An approach to document fingerprinting

Yunhyong Kim[1,2] and Seamus Ross[2,1]

[1] University of Glasgow, Glasgow, UK.
yunhyong.kim@glasgow.ac.uk
[2] University of Toronto, Toronto, Canada.
seamus.ross@utoronto.ca

Abstract. The nature of an individual document is often defined by its relationship to selected tasks, societal values, and cultural meaning. The identifying features, regardless of whether the document content is textual, aural or visual, are often delineated in terms of descriptions about the document, for example, *intended audience*, *coverage of topics*, *purpose of creation*, *structure of presentation* as well as relationships to other entities expressed by *authorship*, *ownership*, *production process*, and *geographical and temporal markers*. To secure a comprehensive view of a document, therefore, we must draw heavily on cognitive and/or computational resources not only to extract and classify information at multiple scales, but also to interlink these across multiple dimensions in parallel. Here we present a preliminary thought experiment for fingerprinting documents using textual documents visualised and analysed at multiple scales and dimensions to explore patterns on which we might capitalise.

Keywords: text analysis, natural language processing, patterns, readability

1 Introduction

The usefulness and potential of automating appraisal and selection for archival and records management and digital library management has been examined earlier([7], [13]). These and other studies emphasise the availability of multiple classes of metadata if these processes are to be automated ([14]). Further, the process often involves answering a range of questions about the document, addressing information such as *intended audience*, *coverage of topics*, *purpose of creation*, *structure of presentation* as well as relationships to other entities expressed by *authorship*, *ownership*, *production process*, as well as *geographical and temporal markers*. Improving mechanisms for automating these processes has significant implications for the construction of digital libraries and the development of information discovery and access services to support both human and machine users.

In 2013, Kim and Ross ([9]), highlighted the potential of bringing together a variety of language processing approaches in a parallel processing workflow as a means of assessing selection and appraisal criteria[3] such as those suggested by the Digital Curation Centre (DCC). The discussion, however, was limited to a very high level consideration of potential with little exploration as to how this might be done and how parallel processing of multiple information classes could benefit selection and appraisal. Each of the document characteristics that come into focus, however, draw heavily on cognitive and/or computational resources to extract, making precise guidelines for a comprehensive extraction framework difficult to implement. Here we step back, to visualise and explore multi-scale multi-dimensional profiles of documents, a *document fingerprint*, that would allow automatically deriving answers to complex questions such as those asked in relation to appraisal and selection in digital preservation.

Typical formulations of document analysis focus on three aspects: form, content, and relationship to other documents. These are usually interlinked and inseparable. To understand the nature of documents, however, it can be useful to attempt independent examination of these layers in parallel. For example, by taking a step back, initially, from the content of the textual language to access content-free form of the text, focus can be redirected to structural and stylistic patterns, just the same as we might study the techniques of a painter divorced from the subject of their painting. Salient features of content (for example, semantic annotation such as general and domain specific named entities) can be explored afterwards and/or in parallel, supported by language specific concepts (e.g., part-speech, chunking, parsing), as can the document's relationship to other information outwith the document itself, to situate it within its temporal and spatial context.

Here we briefly examine the *content-free* form of text that makes explicit structural organisation and the kinds of information to be derived from such analyses. The aim is to move away from document analysis methods that immediately rely heavily on content analysis. This approach aligns with recent efforts to build language identification approaches that do not rely on access to content ([1], [10]). The structural examination is intended to complement the limitations of the bag-of-words model (e.g. the Okapi model [8]) in document analysis, returning to the original discussion of language as not merely a bag-of-words ([6]).

The paper emphasises the potential of examining form, content and relationships in parallel. The argument for carrying out several tasks in parallel for mutual improvement is not new [3]. The consideration of content-free form as a driving factor in information processing, while not new, has had less attention. It is the contention of this paper that automated appraisal can only be made viable by processing tasks to reflect form, content, and document relationship in parallel. Here we propose new first steps towards achieving this goal.

