Caraterização Estética de Texto

Proposta de Dissertação de Mestrado

Orientador: João Paulo Cordeiro
co-Orientador: Pedro Inácio

1 Objetivos

A vasta variedade literária é certamente uma das maiores riquezas culturais da humanidade, senão mesmo a maior. A escrita não é só um instrumento de comunicação humana. Com ela podemos viajar para o passado, podemos entrar na mente e coração de outras pessoas, assim como podemos deixar importantes legados para as gerações futuras. A escrita tem sido uma poderosa ferramenta de revolução das sociedades e do pensamento em geral.

Na era digital, a variedade literária cresce de forma exponencial. Basta pensar no inúmero volume de texto que quotidianamente é publicado na *Web*, sejam artigos jornalísticos, documentos técnico-científicos, enciclopédias, blogs e livros on-line, dos mais variados géneros e assuntos. Esta realidade mais recente é também caracterizada por um menor escrutínio do processo de publicação — em muitos meios, cada um publica o que quer e da forma que quer. Isto levanta necessariamente um novo desafio, que é o de conhecer a qualidade do texto. Intuitivamente estamos certos que uma obra literária clássica terá necessariamente uma qualidade de escrita bem diferente daquela que encontramos em muitos locais de publicação on-line aberta, como é o caso das redes sociais e blogs.

Deste modo, o trabalho tem como objetivo experimentar modelos matemáticos que permitam caracterizar a qualidade de escrita de um texto. Um desses modelos envolve a análise de estruturas *autossemelhantes* (*fractais*) no próprio texto. Esta propriedade mede, de certa forma, a constância implícita de uma série, estando presente em muitos processos naturais e artificiais (e.g., no tráfego de rede, no preço das ações, etc.). Uma das formas de analisar a auto semelhança é através da estimação de um parâmetro, conhecido por parâmetro de Hurst, que ora se aproxima de 0,5 ou de 1,0 conforme a série de valores analisados comporta-se mais ou menos aleatoriamente, respetivamente.

Pretendemos investigar o que realmente carateriza aquilo que quase intuitivamente se entende como qualidade de escrita de um texto, partindo do princípio que esta possa estar relacionada com o valor da auto semelhança subjacente em certas características do texto, quer de nível léxico, sintático ou mesmo semântico. Deste modo, o(a) aluno(a) irá

trabalhar com um conjunto de ferramentas de processamento de texto e de análise da auto semelhança já disponíveis, e explorará medidas e modelos matemáticos que lhe serão indicados, de modo a implementar um procedimento que permite aferir a qualidade de um texto. Como "matéria prima" (corpora), trabalharemos com uma coleção de textos proveniente de blogs e um conjunto de obras literárias, disponíveis on-line, através do projeto Gutenberg (http://www.gutenberg.org).

2 Tarefas a Realizar

T1 Estudo da área e técnicas subjacentes.

T2 Experiências piloto e planeamento das experiências avançadas.

T3 Implementação do trabalho experimental.

T4 Escrita de um artigo científico.

T5 Escrita da dissertação.

3 Cronograma

T1 - 2 meses	T2 - 1 mês	T3 - 3 meses
T4 - 1 mês	T5 – 2 meses	

7 Referências

[1] C. Collberg and S. Kobourov. Self-plagiarism in Computer Science, Communications of the ACM 48(4): 88 - 94, 2005.

[2] Pedro R. M. Inácio, Study of the Impact of Intensive Attacks in the Self-Similarity Degree of the Network Traffic in Intra-Domain Aggregation Points, Ph.D. Thesis, Ph.D. in Computer Science and Engeneering, Department of Computer Science, University of Beira Interior, Covilhã, December, 2009.

[3] Diogo A. B. Fernandes, Miguel Neto, Liliana F. B. Soares, Mário M. Freire and Pedro R. M. Inácio, A Tool for Estimating the Hurst Parameter and for Generating Self-Similar Sequences, in Proceedings of the 46th Summer Computer Simulation Conference 2014 (SCSC 2014), Monterey, CA, USA, July 6-10, 2014.

[4] Apache Incubator (available on January, 2014). The Apache OpenNLP library: https://opennlp.apache.org

[5] João P. Cordeiro. The HultigLib — Nuggets for Text Processing in Java. On-line reference and tutorial. UBI, March 2012.

8 Contactos

João Paulo Cordeiro, Boco 6 / Gabinete 4.3
Pedro Inácio, Bloco 6 / Gabinete 4.1

UBI, Departamento de Informática
Rua Marquês d'Ávila e Bolama
6201-001 Covilhã