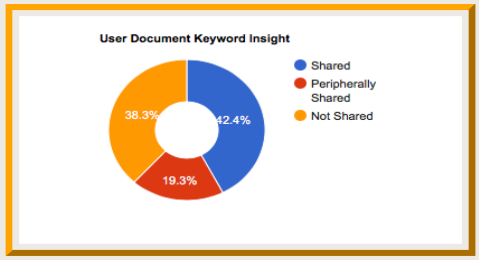
***Note: If the Yule's K and I score is -1, this is due to a "division by 0" error. -1 is the preset score the app gives when this occurs ***

Comparison Score



Compares the calculated scores of the two documents and generates a value based on that comparison. Ideally, if two of the same documents are compared against, a score of 100 should be displayed and a score of 0 will display if two documents are completely dissimilar.

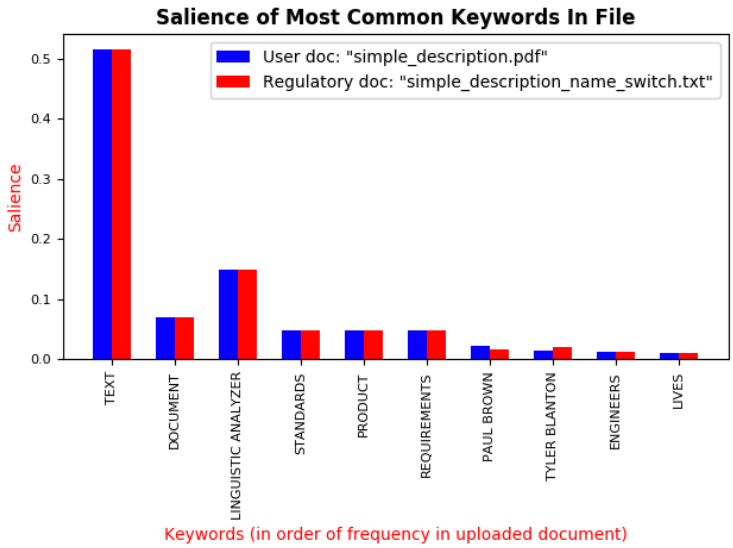
User Document Keyword Insight



This graph displays insight on keyword comparisons between two documents.

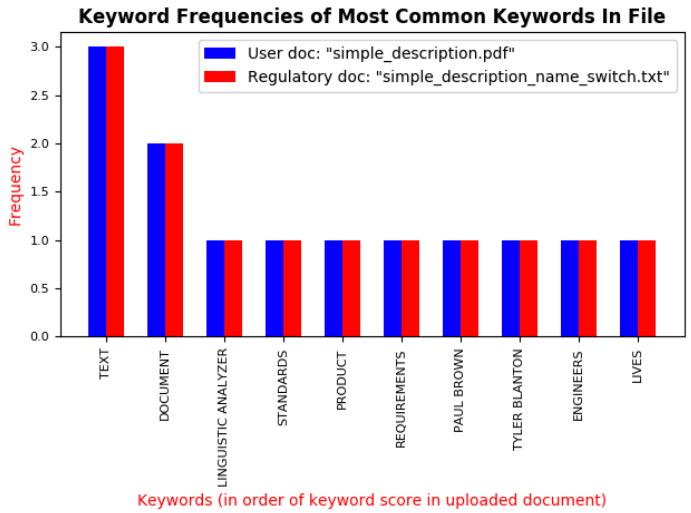
- “Shared” reflects the percentage of keywords shared by both documents.
- “Peripherally shared” displays the percentage of keywords that are contained as part as a grouping. For example, if a user keyword is “Risk”, and the regulatory document keyword is “Risk Analysis”, then this is considered a peripherally shared keyword.
- “Not Shared” refers to the percentage of keywords that the user document does not share with a regulatory document.

View Keyword Salience Graph



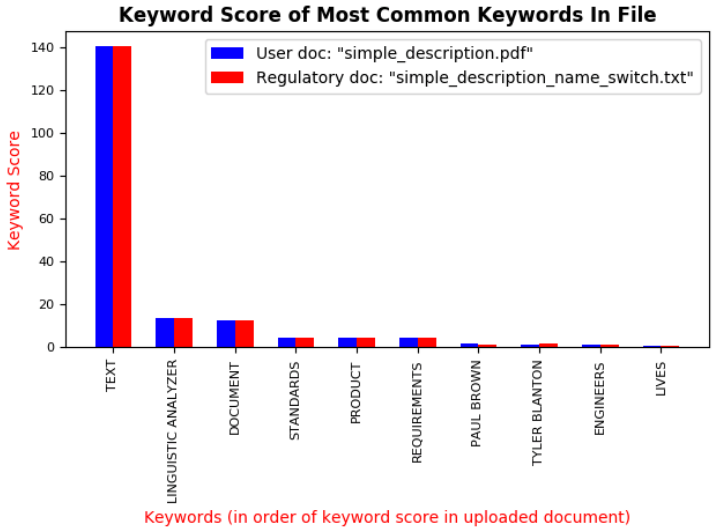
A graph is displayed showing the salience or “weight” of the top 10 frequently used keywords.
The top 10 frequently used keywords from Doc 1 and Doc 2 are compared against each other; if a keyword is shared between the documents, that keyword will be displayed on the graph. If no keywords are shared, a blank graph will be shown with the title “No Common Keywords to Plot”.
0-10 keywords can be displayed.