[3] http://www.dcc.ac.uk/resources/how-guides/appraise-select-data

2 Analysing Text Structure

There are two immediate ways to divorce content from form when dealing with textual information: the statistical analysis of features common to a wide range of documents and languages, and the transformation of the document to a medium which obscures direct access to content as text. We employ both methods in this paper to demonstrate how they can be used to make transparent document structure. We use the NLTK toolkit[4] to segment text and the Stanford NLP tools[5] to annotate text. The text is then transformed to an image based on the segmentation and annotation.

The document structure presented here uses, among other elements, white space to delimit words and fullstops to delimit sentences. The existence of these delimiters are language dependent characteristics, but the concept of segmentation is present in most human languages, implying that similar types of examination can be applied more widely. For example, while it is well known that white space is not used to delimit words in Chinese, the concept of word segmentation is still in operation and, accordingly, tools have been developed to accommodate this (e.g., the Stanford Word Segmenter[6]).

This discussion depends on two assumptions about the target text:

- Text can be extracted from the object of interest without substantial encoding/decoding problems; and,
- There exist conceptual segmentation of the language into related blocks.

In this discussion, we limit the examination to English texts. English texts typically consist of blocks of text which in turn consist of smaller blocks of text (for example, chapters, followed by paragraphs, followed by sentences). Some types of text adhere to this hierarchy more strictly than others (e.g., plays, for instance, do not). Typically, however, the basic text in English might be considered to have a three-story architecture with the notion of words at the basement of the structure. These words are organised into sentences to form the ground floor of the structure. Sentences, in turn, are organised into additional first-floor data structures (chapters, sections, themes, paragraphs), the most simple structure being line changes or blank lines to enforce block layout.

The structural examination presented is agnostic of identities of textual elements: it focuses on notions of lengths, sizes, and distributions. The length of each word can be measured by the number of characters in the word, the length of each sentence measured by the number of words in the sentence, and the lengths of paragraphs, in turn, can be measured by sentences and/or lines. Theoretically speaking, we could start with characters measured by binary bits rather than starting with words (e.g. a method used in [1] and [10] for language identification). The discussion here, however, is limited to structures designed to be accessible to human perception.

[4] http://www.nltk.org/

[5] http://nlp.stanford.edu/software/corenlp.shtml

[6] http://nlp.stanford.edu/software/segmenter.shtml

The computational approach described here is not intended to be perfectly faithful to the concepts of written languages that inspired them. The data are expected to be noisy: the focus is on the potential of numerical patterns in describing textual structure, in particular, those that might help determine higher level concepts mentioned earlier (such as intended audience). For example, sentence lengths and word lengths (often measured by syllables) already play a central role in determining readability[7] (reading ease in relation to your target audience). Understanding structure, could expand this to determine the relationship between structure and readability, which is less understood.

Text segmentation in this paper was carried out with the Python[8] programming language using wordpunct_tokenizer, sent_tokenizer, line_tokenizer, and blankline_tokenizer, as provided by the NLTK toolkit. These tokenizers segment text, constructed with the aim of extracting words (separated from punctuation), sentences, text separated by new lines, and text blocks separated by blank lines (suggestive of paragraphs). These tools will be applied hierarchically: application of higher level tokenisation followed by lower level tokenization. We will take a brief look at the distribution of text block (words, sentences, paragraphs) sizes, and the structural patterns are further presented in a visualisation to make relationships explicit, a process to be explained further in Section 3.

A brief look at two types of named entity recognition and part-of-speech tagging will be included, to show how document structure (and its relationship to genre such as song lyrics and wikipedia articles), named entity recognition, and part-of-speech tagging can be brought together to diagnose errors. A lot of the language processing tools perform at a reasonable standard already on known types of data (often performing at greater than 90% accuracy). Enhancing overall performance across heterogeneous data requires something new. Some suggest correcting training data ([12]). The argument here suggests that parallel processing to capture different types of information (e.g. document structure, syntax, and named entity), could result in improvement of all processes.

