

We claim:

1. A method for automatic ingestion and filing of documents in a database having a plurality of file locations, the method comprising:
 - 5 receiving an electronic file including at least one document;
for each document in the at least one document
identifying text data in the document;
indexing the text data to identify one or more document keywords;
comparing the one or more document keywords to a corpus of stored
10 keywords, the corpus of stored keywords previously generated
based on a plurality of documents in the database, wherein each
of the stored keywords in the corpus has at least one file
location association identifying a file location associated
therewith;
15 generating a plurality of keyword scores based on the comparison of
the one or more document keywords and the corpus; and
generating a plurality of suggested file locations using the keyword
scores and file location associations.
- 20 2. The method of claim 1, further comprising
identifying a plurality of pages in the file;
determining a plurality of page markers for each page;
determining that the at least one document in the file comprises a plurality of
distinct documents; and
25 assigning each page to one of the distinct documents by grouping the plurality
of pages into the distinct documents by comparing the page markers
for the plurality of pages.
3. The method of claim 2, wherein the page markers comprise image-based page
30 markers derived from a visual appearance of the page.
4. The method of any one of claims 2 to 3, wherein the page markers comprise text-
based page markers derived from the text data in the document.

5. The method of any one of claims 1 to 4 further comprising:
for each stored keyword in the corpus of stored keywords, determining a
location-specific weighting for each file location association; and
generating the plurality of suggested file locations by weighting the plurality of
5 keyword scores using the location-specific weightings.
6. The method of claim 5, wherein:
the database is arranged into a file directory having a plurality of folder levels
with each file location in the plurality of file locations associated with a
10 particular folder level, and
the location-specific weighting for each file location association is determined
using the folder level of the file location corresponding to that file
location association.
- 15 7. The method of any one of claims 1 to 6 further comprising, for each document in
the at least one document:
determining a keyword coefficient for each of the document keywords in the
text data, each keyword coefficient indicating a measure of importance
of the corresponding document keyword to the document; and
20 generating the plurality of keyword scores using the keyword coefficient.
8. The method of claim 7, wherein the measure of importance of the corresponding
document keyword to the document is determined by:
identifying keyword text attributes for the document keyword, the keyword text
25 attributes including at least one of a text size, a text location and a text
format; and
determining the keyword coefficient for the document keyword in the text data
based on the keyword text attributes.
- 30 9. The method of any one of claims 1 to 8, wherein identifying the text data
comprises performing optical character recognition on the document to identify the
text data.

10. The method of any one of claims 1 to 9, further comprising, determining a recommended file name for one of the received documents by:
- 5 determining a keyword coefficient for each of the document keywords in the text data, each keyword coefficient indicating a measure of importance of the corresponding document keyword to the document; and
 - determining the recommended file name using the keyword coefficients of the document keywords.
11. A computer program product for automatic ingestion and filing of documents in a database, the computer program product comprising a non-transitory computer readable storage medium and computer-executable instructions stored on the computer readable storage medium, the instructions for configuring a processor to:
- 15 receive an electronic file including at least one document;
 - for each document in the at least one document
 - 15 identify text data in the document;
 - index the text data to identify one or more document keywords;
 - compare the one or more document keywords to a corpus of stored keywords, the corpus of stored keywords previously generated based on a plurality of documents in the database, wherein each
 - 20 of the stored keywords in the corpus has at least one file location association identifying a file location associated therewith;
 - generate a plurality of keyword scores based on the comparison of the one or more document keywords and the corpus; and
 - 25 generate a plurality of suggested file locations using the keyword scores and file location associations.
12. The computer program product of claim 11, further comprising instructions for configuring the processor to:
- 30 identify a plurality of pages in the file;
 - determine a plurality of page markers for each page;
 - determine that the at least one document in the file comprises a plurality of distinct documents; and

assign each page to one of the distinct documents by grouping the plurality of pages into the distinct documents by comparing the page markers for the plurality of pages.

5 13. The computer program product of claim 12, wherein the page markers comprise image-based page markers derived from a visual appearance of the page.

14. The computer program product of any one of claims 12 to 13, wherein the page markers comprise text-based page markers derived from the text data in the
10 document.

15. The computer program product of any one of claims 11 to 14 further comprising instructions for configuring the processor to:

15 for each stored keyword in the corpus of stored keywords, determine a location-specific weighting for each file location association; and generate the plurality of suggested file locations by weighting the plurality of keyword scores using the location-specific weightings.

16. The computer program product of claim 15, wherein:

20 the database is arranged into a file directory having a plurality of folder levels with each file location in the plurality of file locations associated with a particular folder level, and
the computer program product further comprises instructions for configuring the processor to determine the location-specific weighting for each file
25 location association using the folder level of the file location corresponding to that file location association.

17. The computer program product of any one of claims 11 to 16 further comprising instructions for configuring the processor to, for each document in the at least one
30 document:

determine a keyword coefficient for each of the document keywords in the text data, each keyword coefficient indicating a measure of importance of the corresponding document keyword to the document; and generate the plurality of keyword scores using the keyword coefficient.

18. The computer program product of claim 17, further comprising instructions for configuring the processor to determine the measure of importance of the corresponding document keyword to the document by:

- 5 identifying keyword text attributes for the document keyword, the keyword text attributes including at least one of a text size, a text location and a text format; and
- determining the keyword coefficient for the document keyword in the text data based on the keyword text attributes.

10

19. The computer program product of any one of claims 11 to 18, further comprising instructions for configuring the processor to identify the text data by performing optical character recognition on the document.

15

20. The computer program product of any one of claims 11 to 19, further comprising instructions for configuring the processor to determine a recommended file name for one of the received documents by:

- determining a keyword coefficient for each of the document keywords in the text data, each keyword coefficient indicating a measure of importance
- 20 of the corresponding document keyword to the document; and
- determining the recommended file name using the keyword coefficients of the document keywords.