View Keyword Frequency Graph



A graph is displayed showing the top 10 most frequently used keywords with the frequency value of each word.
The top 10 frequently used keywords from Doc 1 and Doc 2 are compared against each other; if a keyword is shared between the documents, that keyword will be displayed on the graph. If no keywords are shared, a blank graph will be shown with the title “No Common Keywords to Plot”.
0-10 keywords can be displayed.

View Keyword Scores Graph



A graph is displayed showing the top 10 keywords and their associated scores.
The top 10 highest keyword scores from Doc 1 and Doc 2 are compared against each other; if a keyword is shared between the documents, that keyword will be

displayed on the graph. If no keywords are shared, a blank graph will be shown with the title “No Common Keywords to Plot”.
0-10 keywords can be displayed.

View Session Log

Logging statements are provided for each instance use of the app. Up to 3 instances are saved before the log is refreshed.
The log can also be accessed if an “Unknown Error has Occurred” page is displayed.

View User Document Keywords

```
140.48,71.19,16.89
TEXT,0.51400685,3,140.18368636363638
LINGUISTIC ANALYZER,0.1484705,1,13.497318181818182
DOCUMENT,0.07096549,2,12.902816363636365
STANDARDS,0.048485473,1,4.407770272727273
PRODUCT,0.048485473,1,4.407770272727273
REQUIREMENTS,0.048485473,1,4.407770272727273
PAUL BROWN,0.0233815,1,2.125590909090909
TYLER BLANTON,0.0148758795,1,1.3523526818181817
ENGINEERS,0.012385857,1,1.1259869999999998
LIVES,0.01105707,1,1.0051881818181818
MEDTRONIC PLC,0.004180418,1,0.38003800000000004
```

This will display the keywords of a user submitted document. These keywords are the basis of the app. The format reads as:

Line 1:
Yule’s K score, Yule’s I score, Average score

Remaining lines:
Keyword, Salience score, Frequency, Keyword score

View Comparison Document Keywords

```
140.48,71.19,16.9
TEXT,0.5142781,3,140.25766363636365
LINGUISTIC ANALYZER,0.14854886,1,13.504441818181817
DOCUMENT,0.07100294,2,12.909625454545454
STANDARDS,0.04851106,1,4.410096363636364
PRODUCT,0.04851106,1,4.410096363636364
REQUIREMENTS,0.04851106,1,4.410096363636364
TYLER BLANTON,0.021698792,1,1.9726174545454545
PAUL BROWN,0.01605103,1,1.4591845454545456
ENGINEERS,0.012392394,1,1.1265812727272726
LIVES,0.011062906,1,1.0057187272727273
MEDTRONIC PLC,0.004182624,1,0.3802385454545455
```

This will display the keywords of a selected regulatory/comparison document. These keywords are the basis of the app. The format reads as:

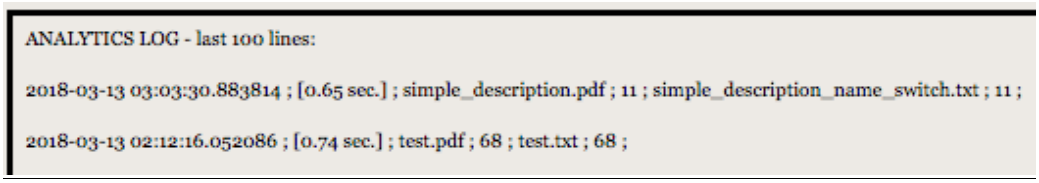
Line 1:
Yule's K score, Yule's I score, Average score

Remaining lines:
Keyword, Salience score, Frequency, Keyword score

Generate Bubble Chart

This graphic is an experimental way of displaying common keywords and providing combined information including Keyword, Salience score, and Frequency. Hover over a particular keyword to display the information.

Analytics Log



This log will display the last 100 instances of the app. It provides the following information:

Date Time; [total process time]; user doc file name; # of keywords; reg doc file name; # of keywords;