3 From Text to Image

In the first instance, the examination is limited to wikipedia articles, poetry, lyrics, and a tagged PubMed[9] MEDLINE abstracts used in the BioNLP/NLPBA 2004 named entity recognition task[10]. Including an article from the dataset of an information extraction task may seem odd. This article, however, is a great example of structured text. The numbers in Table 3 reflects the number of text blocks extracted using the NLTK tokeniser blankline_tokenize, line_tokenize, sent_tokenize, and, wordpunct_tokenize. The last row of the table presents the number of sentences extracted if the text is not segmented first using lines and

[7] https://en.wikipedia.org/wiki/Readability
[8] https://www.python.org/
[9] http://www.ncbi.nlm.nih.gov/pubmed
[10] http://www.nactem.ac.uk/tsujii/GENIA

blank lines. There is a clear discrepancy between the number of sentences extracted using the two methods, and, in the case of the poem, lyrics, and dataset article, the difference is enormous. This confirms what we already know as being common practice: new lines are used everywhere to format these latter types of documents. In Figures 1, 2 and 3, we present, respectively, the graphs for the Epic Poem, Wikipedia article 2, and the abstract from the BioNLP 2004 dataset, showing the number of sentences (y-axis) for a given length measured by the number of words (x-axis). The epic poem has been truncated to the first 171 sentences to make it more comparable to the article in the BioNLP 2004 dataset. The figures suggest that a poem is, in some ways, more similar to an abstract tagged and structured to be part of a dataset than it is to Wikipedia article written in prose. This is not too surprising, especially since the poem in this example is a blank verse, i.e., poetry expressed in regular metrical unrhymed lines,

Table 1. Number of blocks, lines, sentences, words in each document

Segment	**Wikipedia 1**	**Wikipedia 2**	**Epic Poem**	**Lyrics 1**	**Lyrics 2**	**BioNLP dataset abstract**
blankline	55	49	29	5	5	5
line	133	115	10,572	20	33	166
sentence	404	347	11,266	20	44	171
word punct	9,022	10,128	96,827	143	188	416
sentence (from raw)	363	306	1835	6	12	6

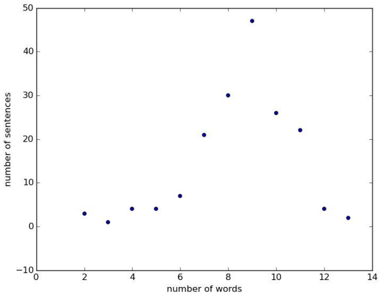

Fig. 1. Epic Poem: graph showing number of sentences (y-axis) for a given length in number of word s (x-axis).

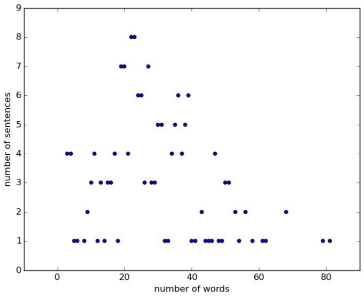

Fig. 2. Wikipedia 2: graph showing number of sentences (y-axis) for a given length in number of word s (x-axis).

almost always iambic pentameter. In fact, the regularity of sentence length distribution in the poem is clear in Figure 4, a graph produced based on an analysis of the entire 96,827 words. In Figures 5 & 6, we have presented a more comprehensive visualisation of the article Wikipedia 1 and the Epic Poem revealing the three-story architecture described in Section 2. On the top, we have words represented as lines, their lengths reflecting the number of characters in the words. In the middle, we have sentences represented as rectangles, their widths representing the number of words in the sentences. Finally, on the bottom, we have paragraphs represented, again by rectangles, their widths mirroring the number of sentences in the paragraph. The representation is only based on the first 2,000 words. The figures illustrate immediately that, word lengths vary more widely in the Wikipedia article than they do in the epic poem (maximum word lengths are twenty-one and sixteen, respectively). The situation is similar for sentences. It is, however, also noticeable that paragraphing is used regularly

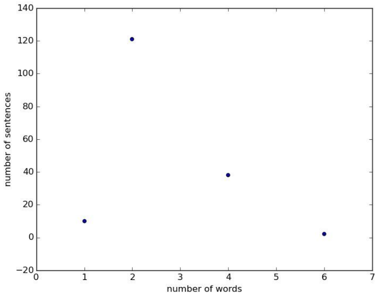

Fig. 3. Tagged Medline abstract: graph showing number of sentences (y-axis) for a given length in number of word s (x-axis).

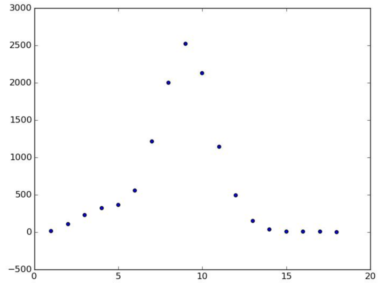

Fig. 4. Epic Poem: graph showing number of sentences (y-axis) for a given length in number of word s (x-axis) for the entire poem.

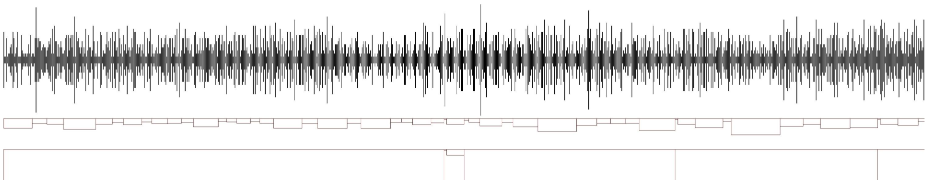

Fig. 5. Wikipedia 1: structural representation of the article with words on the top (lengths of lines corresponding to number of characters), sentences in the middle (size of rectangles reflect number of words in the sentence), and paragraphs on the bottom.

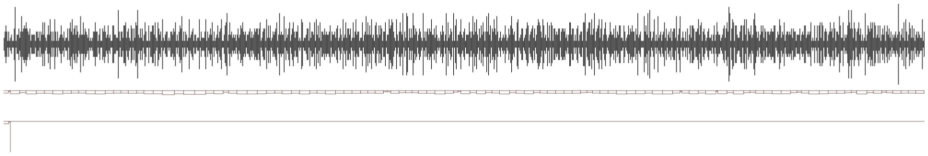

Fig. 6. Epic Poem: structural representation of the poem with words on the top (lengths of lines corresponding to number of characters), sentences in the middle (size of rectangles reflect number of words in the sentence), and paragraphs on the bottom.

throughout the Wikipedia article, whereas, there are hardly any paragraphs in the epic poem. In fact, there are fifty-five blocks of text separated by a blank line in the Wikipedia article consisting of less than 10,000 words compared to twenty-nine blocks in the epic poem across more than 96,000 words. The lack of basement structure poses barriers to readability: for example, the Flesh-Kincaid readability score contrasts the two texts with a beginning university grade audience for the Wikipedia article and a graduate school level audience for the epic poem.

For comparison, in Figure 7 & 8, we present similar visualisation for the MEDLINE article from the BioNLP 2004 dataset (tagged with named enties and structured as trainging data), and song lyrics[11], respectively. The visualisation of words show the MEDLINE article to be very regular with short words and longer words alternating from one extreme to the other as if by rule. There is a frequent stream of short words throughout forming the dark belt in the middle. The lyrics, in contrast, has a wider variety of word lengths, with wavelike hills in many places as words get longer and shorter in increasing steps. Both visualisations show a fair amount of regularity at the sentence and paragraph level, where lines are more regular for the article from the dataset while paragraphs are more regular for song lyrics. The visualisation for the words in natural language text are

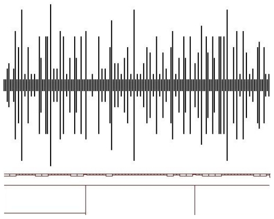

Fig. 7. Visualisation of article from the BioNLP 2004 dataset (top row: words, middle row: sentences, bottom row, paragraphs).

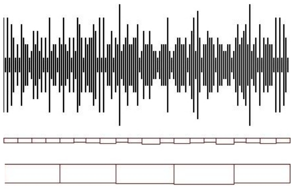

Fig. 8. Visualisation of song lyrics (top row: words, middle row: sentences, bottom row: paragraphs).

[11] Five verse version of “Twinkle Twinkle Little Star”

reminiscent of sound waves, but not so much so for the dataset. By translating the frequency of different lengths into sound frequencies after recalibrating to allow a frequency of one to be 20Hz, this stream can be played as music. The result produces a repeated beat stream, including a constant beat just at the edge of the human hearing range. This is representative of the frequent short length words prominent in the image of Figures 5 & 6 as a black band in the middle.

4 Adding Some Colour

So far, in our discussion we have ignored specific document content and/or classes. Named entity recognition is one way of enriching message understanding. Named entities can be generic, for example, labelling words as instances of location, date, time, person and organisation, or specific to a specialist subject area (e.g. biomedical named entities). So, depending on the recogniser it could provide the reader with a quick summary of topics covered by a document and/or a list of possible candidates to be attributed with authorship, ownership, and geographic and temporal markers.

Just as a small experiment, the first 100 sentences of the article Wikipedia 1 and the Epic Poem were tagged using the Stanford named entity tagger[12] to distinguish four biomedical named entities (DNA, RNA, PROTEIN, CELL TYPE, and CELL LINE). The results are displayed in Figures 9 & 10 as coloured lines in the document word visualisation. The first 100 sentences of the Wikipedia article contained 2013 words, and 0.028% were tagged as biomedical entities, while the first 100 sentences of the poem contained 864 words and 0.0007% of these were returned as instances of biomedical entities. Most of the entities (96.5%) in the Wikipedia article were in the second half of the text, and selected paragraphs seemed to be densely populated with the entities, while 66.6% of words tagged as entities in the poem were in the first half of the poem and seemed to be distributed uniformly across the first half of the text.

Most likely the words tagged as entities in the poem[13] were incorrectly labelled as biomedical entities (in fact, the words were *Eden*, *Man* and *God*[14] all labelled as PROTEIN). While there are incorrect labels in the Wikipedia article[15] (e.g. “economic elements” labelled as protein; proteins labelled as DNA and vice versa), there were also plenty of correct labels (e.g. amyloid precursor protein labelled correctly as PROTEIN). This little experiment suggests that: 1) knowing the genre of the document (for example, poem versus article) can help us predict the accuracy of named entity recognition; and, 2) the way the named entity recogniser labels the document (number of entities returned; the distribution of entities), even if the labelling is inaccurate could inform us about document type.

[12] http://nlp.stanford.edu/software/CRF-NER.shtml

[13] John Milton’s “Paradise Lost”, available from Project Gutenberg.

[14] Capitals retained from original text.

[15] A page on “Alzheimer’s disease.”

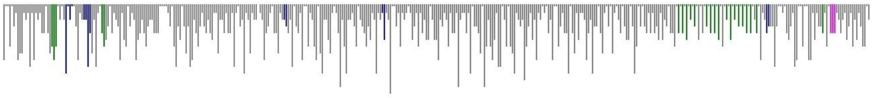

Fig. 9. Wikipedia 1: named entity visualisation (blue: DNA, red: RNA, green: protein, magenta: cell type, and yellow: cell line).

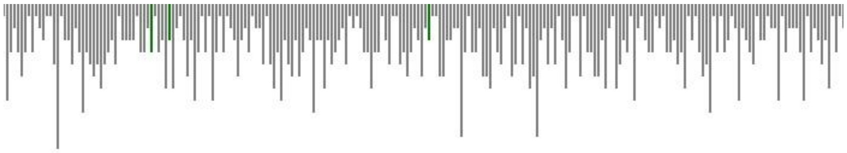

Fig. 10. Epic Poem: named entity visualisation (blue: DNA, red: RNA, green: protein, magenta: cell type, and yellow: cell line).

In Figures 11 & 12, we present a visualisation of generic named entity tagging (for LOCATION, PERSON, ORGANIATION, DATE, and TIME) on the same data. With this tagging the tables are turned on the two texts: only 0.006% of the words in Wikipedia 1 are returned with a named entity tag in the first 100 sentences, while, 0.023% of words in the poem are retuned as a named entity. Despite the change in percentage, the labels on the Wikipedia article still appear to be more plausible (53.8%) than that on the poem (20%). This is most likely because the training data for named entity tagging is almost never a set of poems. This raises the conjecture that precision of tagging performance could be boosted by considering genre coverage in training data[16]. Using the taggers in tandem could also improve the performance of both taggers. For example, closer examination shows that the two types of named entity taggers labelled the same entity APP as DNA and as ORGANISATION, respectively. Since an entity is unlikely to be both DNA and ORGANISATION, this suggests immediately that

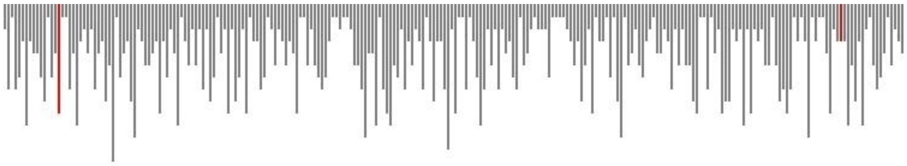

Fig. 11. Wikipedia 1: named entity visualisation (blue: PERSON, red: ORGANISATION, green: TIME, magenta: LOCATION, and yellow: DATE).

[16] Lack of cross-genre applicability in the literature also observed by Nadeau, D. and Sekine, S. A survey of named entity recognition and classification. phLingvisticae Investigationes, 30 (1): 3–26, 2007

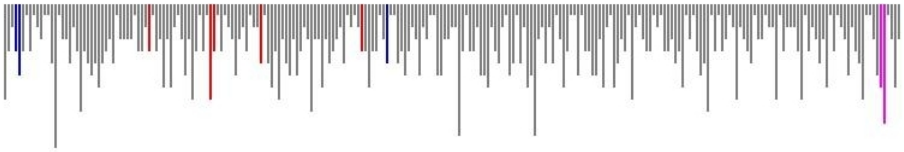

Fig. 12. Epic Poem: named entity visualisation (blue: PERSON, red: ORGANISATION, green: TIME, magenta: LOCATION, and yellow: DATE).

one or both of the taggers has labelled the entity incorrectly (in fact, the former is correct: APP on its own stands for Amyloid Precursor Protein, hence a type of protein but, here, it is used to denote the gene that encodes the protein). Likewise, in the poem, the two independent taggers (incorrectly) tagged Eden as PROTEIN and ORGANISATION, respectively, again signalling probable error.

In addition to named entities, syntactic tagging and text chunking can provide valuable information in identifying salient concepts. Chunks identify linguistic constituents, and could, potentially be used to re-assess reliability of named entity labelling processes, by making explicit which parts of the sentence can belong together in a phrase to specify conceptual entities. The Stanford POS tagger was used to tag the first 100 sentences of the article Wikipedia 1 and Epic Poem, respectively (Figure 13 & 14). There seem to be differences in relative numbers of words in each class distinguishing the two textual classes (the poem seems to have more pronominal tags). This would need further study to validate but other studies have suggested a relationship between genre and POS tags [5]. The tags can be used to identify the longest adjective (blue), noun (green) and verb (pink). For the article, these were neurodegenerative, disorientation, and understood, respectively. For the poem, these were adventurous, disobedience, and unattempted. Combining structural elements (such as size and location) with functional elements (such as POS tag information) can serve to offer a more comprehensive understanding of topical coverage. More immedi-

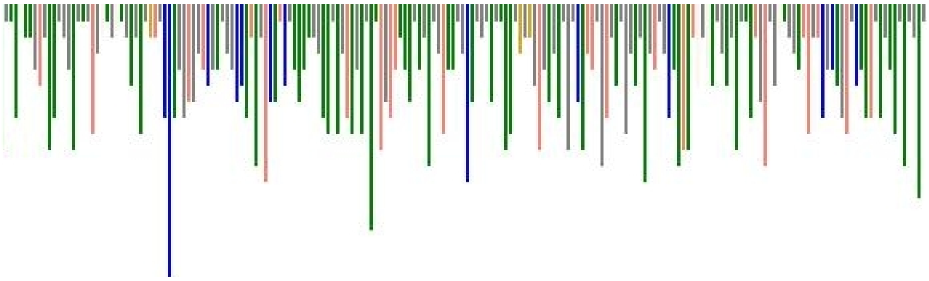

Fig. 13. Wikipedia 1: POS tags visualised (green: Noun, blue: adjective, pink: Verb, gold: Personal Pronouns).

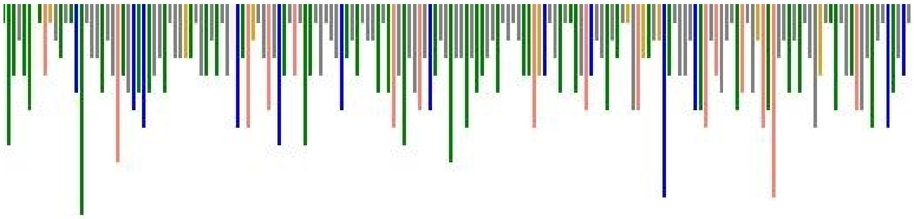

Fig. 14. Epic Poem: POS tags visualised (green: Noun, blue: adjective, pink: Verb, gold: Personal Pronouns).

ately, however, analysing part-of-speech tagging errors can assist simultaneously with resolving errors in named entity recognition. For example, in the poem, *Heavenly* was tagged as a proper noun and ORGANISATION, suggesting the same features can lead to concurring errors.

5 Conclusion

In this research we took a brief look at document structure based on a three-layered architecture consisting of words, sentences and paragraphs. A parallel visualisation of these three levels was explored (in Section 3 above) to highlight possible correlations with document presentation structure, genre, and readability. A substantial amount of information can be gleaned from the documents without analysing their content, which could potentially complement other information extraction tasks, for example, by providing genre information to boost named entity recognition. This analysis has given us confidence to pursue more detailed investigation of the conjecture that many errors resulting from automated content labelling are correlated to document structure and genre. Simple numbers from sentences, words and paragraphs can reveal characteristics of selected text types, such as poems, articles and data structure to enable the first step.

A proper understanding of document structure is only possible by moving away from the bag-of-words model to an approach that considers multiple structures and processes it in parallel. Document structure is integral to understanding document genre, and determining the purpose of creation and use. The changing structure over time tells a story of its own about purpose (e.g., Jane Austen's *Emma* will exhibit a different structure depending on whether it is an edition intended for human consumption or it is part of a computational linguistics corpus).

At the same time named entity recognition captures candidate entities that identify authorship, ownership, affiliations and geographical and temporal markers. Used with entities of specialist domains, named entity recognition can provide a description of document topics. In this paper we saw that tags from one tagger could potentially assist another tagger in *self-assessing* possibility of error. Further, errors can concur around the same area with respect to independent taggers. The system can identify areas of text that might pose an increased level of difficulty for taggers by having access to the performance of several taggers.

The visualisation for the words is reminiscent of sound waves and so it should be because document content is no different from other signals of information. By translating the frequency of different lengths into sound frequencies, this stream can be played as music to reveal a repeating background beat with no specific melody. As a poetic twist we might even conjecture that it is the content analysis that introduces melody to text; characterising textual melodiousness might provide a new metric for genre classification.

6 Next Steps

This research illuminates the potential of visualising and analysing multi-scale and multi-dimensional document characteristics in parallel. These results show that more research will be required if a clear path for document fingerprinting is to be established. Our underlying research is limited to aspects that might lead to answering questions related to document characteristics (e.g., intended audience, geographical markers) discussed in Section 1, and, even in this we only examine some of the information presentation, extraction and classification issues that might be involved; going forward a wider range should be explored.

The discussion is further limited to textual documents. The general concepts, however, are likely to apply equally well to non-textual content as long as it has a natural hierarchical segmentation (e.g. in the case of images, "components" that form "objects" that form "scenes"). The discussed features (e.g. intended audience, topic coverage, authorships) still apply in aural, visual contexts. This would clearly be one of the next targets for research and may unearth further patterns across document types on which we might capitalise.

This research is intended as ground work for multiple other applications. Effective pipelining of automation and the application of multiple techniques in tandem provides the most viable method for addressing resource management in archival and records-based digital libraries. A variety of approaches have been proposed to support this automation and our own investigations have in the past focused on document analysis. Here we have shown that other methods have potential to enhance the precision and recall of these processes.*Document fingerprinting* has implications for areas such as author attribution([4]), near duplicate detection ([11]), and plagiarism detection ([2]). These, however, will also depend on the scalability and efficiency of the approach, suggesting the necessity for testing performance on larger datasets.

7 Acknowledgement

This research was supported in part by the Universities of Glasgow and Toronto, and the European Commission through Blogforever (FP7-ICT-2009-6-269963). We benefited from insightful observations by anonymous ICADL2015 reviewers.

References

1. Baldwin, T., Lui, M.: Language identification: The long and the short of the matter. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 229–237. HLT '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), http://dl.acm.org/citation.cfm?id=1857999.1858026
2. Barrón-Cedeño, A., Vila, M., Martí, M., Rosso, P.: Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. Comput. Linguist. 39(4), 917–947 (Dec 2013), http://dx.doi.org/10.1162/COLI_a_00153
3. Cohen, H., Crammer, K.: Learning multiple tasks in parallel with a shared annotator. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) NIPS. pp. 1170–1178 (2014), http://dblp.uni-trier.de/db/conf/nips/nips2014.html#CohenC14
4. Donais, J.A., Frost, R.A., Peelar, S.M., Roddy, R.A.: A system for the automated author attribution of text and instant messages. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 1484–1485. ASONAM '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2492517.2500308
5. Fang, A.C., Cao, J.: Enhanced genre classification through linguistically fine-grained pos tags. In: Otoguro, R., Ishikawa, K., Umemoto, H., Yoshimoto, K., Harada, Y. (eds.) PACLIC. pp. 85–94. Institute for Digital Enhancement of Cognitive Development, Waseda University (2010)
6. Harris, Z.: Distributional structure. Word 10(23), 146–162 (1954)
7. Harvey, R.: Appraisal and selection. In: Curation Reference Manual. Digital Curation Center, http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/appraisal-and-selection (2007)
8. Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments - part 1. Inf. Process. Manage. 36(6), 779–808 (2000), http://dblp.uni-trier.de/db/journals/ipm/ipm36.html#JonesWR00
9. Kim, Y., Ross, S.: Closing the loop: Assisting archival appraisal and information retrieval in one sweep. In: Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries. pp. 16:1–16:10. ASIST '13, American Society for Information Science, Silver Springs, MD, USA (2013), http://dl.acm.org/citation.cfm?id=2655780.2655796
10. Lui, M., Lau, J.H., Baldwin, T.: Automatic detection and language identification of multilingual documents. TACL 2, 27–40 (2014)
11. Manku, G.S., Jain, A., Das Sarma, A.: Detecting near-duplicates for web crawling. In: Proceedings of the 16th International Conference on World Wide Web. pp. 141–150. WWW '07, ACM, New York, NY, USA (2007), http://doi.acm.org/10.1145/1242572.1242592
12. Manning, C.D.: Part-of-speech tagging from 97linguistics? In: Proceedings 12th International Conference on Computational Linguistics and Intelligent Text Processing. pp. 171–189. CICLing'11, Springer-Verlag, Berlin, Heidelberg (2011), http://nlp.stanford.edu/ manning/papers/CICLing2011-manning-tagging.pdf
13. Oliver, G., Ross, S., Guercio, M., Pala, C.: Report on automated re-appraisal: Managing archives in digital libraries (2008)
14. Oliver, G., Kim, Y., Ross, S.: Documentary genre and digital recordkeeping: red herring or a way forward? Archival Science 8, 295–305 (2008)